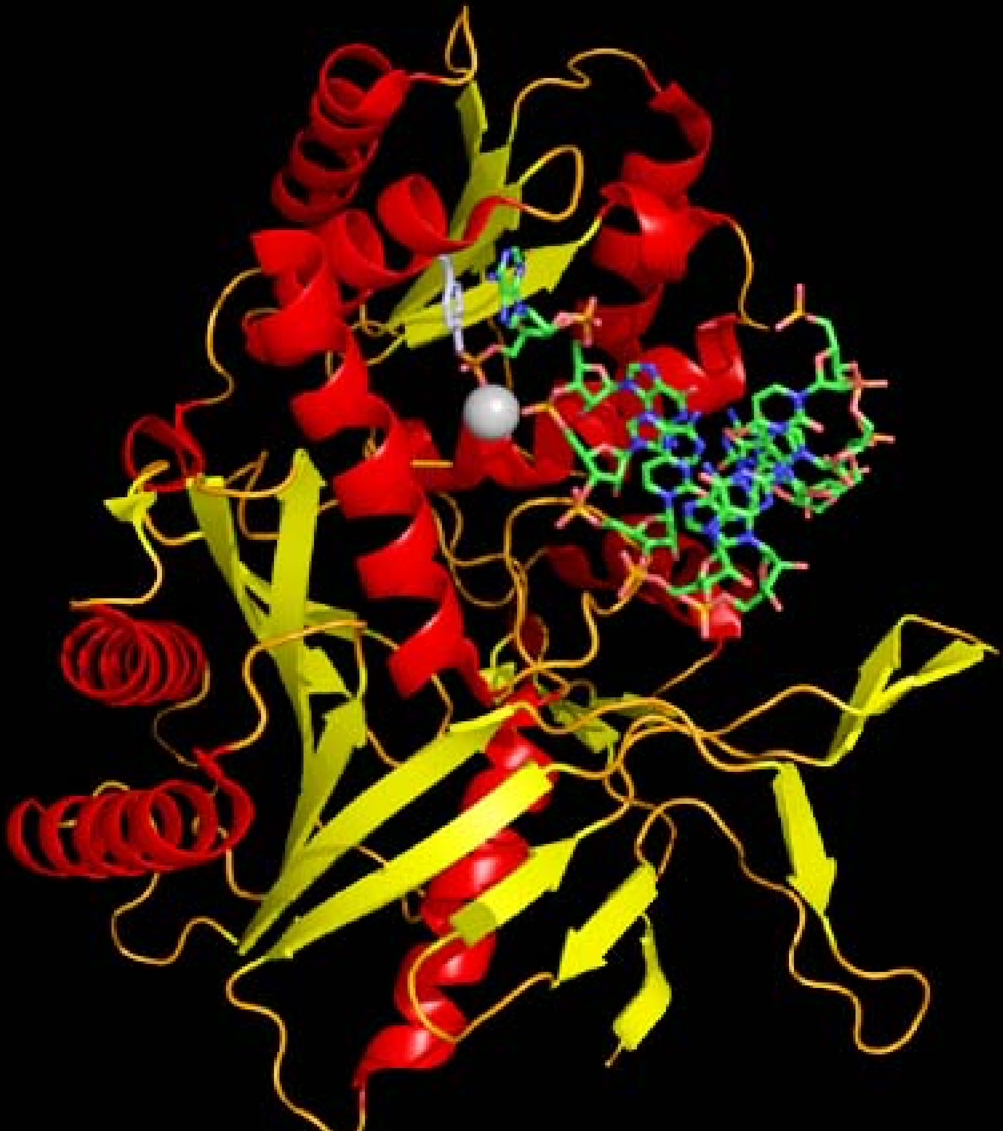
Handbook of RNA Biology



Burl Guillen

First Edition, 2012

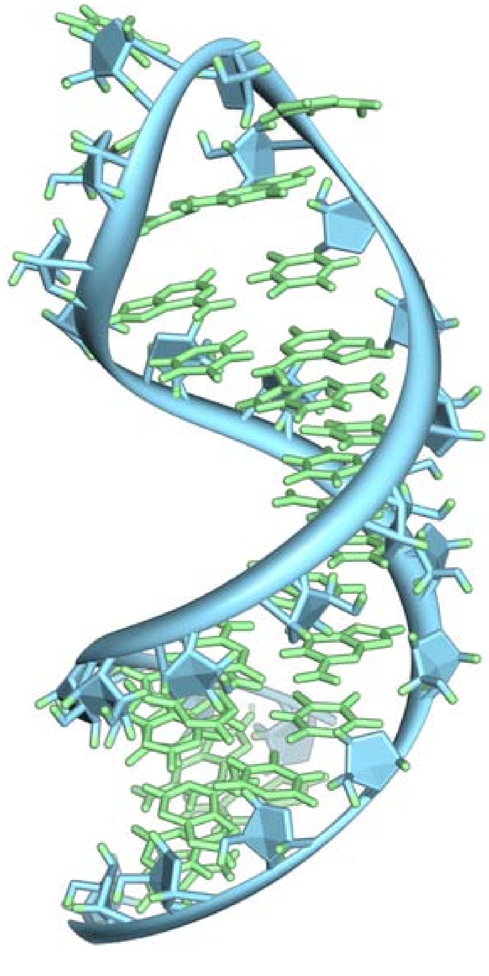
ISBN 978-81-323-3300-5

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

RNA



A hairpin loop from a pre-mRNA. Highlighted are the nucleobases (green) and the ribose-phosphate backbone (blue).

Ribonucleic acid (**RNA**) is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life.

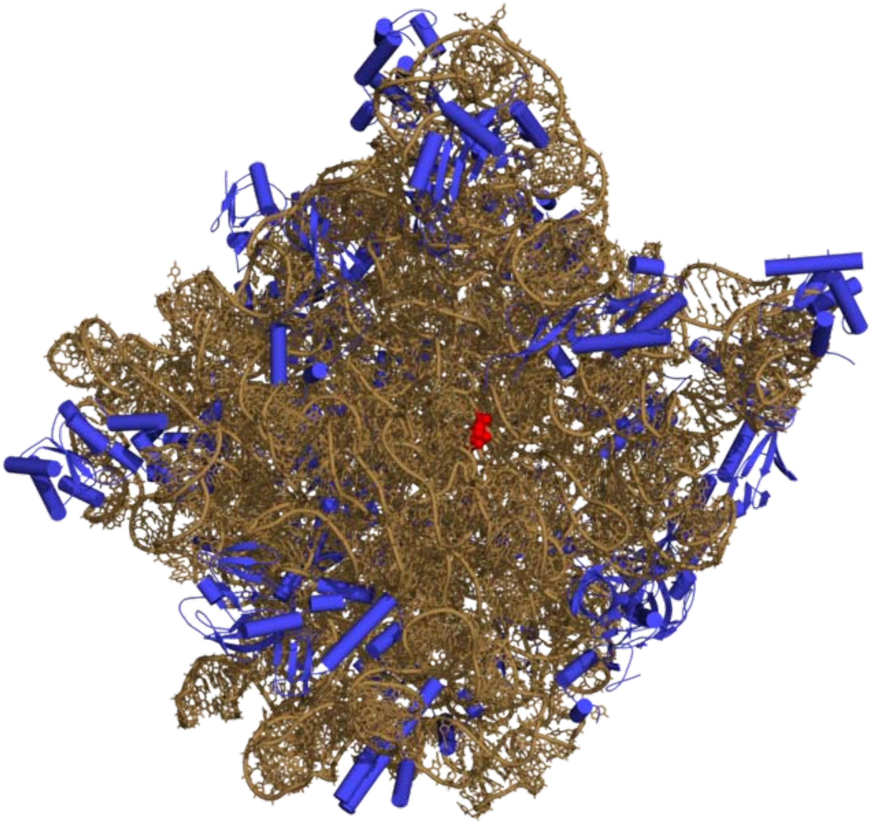
Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase (sometimes called a nitrogenous base), a ribose sugar, and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. For example, some viruses use RNA instead of DNA as their genetic material, and all organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins.

Like proteins, some RNA molecules play an active role in cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these active processes is protein synthesis, a universal function whereby mRNA molecules direct the assembly of proteins on ribosomes. This process uses transfer RNA (tRNA) molecules to deliver amino acids to the ribosome, where ribosomal RNA (rRNA) links amino acids together to form proteins.

The chemical structure of RNA is very similar to that of DNA, with two differences--(a) RNA contains the sugar ribose while DNA contains the slightly different sugar deoxyribose (a type of ribose that lacks one oxygen atom), and (b) RNA has the nucleobase uracil while DNA contains thymine (uracil and thymine have similar base-pairing properties).

Unlike DNA, most RNA molecules are single-stranded. Single-stranded RNA molecules adopt very complex three-dimensional structures, since they are not restricted to the repetitive double-helical form of double-stranded DNA. RNA is made within living cells by RNA polymerases, enzymes that act to copy a DNA or RNA template into a new RNA strand through processes known as transcription or RNA replication, respectively.

Comparison with DNA

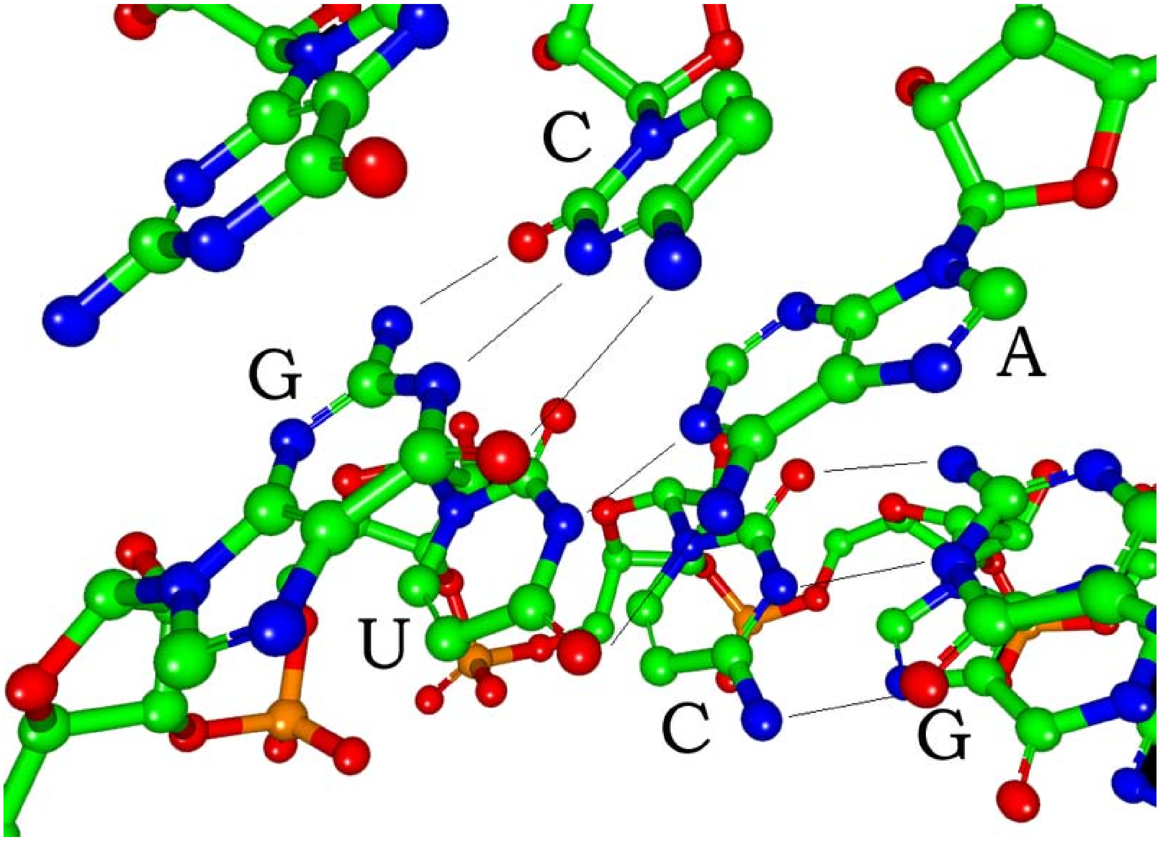


Three-dimensional representation of the 50S ribosomal subunit. RNA is in ochre, protein in blue. The active site is in the middle (red).

RNA and DNA are both nucleic acids, but differ in three main ways. First, unlike DNA, which is, in general, double-stranded, RNA is a single-stranded molecule in many of its biological roles and has a much shorter chain of nucleotides. Second, while DNA contains *deoxyribose*, RNA contains *ribose* (in deoxyribose there is no hydroxyl group attached to the pentose ring in the 2' position). These hydroxyl groups make RNA less stable than DNA because it is more prone to hydrolysis. Third, the complementary base to adenine is not thymine, as it is in DNA, but rather uracil, which is an unmethylated form of thymine.

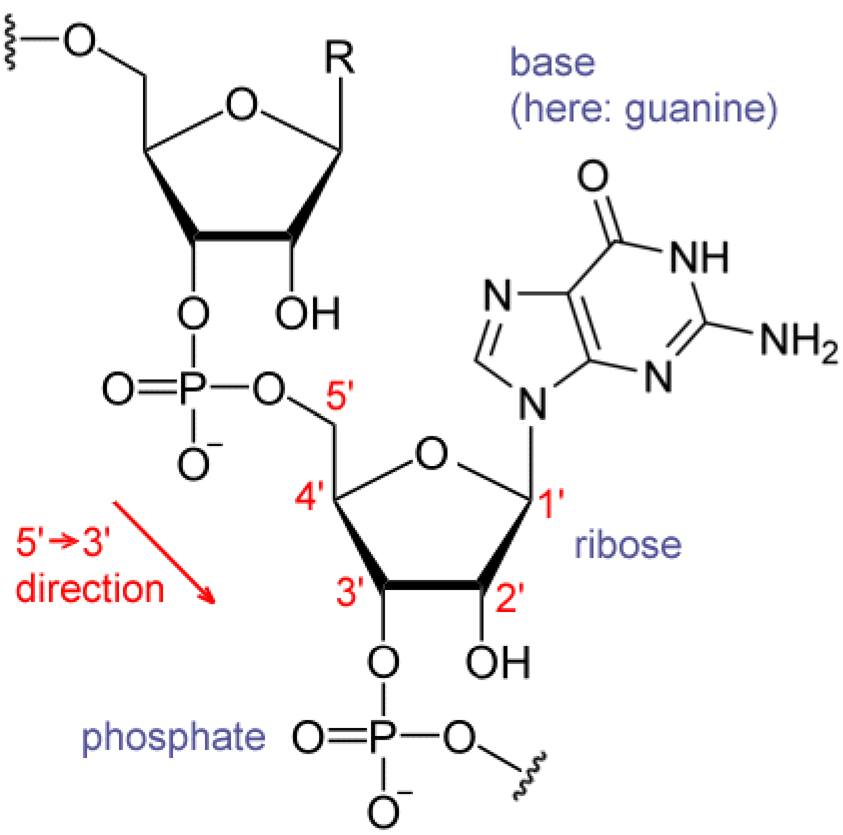
Like DNA, most biologically active RNAs, including mRNA, tRNA, rRNA, snRNAs, and other non-coding RNAs, contain self-complementary sequences that allow parts of the RNA to fold and pair with itself to form double helices. Structural analysis of these RNAs has revealed that they are highly structured. Unlike DNA, their structures do not consist of long double helices but rather collections of short helices packed together into structures akin to proteins. In this fashion, RNAs can achieve chemical catalysis, like enzymes. For instance, determination of the structure of the ribosome—an enzyme that catalyzes peptide bond formation—revealed that its active site is composed entirely of RNA.

Structure



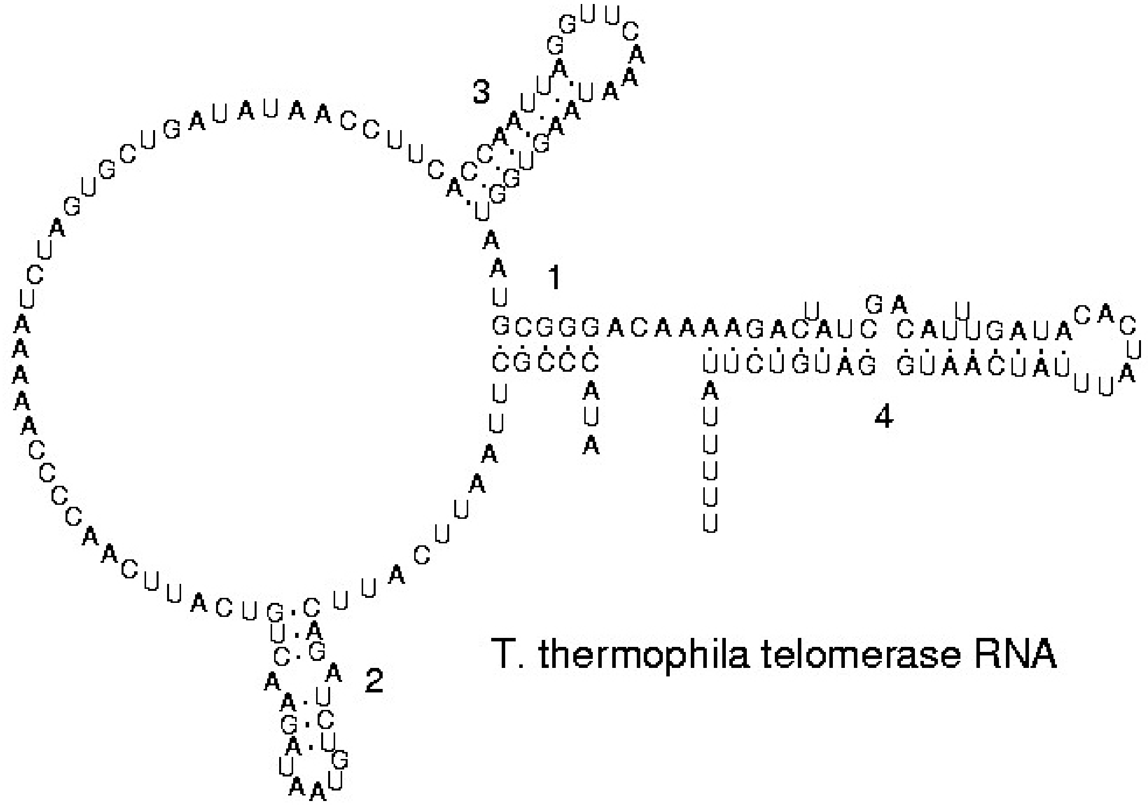
Watson-Crick base pairs in a siRNA (hydrogen atoms are not shown)

Each nucleotide in RNA contains a ribose sugar, with carbons numbered 1' through 5'. A base is attached to the 1' position, in general, adenine (A), cytosine (C), guanine (G), or uracil (U). Adenine and guanine are purines, cytosine, and uracil are pyrimidines. A phosphate group is attached to the 3' position of one ribose and the 5' position of the next. The phosphate groups have a negative charge each at physiological pH, making RNA a charged molecule (polyanion). The bases may form hydrogen bonds between cytosine and guanine, between adenine and uracil and between guanine and uracil. However, other interactions are possible, such as a group of adenine bases binding to each other in a bulge, or the GNRA tetraloop that has a guanine–adenine base-pair.



Chemical structure of RNA

An important structural feature of RNA that distinguishes it from DNA is the presence of a hydroxyl group at the 2' position of the ribose sugar. The presence of this functional group causes the helix to adopt the A-form geometry rather than the B-form most commonly observed in DNA. This results in a very deep and narrow major groove and a shallow and wide minor groove. A second consequence of the presence of the 2'-hydroxyl group is that in conformationally flexible regions of an RNA molecule (that is, not involved in formation of a double helix), it can chemically attack the adjacent phosphodiester bond to cleave the backbone.



Secondary structure of a telomerase RNA

RNA is transcribed with only four bases (adenine, cytosine, guanine and uracil), but these bases and attached sugars can be modified in numerous ways as the RNAs mature. Pseudouridine (Ψ), in which the linkage between uracil and ribose is changed from a C–N bond to a C–C bond, and ribothymidine (T) are found in various places (the most notable ones being in the TΨC loop of tRNA). Another notable modified base is hypoxanthine, a deaminated adenine base whose nucleoside is called inosine (I). Inosine plays a key role in the wobble hypothesis of the genetic code.

There are nearly 100 other naturally occurring modified nucleosides, of which pseudouridine and nucleosides with 2'-O-methylribose are the most common. The specific roles of many of these modifications in RNA are not fully understood. However, it is notable that, in ribosomal RNA, many of the post-transcriptional modifications occur in highly functional regions, such as the peptidyl transferase center and the subunit interface, implying that they are important for normal function.

The functional form of single stranded RNA molecules, just like proteins, frequently requires a specific tertiary structure. The scaffold for this structure is provided by secondary structural elements that are hydrogen bonds within the molecule. This leads to several recognizable "domains" of secondary structure like hairpin loops, bulges, and internal loops. Since RNA is charged, metal ions such as Mg^{2+} are needed to stabilise many secondary and tertiary structures.

Synthesis

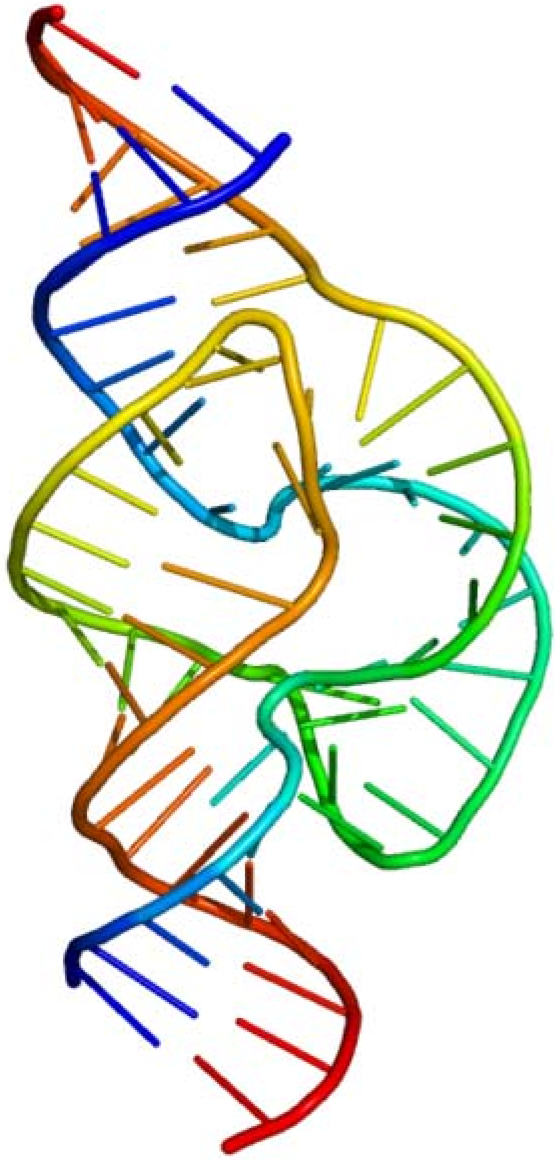
Synthesis of RNA is usually catalyzed by an enzyme—RNA polymerase—using DNA as a template, a process known as transcription. Initiation of transcription begins with the binding of the enzyme to a promoter sequence in the DNA (usually found "upstream" of a gene). The DNA double helix is unwound by the helicase activity of the enzyme. The enzyme then progresses along the template strand in the 3' to 5' direction, synthesizing a complementary RNA molecule with elongation occurring in the 5' to 3' direction. The DNA sequence also dictates where termination of RNA synthesis will occur.

RNAs are often modified by enzymes after transcription. For example, a poly(A) tail and a 5' cap are added to eukaryotic pre-mRNA and introns are removed by the spliceosome.

There are also a number of RNA-dependent RNA polymerases that use RNA as their template for synthesis of a new strand of RNA. For instance, a number of RNA viruses (such as poliovirus) use this type of enzyme to replicate their genetic material. Also, RNA-dependent RNA polymerase is part of the RNA interference pathway in many organisms.

Types of RNA

Overview



Structure of a hammerhead ribozyme, a ribozyme that cuts RNA

Messenger RNA (mRNA) is the RNA that carries information from DNA to the ribosome, the sites of protein synthesis (translation) in the cell. The coding sequence of the mRNA determines the amino acid sequence in the protein that is produced. Many RNAs do not code for protein however (about 97% of the transcriptional output is non-protein-coding in eukaryotes).

These so-called non-coding RNAs ("ncRNA") can be encoded by their own genes (RNA genes), but can also derive from mRNA introns. The most prominent examples of non-coding RNAs are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation. There are also non-coding RNAs involved in gene regulation, RNA processing and other roles. Certain RNAs are able to catalyse chemical reactions such as cutting and ligating other RNA molecules, and the catalysis of peptide bond formation in the ribosome; these are known as ribozymes.

In translation

Messenger RNA (mRNA) carries information about a protein sequence to the ribosomes, the protein synthesis factories in the cell. It is coded so that every three nucleotides (a codon) correspond to one amino acid. In eukaryotic cells, once precursor mRNA (pre-mRNA) has been transcribed from DNA, it is processed to mature mRNA. This removes its introns—non-coding sections of the pre-mRNA. The mRNA is then exported from the nucleus to the cytoplasm, where it is bound to ribosomes and translated into its corresponding protein form with the help of tRNA. In prokaryotic cells, which do not have nucleus and cytoplasm compartments, mRNA can bind to ribosomes while it is being transcribed from DNA. After a certain amount of time the message degrades into its component nucleotides with the assistance of ribonucleases.

Transfer RNA (tRNA) is a small RNA chain of about 80 nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. It has sites for amino acid attachment and an anticodon region for codon recognition that binds to a specific sequence on the messenger RNA chain through hydrogen bonding.

Ribosomal RNA (rRNA) is the catalytic component of the ribosomes. Eukaryotic ribosomes contain four different rRNA molecules: 18S, 5.8S, 28S and 5S rRNA. Three of the rRNA molecules are synthesized in the nucleolus, and one is synthesized elsewhere. In the cytoplasm, ribosomal RNA and protein combine to form a nucleoprotein called a ribosome. The ribosome binds mRNA and carries out protein synthesis. Several ribosomes may be attached to a single mRNA at any time. rRNA is extremely abundant and makes up 80% of the 10 mg/ml RNA found in a typical eukaryotic cytoplasm.

Transfer-messenger RNA (tmRNA) is found in many bacteria and plastids. It tags proteins encoded by mRNAs that lack stop codons for degradation and prevents the ribosome from stalling.

Regulatory RNAs

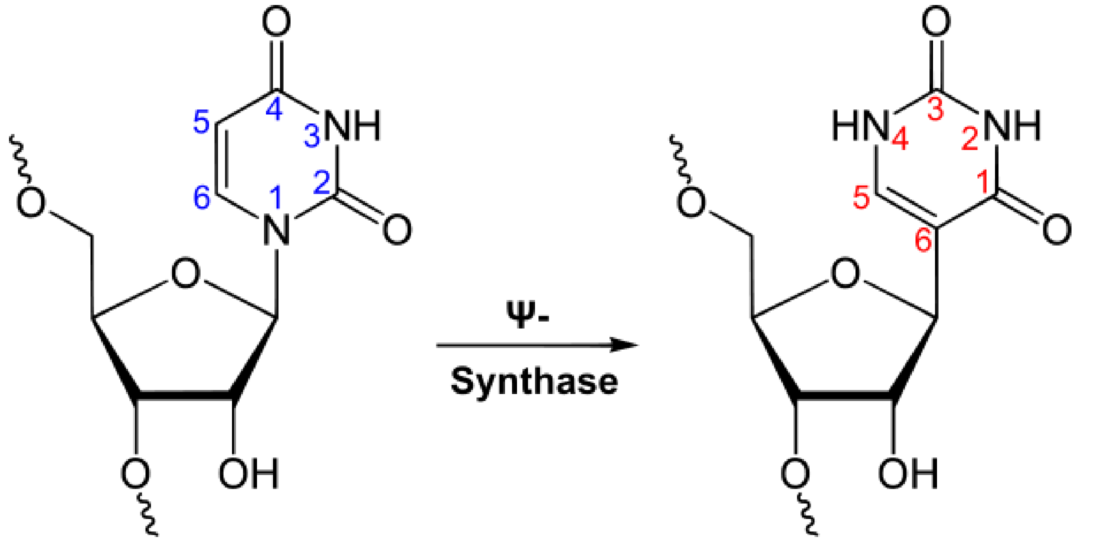
Several types of RNA can downregulate gene expression by being complementary to a part of an mRNA or a gene's DNA. MicroRNAs (miRNA; 21-22 nt) are found in eukaryotes and act through RNA interference (RNAi), where an effector complex of miRNA and enzymes can break down mRNA to which the miRNA is complementary, block the mRNA from being translated, or accelerate its degradation. While small interfering RNAs (siRNA; 20-25 nt) are often produced by breakdown of viral RNA, there are also endogenous sources of siRNAs.

siRNAs act through RNA interference in a fashion similar to miRNAs. Some miRNAs and siRNAs can cause genes they target to be methylated, thereby decreasing or increasing transcription of those genes. Animals have Piwi-interacting RNAs (piRNA; 29-30 nt) which are active in germline cells and are thought to be a defense against transposons and play a role in gametogenesis.

Many prokaryotes have CRISPR RNAs, a regulatory system similar to RNA interference. Antisense RNAs are widespread; most downregulate a gene, but a few are activators of transcription. One way antisense RNA can act is by binding to an mRNA, forming double-stranded RNA that is enzymatically degraded. There are many long noncoding RNAs that regulate genes in eukaryotes, one such RNA is Xist, which coats one X chromosome in female mammals and inactivates it.

An mRNA may contain regulatory elements itself, such as riboswitches, in the 5' untranslated region or 3' untranslated region; these cis-regulatory elements regulate the activity of that mRNA. The untranslated regions can also contain elements that regulate other genes.

In RNA processing



Uridine to pseudouridine is a common RNA modification

Many RNAs are involved in modifying other RNAs. Introns are spliced out of pre-mRNA by spliceosomes, which contain several small nuclear RNAs (snRNA), or the introns can be ribozymes that are spliced by themselves. RNA can also be altered by having its nucleotides modified to other nucleotides than A, C, G and U. In eukaryotes, modifications of RNA nucleotides are generally directed by small nucleolar RNAs (snoRNA; 60-300 nt), found in the nucleolus and cajal bodies. snoRNAs associate with enzymes and guide them to a spot on an RNA by basepairing to that RNA. These enzymes then perform the nucleotide modification. rRNAs and tRNAs are extensively modified, but snRNAs and mRNAs can also be the target of base modification.

RNA genomes

Like DNA, RNA can carry genetic information. RNA viruses have genomes composed of RNA, and a variety of proteins encoded by that genome. The viral genome is replicated by some of those proteins, while other proteins protect the genome as the virus particle moves to a new host cell. Viroids are another group of pathogens, but they consist only of RNA, do not encode any protein and are replicated by a host plant cell's polymerase.

In reverse transcription

Reverse transcribing viruses replicate their genomes by reverse transcribing DNA copies from their RNA; these DNA copies are then transcribed to new RNA. Retrotransposons also spread by copying DNA and RNA from one another, and telomerase contains an RNA that is used as template for building the ends of eukaryotic chromosomes.

Double-stranded RNA

Double-stranded RNA (dsRNA) is RNA with two complementary strands, similar to the DNA found in all cells. dsRNA forms the genetic material of some viruses (double-stranded RNA viruses). Double-stranded RNA such as viral RNA or siRNA can trigger RNA interference in eukaryotes, as well as interferon response in vertebrates.

Key discoveries in RNA biology

Research on RNA has led to many important biological discoveries and numerous Nobel Prizes. Nucleic acids were discovered in 1868 by Friedrich Miescher, who called the material 'nuclein' since it was found in the nucleus. It was later discovered that prokaryotic cells, which do not have a nucleus, also contain nucleic acids. The role of RNA in protein synthesis was suspected already in 1939. Severo Ochoa won the 1959 Nobel Prize in Medicine (shared with Arthur Kornberg) after he discovered an enzyme that can synthesize RNA in the laboratory. Ironically, the enzyme discovered by Ochoa (polynucleotide phosphorylase) was later shown to be responsible for RNA degradation, not RNA synthesis.

The sequence of the 77 nucleotides of a yeast tRNA was found by Robert W. Holley in 1965, winning Holley the 1968 Nobel Prize in Medicine (shared with Har Gobind Khorana and Marshall Nirenberg). In 1967, Carl Woese hypothesized that RNA might be catalytic and suggested that the earliest forms of life (self-replicating molecules) could have relied on RNA both to carry genetic information and to catalyze biochemical reactions—an RNA world.

During the early 1970s, retroviruses and reverse transcriptase were discovered, showing for the first time that enzymes could copy RNA into DNA (the opposite of the usual route for transmission of genetic information). For this work, David Baltimore, Renato Dulbecco and Howard Temin were awarded a Nobel Prize in 1975. In 1976, Walter Fiers and his team determined the first complete nucleotide sequence of an RNA virus genome, that of bacteriophage MS2.

In 1977, introns and RNA splicing were discovered in both mammalian viruses and in cellular genes, resulting in a 1993 Nobel to Philip Sharp and Richard Roberts. Catalytic RNA molecules (ribozymes) were discovered in the early 1980s, leading to a 1989 Nobel award to Thomas Cech and Sidney Altman. In 1990 it was found in petunia that introduced genes can silence similar genes of the plant's own, now known to be a result of RNA interference.

At about the same time, 22 nt long RNAs, now called microRNAs, were found to have a role in the development of *C. elegans*. Studies on RNA interference gleaned a Nobel Prize for Andrew Fire and Craig Mello in 2006, and another Nobel was awarded for studies on transcription of RNA to Roger Kornberg in the same year. The discovery of gene regulatory RNAs has led to attempts to develop drugs made of RNA, such as siRNA, to silence genes.

Chapter- 2

History of RNA Biology

Numerous key discoveries in biology have emerged from studies of RNA (ribonucleic acid), including seminal work in the fields of biochemistry, genetics, microbiology, molecular biology, molecular evolution and structural biology. As of 2010, 30 scientists have been awarded Nobel Prizes for experimental work that includes studies of RNA.

1930 - 1950

RNA and DNA have distinct chemical properties

When first studied in the early 1900s, the chemical and biological differences between RNA and DNA were not apparent, and they were named after the materials from which they were isolated; RNA was initially known as "yeast nucleic acid" and DNA was "pancreas nucleic acid". Using diagnostic chemical tests, carbohydrate chemists showed that the two nucleic acids contained different sugars, whereupon the common name for RNA became "ribose nucleic acid". Other early biochemical studies showed that RNA was readily broken down at high pH, while DNA was stable (although denatured) in alkali. Nucleoside composition analysis showed first that RNA contained similar nucleobases to DNA, with uracil instead of thymine, and that RNA contained a number of minor nucleobase components, e.g. small amounts of pseudouridine and dimethylguanine.

1951 - 1965

Messenger RNA (mRNA) carries genetic information that directs protein synthesis

The concept of messenger RNA emerged during the late 1950s, and is associated with Crick's description of his "Central Dogma of Molecular Biology", which asserted that DNA led to the formation of RNA, which in turn led to the synthesis of proteins. During the early 1960s, sophisticated genetic analysis of mutations in the lac operon of E. coli and in the rII locus of bacteriophage T4 were instrumental in defining the nature of both messenger RNA and the genetic code. The short-lived nature of bacterial RNAs, together with the highly complex nature of the cellular mRNA population, made the biochemical isolation of mRNA very challenging. This problem was overcome in the 1960s by the use of reticulocytes in vertebrates, which produce large quantities of mRNA that are highly enriched in RNA encoding alpha- and beta-globin (the two major protein chains of hemoglobin).

Ribosomes make proteins

In the 1950s, results of labeling experiments in rat liver showed that radioactive amino acids were found to be associated with "microsomes" (later redefined as ribosomes) very rapidly after administration, and before they became widely incorporated into cellular proteins. Ribosomes were first visualized using electron microscopy, and their ribonucleoprotein components were identified by biophysical methods, chiefly sedimentation analysis within ultracentrifuges capable of generating very high accelerations (equivalent to hundreds of thousands times gravity). Polysomes (multiple ribosomes moving along a single mRNA molecule) were identified in the early 1960s, and their study led to an understanding of how ribosomes proceed to read the mRNA in a 5' to 3' direction, processively generating proteins as they do so.

Transfer RNA (tRNA) is the physical link between RNA and protein

Biochemical fractionation experiments showed that radioactive amino acids were rapidly incorporated into small RNA molecules that remained soluble under conditions where larger RNA-containing particles would precipitate. These molecules were termed soluble (sRNA) and were later renamed transfer RNA (tRNA). Subsequent studies showed that (i) every cell has multiple species of tRNA, each of which is associated with a single specific amino acid, (ii) that there are a matching set of enzymes responsible for linking tRNAs with the correct amino acids, and (iii) that tRNA anticodon sequences form a specific decoding interaction with mRNA codons.

The genetic code is solved

The genetic code consists of the translation of particular nucleotide sequences in mRNA to specific amino acid sequences in proteins (polypeptides). The ability to work out out the genetic code emerged from the convergence of three different areas of study--(i) new methods to generate synthetic RNA molecules of defined composition to serve as artificial mRNAs, (ii) development of *in vitro* translation systems that could used to translate the synthetic mRNAs into protein, and (iii) experimental and theoretical genetic work which established that the code was written in three letter "words" (codons). Today, our understanding of the genetic code permits the prediction of the amino sequence of the protein products of the tens of thousands of genes whose sequences are being determined in genome studies.

RNA polymerase is purified

The biochemical purification and characterization of RNA polymerase from the bacterium Escherichia coli enabled the understanding of the mechanisms through which RNA polymerase initiates and terminates transcription, and how those processes are regulated to regulate gene expression (i.e. turn genes on and off). Following the isolation of E. coli RNA polymerase, the three RNA polymerases of the eukaryotic nucleus were identified, as well as those associated with viruses and organelles. Studies of transcription also led to the identification of many protein factors that influence transcription, including repressors, activators and enhancers. The availability of purified preparations of RNA polymerase permitted investigators to develop a wide range of novel methods for studying RNA in the test tube, and led directly to many of the subsequent key discoveries in RNA biology.

1966 - 1975

First complete nucleotide sequence of a biological nucleic acid molecule

Although determining the sequence of proteins was becoming somewhat routine, methods for sequencing of nucleic acids were not available until the mid-1960s. In this seminal work, a specific tRNA was purified in substantial quantities, and then sliced into overlapping fragments using a variety of ribonucleases. Analysis of the detailed nucleotide composition of each fragment provided the information necessary to deduce the sequence of the tRNA. Today, the sequence analysis of much larger nucleic acid molecules is highly-automated and enormously faster.

Evolutionary variation of homologous RNA sequences reveals folding patterns

Additional tRNA molecules were purified and sequenced. The first comparative sequence analysis was done and revealed that the sequences varied through evolution in such a way that all of the tRNAs could fold into very similar secondary structures (two-dimensional structures) and had identical sequences at numerous positions (e.g. CCA at the 3' end). The radial four-arm structure of tRNA molecules is termed the 'cloverleaf structure', and results from the evolution of sequences with common ancestry and common biological function. Since the discovery of the tRNA cloverleaf, comparative analysis of numerous other homologous RNA molecules has led to the identification of common sequences and folding patterns.

First complete genomic nucleotide sequence

The 3569 nucleotide sequence of all of the genes of the RNA bacteriophage MS2 was determined by a large team of researchers over several years, and was reported in a series of scientific papers. These results enabled the analysis of the first complete genome, albeit an extremely tiny one by modern standards. Several surprising features were identified, including genes that partially overlap one another and the first clues that different organisms might have slightly different codon usage patterns.

Reverse transcriptase can copy RNA into DNA

Retroviruses were shown to have a single-stranded RNA genome and to replicate via a DNA intermediate, the reverse of the usual DNA-to-RNA transcription pathway. They encode a RNA-dependent DNA polymerase (reverse transcriptase) that is essential for this process. Some retroviruses can cause diseases, including several that are associated with cancer, and HIV-1 which causes AIDS. Reverse transcriptase has been widely used as an experimental tool for the analysis of RNA molecules in the laboratory, in particular the conversion of RNA molecules into DNA prior to molecular cloning and/or polymerase chain reaction (PCR).

RNA replicons evolve rapidly

Biochemical and genetic analyses showed that the enzyme systems that replicate viral RNA molecules (reverse transcriptases and RNA replicases) lack molecular proofreading (3' to 5' exonuclease) activity, and that RNA sequences do not benefit from extensive repair systems analogous to those that exist for maintaining and repairing DNA sequences. Consequently, RNA genomes appear to be subject to significantly higher mutation rates than DNA genomes. For example, mutations in HIV-1 that lead to the emergence of viral mutants that are insensitive to antiviral drugs are common, and constitute a major clinical challenge.

Ribosomal RNA (rRNA) sequences provide a record of the evolutionary history of all life forms

Analysis of ribosomal RNA sequences from a large number of organisms demonstrated that all extant forms of life on Earth share common structural and sequence features of the ribosomal RNA, reflecting a common ancestry. Mapping the similarities and differences among rRNA molecules from different sources provides clear and quantitative information about the phylogenetic (i.e. evolutionary) relationships among organisms. Analysis of rRNA molecules led to the identification of a third major kingdom of organisms, the archaea, in addition to the prokaryotes and eukaryotes.

Non-encoded nucleotides are added to the ends of RNA molecules

Molecular analysis of mRNA molecules showed that, following transcription, mRNAs have non-DNA-encoded nucleotides added to both their 5' and 3' ends (guanosine caps and poly-A, respectively). Enzymes were also identified that add and maintain the universal CCA sequence on the 3' end of tRNA molecules. These events are among the first discovered examples of RNA processing, a complex series of reactions that are needed to convert RNA primary transcripts into biologically active RNA molecules.

1976 - 1985

Small RNA molecules are abundant in the eukaryotic nucleus

Small nuclear RNA molecules (snRNAs) were identified in the eukaryotic nucleus using immunological studies with autoimmune antibodies, which bind to small nuclear ribonucleoprotein complexes (snRNPs; complexes of the snRNA and protein). Subsequent biochemical, genetic, and phylogenetic studies established that many of these molecules play key roles in essential RNA processing reactions within the nucleus and nucleolus, including RNA splicing, polyadenylation, and the maturation of ribosomal RNAs.

RNA molecules require a specific, complex three-dimensional structure for activity

The detailed three-dimensional structure of tRNA molecules was determined using X-ray crystallography, and revealed highly complex, compact three dimensional structures consisting of tertiary interactions laid upon the basic cloverleaf secondary structure. Key features of tRNA tertiary structure include the coaxial stacking of adjacent helices and non-Watson-Crick interactions among nucleotides within the apical loops. Additional crystallographic studies showed that a wide range of RNA molecules (including ribozymes, riboswitches and ribosomal RNA) also fold into specific structures containing a variety of 3D structural motifs. The ability of RNA molecules to adopt specific tertiary structures is essential for their biological activity, and results from the single-stranded nature of RNA. In many ways, RNA folding is more highly analogous to the folding of proteins rather than to the highly repetitive folded structure of the DNA double helix.

Genes are commonly interrupted by introns that must be removed by RNA splicing

Analysis of mature eukaryotic messenger RNA molecules showed that they are often much smaller than the DNA sequences that encode them. The genes were shown to be discontinuous, composed of sequences that are not present in the final mature RNA (introns), located between sequences that are retained in the mature RNA (exons). Introns were shown to be removed after transcription through a process termed RNA splicing. Splicing of RNA transcripts requires a highly precise and coordinated sequence of molecular events, consisting of (a) definition of boundaries between exons and introns, (b) RNA strand cleavage at exactly those sites, and (c) covalent linking (ligation) of the RNA exons in the correct order. The discovery of discontinuous genes and RNA splicing was entirely unexpected by the community of RNA biologists, and stands as one of the most shocking findings in molecular biology research.

Alternative pre-mRNA splicing generates multiple proteins from a single gene

The great majority of protein-coding genes encoded within the nucleus of metazoan cells contain multiple introns. In many cases, these introns were shown to be processed in more than one pattern, thus generating a family of related mRNAs that differ, for example, by the inclusion or exclusion of particular exons. The end result of alternative splicing is that a single gene can encode a number of different protein isoforms that can exhibit a variety of (usually related) biological functions. Indeed, most of the proteins encoded by the human genome are generated by alternative splicing.

Discovery of catalytic RNA (ribozymes)

An experimental system was developed in which an intron-containing rRNA precursor from the nucleus of the ciliated protozoan Tetrahymena could be spliced *in vitro*. Subsequent biochemical analysis shows that this group I intron was self-splicing; that is, the precursor RNA is capable of carrying out the complete splicing reaction in the absence of proteins. In separate work, the RNA component of the bacterial enzyme ribonuclease P (a ribonucleoprotein complex) was shown to catalyze its tRNA-processing reaction in the absence of proteins. These experiments represented landmarks in RNA biology, since they revealed that RNA could play an active role in cellular processes, by catalyzing specific biochemical reactions. Before these discoveries, it was believed that biological catalysis was solely the realm of protein enzymes.

RNA was likely critical for prebiotic evolution

The discovery of catalytic RNA (ribozymes) showed that RNA could both encode genetic information (like DNA) and catalyze specific biochemical reactions (like protein enzymes). This realization led to the RNA World Hypothesis, a proposal that RNA may have played a critical role in prebiotic evolution at a time before the molecules with more specialized functions (DNA and proteins) came to dominate biological information coding and catalysis. Although it is not possible for us to know the course of prebiotic evolution with any certainty, the presence of functional RNA molecules with common ancestry in all modern-day life forms is a strong argument that RNA was widely present at the time of the last common ancestor.

Introns can be mobile genetic elements

Some self-splicing introns can spread through a population of organisms by "homing", inserting copies of themselves into genes at sites that previously lacked an intron. Because they are self-splicing (that is, they remove themselves at the RNA level from genes into which they have inserted), these sequences represent transposons that are genetically silent, i.e. they do not interfere with the expression of the gene into which they become inserted. These introns can be regarded as examples of selfish DNA. Some mobile introns encode homing endonucleases, enzymes that initiate the homing process by specifically cleaving double-stranded DNA at or near the intron-insertion site of alleles lacking an intron. Mobile introns are frequently members of either the group I or group II families of self-splicing introns.

Spliceosomes mediate nuclear pre-mRNA splicing

Introns are removed from nuclear pre-mRNAs by splicesomes, large ribonucleoprotein complexes made up of snRNA and protein molecules whose composition and molecular interactions change during the course of the RNA splicing reactions. Spliceosomes assemble on and around splice sites (the boundaries between introns and exons in the unspliced pre-mRNA) in mRNA precursors and use RNA-RNA interactions to identify critical nucleotide sequences and, probably, to catalyze the splicing reactions. Nuclear pre-mRNA introns and spliceosome-associated snRNAs show similar structural features to self-splicing group II introns. In addition, the splicing pathway of nuclear pre-mRNA introns and group II introns shares a similar reaction pathway. These similarities have led to the hypothesis that these molecules may share a common ancestor.

1986 - 2000

RNA sequences can be edited within cells

Messenger RNA precursors from a wide range of organisms can be edited before being translated into protein. In this process, non-encoded nucleotides may be inserted into specific sites in the RNA, and encoded nucleotides may be removed or replaced. RNA editing was first discovered within the mitochondria of kinetoplastid protozoans, where it has been shown to be extensive. For example, some protein-coding genes encode fewer than 50% of the nucleotides found within the mature, translated mRNA. Other RNA editing events are found in mammals, plants, bacteria and viruses. These latter editing events involve fewer nucleotide modifications, insertions and deletions than the events within kinetoplast DNA, but still have high biological significance for gene expression and its regulation.

Telomerase uses a built-in RNA template to maintain chromosome ends

Telomerase is an enzyme that is present in all eukaryotic nuclei which serves to maintain the ends of the linear DNA in the linear chromosomes of the eukaryotic nucleus, through the addition of terminal sequences that are lost in each round of DNA replication. Before telomerase was identified, its activity was predicted on the basis of a molecular understanding of DNA replication, which indicated that the DNA polymerases known at that time could not replicate the 3' end of a linear chromosome, due to the absence of a template strand. Telomerase was shown to be a ribonucleoprotein enzyme that contains an RNA component that serves as a template strand, and a protein component that has reverse transcriptase activity and adds nucleotides to the chromosome ends using the internal RNA template.

Ribosomal RNA catalyzes peptide bond formation

For years, scientists had worked to identify which protein(s) within the ribosome were responsible for peptidyl transferase function during translation, because the covalent linking of amino acids represents one of the most central chemical reactions in all of biology. Careful biochemical studies showed that extensively-deproteinized large ribosomal subunits could still catalyze peptide bond formation, thereby implying that the sought-after activity might lie within ribosomal RNA rather than ribosomal proteins. Structural biologists, using X-ray crystallography, localized the peptidyl transferase center of the ribosome to a highly-conserved region of the large subunit ribosomal RNA (rRNA) that is located at the place within the ribosome where the amino-acid-bearing ends of tRNA bind, and where no proteins are present. These studies led to the conclusion that the ribosome is a ribozyme. The rRNA sequences that make up the ribosomal active site represent some of the most highly conserved sequences in the biological world. Together, these observations indicate that peptide bond formation catalyzed by RNA was a feature of the last common ancestor of all known forms of life.

Combinatorial selection of RNA molecules enables in vitro evolution

Experimental methods were invented that allowed investigators to use large, diverse populations of RNA molecules to carry out in vitro molecular experiments that utilized powerful selective replication strategies used by geneticists, and which amount to evolution in the test tube. These experiments have been described using different names, the most common of which are "combinatorial selection", "in vitro selection", and SELEX (for Systematic Evolution of Ligands by Exponential Enrichment). These experiments have been used for isolating RNA molecules with a wide range of properties, from binding to particular proteins, to catalyzing particular reactions, to binding low molecular weight organic ligands. They have equal applicability to elucidating interactions and mechanisms that are known properties of naturally-occurring RNA molecules to isolating RNA molecules with biochemical properties that are not known in nature. In developing in vitro selection technology for RNA, laboratory systems for synthesizing complex populations of RNA molecules were established, and used in conjunction with the selection of molecules with user-specified biochemical activities, and in vitro schemes for RNA replication. These steps can be viewed as (a) mutation, (b) selection, and (c) replication. Together, then, these three processes enable in vitro molecular evolution.

2001 - present

Many mobile DNA elements use an RNA intermediate

Transposable genetic elements (transposons) are found which can replicate via transcription into an RNA intermediate which is subsequently converted to DNA by reverse transcriptase. These sequences, many of which are likely related to retroviruses, constitute much of the DNA of the eukaryotic nucleus, especially so in plants. Genomic sequencing shows that retrotransposons make up 36% of the human genome and over half of the genome of major cereal crops (wheat and maize).

Riboswitches bind cellular metabolites and control gene expression

Segments of RNA, typically embedded within the 5'-untranslated region of a vast number of bacterial mRNA molecules, have a profound effect on gene expression through a previously-undiscovered mechanism that does not involve the participation of proteins. In many cases, riboswitches change their folded structure in response to environmental conditions (e.g. ambient temperature or concentrations of specific metabolites), and the structural change controls the translation or stability of the mRNA in which the riboswitch is embedded. In this way, gene expression can be dramatically regulated at the post-transcriptional level.

Small RNA molecules regulate gene expression by post-transcriptional gene silencing

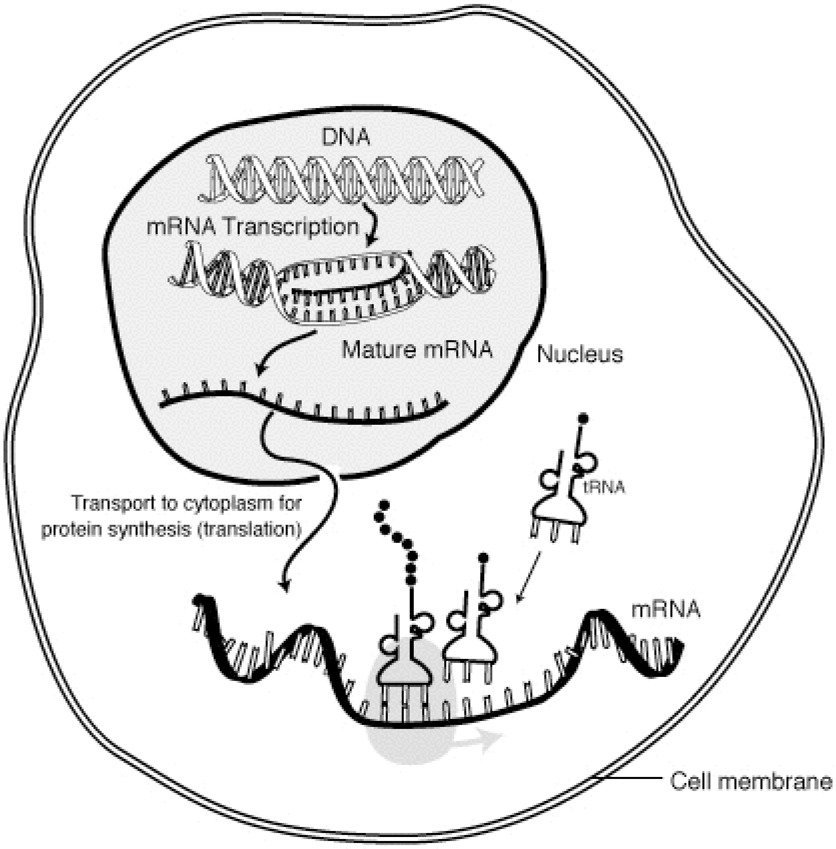
Another previously unknown mechanism by which RNA molecules are involved in genetic regulation was discovered in the 1990s. Small RNA molecules termed microRNA (miRNA) and small interfering RNA (siRNA) are abundant in eukaryotic cells and exert post-transcriptional control over mRNA expression. They function by binding to specific sites within the mRNA and inducing cleavage of the mRNA via a specific silencing-associated RNA degradation pathway.

Noncoding RNA controls epigenetic phenomena

In addition to their well-established roles in translation and splicing, members of noncoding RNA (ncRNA) families have recently been found to function in genome defense and chromosome inactivation. For example, piwi-interacting RNAs (piRNAs) prevent genome instability in germ line cells, while Xist (X-inactive-specific-transcript) is essential for X-chromosome inactivation in mammals.

Chapter- 3

Messenger RNA



The "life cycle" of an **mRNA** in a eukaryotic cell. RNA is transcribed in the nucleus; processed, it is transported to the cytoplasm and translated by the ribosome. At the end of its life, the mRNA is degraded.

Messenger RNA (**mRNA**) is a molecule of RNA encoding a chemical "blueprint" for a protein product. mRNA is transcribed from a DNA template, and carries coding information to the sites of protein synthesis: the ribosomes. Here, the nucleic acid polymer is translated into a polymer of amino acids: a protein. In mRNA as in DNA, genetic information is encoded in the sequence of nucleotides arranged into codons consisting of three bases each. Each codon encodes for a specific amino acid, except the stop codons that terminate protein synthesis. This process requires two other types of RNA: transfer RNA (tRNA) mediates recognition of the codon and provides the corresponding amino acid, while ribosomal RNA (rRNA) is the central component of the ribosome's protein manufacturing machinery.

Synthesis, processing, and function

The brief existence of an mRNA molecule begins with transcription and ultimately ends in degradation. During its life, an mRNA molecule may also be processed, edited, and transported prior to translation. Eukaryotic mRNA molecules often require extensive processing and transport, while prokaryotic molecules do not.

Transcription

During transcription, RNA polymerase makes a copy of a gene from the DNA to mRNA as needed. This process is similar in eukaryotes and prokaryotes. One notable difference, however, is that prokaryotic RNA polymerase associates with mRNA processing enzymes during transcription so that processing can proceed quickly after the start of transcription. The short-lived, unprocessed or partially processed, product is termed *pre-mRNA*; once completely processed, it is termed *mature mRNA*.

Eukaryotic pre-mRNA processing

Processing of mRNA differs greatly among eukaryotes, bacteria and archea. Non-eukaryotic mRNA is essentially mature upon transcription and requires no processing, except in rare cases. Eukaryotic pre-mRNA, however, requires extensive processing.

5' cap addition

A *5′ cap* (also termed an RNA cap, an RNA 7-methylguanosine cap or an RNA m^7G cap) is a modified guanine nucleotide that has been added to the "front" or 5' end of a eukaryotic messenger RNA shortly after the start of transcription. The 5' cap consists of a terminal 7-methylguanosine residue which is linked through a 5'-5'-triphosphate bond to the first transcribed nucleotide. Its presence is critical for recognition by the ribosome and protection from RNases.

Cap addition is coupled to transcription, and occurs co-transcriptionally, such that each influences the other. Shortly after the start of transcription, the 5' end of the mRNA being synthesized is bound by a cap-synthesizing complex associated with RNA polymerase. This enzymatic complex catalyzes the chemical reactions that are required for mRNA capping. Synthesis proceeds as a multi-step biochemical reaction.

Splicing

Splicing is the process by which pre-mRNA is modified to remove certain stretches of non-coding sequences called introns; the stretches that remain include protein-coding sequences and are called exons. Sometimes pre-mRNA messages may be spliced in several different ways, allowing a single gene to encode multiple proteins. This process is called alternative splicing. Splicing is usually performed by an RNA-protein complex called the spliceosome, but some RNA molecules are also capable of catalyzing their own splicing.

Editing

In some instances, an mRNA will be edited, changing the nucleotide composition of that mRNA. An example in humans is the apolipoprotein B mRNA, which is edited in some tissues, but not others. The editing creates an early stop codon, which upon translation, produces a shorter protein.

Polyadenylation

Polyadenylation is the covalent linkage of a polyadenylyl moiety to a messenger RNA molecule. In eukaryotic organisms, most messenger RNA (mRNA) molecules are polyadenylated at the 3' end. The poly(A) tail and the protein bound to it aid in protecting mRNA from degradation by exonucleases. Polyadenylation is also important for transcription termination, export of the mRNA from the nucleus, and translation. mRNA can also be polyadenylated in prokaryotic organisms, where poly(A) tails act to facilitate, rather than impede, exonucleolytic degradation.

Polyadenylation occurs during and immediately after transcription of DNA into RNA. After transcription has been terminated, the mRNA chain is cleaved through the action of an endonuclease complex associated with RNA polymerase. After the mRNA has been cleaved, around 250 adenosine residues are added to the free 3' end at the cleavage site. This reaction is catalyzed by polyadenylate polymerase. Just as in alternative splicing, there can be more than one polyadenylation variant of a mRNA.

Transport

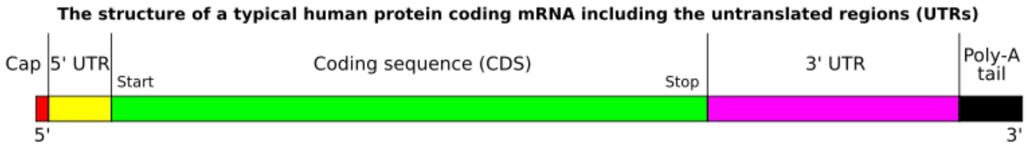
Another difference between eukaryotes and prokaryotes is mRNA transport. Because eukaryotic transcription and translation is compartmentally separated, eukaryotic mRNAs must be exported from the nucleus to the cytoplasm. Mature mRNAs are recognized by their processed modifications and then exported through the nuclear pore. In neurons mRNA must be transported from the soma to the dendrites where local translation occurs in response to external stimuli. Many messages are marked with so-called "zip codes" which targets their transport to a specific location.

Translation

Because prokaryotic mRNA does not need to be processed or transported, translation by the ribosome can begin immediately after the end of transcription. Therefore, it can be said that prokaryotic translation is *coupled* to transcription and occurs *co-transcriptionally*.

Eukaryotic mRNA that has been processed and transported to the cytoplasm (i.e. mature mRNA) can then be translated by the ribosome. Translation may occur at ribosomes free-floating in the cytoplasm, or directed to the endoplasmic reticulum by the signal recognition particle. Therefore, unlike in prokaryotes, eukaryotic translation *is not* directly coupled to transcription.

Structure


The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
5'
3'

The structure of a mature eukaryotic mRNA. A fully processed mRNA includes a 5' cap, 5' UTR, coding region, 3' UTR, and poly(A) tail.

5' cap

The *5′ cap* is a modified guanine nucleotide added to the "front" (5' end) of the pre-mRNA using a 5'-5'-triphosphate linkage. This modification is critical for recognition and proper attachment of mRNA to the ribosome, as well as protection from 5' exonucleases. It may also be important for other essential processes, such as splicing and transport.

Coding regions

Coding regions are composed of codons, which are decoded and translated (in eukaryotes usually into one and in prokaryotes usually into several) proteins by the ribosome. Coding regions begin with the start codon and end with a stop codon. Generally, the start codon is an AUG triplet and the stop codon is UAA, UAG, or UGA. The coding regions tend to be stabilised by internal base pairs, this impedes degradation. In addition to being protein-coding, portions of coding regions may serve as regulatory sequences in the pre-mRNA as exonic splicing enhancers or exonic splicing silencers.

Untranslated regions

Untranslated regions (UTRs) are sections of the mRNA before the start codon and after the stop codon that are not translated, termed the five prime untranslated region (5' UTR) and three prime untranslated region (3' UTR), respectively. These regions are transcribed with the coding region and thus are exonic as they are present in the mature mRNA. Several roles in gene expression have been attributed to the untranslated regions, including mRNA stability, mRNA localization, and translational efficiency. The ability of a UTR to perform these functions depends on the sequence of the UTR and can differ between mRNAs.

The stability of mRNAs may be controlled by the 5' UTR and/or 3' UTR due to varying affinity for RNA degrading enzymes called ribonucleases and for ancillary proteins that can promote or inhibit RNA degradation.

Translational efficiency, including sometimes the complete inhibition of translation, can be controlled by UTRs. Proteins that bind to either the 3' or 5' UTR may affect translation by influencing the ribosome's ability to bind to the mRNA. MicroRNAs bound to the 3' UTR also may affect translational efficiency or mRNA stability.

Cytoplasmic localization of mRNA is thought to be a function of the 3' UTR. Proteins that are needed in a particular region of the cell can actually be translated there; in such a case, the 3' UTR may contain sequences that allow the transcript to be localized to this region for translation.

Some of the elements contained in untranslated regions form a characteristic secondary structure when transcribed into RNA. These structural mRNA elements are involved in regulating the mRNA. Some, such as the SECIS element, are targets for proteins to bind. One class of mRNA element, the riboswitches, directly bind small molecules, changing their fold to modify levels of transcription or translation. In these cases, the mRNA regulates itself.

Poly(A) tail

The 3' poly(A) tail is a long sequence of adenine nucleotides (often several hundred) added to the 3' end of the pre-mRNA. This tail promotes export from the nucleus and translation, and protects the mRNA from degradation.

Monocistronic versus polycistronic mRNA

An mRNA molecule is said to be monocistronic when it contains the genetic information to translate only a single protein. This is the case for most of the eukaryotic mRNAs. On the other hand, polycistronic mRNA carries the information of several genes, which are translated into several proteins. These proteins usually have a related function and are grouped and regulated together in an operon. Most of the mRNA found in bacteria and archea are polycistronic. Dicistronic or bicistronic is the term used to describe an mRNA that encodes only two proteins.

mRNA circularization

In eukaryotes it is thought that mRNA molecules form circular structures due to an interaction between the cap binding complex and poly(A)-binding protein. Circularization is thought to promote recycling of ribosomes on the same message leading to efficient translation.

Degradation

Different mRNAs within the same cell have distinct lifetimes (stabilities). In bacterial cells, individual mRNAs can survive from seconds to more than an hour; in mammalian cells, mRNA lifetimes range from several minutes to days. The greater the stability of an mRNA, the more protein may be produced from that mRNA. The limited lifetime of mRNA enables a cell to alter protein synthesis rapidly in response to its changing needs. There are many mechanisms that lead to the destruction of a mRNA, some of which are described below.

Prokaryotic mRNA degradation

In prokaryotes the lifetime of mRNA is generally much shorter than in eukaryotes. Prokaryotes degrade messages by using a combination of ribonucleases, including endonucleases, 3' exonucleases, and 5' exonucleases. In some instances, small RNA molecules (sRNA) tens to hundreds of nucleotides long can stimulate the degradation of specific mRNAs by base pairing with complementary sequences and facilitating ribonuclease cleavage. It was recently shown that bacteria also have a sort of 5' cap consisting of a triphosphate on the 5' end. Removal of two of the phosphates leaves a 5' monophosphate, causing the message to be destroyed by the endonuclease RNase E.

Eukaryotic mRNA turnover

Inside eukaryotic cells there is a balance between the processes of translation and mRNA decay. Messages that are being actively translated are bound by ribosomes, the eukaryotic initiation factors eIF-4E and eIF-4G, and poly(A)-binding protein. eIF-4E and eIF-4G block the decapping enzyme (DCP2), and poly(A)-binding protein blocks the exosome complex, protecting the ends of the message. The balance between translation and decay is reflected in the size and abundance of cytoplasmic structures known as P-bodies The poly(A) tail of the mRNA is shortened by specialized exonucleases that are targeted to specific messenger RNAs by a combination of cis-regulatory sequences on the RNA and trans-acting RNA-binding proteins. Poly(A) tail removal is thought to disrupt the circular structure of the message and destabilize the cap binding complex. The message is then subject to degradation by either the exosome complex or the decapping complex. In this way, translationally inactive messages can be destroyed quickly, while active messages remain intact. The mechanism by which translation stops and the message is handed-off to decay complexes is not understood in detail.

AU-rich element decay

The presence of AU-rich elements in some mammalian mRNAs tends to destabilize those transcripts through the action of cellular proteins that bind these sequences and stimulate poly(A) tail removal. Loss of the poly(A) tail is thought to promote mRNA degradation by facilitating attack by both the exosome complex and the decapping complex. Rapid mRNA degradation via AU-rich elements is a critical mechanism for preventing the overproduction of potent cytokines such as tumor necrosis factor (TNF) and granulocytemacrophage colony stimulating factor (GM-CSF). AU-rich elements also regulate the biosynthesis of proto-oncogenic transcription factors like c-Jun and c-Fos.

Nonsense mediated decay

Eukaryotic messages are subject to surveillance by nonsense mediated decay (NMD), which checks for the presence of premature stop codons (nonsense codons) in the message. These can arise via incomplete splicing, V(D)J recombination in the adaptive immune system, mutations in DNA, transcription errors, leaky scanning by the ribosome causing a frame shift, and other causes. Detection of a premature stop codon triggers mRNA degradation by 5' decapping, 3' poly(A) tail removal, or endonucleolytic cleavage.

Small interfering RNA (siRNA)

In metazoans, small interfering RNAs (siRNAs) processed by Dicer are incorporated into a complex known as the RNA-induced silencing complex or RISC. This complex contains an endonuclease that cleaves perfectly complementary messages to which the siRNA binds. The resulting mRNA fragments are then destroyed by exonucleases. siRNA is commonly used in laboratories to block the function of genes in cell culture. It is thought to be part of the innate immune system as a defense against double-stranded RNA viruses.

MicroRNA (miRNA)

MicroRNAs (miRNAs) are small RNAs that typically are partially complementary to sequences in metazoan messenger RNAs. Binding of a miRNA to a message can repress translation of that message and accelerate poly(A) tail removal, thereby hastening mRNA degradation. The mechanism of action of miRNAs is the subject of active research.

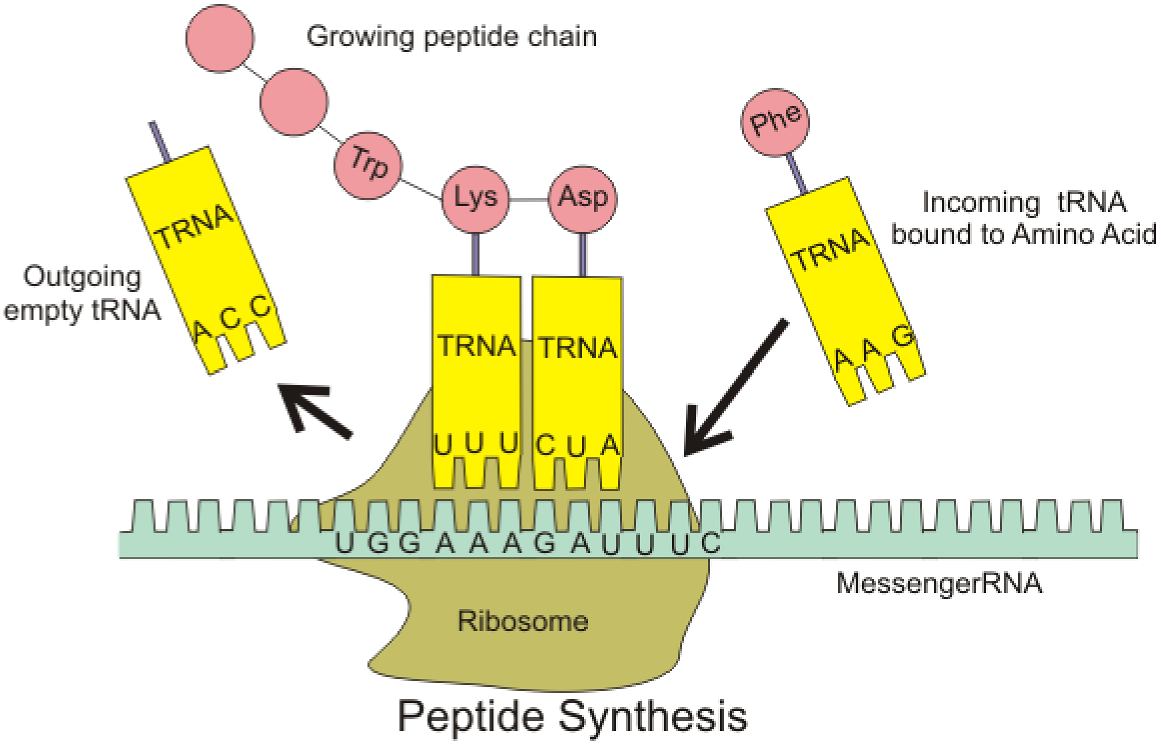
Other decay mechanisms

There are other ways in which messages can be degraded, including non-stop decay, silencing by Piwi-interacting RNA (piRNA), and surely other means.

Chapter- 4

Transfer RNA and Ribosomal RNA

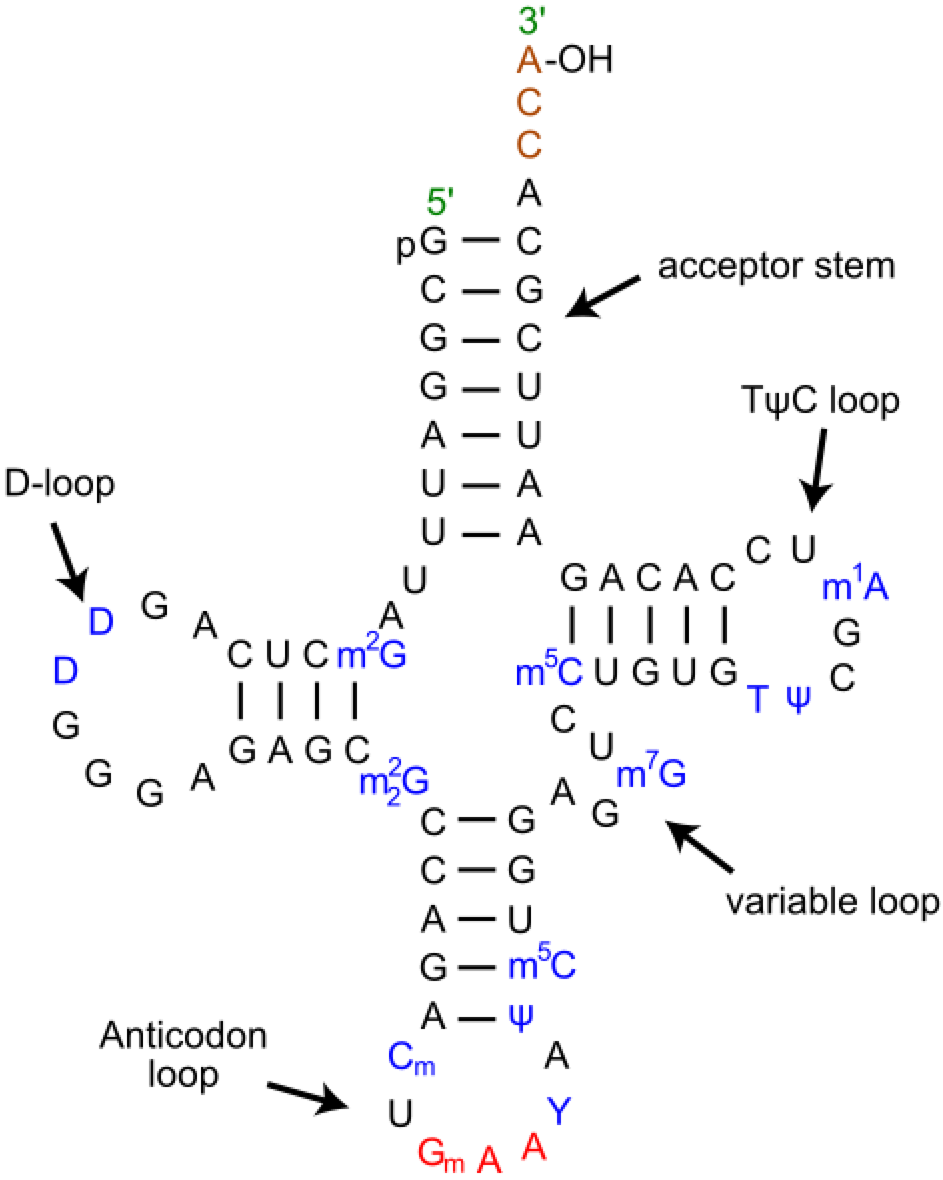
Transfer RNA



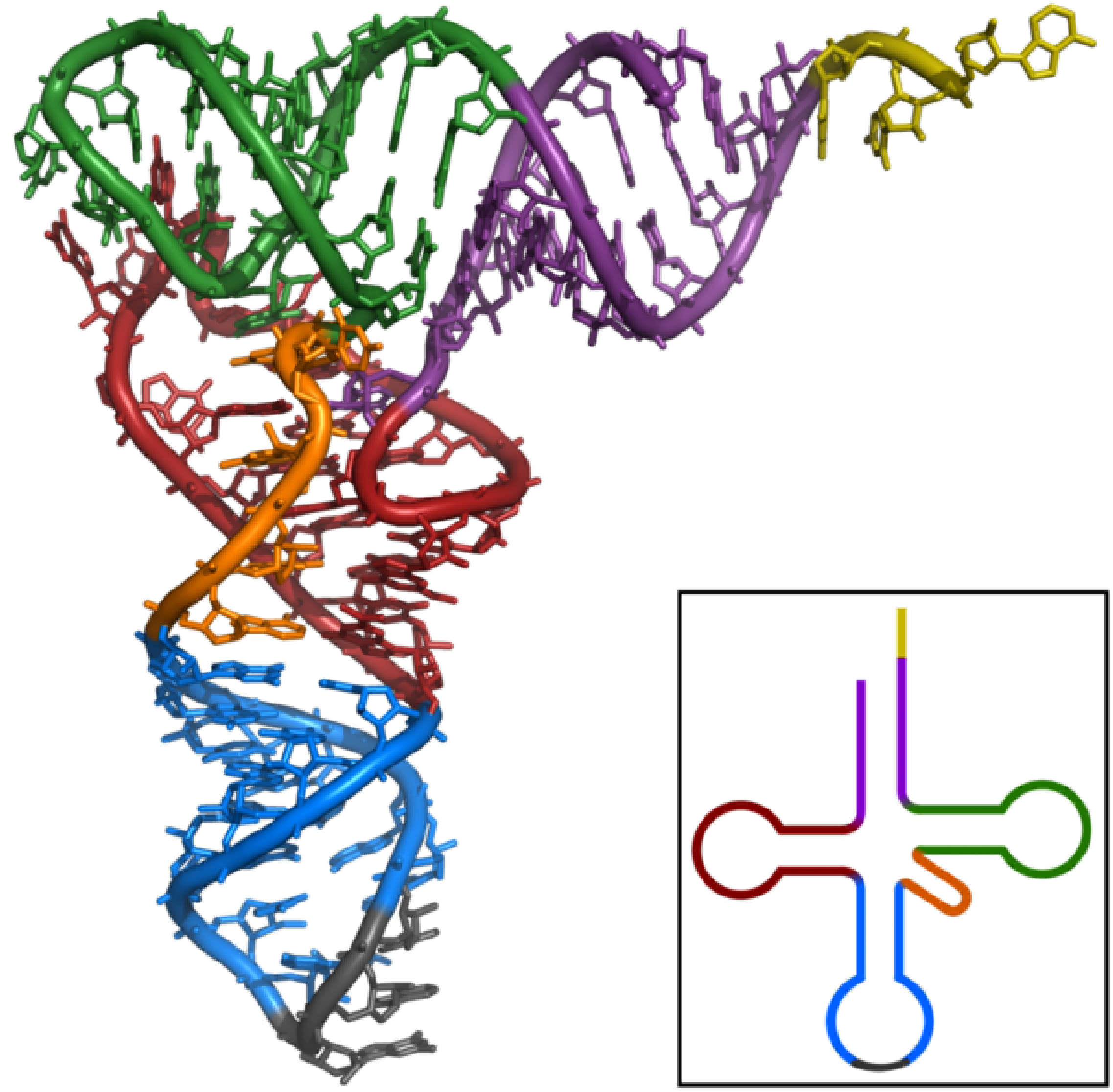
The interaction of tRNA and mRNA in protein synthesis

Transfer RNA (**tRNA**) is a small RNA molecule (usually about 73-95 nucleotides) that transfers a specific active amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. It has a 3' terminal site for amino acid attachment. This covalent linkage is catalyzed by an aminoacyl tRNA synthetase. It also contains a three base region called the anticodon that can base pair to the corresponding three base codon region on mRNA. Each type of tRNA molecule can be attached to only one type of amino acid, but because the genetic code contains multiple codons that specify the same amino acid, tRNA molecules bearing different anticodons may also carry the same amino acid.

Structure



Secondary *cloverleaf structure* of tRNA^{Phe} from yeast



Tertiary structure of tRNA. *CCA tail* in orange, *Acceptor stem* in purple, *D arm* in red, *Anticodon arm* in blue with *Anticodon* in black, *T arm* in green.

The structure of tRNA can be decomposed into its primary structure, its secondary structure (usually visualized as the *cloverleaf structure*), and its tertiary structure (all tRNAs have a similar L-shaped 3D structure that allows them to fit into the P and A sites of the ribosome). The cloverleaf structure becomes the 3D L-shaped structure through coaxial stacking of the helices which is a common RNA Tertiary Structure motif.

1. The 5'-terminal phosphate group.
2. The acceptor stem is a 7-base pair (bp) stem made by the base pairing of the 5'-terminal nucleotide with the 3'-terminal nucleotide (which contains the CCA 3'-terminal group used to attach the amino acid). The acceptor stem may contain non-Watson-Crick base pairs.

3. The CCA tail is a cytosine-cytosine-adenine sequence at the 3' end of the tRNA molecule. This sequence is important for the recognition of tRNA by enzymes critical in translation. In prokaryotes, the CCA sequence is transcribed in some tRNA sequences. In most prokaryotic tRNAs and eukaryotic tRNAs, the CCA sequence is added during processing and therefore does not appear in the tRNA gene.
4. The D arm is a 4 bp stem ending in a loop that often contains dihydrouridine.
5. The anticodon arm is a 5-bp stem whose loop contains the anticodon.
6. The T arm is a 5 bp stem containing the sequence TΨC where Ψ is a pseudouridine.
7. Bases that have been modified, especially by methylation, occur in several positions throughout the tRNA. The first anticodon base, or wobble-position, is sometimes modified to inosine (derived from adenine), pseudouridine (derived from uracil) or lysidine (derived from cytosine).

Anticodon

An **anticodon** is a unit made up of three nucleotides that correspond to the three bases of the codon on the mRNA. Each tRNA contains a specific anticodon triplet sequence that can base-pair to one or more codons for an amino acid. For example, the codon for lysine is AAA; the anticodon of a lysine tRNA might be UUU. Some anticodons can pair with more than one codon due to a phenomenon known as wobble base pairing. Frequently, the first nucleotide of the anticodon is one of two not found on mRNA: inosine and pseudouridine, which can hydrogen bond to more than one base in the corresponding codon position. In the genetic code, it is common for a single amino acid to be specified by all four third-position possibilities, or at least by both Pyrimidines and Purines; for example, the amino acid glycine is coded for by the codon sequences GGU, GGC, GGA, and GGG.

To provide a one-to-one correspondence between tRNA molecules and codons that specify amino acids, 61 types of tRNA molecules would be required per cell. However, many cells contain fewer than 61 types of tRNAs because the wobble base is capable of binding to several, though not necessarily all, of the codons that specify a particular amino acid. A minimum of 31 tRNA are required to translate, unambiguously, all 61 sense codons of the standard genetic code.

Aminoacylation

Aminoacylation is the process of adding an aminoacyl group to a compound. It produces tRNA molecules with their CCA 3' ends covalently linked to an amino acid.

Each tRNA is aminoacylated (or *charged*) with a specific amino acid by an aminoacyl tRNA synthetase. There is normally a single aminoacyl tRNA synthetase for each amino acid, despite the fact that there can be more than one tRNA, and more than one anticodon, for an amino acid. Recognition of the appropriate tRNA by the synthetases is not mediated solely by the anticodon, and the acceptor stem often plays a prominent role.

Reaction:

1. amino acid + ATP → aminoacyl-AMP + PPi
2. aminoacyl-AMP + tRNA → aminoacyl-tRNA + AMP

Sometimes, certain organisms can have one or more aminoacyl tRNA synthetases missing. This leads to mischarging of the tRNA by a chemically related amino acid. The correct amino acid is made by enzymes that modify the mischarged amino acid to the correct one.

For example, *Helicobacter pylori* has glutaminyl tRNA synthetase missing. Thus, glutamate tRNA synthetase mischarges tRNA-glutamine(tRNA-Gln) with glutamate. An amidotransferase then converts the acid side chain of the glutamate to the amide, forming the correctly charged gln-tRNA-Gln.

Binding to ribosome

The ribosome has three binding sites for tRNA molecules: the A (aminoacyl), P (peptidyl), and E (exit) sites. During translation the A site binds an incoming aminoacyl-tRNA as directed by the codon currently occupying this site. This codon specifies the next amino acid to be added to the growing peptide chain. The A site only works after the first aminoacyl-tRNA has attached to the P site. The P-site codon is occupied by peptidyl-tRNA that is a tRNA with multiple amino acids attached as a long chain. The P site is actually the first to bind to aminoacyl tRNA. This tRNA in the P site carries the chain of amino acids that has already been synthesized. The E site is occupied by the empty tRNA as it's about to exit the ribosome.

tRNA genes

Organisms vary in the number of tRNA genes in their genome. The nematode worm *C. elegans*, a commonly used model organism in genetics studies, has 29,647 genes in its nuclear genome, of which 620 code for tRNA. The budding yeast *Saccharomyces cerevisiae* has 275 tRNA genes in its genome. In the human genome, which according to current estimates has about 27,161 genes in total, there are about 4,421 non-coding RNA genes, which include tRNA genes. There are 22 mitochondrial tRNA genes; 497 nuclear genes encoding cytoplasmic tRNA molecules and there are 324 tRNA-derived putative pseudogenes.

Cytoplasmic tRNA genes can be grouped into 49 families according to their anticodon features. These genes are found on all chromosomes, except 22 and Y chromosome. High clustering on 6p is observed (140 tRNA genes), as well on 1 chromosome.

tRNA biogenesis

In eukaryotic cells, tRNAs are transcribed by RNA polymerase III as pre-tRNAs in the nucleus. RNA polymerase III recognizes two internal promoter sequences (A-box B internal promoter) inside tRNA genes. The first promoter begins at nucleotide 8 of mature tRNAs and the second promoter is located 30-60 nucleotides downstream of the first promoter. The transcription terminates after a strech of four or more thymidines.

Pre-tRNAs undergo extensive modifications inside the nucleus. Some pre-tRNAs contain introns; in bacteria these self-splice, whereas in eukaryotes and archaea they are removed by tRNA splicing endonuclease. The 5' sequence is removed by RNase P, whereas the 3' end is removed by the tRNase Z enzyme. A notable exception is in the archaeon *Nanoarchaeum equitans* which does not possess an RNase P enzyme and has a promoter placed such that transcription starts at the 5' end of the mature tRNA.. The non-templated 3' CCA tail is added by a nucleotidyl transferase. Before tRNAs are exported into the cytoplasm by Los1/Xpo-t, tRNAs are aminoacylated. The order of the processing events is not conserved. For example in yeast, the splicing is not carried out in the nucleus but at the cytoplasmic side of mitochondrial membranes.

History

The existence of tRNA was first hypothesized by Francis Crick, based on the assumption that there must exist an adapter molecule capable of mediating the translation of the RNA alphabet into the protein alphabet. Significant research on structure was conducted in the early 1960s by Alex Rich and Don Caspar, two researchers in Boston, the Jacques Fresco group in Princeton University and a United Kingdom group at King's College London. In 1965, a publication by Robert W. Holley reported the primary structure and suggested three secondary structures. The cloverleaf structure was ascertained by several other studies in the following years and was finally confirmed using X-ray crystallography studies in 1974. Two independent groups, Kim Sung-Hou working under Alexander Rich and a British group headed by Aaron Klug, published the same crystallography findings within a year.

Ribosomal RNA

Ribosomal ribonucleic acid (**rRNA**) is the RNA component of the ribosome, the protein manufacturing organelle of all living cells. Ribosomal RNA provides a mechanism for decoding mRNA into amino acids and interacts with tRNAs during translation by providing peptidyl transferase activity. The tRNAs bring the necessary amino acids corresponding to the appropriate mRNA codon.

Inside the ribosome

The ribosomal RNAs form two subunits, the large subunit (LSU) and small subunit (SSU). mRNA is sandwiched between the small and large subunits and the ribosome catalyzes the formation of a peptide bond between the 2 amino acids that are contained in the rRNA.

A ribosome also has 3 binding sites called A, P, and E.

- The A site in the ribosome binds to an aminoacyl-tRNA (a tRNA bound to an amino acid).
- The amino (NH_2) group of the aminoacyl-tRNA, which contains the new amino acid, attacks the ester linkage of peptidyl-tRNA (contained within the P site), which contains the last amino acid of the growing chain, forming a new peptide bond. This reaction is catalyzed by peptidyl transferase.
- The tRNA that was holding on the last amino acid is moved to the E site, and what used to be the aminoacyl-tRNA is the peptidyl-tRNA.

A single mRNA can be translated simultaneously by multiple ribosomes.

Prokaryotes vs. Eukaryotes

Both prokaryotic and eukaryotic ribosomes can be broken down into two subunits (the S in 16S represents Svedberg units):

Type	**Size**	**Large subunit**	**Small subunit**
prokaryotic	70S	50S (5S, 23S)	30S (16S)
eukaryotic	80S	60S (5S, 5.8S, 28S)	40S (18S)

Note that the S units of the subunits cannot simply be added because they represent measures of sedimentation rate rather than of mass. The sedimentation rate of each subunit is affected by its shape, as well as by its mass.

Prokaryotes

In prokaryotes a small 30S ribosomal subunit contains the 16S rRNA.

The large 50S ribosomal subunit contains two rRNA species (the 5S and 23S rRNAs).

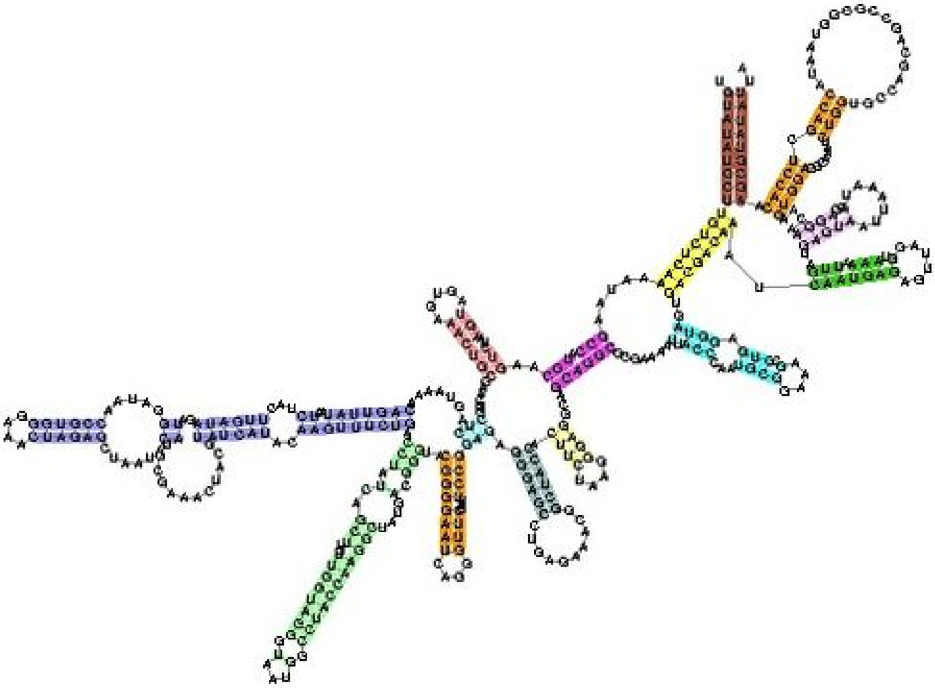
Bacterial 16S, 23S, and 5S rRNA genes are typically organized as a co-transcribed operon.

There may be one or more copies of the operon dispersed in the genome (for example, *Escherichia coli* has seven).

Archaea contains either a single rDNA operon or multiple copies of the operon.

The 3' end of the 16S rRNA (in a ribosome) binds to a sequence on the 5' end of mRNA called the Shine-Dalgarno sequence.

Eukaryotes



Small subunit ribosomal RNA, 5' domain taken from the Rfam database. This example is RF00177

In contrast, eukaryotes generally have many copies of the rRNA genes organized in tandem repeats; in humans approximately 300–400 rDNA repeats are present in five clusters (on chromosomes 13, 14, 15, 21 and 22).

The 18S rRNA in most eukaryotes is in the small ribosomal subunit, and the large subunit contains three rRNA species (the 5S, 5.8S and 28S rRNAs).

Mammalian cells have 2 mitochondrial (12S and 16S) rRNA molecules and 4 types of cytoplasmic rRNA (28S, 5.8S, 5S (large ribosome subunit) and 18S (small subunit)). 28S, 5.8S, and 18S rRNAs are encoded by a single transcription unit (45S) separated by 2 internally transcribed spacers. The 45S rDNA organized into 5 clusters (each has 30-40 repeats) on chromosomes 13, 14, 15, 21, and 22. These are transcribed by RNA polymerase I. 5S occurs in tandem arrays (~200-300 true 5S genes and many dispersed pseudogenes), the largest one on the chromosome 1q41-42. 5S rRNA is transcribed by RNA polymerase III.

The tertiary structure of the small subunit ribosomal RNA (SSU rRNA) has been resolved by X-ray crystallography. The secondary structure of SSU rRNA contains 4 distinct domains — the 5', central, 3' major and 3' minor domains. A model of the secondary structure for the 5' domain (500-800 nucleotides) is shown.

Translation

Translation is the net effect of proteins being synthesized by ribosomes, from a copy (mRNA) of the DNA template in the nucleus. One of the components of the ribosome (16S rRNA) base pairs complementary to a sequence upstream of the start codon in mRNA.

Importance of rRNA

Ribosomal RNA characteristics are important in medicine and in evolution.

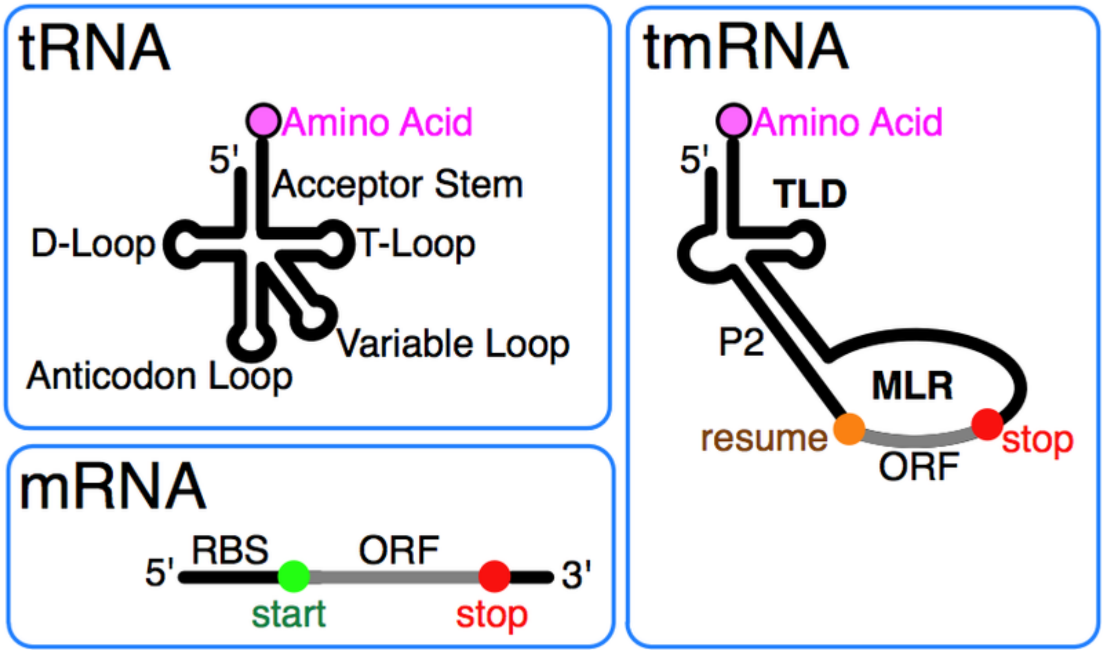
- rRNA is the target of several clinically relevant antibiotics: chloramphenicol, erythromycin, kasugamycin, micrococcin, paromomycin, ricin, sarcin, spectinomycin, streptomycin, and thiostrepton.

- rRNA is the one of the only genes present in all cells. For this reason, genes that encode the rRNA (rDNA) are sequenced to identify an organism's taxonomic group, calculate related groups, and estimate rates of species divergence. For this reason many thousands of rRNA sequences are known and stored in specialized databases such as RDP-II and SILVA.

Chapter- 5

Transfer-Messenger RNA

Transfer-messenger RNA (abbreviated **tmRNA**, also known as **10Sa RNA** and by its genetic name **SsrA**) is a bacterial RNA molecule with dual tRNA-like and messenger RNA-like properties. The tmRNA forms a ribonucleoprotein complex (**tmRNP**) together with Small Protein B (SmpB), Elongation Factor Tu (EF-Tu), and ribosomal protein S1. In *trans*-translation, tmRNA and its associated proteins bind to bacterial ribosomes which have stalled in the middle of protein biosynthesis, for example when reaching the end of a messenger RNA which has lost its stop codon. The tmRNA is remarkably versatile: it recycles the stalled ribosome, adds a proteolysis-inducing tag to the unfinished polypeptide, and facilitates the degradation of the aberrant messenger RNA. In the majority of bacteria these functions are carried out by standard one-piece tmRNAs. In other bacterial species, a permuted *ssrA* gene produces a two-piece tmRNA in which two separate RNA chains are joined by base-pairing.



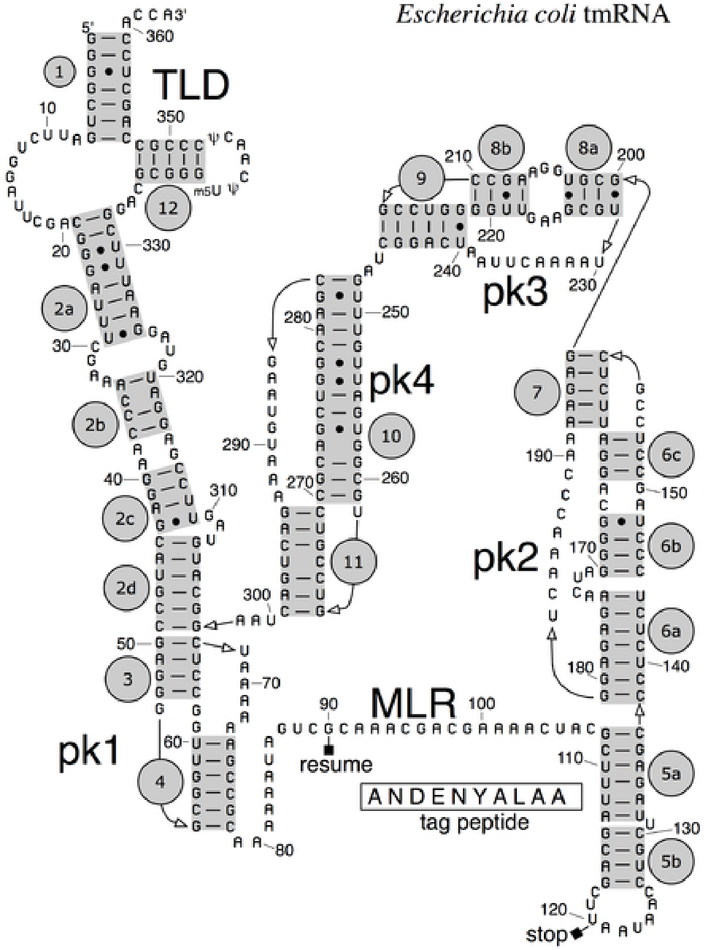
tmRNA combines features of tRNA and mRNA

Discovery of tmRNA and early work

tmRNA was first designated 10Sa RNA after a mixed “10S” electrophoretic fraction of *Escherichia coli* RNA was further resolved into tmRNA and the similarly-sized RNase P RNA (10Sb). The presence of pseudouridine in the mixed 10S RNA hinted that tmRNA has modified bases found also in tRNA. The similarity at the 3' end of tmRNA to the T stem-loop of tRNA was first recognized upon sequencing *ssrA* from *Mycobacterium tuberculosis*. Subsequent sequence comparison revealed the full tRNA-like domain (TLD) formed by the 5' and 3' ends of tmRNA, including the acceptor stem with elements like those in alanine tRNA that promote its aminoacylation by alanine-tRNA ligase. It also revealed differences from tRNA: the anticodon arm is missing in tmRNA, and the D arm region is a loop without base pairs.

tmRNA structure

Secondary structure of the standard one-piece tmRNAs



Secondary structure of *E. coli* tmRNA. Shown are the 5' and 3' ends of the 363-nucleotide RNA chain numbered in increments of ten. Short lines indicate Watson-Crick pairings (G-C and A-U); dots are G-U pairings. Prominent are the tRNA-like domain (TLD), the messenger RNA-like region (MLR), and the four pseudoknots (pk1 to pk4). The MLR encodes the tag peptide between resume and stop codons. RNA helices (numbered one to 12) and their sections (letters) are gray.

The complete *E. coli* tmRNA secondary structure was elucidated by comparative sequence analysis and structural probing. Watson-Crick and G-U base pairs were identified by comparing the bacterial tmRNA sequences using automated computational methods in combination with manual alignment procedures. The accompanying figure shows the base pairing pattern of this prototypical tmRNA, which is organized into 12 phylogenetically supported helices (also called pairings P1 to P12), some divided into helical segments.

A prominent feature of every tmRNA is the conserved tRNA-like domain (TLD), composed of helices 1, 12, and 2a (analogs of the tRNA acceptor stem, T-stem and variable stem, respectively), and containing the 5' monophosphate and alanylatable 3' CCA ends. The mRNA-like region (MLR) is in standard tmRNA a large loop containing pseudoknots and a coding sequence (CDS) for the tag peptide, marked by the resume codon and the stop codon. The encoded tag peptide (ANDENYALAA in *E. coli*) varies among bacteria, perhaps depending on the set of proteases and adaptors available.

tmRNAs typically contain four pseudoknots, one (pk1) upstream of the tag peptide CDS, and the other three pseudoknots (pk2 to pk4) downstream of the CDS. The pseudoknot regions, although generally conserved, are evolutionarily plasic. For example, in the (one-piece) tmRNAs of cyanobacteria, pk4 is substituted with two tandemly arranged smaller pseudoknots. This suggests that tmRNA folding outside the TLD can be important, yet the pseudoknot region lacks conserved residues and pseudoknots are among the first structures to be lost as *ssrA* sequences diverge in plastid and endosymbiont lineages. Base pairing in the three-pseudoknot region of *E. coli* tmRNA is disrupted during *trans*-translation.

Two-piece tmRNAs

Circularly permuted *ssrA* has been reported in three major lineages: i) all alphaproteobacteria and the primitive mitochondria of jakobid protists, ii) two disjoint groups of cyanobacteria (*Gloeobacter* and a clade containing *Prochlorococcus* and many *Synechococcus*), and iii) some members of the betaproteobacteria (*Cupriavidus* and some Rhodocyclales). All produce the same overall two-piece (acceptor and coding pieces) form, equivalent to the standard form nicked downstream of the reading frame. None retain more than two pseudoknots compared to the four (or more) of standard tmRNA.

Alphaproteobacteria have two signature sequences: replacement of the typical T-loop sequence TΨCRANY with GGCRGUA, and the sequence AACAGAA in the large loop of the 3′-terminal pseudoknot. In mitochondria, the MLR has been lost, and a remarkable re-permutation of mitochondrial *ssrA* results in a small one-piece product in *Jakoba libera*.

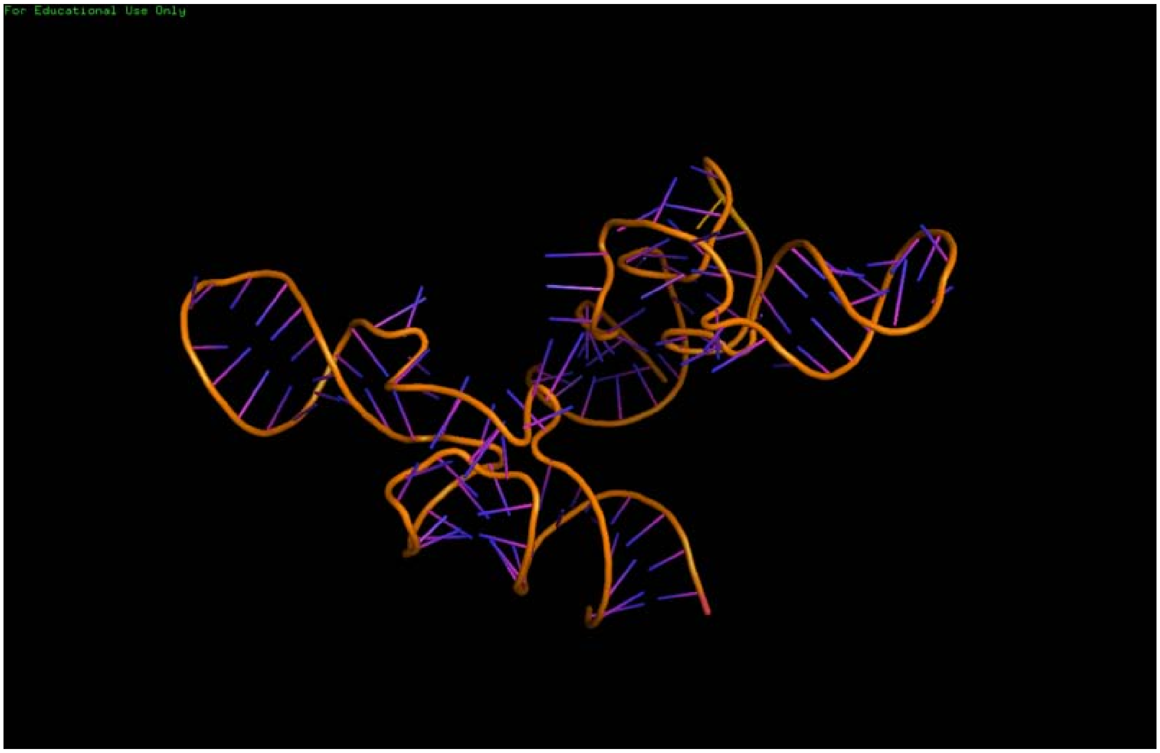
The cyanobacteria provide the most plausible case for evolution of a permuted gene from a standard gene, due to remarkable sequence similarities between the two gene types as they occur in different *Synechococcus* strains.

tmRNA processing

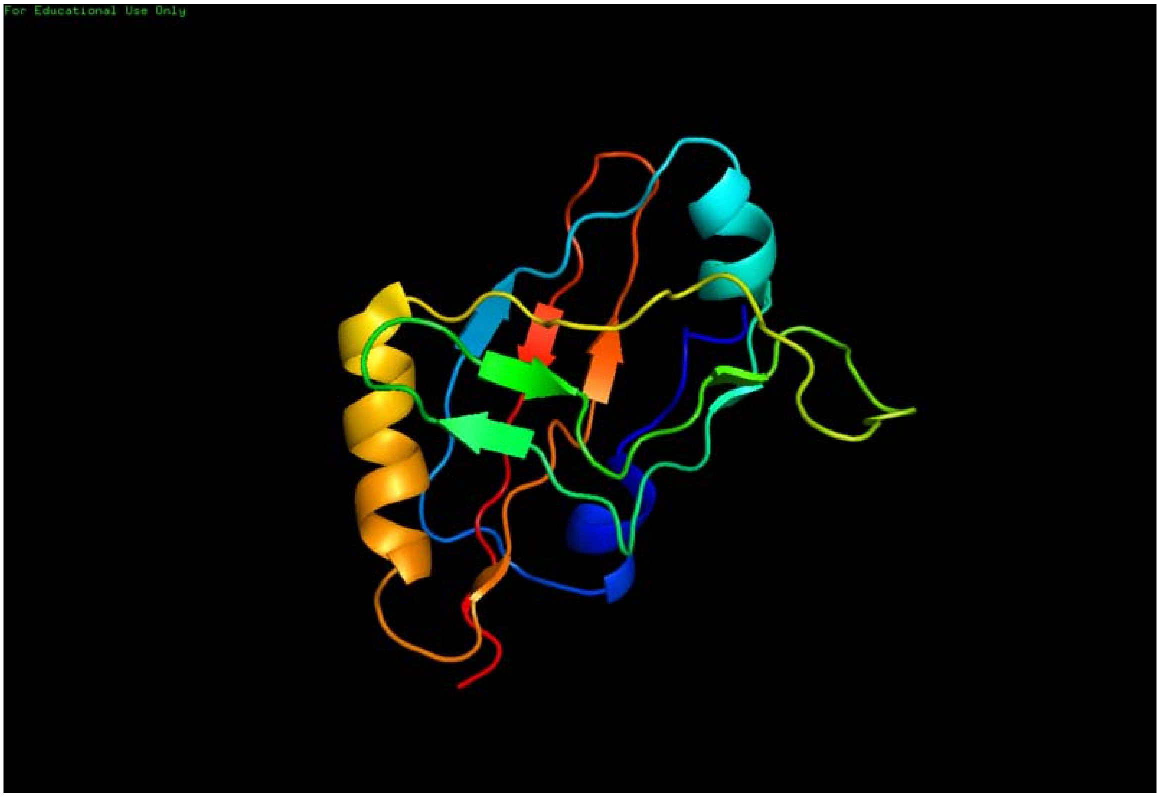
Most tmRNAs are transcribed as larger precursors which are processed much like tRNA. Cleavage at the 5´ end is by ribonuclease P. Multiple exonucleases can participate in the processing of the 3´ end of tmRNA, although RNase T and RNase PH are most effective. Depending on the bacterial species, the 3'-CCA is either encoded or added by tRNA nucleotidyltransferase.

Similar processing at internal sites of permuted precursor tmRNA explains its physical splitting into two pieces. The two-piece tmRNAs have two additional ends whose processing must be considered. For alphaproteobacteria, one 5´ end is the unprocessed start site of transcription. The far 3´ end may in some cases be the result of rho-independent termination.

Three-dimensional structures



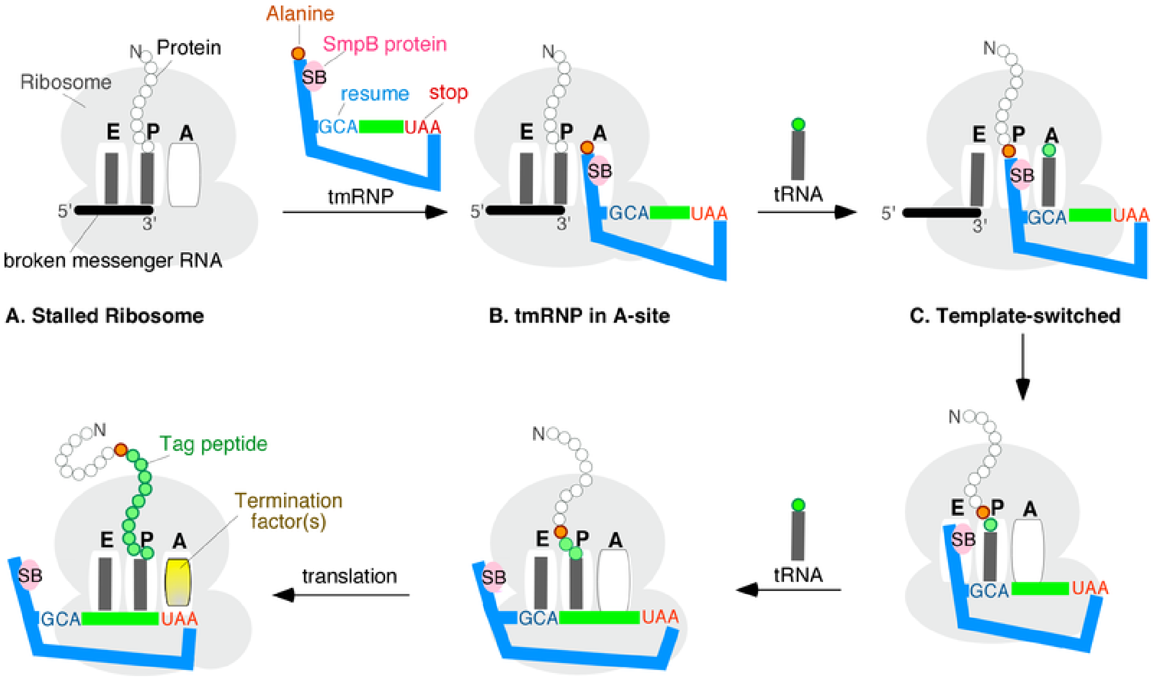
Cartoon ribbon structure of the tRNA-like domain of tmRNA. The domain consists of the 3' and 5' ends of the tmRNA. Image was created using Pymol molecular imaging software for students and data obtained from the RCSB Protein Data Bank file for structure 1J1H



Cartoon ribbon structure of the tmRNA dedicated binding protein, SmpB. Image was created using Pymol molecular imaging software for students and data obtained from the RCSB Protein Data Bank file for structure 1CZJ

High-resolution structures of the complete tmRNA molecules are currently unavailable and may be difficult to obtain due the inherent flexibility of the MLR. In 2007, the crystal structure of the *Thermus thermophilus* TLD bound to the SmpB protein was obtained at 3 Å resolution. This structure shows that SmpB mimics the D stem and the anticodon of a canonical tRNA whereas helical section 2a of tmRNA corresponds to the variable arm of tRNA. A cryo-electron microscopy study of tmRNA at an early stage of *trans*-translation shows the spatial relationship between the ribosome and the tmRNP (tmRNA bound to the EF-Tu protein). The TLD is located near the GTPase-associated center in the 50S ribosomal subunit; helix 5 and pseudoknots pk2 to pk4 form an arc around the beak of the 30S ribosomal subunit.

Trans-*translation*



A. Stalled Ribosome

B. tmRNP in A-site

C. Template-switched

F. Termination

E. Stable

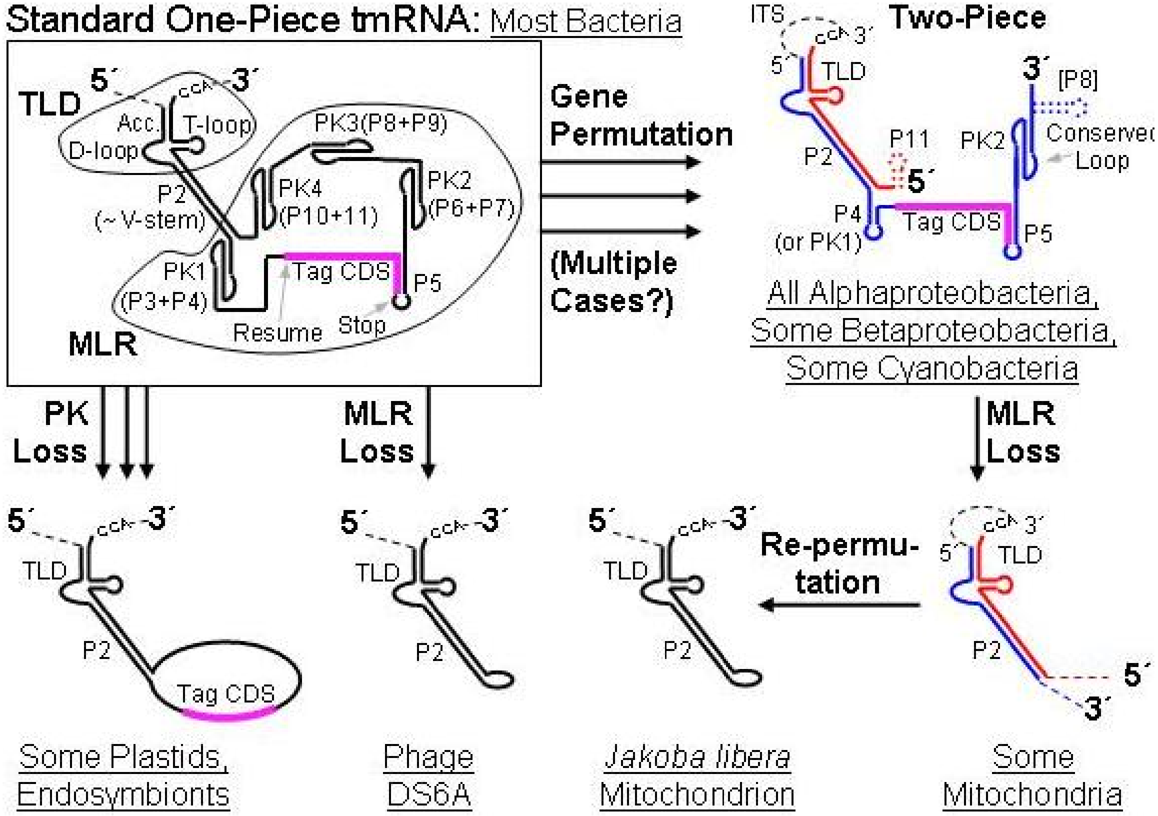
D. Slip-prone

trans-Translation stages **A** through **F**. A ribosome with its RNA binding sites, designated E, P, and A, is stuck near the 3' end of a broken mRNA. The tmRNP binds to the A-site, allowing the ribosome to switch templates from the broken message onto the open reading frame of the tmRNA via the resume codon (blue GCA). Regular translation eventually resumes. Upon reaching the tmRNA stop codon (red UAA), a hybrid protein with a proteolysis tag (green beads) is released.

Coding by tmRNA was discovered in 1995 when Simpson and coworkers overexpressed a mouse cytokine in *E. coli* and found several truncated cytokine-derived peptides each tagged at the carboxyl termini with the same 11-amino acid residue extension (A)ANDENYALAA. With the exception of the N-terminal alanine, which comes from the 3' end of tmRNA itself, this tag sequence was traced to a short open reading frame in *E. coli* tmRNA. Recognizing that the tag peptide confers proteolysis, the *trans*-translation model for tmRNA action was proposed.

While details of the *trans*-translation mechanism are under investigation it is generally agreed that tmRNA first occupies the empty A site of the stalled ribosome. Subsequently, the ribosome moves from the 3' end of the truncated messenger RNA onto the resume codon of the MLR, followed by a slippage-prone stage from where translation continues normally until the in-frame tmRNA stop codon is encountered. *Trans-translation* is essential in some bacterial species, whereas other bacteria require tmRNA to survive when subjected to stressful growth conditions. Depending on the organism, the tag peptide may be recognized by a variety of proteases or protease adapters.

Mobile genetic elements and the tmRNA gene



History of *ssrA*. Precursor RNAs are shown, whose dashed portions are excised during maturation. The permuted genes produce both an acceptor piece (red) and coding piece (blue); dotted lines mark secondary structures not always present. Abbreviations: TLD, tRNA-like domain; MLR, mRNA-like region; ITS, internal transcribed spacer; P, paired region; PK, pseudoknot; RF, reading frame.

ssrA is both a target for some mobile DNAs and a passenger on others. It has been found interrupted by three types of mobile elements. By different strategies none of these disrupt gene function: group I introns remove themselves by self-splicing, rickettsial palindromic elements (RPEs) insert in innocuous sites, and integrase-encoding genomic islands split their target *ssrA* yet restore the split-off portion.

Non-chromosomal *ssrA* was first detected in a genomic survey of mycobacteriophages (in 10% of the phages). Other mobile elements including plasmids and genomic islands have been found bearing *ssrA*. One interesting case is *Rhodobacter sphaeroides* ATCC 17025, whose native tmRNA gene is disrupted by a genomic island; unlike all other genomic islands in tmRNA (or tRNA) genes this island has inactivated the native target gene without restoration, yet compensates by carrying its own tmRNA gene. A very unusual relative of *ssrA* is found in the lytic mycobacteriophage DS6A, that encodes little more that the TLD.

Chapter- 6

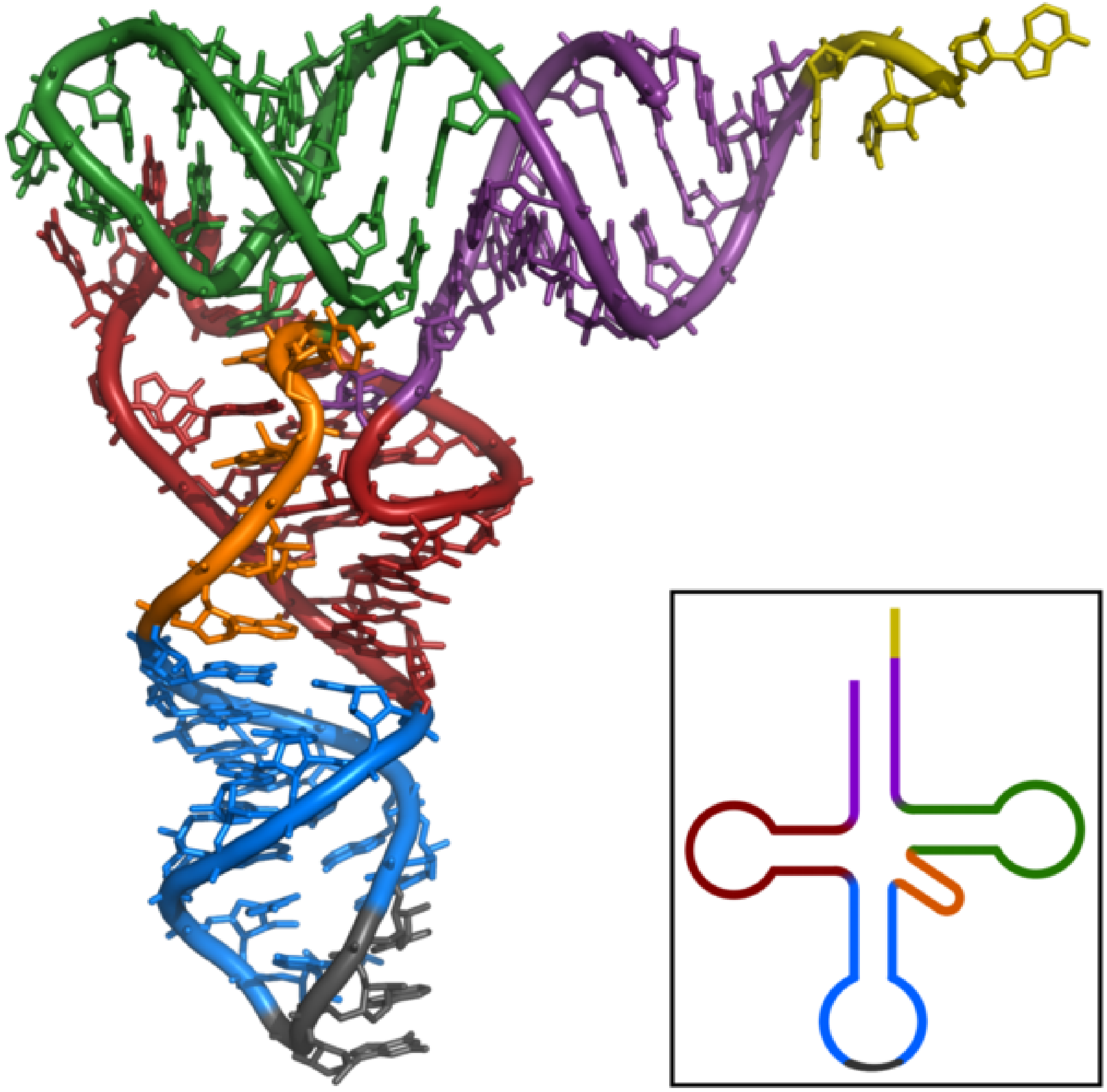
Non-Coding RNA

A **non-coding RNA** (**ncRNA**) is a functional RNA molecule that is not translated into a protein. Less-frequently used synonyms are non-protein-coding RNA (npcRNA), non-messenger RNA (nmRNA), small non-messenger RNA (snmRNA) and functional RNA (fRNA). The term **small RNA** (**sRNA**) is often used for small bacterial ncRNAs. The DNA sequence from which a non-coding RNA is transcribed as the end product is often called an **RNA gene** or non-coding RNA gene.

Non-coding RNA genes include highly abundant and functionally important RNAs such as transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as RNAs such as snoRNAs, microRNAs, siRNAs and piRNAs and the long ncRNAs that include examples such as Xist and HOTAIR. The number of ncRNAs encoded within the human genome is unknown, however recent transcriptomic and bioinformatic studies suggest the existence of thousands of ncRNAs. Since many of the newly identified ncRNAs have not been validated for their function, it is possible that many are non-functional.

History and discovery

Nucleic acids were first discovered in 1868 by Friedrich Miescher and by 1939 RNA had been implicated in protein synthesis. Two decades later, Francis Crick predicted a functional RNA component which mediated translation; he reasoned that RNA is better suited to base-pair with the mRNA transcript than a pure polypeptide.



The cloverleaf structure of Yeast tRNA^{Phe} (*inset*) and the 3D structure determined by X-ray analysis.

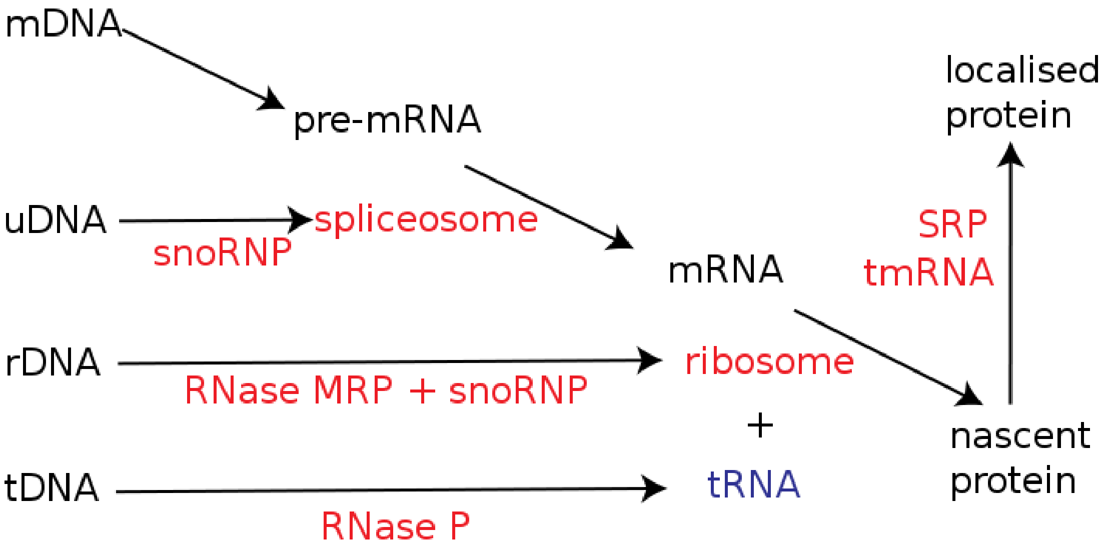
The first non-coding RNA to be characterised was an alanine tRNA found in baker's yeast, its structure was published in 1965. To produce a purified alanine tRNA sample, Robert W. Holley *et al.* used 140kg of commercial baker's yeast to give just 1g of purified tRNA^{Ala} for analysis. The 80 nucleotide tRNA was sequenced by first being digested with Pancreatic ribonuclease (producing fragements ending in Cytosine or Uridine) and then with takadiastase ribonuclease Tl (producing fragments which finished with Guanosine). Chromatography and identification of the 5' and 3' ends then helped arrange the fragments to establish the RNA sequence. Of the three structures originally proposed for this tRNA, the 'cloverleaf' structure was independently proposed in several following publications. The cloverleaf secondary structure was finalised following X-ray crystallography anaylsis performed by two independent research groups in 1974.

Ribosomal RNA was next to be discovered, followed by URNA in the early 1980s. Since then, the discovery of new non-coding RNAs has continued with snoRNAs, Xist, CRISPR and many more. Recent notable additions include riboswitches and miRNA, the discovery of the RNAi mechanism associated with the latter earned Craig C. Mello and Andrew Fire the 2006 Nobel Prize in Physiology or Medicine.

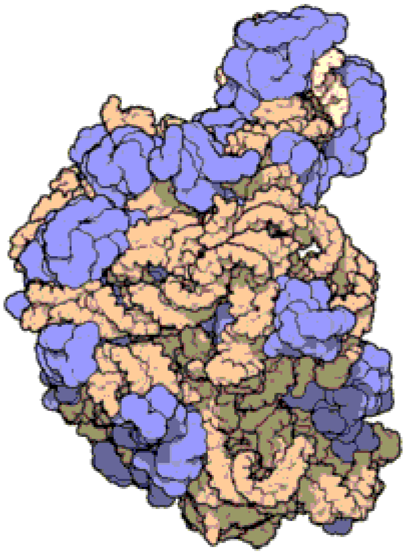
Biological roles of ncRNA

Noncoding RNAs belong to several groups and are involved in many cellular processes. These range from ncRNAs of central importance that are conserved across all or most cellular life through to more transient ncRNAs specific to one or a few closely related species. The more conserved ncRNAs are thought to be molecular fossils or relics from LUCA and the RNA world.

ncRNAs in translation



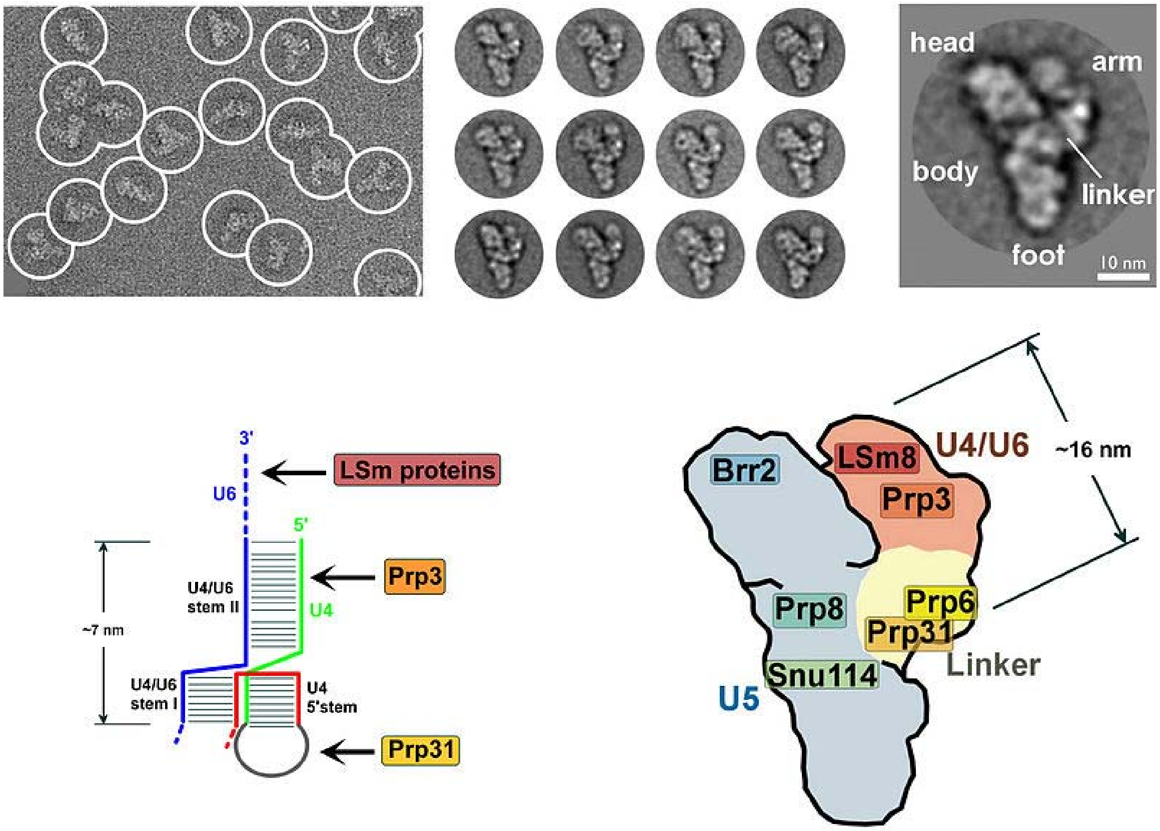
An illustration of the central dogma of molecular biology annotated with the processes ncRNAs are involved in. RNPs are shown in red, ncRNAs are shown in blue.



Atomic structure of the 50S Subunit from *Haloarcula marismortui*. Proteins are shown in blue and the two RNA strands in orange and yellow. The small patch of green in the center of the subunit is the active site.

Many of the conserved, essential and abundant ncRNAs are involved in translation. Ribonucleoprotein (RNP) particles called ribosomes are the 'factories' where translation takes place in the cell. The ribosome consists of more than 60% ribosomal RNA, these are made up of 3 ncRNAs in prokaryotes and 4 ncRNAs in eukaryotes. Ribosomal RNAs catalyse the translation of nucleotide sequences to protein. Another set of ncRNAs, Transfer RNAs, form an 'adaptor molecule' between mRNA and protein. The H/ACA box and C/D box snoRNAs are ncRNAs found in archaea and eukaryotes, RNase MRP is restricted to eukaryotes, both groups of ncRNA are involved in the maturation of rRNA. The snoRNAs guide covalent modifications of rRNA, tRNA and snRNAs, RNase MRP cleaves the internal transcribed spacer 1 between 18S and 5.8S rRNAs. The ubiquitous ncRNA, RNase P, is an evolutionary relative of RNase MRP. RNase P matures tRNA sequences by generating mature 5'-ends of tRNAs through cleaving the 5'-leader elements of precursor-tRNAs. Another ubiquitous RNP called SRP recognizes and transports specific nascent proteins to the endoplasmic reticulum in eukaryotes and the plasma membrane in prokaryotes. In bacteria Transfer-messenger RNA (tmRNA) is an RNP involved in rescuing stalled ribosomes, tagging incomplete polypeptides and promoting the degradation of aberrant mRNA.

ncRNAs in RNA splicing



Electron microscopy images of the yeast spliceosome. Note the bulk of the complex is in fact ncRNA.

In eukaryotes the spliceosome performs the splicing reactions essential for removing intron sequences, this process is required for the formation of mature mRNA. The spliceosome is another RNP often also known as the snRNP or tri-snRNP. There are two different forms of the spliceosome, the major and minor forms. The ncRNA components of the major spliceosome are U1, U2, U4 and U5. The ncRNA components of the minor spliceosome are U11, U12, U5, U4atac and U6atac.

Another group of introns can catalyse their own removal from host transcripts, these are called self-splicing RNAs. There are two main groups of self-splicing RNAs, these are the group I catalytic intron and group II catalytic intron. These ncRNAs catalyze their own excision from mRNA, tRNA and rRNA precursors in a wide range of organisms.

In mammals it has been found that snoRNAs can also regulate the alternative splicing of mRNA, for example snoRNA HBII-52 regulates the splicing of serotonin receptor 2C.

In nematodes the SmY ncRNA appears to be involved in mRNA trans-splicing.

ncRNAs in gene regulation

The expression of many thousands of genes are regulated by ncRNAs. This regulation can occur in trans or in cis.

trans-acting ncRNAs

In higher eukaryotes microRNAs regulate gene expression. A single miRNA can reduce the expression levels of hundreds of genes. The mechanism by which mature miRNA molecules act is through partial complementary to one or more messenger RNA (mRNA) molecules, generally in 3' UTRs. The main function of miRNAs is to down-regulate gene expression.

The ncRNA RNase P has also been shown to influence gene expression. In the human nucleus RNase P is required for the normal and efficient transcription of various ncRNAs transcribed by RNA polymerase III. These include tRNA, 5S rRNA, SRP RNA and U6 snRNA genes. RNase P exerts its role in transcription through association with Pol III and chromatin of active tRNA and 5S rRNA genes.

It has been shown that 7SK RNA, a metazoan ncRNA, acts as a negative regulator of the RNA polymerase II elongation factor P-TEFb, and that this activity is influenced by stress response pathways.

The bacterial ncRNA, 6S RNA, specifically associates with RNA polymerase holoenzyme containing the sigma70 specificity factor. This interaction represses expression from a sigma70-dependent promoter during stationary phase.

Another bacterial ncRNA, OxyS RNA represses translation by binding to Shine-Dalgarno sequences thereby occluding ribosome binding. OxyS RNA is induced in response to oxidative stress in Escherichia coli.

The B2 RNA is a small noncoding RNA polymerase III transcript that represses mRNA transcription in response to heat shock in mouse cells. B2 RNA inhibits transcription by binding to core Pol II. Through this interaction, B2 RNA assembles into preinitiation complexes at the promoter and blocks RNA synthesis.

A recent study has shown that just the act of transcription of ncRNA sequence can have an influence on gene expression. RNA polymerase II transcription of ncRNAs is required for chromatin remodelling in the Schizosaccharomyces pombe. Chromatin is progressively converted to an open configuration, as several species of ncRNAs are transcribed.

cis-acting ncRNAs

A number of ncRNAs are embedded in the 5' UTRs of protein coding genes and influence their expression in various ways. For example, a riboswitch can directly bind a small target molecule, the binding of the target affects the gene's activity.

RNA leader sequences are found upstream of the first gene of in amino acid biosynthetic operons. These RNA elements form one of two possible structures in regions encoding very short peptide sequences that are rich in the end product amino acid of the operon. A terminator structure forms when there is an excess of the regulatory amino acid and ribosome movement over the leader transcript is not impeded. When there is a deficiency of the charged tRNA of the regulatory amino acid the ribosome translating the leader peptide stalls and the antiterminator structure forms. This allows RNA polymerase to transcribe the operon. Known RNA leaders are Histidine operon leader, Leucine operon leader, Threonine operon leader and the Tryptophan operon leader.

Iron response elements (IRE) are bound by iron response proteins (IRP). The IRE is found in UTRs (Untranslated Regions) of various mRNAs whose products are involved in iron metabolism. When iron concentration is low, IRPs bind the ferritin mRNA IRE leading to translation repression.

Internal ribosome entry sites (IRES) are a RNA structure that allow for translation initiation in the middle of a mRNA sequence as part of the process of protein synthesis.

ncRNAs and genome defense

Piwi-interacting RNAs (piRNAs) expressed in mammalian testes and somatic cells, they form RNA-protein complexes with Piwi proteins. These piRNA complexes (piRCs) have been linked to transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells, particularly those in spermatogenesis.

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are repeats found in the DNA of many bacteria and archaea. The repeats are separated by spacers of similar length. It has been demonstrated that these spacers can be derived from phage and subsequently help protect the cell from infection.

ncRNAs and chromosome structure

Telomerase is an RNP enzyme that adds specific DNA sequence repeats ("TTAGGG" in vertebrates) to telomeric regions, which are found at the ends of eukaryotic chromosomes. The telomeres contain condensed DNA material, giving stability to the chromosomes. The enzyme is a reverse transcriptase that carries Telomerase RNA, which is used as a template when it elongates telomeres, which are shortened after each replication cycle.

Xist (X-inactive-specific transcript) is an long ncRNA gene on the X chromosome of the placental mammals that acts as major effector of the X chromosome inactivation process forming Barr bodies. An antisense RNA, Tsix, is a negative regulator of Xist. X chromosomes lacking Tsix expression (and thus having high levels of Xist transcription) are inactivated more frequently than normal chromosomes. In drosophilids, which also use an XY sex-determination system, the roX (RNA on the X) RNAs are involved in dosage compensation. Both Xist and roX operate by epigenetic regulation of transcription through the recruitment of histone-modifying enzymes.

Bifunctional RNA

Bifunctional RNAs are RNAs that have two distinct functions, these are also known as dual function RNAs. The majority of the known bifunctional RNAs are both mRNAs that encode a protein and ncRNAs. However there are also a growing number of ncRNAs that fall into two different ncRNA categories e.g. H/ACA box snoRNA and miRNA.

Two well known examples of bifunctional RNAs are SgrS RNA and RNAIII. However, a handful of other bifunctional RNAs are known to exist, e.g. SRA (Steroid Receptor Activator), VegT RNA, Oskar RNA and ENOD40.

ncRNAs and disease

As with proteins, mutations or imbalances in the ncRNA repertoire within the body can cause a variety of diseases.

Cancer

Many ncRNAs show abnormal expression patterns in cancerous tissues. These include miRNAs, long mRNA-like ncRNAs, GAS5, SNORD50, telomerase RNA and Y RNAs. The miRNAs are involved in the large scale regulation of many protein coding genes, the Y RNAs are important for the initiation of DNA replication, telomerase RNA that serves as a primer for telomerase, an RNP that extends telomeric regions at chromosome ends. The direct function of the long mRNA-like ncRNAs is less clear.

Germ-line mutations in miR-16-1 and miR-15 primary precursors have been shown to be much more frequent in patients with chronic lymphocytic leukemia compared to control populations.

It has been suggested that a rare SNP (rs11614913) that overlaps hsa-mir-196a2 has been found to be associated with non-small cell lung carcinoma. Likewise, a screen of 17 miRNAs that have been predicted to regulate a number of breast cancer associated genes found variations in the microRNAs miR-17 and miR-30c-1, these patients were noncarriers of BRCA1 or BRCA2 mutations, lending the possibility that familial breast cancer may be caused by variation in these miRNAs.

Prader–Willi syndrome

The deletion of the 48 copies of the C/D box snoRNA SNORD116 has been shown to be the primary cause of Prader–Willi syndrome. Prader–Willi is a developmental disorder associated with over-eating and learning difficulties. SNORD116 has potential target sites within a number of protein-coding genes, and could have a role in regulating alternative splicing.

Autism

The chromosomal locus containing the small nucleolar RNA SNORD115 gene cluster has been duplicated in approximately 5% of individuals with autistic traits. A mouse model engineered to have a duplication of the SNORD115 cluster displays autistic-like behaviour.

Cartilage-hair hypoplasia

Mutations within RNase MRP have been shown to cause cartilage-hair hypoplasia, a disease associated with an array of symptoms such as short stature, sparse hair, skeletal abnormalities and a suppressed immune system that is frequent among Amish and Finnish. The best characterised variant is an A-to-G transition at nucleotide 70 that is in a loop region two bases 5' of a conserved pseudoknot. However, many other mutations within RNase MRP also cause CHH.

Alzheimer's disease

The antisense RNA, BACE1-AS is transcribed from the opposite strand to BACE1 and is upregulated in patients with Alzheimer's disease. BACE1-AS regulates the expression of BACE1 by increasing BACE1 mRNA stability and generating additional BACE1 through a post-transcriptional feed-forward mechanism. By the same mechanism it also raises concentrations of beta amyloid, the main constituent of senile plaques. BACE1-AS concentrations are elevated in subjects with Alzheimer's disease and in amyloid precursor protein transgenic mice.

miR-96 and hearing loss

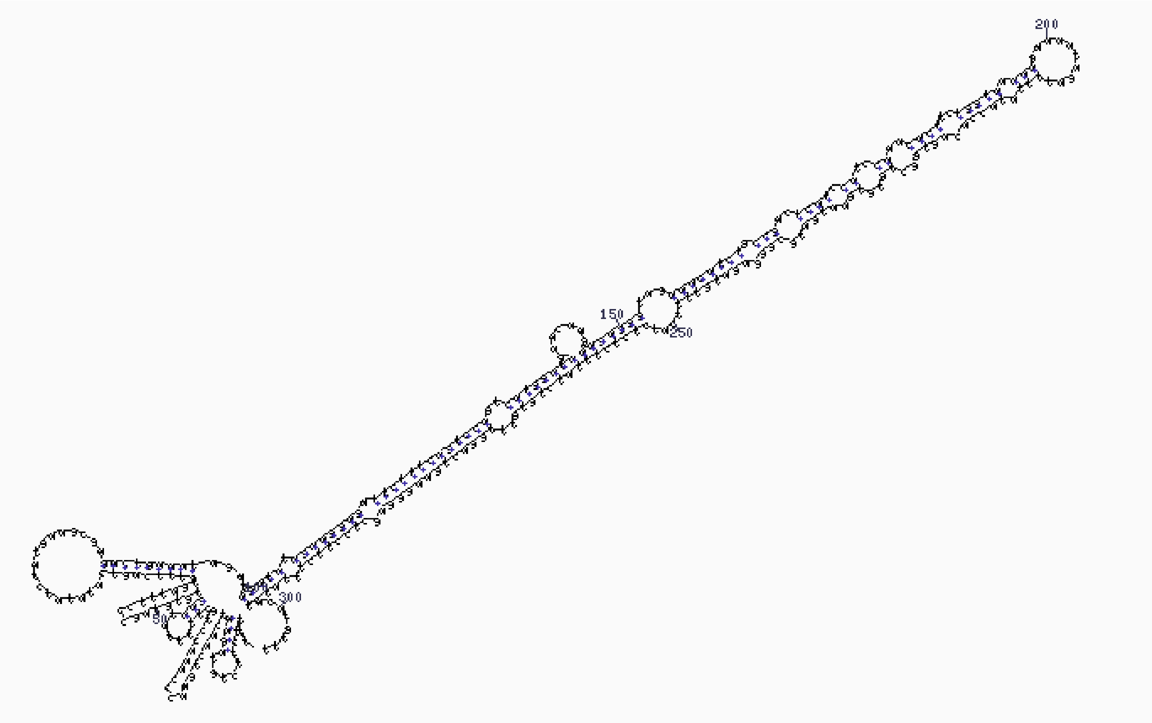
Variation within the seed region of mature miR-96 has been associated with autosomal dominant, progressive hearing loss in humans and mice. The homozygous mutant mice were profoundly deaf, showing no cochlear responses. Heterozygous mice and humans progressively lose the ability to hear.

Distinction between functional RNA (fRNA) and ncRNA

Several publications have started using the term **functional RNA (fRNA)**, as opposed to ncRNA, to describe regions functional at the RNA level that may or may not be stand-alone RNA transcripts. Therefore, every ncRNA is a fRNA, but there exist fRNA (such as riboswitches, SECIS elements, and other cis-regulatory regions) that are not ncRNA. Yet the term fRNA could also include mRNA as this is RNA coding for protein and hence is functional. Additionally artificially evolved RNAs also fall under the fRNA umbrella term. Some publications state that the terms *ncRNA* and *fRNA* are nearly synonymous.

Chapter- 7

MicroRNA



The stem-loop secondary structure of a pre-microRNA from *Brassica oleracea*

MicroRNAs (miRNAs) are short ribonucleic acid (RNA) molecules, on average only 22 nucleotides long and are found in all eukaryotic cells. miRNAs are post-transcriptional regulators that bind to complementary sequences on target messenger RNA transcripts (mRNAs), usually resulting in translational repression and gene silencing. The human genome may encode over 1000 miRNAs, which may target about 60% of mammalian genes and are abundant in many human cell types.

miRNAs show very different characteristics between plants and metazoans. In plants the miRNA complementarity to its mRNA target is nearly perfect, with no or few mismatched bases. In metazoans on the other hand miRNA complementarity is far from perfect and one miRNA can target many different sites on the same mRNA or on many different mRNAs. Another difference is the location of target sites on mRNAs. In metazoans the miRNA target sites are in the three prime untranslated regions (3'UTR) of the mRNA. In plants targets can be located in the 3' UTR but are more often in the coding region itself. MiRNAs are well conserved in eukaryotic organism and are thought to be a vital and evolutionary ancient component of genetic regulation.

The first miRNAs were characterized in the early 1990s, but miRNAs were not recognized as a distinct class of biologic regulators with conserved functions until the early 2000s. Since then, miRNA research has revealed multiple roles in negative regulation (transcript degradation and sequestering, translational suppression) and possible involvement in positive regulation (transcriptional and translational activation). By affecting gene regulation, miRNAs are likely to be involved in most biologic processes. Different sets of expressed miRNAs are found in different cell types and tissues.

Aberrant expression of miRNAs has been implicated in numerous disease states, and miRNA-based therapies are under investigation.

History

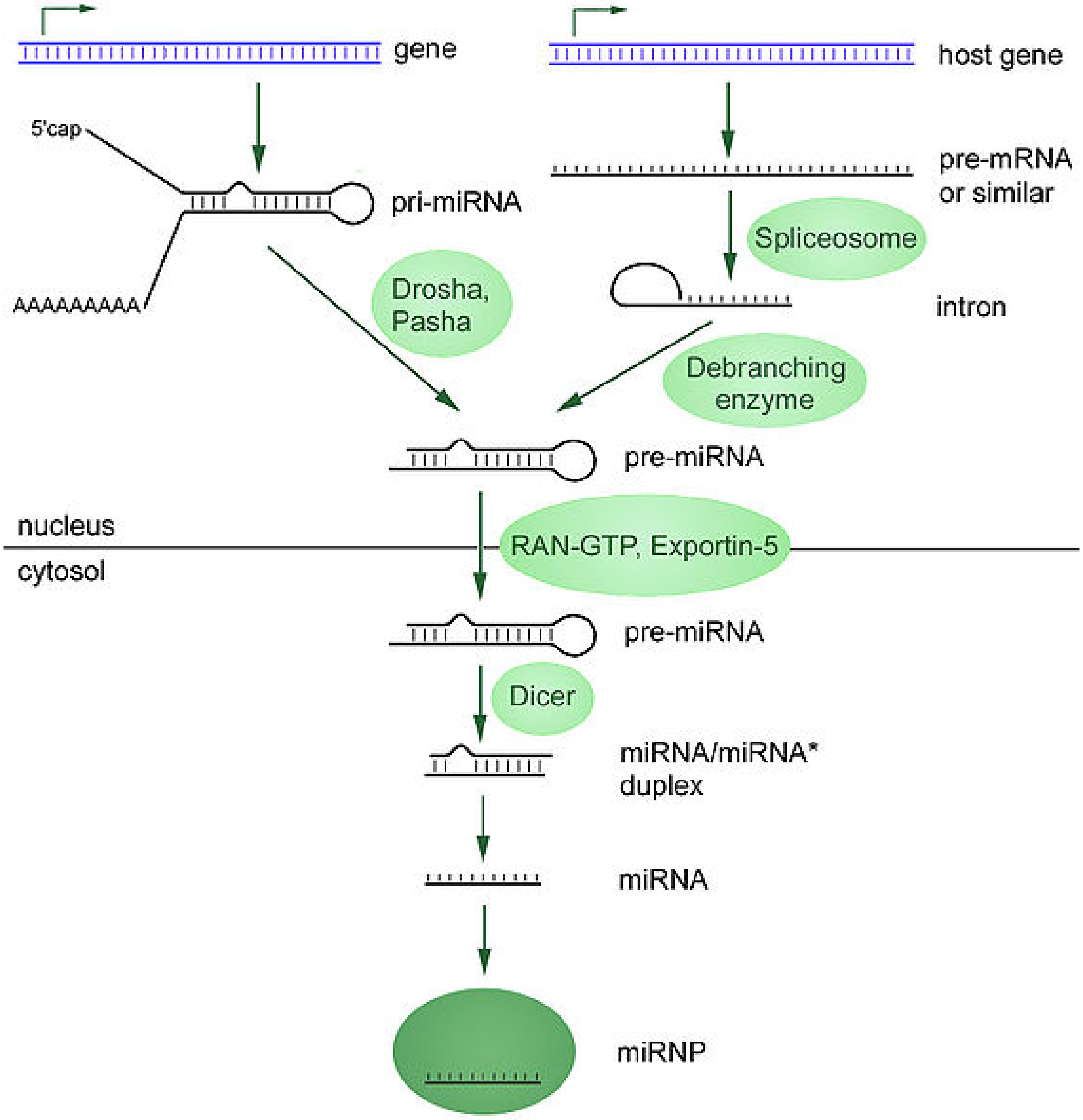
MicroRNAs were discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum during a study of the gene *lin-14* in *C. elegans* development. They found that LIN-14 protein abundance was regulated by a short RNA product encoded by the *lin-4* gene. A 61 nucleotide precursor from *lin-4* gene matured to a 22 nucleotide RNA containing sequences partially complementary to multiple sequences in the 3’ UTR of the *lin-14* mRNA. This complementarity was sufficient and necessary to inhibit the translation of *lin-14* mRNA into LIN-14 protein. Retrospectively, the *lin-4* small RNA was the first microRNA to be identified, though at the time, it was thought to be a nematode idiosyncrasy. Only in 2000 was a second RNA characterized: let-7, which repressed *lin-41*, *lin-14*, *lin-28*, *lin-42*, and *daf-12* expression during developmental stage transitions in *C. elegans*. let-7 was soon found to be conserved in many species, indicating the existence of a wider phenomenon.

Nomenclature

Under a standard nomenclature system, names are assigned to experimentally confirmed miRNAs before publication of their discovery. The prefix "mir" is followed by a dash and a number, the latter often indicating order of naming. For example, mir-123 was named and likely discovered prior to mir-456. The uncapitalized "mir-" refers to the pre-miRNA, while a capitalized "miR-" refers to the mature form. miRNAs with nearly identical sequences bar one or two nucleotides are annotated with an additional lower case letter. For example, miR-123a would be closely related to miR-123b. Pre-miRNAs that lead to 100% identical mature miRNAs but that are located at different places in the genome are indicated with an additional dash-number suffix. For example, the pre-miRNAs hsa-mir-194-1 and hsa-mir-194-2 lead to an identical mature miRNA (hsa-miR-

194) but are located in different regions of the genome. Species of origin is designated with a three-letter prefix, e.g., hsa-miR-123 would be from human (*Homo sapiens*) and oar-miR-123 would be a sheep (*Ovis aries*) miRNA. Other common prefixes include 'v' for viral (miRNA encoded by a viral genome) and 'd' for *Drosophila* miRNA (a fruit fly commonly studied in genetic research). When two mature microRNAs originate from opposite arms of the same pre-miRNA, they are denoted with a -3p or -5p suffix. (In the past, this distinction was also made with 's' (sense) and 'as' (antisense)). When relative expression levels are known, an asterisk following the name indicates an miRNA expressed at low levels relative to the miRNA in the opposite arm of a hairpin. For example, miR-123 and miR-123* would share a pre-miRNA hairpin, but more miR-123 would be found in the cell.

Biogenesis



MicroRNAs are produced from either their own genes or from introns.

Most microRNA genes are found in intergenic regions or in anti-sense orientation to genes and contain their own miRNA gene promoter and regulatory units. As much as 40% of miRNA genes may lie in the introns of protein and non-protein coding genes or even in exons. These are usually, though not exclusively, found in a sense orientation. and thus usually are regulated together with their host genes. Other miRNA genes showing a common promoter include the 42-48% of all miRNAs originating from polycistronic units containing 2-7 discrete loops from which mature miRNAs are processed, although this does not necessarily mean the mature miRNAs of a family will be homologous in structure and function. The promoters mentioned have been shown to have some similarities in their motifs to promoters of other genes transcribed by RNA polymerase II such as protein coding genes. The DNA template is not the final word on mature miRNA production: 6% of human miRNAs show RNA editing, the site-specific modification of RNA sequences to yield products different from those encoded by their DNA. This increases the diversity and scope of miRNA action beyond that implicated from the genome alone.

Transcription

miRNA genes are usually transcribed by RNA polymerase II (Pol II). The polymerase often binds to a promoter found near the DNA sequence encoding what will become the hairpin loop of the pre-miRNA. The resulting transcript is capped with a specially-modified nucleotide at the 5’ end, polyadenylated with multiple adenosines (a poly(A) tail), and spliced. The product, called a primary miRNA (pri-miRNA), may be hundreds or thousands of nucleotides in length and contain one or more miRNA stem loops. When a stem loop precursor is found in the 3’ UTR, a transcript may serve as a pri-miRNA and a mRNA. RNA polymerase III (Pol III) transcribes some miRNAs, especially those with upstream Alu sequences, transfer RNAs (tRNAs), and mammalian wide interspersed repeat (MWIR) promoter units.

Nuclear processing

A single pri-miRNA may contain from one to six miRNA precursors. These hairpin loop structures are composed of about 70 nucleotides each. Each hairpin is flanked by sequences necessary for efficient processing. The double-stranded RNA structure of the hairpins in a pri-miRNA is recognized by a nuclear protein known as DiGeorge Syndrome Critical Region 8 (DGCR8 or "Pasha" in invertebrates), named for its association with DiGeorge Syndrome. DGCR8 associates with the enzyme Drosha, a protein that cuts RNA, to form the "Microprocessor" complex. In this complex, DGCR8 orients the catalytic RNase III domain of Drosha to liberate hairpins from pri-miRNAs by cleaving RNA about eleven nucleotides from the hairpin base (two helical RNA turns into the stem). The resulting hairpin, known as a pre-miRNA (precursor-miRNA), has a two-nucleotide overhang at its 3’ end; it has 3' hydroxyl and 5' phosphate groups.

pre-miRNAs that are spliced directly out of introns, bypassing the Microprocessor complex, are known as "mirtrons." Originally thought to exist only in *Drosophila* and *C. elegans*, mirtrons have now been found in mammals.

Perhaps as many as 16% of pri-miRNAs may be altered through nuclear RNA editing. Most commonly, enzymes known as adenosine deaminases acting on RNA (ADARs) catalyze adenosine to inosine (A to I) transitions. RNA editing can halt nuclear processing (for example, of pri-miR-142, leading to degradation by the ribonuclease Tudor-SN) and alter downstream processes including cytoplasmic miRNA processing and target specificity (e.g., by changing the seed region of miR-376 in the central nervous system).

Nuclear export

pre-miRNA hairpins are exported from the nucleus in a process involving the nucleocytoplasmic shuttle Exportin-5. This protein, a member of the *karyopherin* family, recognizes a two-nucleotide overhang left by the RNase III enzyme Drosha at the 3' end of the pre-miRNA hairpin. Exportin-5-mediated transport to the cytoplasm is energy-dependent, using GTP bound to the Ran protein.

Cytoplasmic processing

In the cytoplasm, the pre-miRNA hairpin is cleaved by the RNase III enzyme Dicer. This endoribonuclease interacts with the 3' end of the hairpin and cuts away the loop joining the 3' and 5' arms, yielding an imperfect miRNA:miRNA* duplex about 22 nucleotides in length. Overall hairpin length and loop size influence the efficiency of Dicer processing, and the imperfect nature of the miRNA:miRNA* pairing also affects cleavage. Although either strand of the duplex may potentially act as a functional miRNA, only one strand is usually incorporated into the RNA-induced silencing complex (RISC) where the miRNA and its mRNA target interact.

Biogenesis in plants

miRNA biogenesis in plants differs from metazoan biogenesis mainly in the steps of nuclear processing and export. Instead of being cleaved by two different enzymes, once inside and once outside the nucleus, both cleavages of the plant miRNA is performed by a Dicer homolog, called Dicer-like1 (DL1). DL1 is only expressed in the nucleus of plant cells, which indicates that both reactions take place inside the nucleus. Before plant miRNA:miRNA* duplexes are transported out of the nucleus its 3' overhangs are methylated by a RNA methyltransferaseprotein called Hua-Enhancer1 (HEN1). The duplex is then transported out of the nucleus to the cytoplasm by a protein called Hasty (HST), an Exportin 5 homolog, where they disassemble and the mature miRNA is incorporated into the RISC.

The RNA-induced silencing complex

The mature miRNA is part of an active RNA-induced silencing complex (RISC) containing Dicer and many associated proteins. RISC is also known as a microRNA ribonucleoprotein complex (miRNP); RISC with incorporated miRNA is sometimes referred to as "miRISC."

Dicer processing of the pre-miRNA is thought to be coupled with unwinding of the duplex. Generally, only one strand is incorporated into the miRISC, selected on the basis of its thermodynamic instability and weaker base-pairing relative to the other strand. The position of the stem-loop may also influence strand choice. The other strand, called the passenger strand due to its lower levels in the steady state, is denoted with an asterisk (*) and is normally degraded. In some cases, both strands of the duplex are viable and become functional miRNA that target different mRNA populations.

Members of the argonaute (Ago) protein family are central to RISC function. Argonautes are needed for miRNA-induced silencing and contain two conserved RNA binding domains: a PAZ domain that can bind the single stranded 3’ end of the mature miRNA and a PIWI domain that structurally resembles ribonuclease-H and functions to interact with the 5’ end of the guide strand. They bind the mature miRNA and orient it for interaction with a target mRNA. Some argonautes, for example human Ago2, cleave target transcripts directly; argonautes may also recruit additional proteins to achieve translational repression. The human genome encodes eight argonaute proteins divided by sequence similarities into two families: AGO (with four members present in all mammalian cells and called E1F2C/hAgo in humans), and PIWI (found in the germ line and hematopoietic stem cells).

Additional RISC components include TRBP [human immunodeficiency virus (HIV) transactivating response RNA (TAR) binding protein], PACT (protein activator of the interferon induced protein kinase (PACT), the SMN complex, fragile X mental retardation protein (FMRP), and Tudor staphylococcal nuclease-domain-containing protein (Tudor-SN).

Mode of Silencing

Gene silencing may occur either via mRNA degradation or preventing mRNA from being translated. It has been demonstrated that if there is complete complementation between the miRNA and target mRNA sequence, Ago2 can cleave the mRNA and lead to direct mRNA degradation. Yet, if there isn't complete complementation the silencing is achieved by preventing translation.

miRNA turnover

Turnover of mature miRNA is needed for rapid changes in miRNA expression profiles. During miRNA maturation in the cytoplasm, uptake by the Argonaute protein is thought to stabilize the guide strand, while the opposite (* or "passenger") strand is preferentially destroyed. In what has been called a "Use it or lose it" strategy, Argonaute may preferentially retain miRNAs with many targets over miRNAs with few or no targets, leading to degradation of the non-targeting molecules.

Decay of mature miRNAs in animals is mediated by the 5´-to-3´ exoribonuclease XRN2, also known as Rat1p. In plants, SDN (small RNA degrading nuclease) family members degrade miRNAs in the opposite (3'-to-5') direction. Similar enzymes are encoded in animal genomes, but their roles have not yet been described.

Several miRNA modifications affect miRNA stability. As indicated by work in the model organism *Arabidopsis thaliana* (thale cress), mature plant miRNAs appear to be stabilized by the addition of methyl moieties at the 3' end. The 2'-O-conjugated methyl groups block the addition of uracil (U) residues by uridyltransferase enzymes, a modification that may be associated with miRNA degradation. However, uridylation may also protect some miRNAs; the consequences of this modification are incompletely understood. Uridylation of some animal miRNAs has also been reported. Both plant and animal miRNAs may be altered by addition of adenine (A) residues to the 3' end of the miRNA. An extra A added to the end of mammalian miR-122, a liver-enriched miRNA important in Hepatitis C, stabilizes the molecule, and plant miRNAs ending with an adenine residue have slower decay rates.

Cellular functions

The function of miRNAs appears to be in gene regulation. For that purpose, a miRNA is complementary to a part of one or more messenger RNAs (mRNAs). Animal miRNAs are usually complementary to a site in the 3' UTR whereas plant miRNAs are usually complementary to coding regions of mRNAs. Perfect or near perfect base pairing with the target RNA promotes cleavage of the RNA. This is the primary mode of plant microRNAs. In animals, microRNAs more often only partially base pair and inhibit protein translation of the target mRNA (this exists in plants as well but is less common). MicroRNAs that are partially complementary to the target can also speed up deadenylation, causing mRNAs to be degraded sooner. For partially complementary microRNA to recognise their targets, the nucleotides 2–7 of the miRNA ('seed region') still have to be perfectly complementary. miRNAs occasionally also causes histone modification and DNA methylation of promoter sites and therefore affecting the expression of targeted genes.

Animal microRNAs target in particular developmental genes. In contrast, genes involved in functions common to all cells, such as gene expression, have very few microRNA target sites and seem to be under selection to avoid targeting by microRNAs.

dsRNA can also activate gene expression, a mechanism that has been termed "small RNA-induced gene activation" or RNAa. dsRNAs targeting gene promoters can induce potent transcriptional activation of associated genes. This was demonstrated in human cells using synthetic dsRNAs termed small activating RNAs (saRNAs), but has also been demonstrated for endogenous microRNA.

Evolution

MicroRNAs are significant phylogenetic markers because of their astonishingly low rate of evolution. Their origin may have permitted the development of morphological innovation, and by making gene expression more specific and 'fine-tunable', permitted the genesis of complex organs and perhaps, ultimately, complex life. Indeed, rapid bursts of morphological innovation are generally associated with a high rate of microRNA accumulation.

MicroRNAs originate predominantly by the random formation of hairpins in "non-coding" sections of DNA (i.e. introns or intergene regions), but also by the duplication and modification of existing microRNAs. The rate of evolution (i.e. nucleotide substitution) in recently-originated microRNAs is comparable to that elsewhere in the non-coding DNA, implying evolution by neutral drift; however, older microRNAs have a much lower rate of change (often less than one substitution per hundred million years), suggesting that once a microRNA gains a function it undergoes extreme purifying selection. At this point, a microRNA is rarely lost from an animal's genome, although microRNAs which are more recently derived (and thus presumably non-functional) are frequently lost. This makes them a valuable phylogenetic marker, and they are being looked upon as a possible solution to such outstanding phylogenetic problems as the relationships of arthropods.

MicroRNAs feature in the genomes of most eukaryotic organisms, from the brown algae to the metazoa. Across all species, in excess of 5000 had been identified by March 2010. Whilst short RNA sequences (50 – hundreds of base pairs) of a broadly comparable function occur in bacteria, bacteria lack true microRNAs.

Experimental detection and manipulation of miRNA

MicroRNA expression can be quantified in a two-step polymerase chain reaction process of modified RT-PCR followed by quantitative real-time PCR. Variations of this method achieve absolute or relative quantification. miRNAs can also be hybridized to microarrays, slides or chips with probes to hundreds or thousands of miRNA targets, so that relative levels of miRNAs can be determined in different samples. MicroRNAs can be both discovered and profiled by high-throughput sequencing methods. The activity of an miRNA can be experimentally inhibited using a locked nucleic acid (LNA) oligo, a Morpholino oligo or a 2'-O-methyl RNA oligo. MicroRNA maturation can be inhibited at several points by steric-blocking oligos. The miRNA target site of an mRNA transcript can also be blocked by a steric-blocking oligo. Additionally, a specific miRNA can be silenced by a complementary antagomir. For the “in situ” detection of miRNA, the use of LNA is currently the only efficient method. The locked conformation of LNA results in enhanced hybridization properties and increases sensitivity and selectivity, making it ideal for detection of short miRNA.

miRNA and disease

Just as miRNA is involved in the normal functioning of eukaryotic cells, so has dysregulation of miRNA been associated with disease. A manually curated, publicly available database miR2Disease documents known relationships between miRNA dysregulation and human disease.

miRNA and cancer

Several miRNAs have been found to have links with some types of cancer.

A study of mice altered to produce excess c-Myc — a protein with mutated forms implicated in several cancers — shows that miRNA has an effect on the development of cancer. Mice that were engineered to produce a surplus of types of miRNA found in lymphoma cells developed the disease within 50 days and died two weeks later. In contrast, mice without the surplus miRNA lived over 100 days. Leukemia can be caused by the insertion of a viral genome next to the 17-92 array of microRNAs leading to increased expression of this microRNA.

Another study found that two types of miRNA inhibit the E2F1 protein, which regulates cell proliferation. miRNA appears to bind to messenger RNA before it can be translated to proteins that switch genes on and off.

By measuring activity among 217 genes encoding miRNA, patterns of gene activity that can distinguish types of cancers can be discerned. miRNA signatures may enable classification of cancer. This will allow doctors to determine the original tissue type which spawned a cancer and to be able to target a treatment course based on the original tissue type. miRNA profiling has already been able to determine whether patients with chronic lymphocytic leukemia had slow growing or aggressive forms of the cancer.

Transgenic mice that over-express or lack specific miRNAs have provided insight into the role of small RNAs in various malignancies.

A novel miRNA-profiling based screening assay for the detection of early-stage colorectal cancer has been developed and is currently in clinical trials. Early results showed that blood plasma samples collected from patients with early, resectable (Stage II) colorectal cancer could be distinguished from those of sex-and age-matched healthy volunteers. Sufficient selectivity and specificity could be achieved using small (less than 1 mL) samples of blood. The test has potential to be a cost-effective, non-invasive way to identify at-risk patients who should undergo colonoscopy.

miRNA and heart disease

The global role of miRNA function in the heart has been addressed by conditionally inhibiting miRNA maturation in the murine heart, and has revealed that miRNAs play an essential role during its development. miRNA expression profiling studies demonstrate that expression levels of specific miRNAs change in diseased human hearts, pointing to their involvement in cardiomyopathies. Furthermore, studies on specific miRNAs in animal models have identified distinct roles for miRNAs both during heart development and under pathological conditions, including the regulation of key factors important for cardiogenesis, the hypertrophic growth response, and cardiac conductance.

miRNA and the nervous system

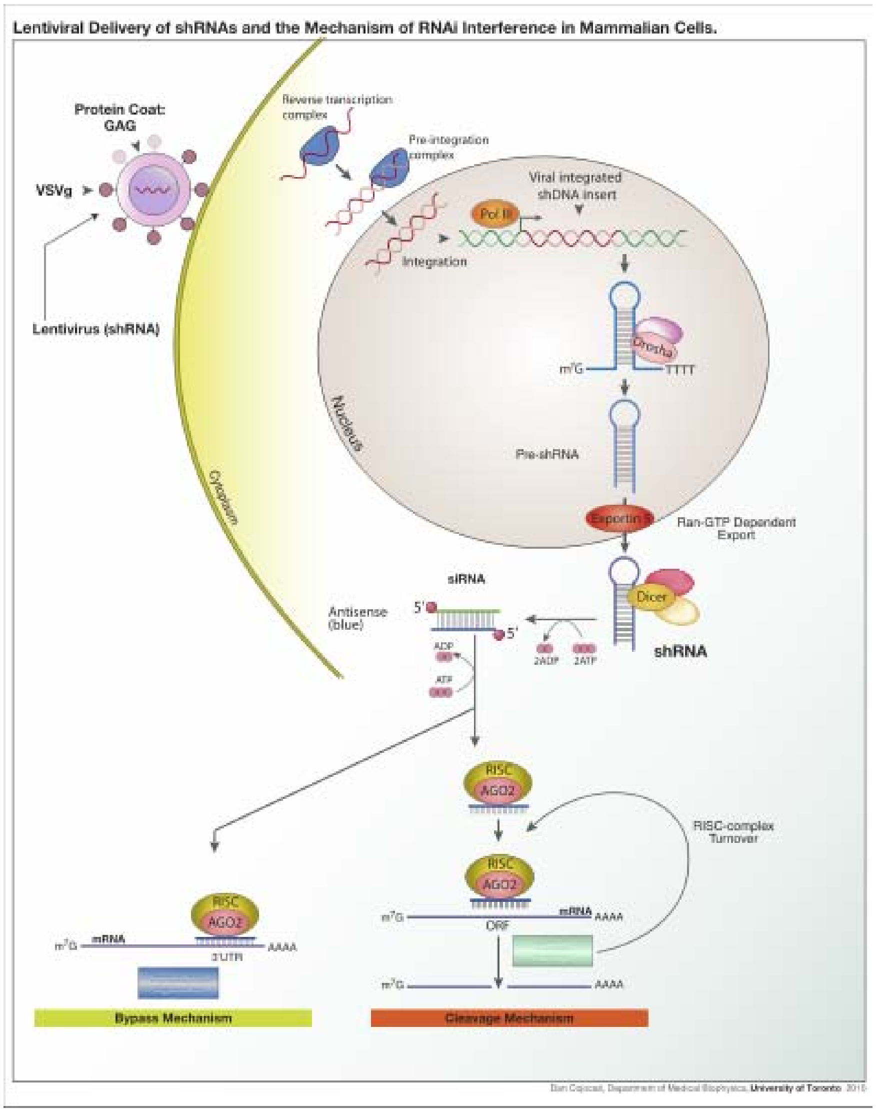
miRNAs appear to regulate the nervous system. Neural miRNAs are involved at various stages of synaptic development, including dendritogenesis (involving miR-132, miR-134 and miR-124), synapse formation and synapse maturation (where miR-134 and miR-138 are thought to be involved). Some studies find altered miRNA expression in schizophrenia.

miRNA and non-coding RNAs

When the human genome project mapped its first chromosome in 1999, it was predicted the genome would contain over 100,000 protein coding genes. However, only around 20,000 were eventually identified (International Human Genome Sequencing Consortium, 2004). Since then, the advent of bioinformatics approaches combined with genome tiling studies examining the transcriptome, systematic sequencing of full length cDNA libraries, and experimental validation (including the creation of miRNA derived antisense oligonucleotides called antagomirs) have revealed that many transcripts are non protein-coding RNA, including several snoRNAs and miRNAs.

Chapter- 8

RNA Interference



Lentiviral Delivery of shRNAs and the Mechanism of RNAi Interference in Mammalian Cells.

Dan Cojocari, Department of Medical Biophysics, University of Toronto 2010

Lentiviral Delivery of designed shRNA's and the mechanism of RNA interference in mammalian cells.

RNA interference (**RNAi**) is a system within living cells that takes part in controlling which genes are active and how active they are. Two types of small RNA molecules – microRNA (miRNA) and small interfering RNA (siRNA) – are central to RNA interference. RNAs are the direct products of genes, and these small RNAs can bind to specific other RNAs and either increase or decrease their activity, for example by preventing a messenger RNA from producing a protein. RNA interference has an important role in defending cells against parasitic genes – viruses and transposons – but also in directing development as well as gene expression in general.

The RNAi pathway is found in many eukaryotes including animals and is initiated by the enzyme Dicer, which cleaves long double-stranded RNA (dsRNA) molecules into short fragments of ~20 nucleotides. The siRNA will be unwound into two ssRNA, namely the passenger strand and the guide strand. The passenger strand will be degraded, and the guide strand is incorporated into the RNA-induced silencing complex (RISC). The most well-studied outcome is post-transcriptional gene silencing, which occurs when the guide strand base pairs with a complementary sequence of a messenger RNA molecule and induces cleavage by Argonaute, the catalytic component of the RISC complex. This process is known to spread systemically throughout the organism despite initially limited molar concentrations of siRNA.

The selective and robust effect of RNAi on gene expression makes it a valuable research tool, both in cell culture and in living organisms because synthetic dsRNA introduced into cells can induce suppression of specific genes of interest. RNAi may also be used for large-scale screens that systematically shut down each gene in the cell, which can help identify the components necessary for a particular cellular process or an event such as cell division. Exploitation of the pathway is also a promising tool in biotechnology and medicine.

Historically, RNA interference was known by other names, including cosuppression, post transcriptional gene silencing, and quelling. Only after these apparently unrelated processes were fully understood did it become clear that they all described the RNAi phenomenon. In 2006, Andrew Fire and Craig C. Mello shared the Nobel Prize in Physiology or Medicine for their work on RNA interference in the nematode worm *C. elegans*, which they published in 1998.

Cellular mechanism

The dicer protein from *Giardia intestinalis*, which catalyzes the cleavage of dsRNA to siRNAs. The RNase domains are colored green, the PAZ domain yellow, the platform domain red, and the connector helix blue.

RNAi is an RNA-dependent gene silencing process that is controlled by the RNA-induced silencing complex (RISC) and is initiated by short double-stranded RNA molecules in a cell's cytoplasm, where they interact with the catalytic RISC component argonaute. When the dsRNA is exogenous (coming from infection by a virus with an RNA genome or laboratory manipulations), the RNA is imported directly into the cytoplasm and cleaved to short fragments by the enzyme Dicer. The initiating dsRNA can also be endogenous (originating in the cell), as in pre-microRNAs expressed from RNA-coding genes in the genome. The primary transcripts from such genes are first processed to form the characteristic stem-loop structure of pre-miRNA in the nucleus, then exported to the cytoplasm to be cleaved by Dicer. Thus, the two dsRNA pathways, exogenous and endogenous, converge at the RISC complex.

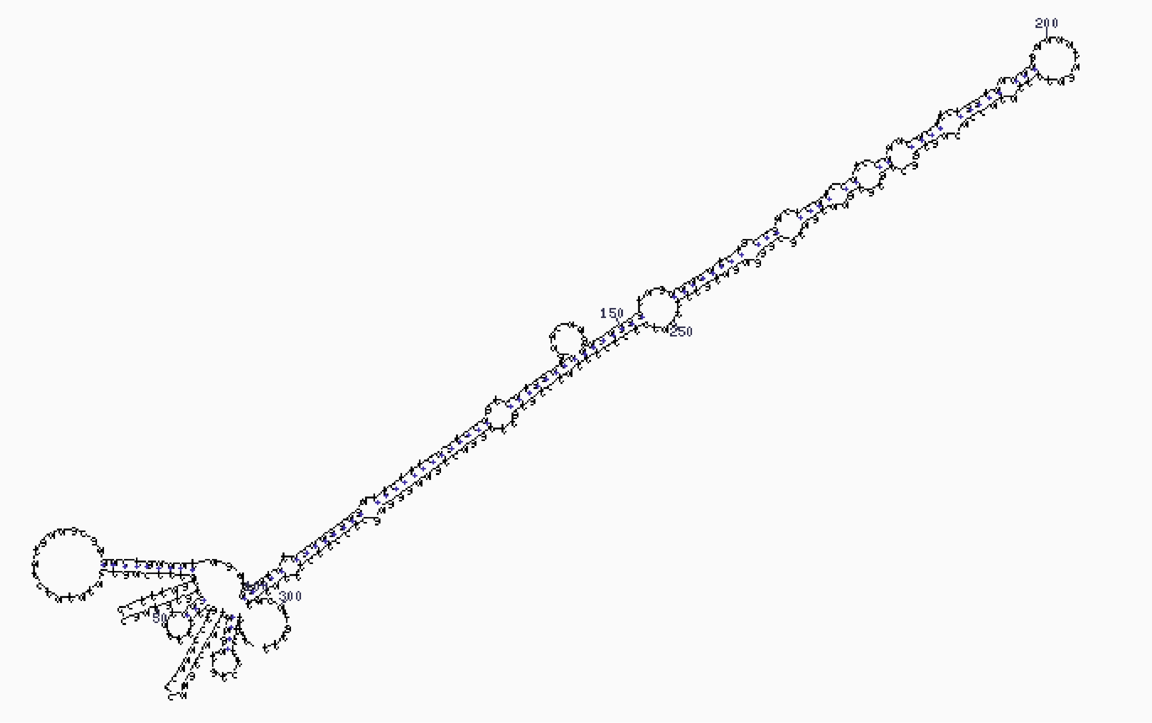
dsRNA cleavage

Endogenous dsRNA initiates RNAi by activating the ribonuclease protein Dicer, which binds and cleaves double-stranded RNAs (dsRNAs) to produce double-stranded fragments of 20–25 base pairs with a 2 nucleotide overhang at 3' end. Bioinformatics studies on the genomes of multiple organisms suggest this length maximizes target-gene specificity and minimizes non-specific effects. These short double-stranded fragments are called small interfering RNAs (siRNAs). These siRNAs are then separated into single strands and integrated into an active RISC complex. After integration into the RISC, siRNAs base-pair to their target mRNA and induce cleavage of the mRNA, thereby preventing it from being used as a translation template.

Exogenous dsRNA is detected and bound by an effector protein, known as RDE-4 in *C. elegans* and R2D2 in *Drosophila*, that stimulates dicer activity. This protein only binds long dsRNAs, but the mechanism producing this length specificity is unknown. These RNA-binding proteins then facilitate transfer of cleaved siRNAs to the RISC complex.

In *C. elegans*, this initiation response is amplified by the cell by the synthesis of a population of 'secondary' siRNAs using the dicer-produced initiating or 'primary' siRNAs as templates. These siRNAs are structurally distinct from dicer-produced siRNAs and appear to be produced by an RNA-dependent RNA polymerase (RdRP).

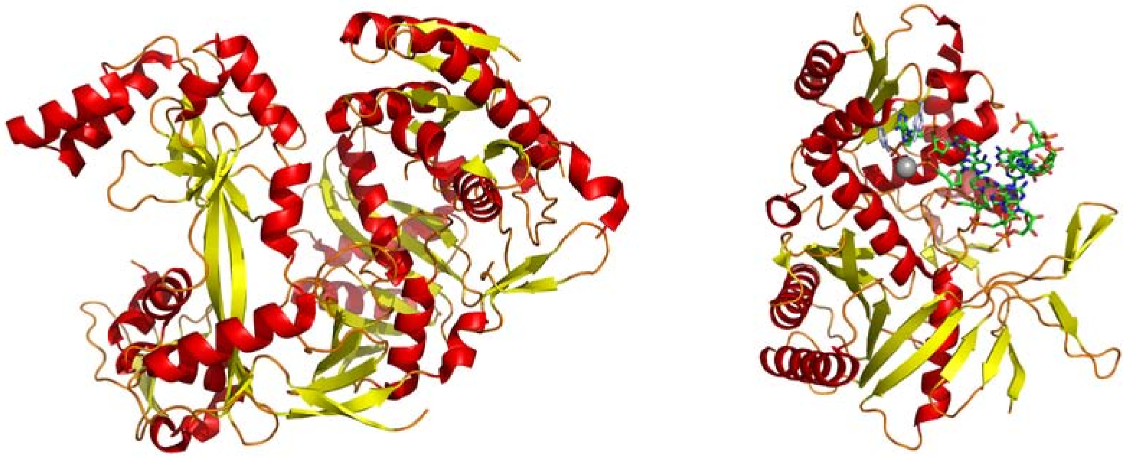
MicroRNA



The stem-loop secondary structure of a pre-microRNA from *Brassica oleracea*

MicroRNAs (miRNAs) are genomically encoded non-coding RNAs that help regulate gene expression, particularly during development. The phenomenon of RNA interference, broadly defined, includes the endogenously induced gene silencing effects of miRNAs as well as silencing triggered by foreign dsRNA. Mature miRNAs are structurally similar to siRNAs produced from exogenous dsRNA, but before reaching maturity, miRNAs must first undergo extensive post-transcriptional modification. An miRNA is expressed from a much longer RNA-coding gene as a primary transcript known as a *pri-miRNA* which is processed, in the cell nucleus, to a 70-nucleotide stem-loop structure called a *pre-miRNA* by the microprocessor complex. This complex consists of an RNase III enzyme called Drosha and a dsRNA-binding protein Pasha. The dsRNA portion of this pre-miRNA is bound and cleaved by Dicer to produce the mature miRNA molecule that can be integrated into the RISC complex; thus, miRNA and siRNA share the same cellular machinery downstream of their initial processing.

The siRNAs derived from long dsRNA precursors differ from miRNAs in that miRNAs, especially those in animals, typically have incomplete base pairing to a target and inhibit the translation of many different mRNAs with similar sequences. In contrast, siRNAs typically base-pair perfectly and induce mRNA cleavage only in a single, specific target. In *Drosophila* and *C. elegans*, miRNA and siRNA are processed by distinct argonaute proteins and dicer enzymes.



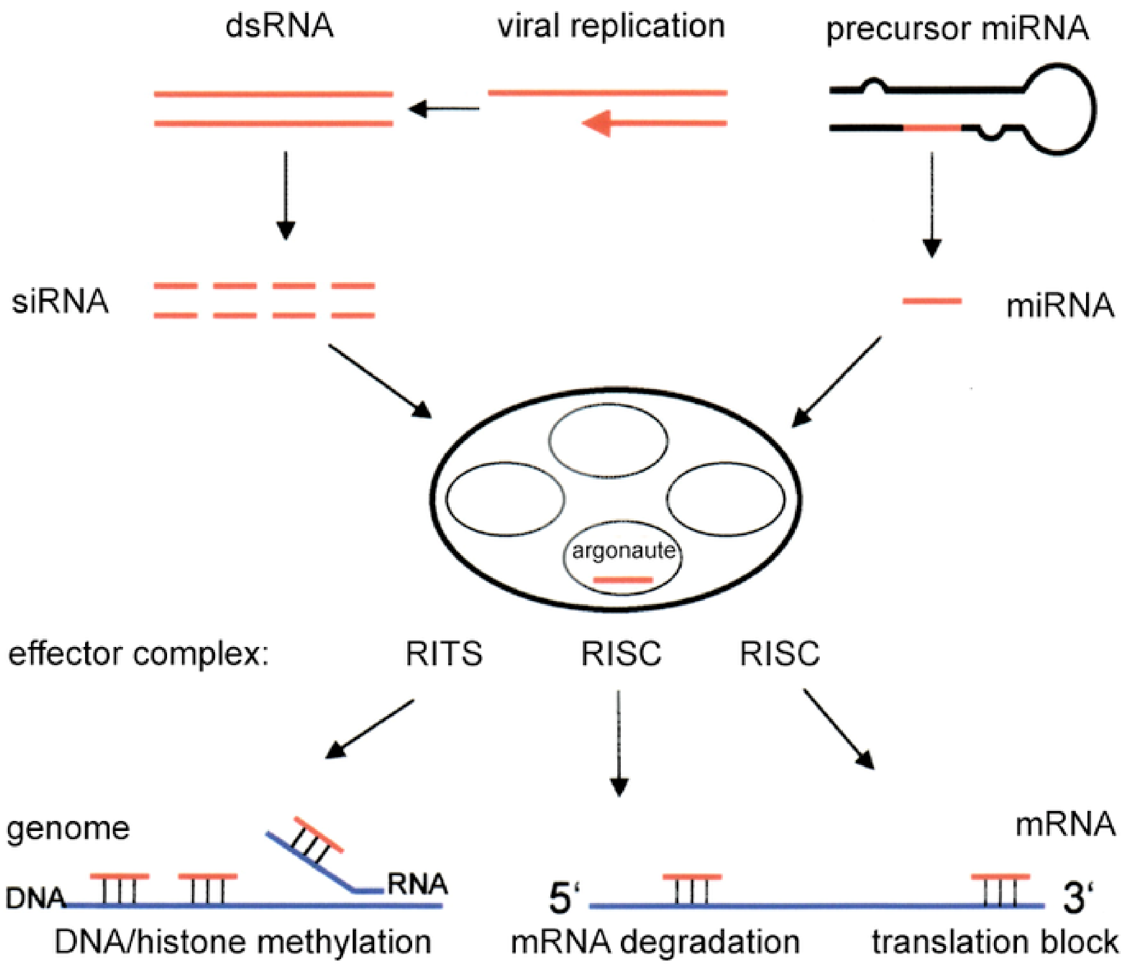
Left: A full-length argonaute protein from the archaea species *Pyrococcus furiosus*.
Right: The PIWI domain of an argonaute protein in complex with double-stranded RNA.

RISC activation and catalysis

The active components of an RNA-induced silencing complex (RISC) are endonucleases called argonaute proteins, which cleave the target mRNA strand complementary to their bound siRNA. As the fragments produced by dicer are double-stranded, they could each in theory produce a functional siRNA. However, only one of the two strands, which is known as the *guide strand*, binds the argonaute protein and directs gene silencing. The other *anti-guide strand* or *passenger strand* is degraded during RISC activation. Although it was first believed that an ATP-dependent helicase separated these two strands, the process is actually ATP-independent and performed directly by the protein components of RISC. The strand selected as the guide tends to be the one whose 5' end is least paired to its complement, but strand selection is unaffected by the direction in which dicer cleaves the dsRNA before RISC incorporation. Instead, the R2D2 protein may serve as the differentiating factor by binding the more-stable 5' end of the passenger strand.

The structural basis for binding of RNA to the argonaute protein was examined by X-ray crystallography of the binding domain of an RNA-bound argonaute protein. Here, the phosphorylated 5' end of the RNA strand enters a conserved basic surface pocket and makes contacts through a divalent cation (an atom with two positive charges) such as magnesium and by aromatic stacking (a process that allows more than one atom to share an electron by passing it back and forth) between the 5' nucleotide in the siRNA and a conserved tyrosine residue. This site is thought to form a nucleation site for the binding of the siRNA to its mRNA target.

It is not understood how the activated RISC complex locates complementary mRNAs within the cell. Although the cleavage process has been proposed to be linked to translation, translation of the mRNA target is not essential for RNAi-mediated degradation. Indeed, RNAi may be more effective against mRNA targets that are not translated. Argonaute proteins, the catalytic components of RISC, are localized to specific regions in the cytoplasm called P-bodies (also cytoplasmic bodies or GW bodies), which are regions with high rates of mRNA decay; miRNA activity is also clustered in P-bodies. Disruption of P-bodies decreases the efficiency of RNA interference, suggesting that they are the site of a critical step in the RNAi process.



The enzyme dicer trims double stranded RNA, to form small interfering RNA or microRNA. These processed RNAs are incorporated into the RNA-induced silencing complex (RISC), which targets messenger RNA to prevent translation.

Transcriptional silencing

Components of the RNA interference pathway are also used in many eukaryotes in the maintenance of the organization and structure of their genomes. Modification of histones and associated induction of heterochromatin formation serves to downregulate genes pre-transcriptionally; this process is referred to as RNA-induced transcriptional silencing (RITS), and is carried out by a complex of proteins called the RITS complex. In fission yeast this complex contains argonaute, a chromodomain protein Chp1, and a protein called Tas3 of unknown function. As a consequence, the induction and spread of heterochromatic regions requires the argonaute and RdRP proteins. Indeed, deletion of these genes in the fission yeast *S. pombe* disrupts histone methylation and centromere formation, causing slow or stalled anaphase during cell division. In some cases, similar processes associated with histone modification have been observed to transcriptionally upregulate genes.

The mechanism by which the RITS complex induces heterochromatin formation and organization is not well understood, and most studies have focused on the mating-type region in fission yeast, which may not be representative of activities in other genomic regions or organisms. In maintenance of existing heterochromatin regions, RITS forms a complex with siRNAs complementary to the local genes and stably binds local methylated histones, acting co-transcriptionally to degrade any nascent pre-mRNA transcripts that are initiated by RNA polymerase. The formation of such a heterochromatin region, though not its maintenance, is dicer-dependent, presumably because dicer is required to generate the initial complement of siRNAs that target subsequent transcripts. Heterochromatin maintenance has been suggested to function as a self-reinforcing feedback loop, as new siRNAs are formed from the occasional nascent transcripts by RdRP for incorporation into local RITS complexes. The relevance of observations from fission yeast mating-type regions and centromeres to mammals is not clear, as heterochromatin maintenance in mammalian cells may be independent of the components of the RNAi pathway.

Crosstalk with RNA editing

The type of RNA editing that is most prevalent in higher eukaryotes converts adenosine nucleotides into inosine in dsRNAs via the enzyme adenosine deaminase (ADAR). It was originally proposed in 2000 that the RNAi and A→I RNA editing pathways might compete for a common dsRNA substrate. Indeed, some pre-miRNAs do undergo A→I RNA editing, and this mechanism may regulate the processing and expression of mature miRNAs. Furthermore, at least one mammalian ADAR can sequester siRNAs from RNAi pathway components. Further support for this model comes from studies on ADAR-null *C. elegans* strains indicating that A→I RNA editing may counteract RNAi silencing of endogenous genes and transgenes.

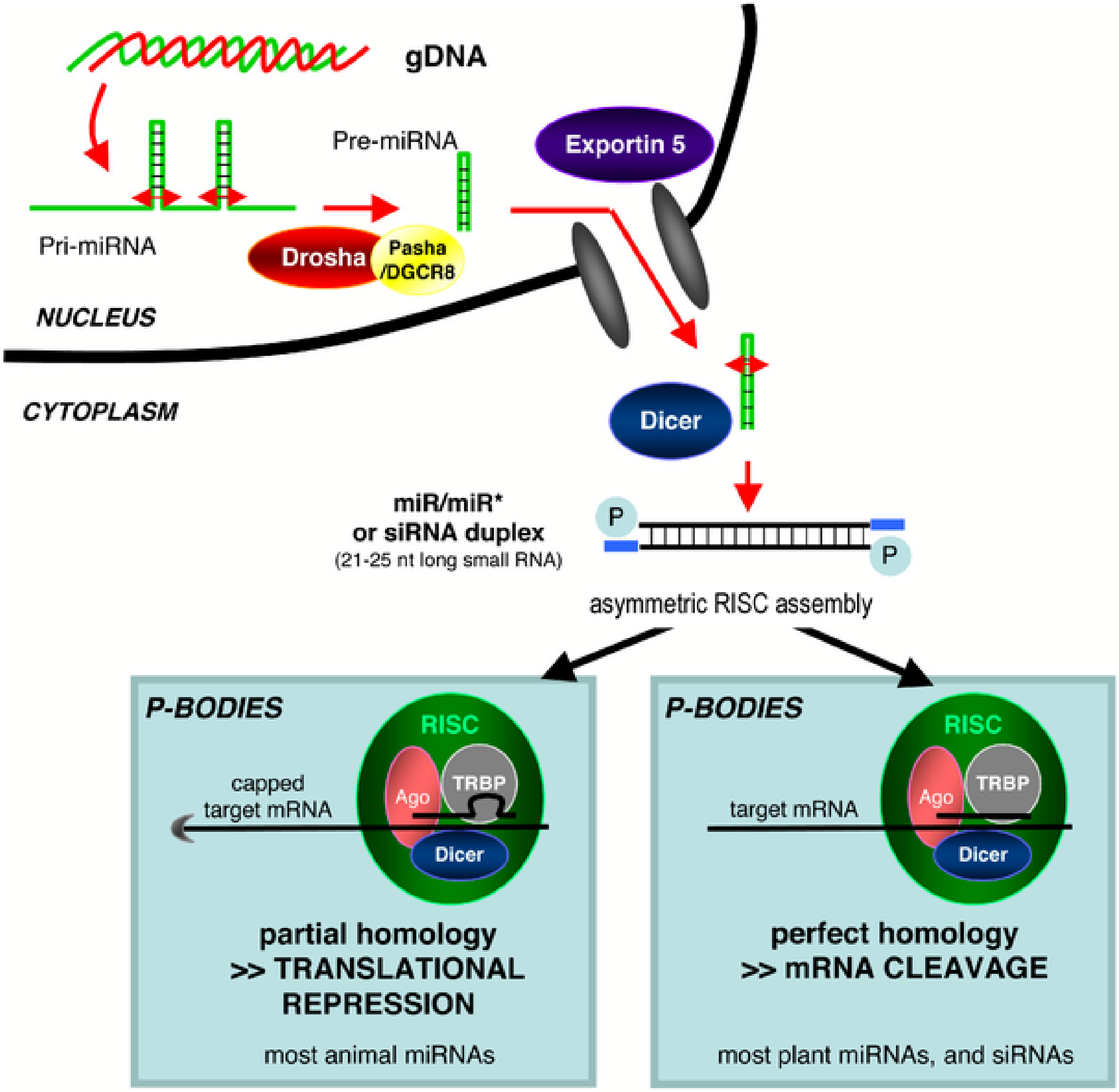


Illustration of the major differences between plant and animal gene silencing. Natively expressed microRNA or exogenous small interfering RNA is processed by dicer and integrated into the RISC complex, which mediates gene silencing.

Variation among organisms

Organisms vary in their ability to take up foreign dsRNA and use it in the RNAi pathway. The effects of RNA interference can be both systemic and heritable in plants and *C. elegans*, although not in *Drosophila* or mammals. In plants, RNAi is thought to propagate by the transfer of siRNAs between cells through plasmodesmata (channels in the cell walls that enable communication and transport). The heritability comes from methylation of promoters targeted by RNAi; the new methylation pattern is copied in each new generation of the cell. A broad general distinction between plants and animals lies in the targeting of endogenously produced miRNAs; in plants, miRNAs are usually perfectly or nearly perfectly complementary to their target genes and induce direct mRNA cleavage by RISC, while animals' miRNAs tend to be more divergent in sequence and induce translational repression. This translational effect may be produced by inhibiting the interactions of translation initiation factors with the messenger RNA's polyadenine tail.

Some eukaryotic protozoa such as *Leishmania major* and *Trypanosoma cruzi* lack the RNAi pathway entirely. Most or all of the components are also missing in some fungi, most notably the model organism *Saccharomyces cerevisiae*. A recent study however reveals the presence of RNAi in other budding yeast species such as *Saccharomyces castellii* and *Candida albicans*, further demonstrating that inducing two RNAi-related proteins from *S. castellii* facilitates RNAi in *S. cerevisiae*. That certain ascomycetes and basidiomycetes are missing RNA interference pathways indicates that proteins required for RNA silencing have been lost independently from many fungal lineages, possibly due to the evolution of a novel pathway with similar function, or to the lack of selective advantage in certain niches.

Related prokaryotic systems

Gene expression in prokaryotes is influenced by an RNA-based system similar in some respects to RNAi. Here, RNA-encoding genes control mRNA abundance or translation by producing a complementary RNA that binds to an mRNA by base pairing. However these regulatory RNAs are not generally considered to be analogous to miRNAs because the dicer enzyme is not involved. It has been suggested that CRISPR interference systems in prokaryotes are analogous to eukaryotic RNA interference systems, although none of the protein components are orthologous.

Biological functions

Immunity

RNA interference is a vital part of the immune response to viruses and other foreign genetic material, especially in plants where it may also prevent self-propagation by transposons. Plants such as *Arabidopsis thaliana* express multiple dicer homologs that are specialized to react differently when the plant is exposed to different types of viruses. Even before the RNAi pathway was fully understood, it was known that induced gene silencing in plants could spread throughout the plant in a systemic effect, and could be transferred from stock to scion plants via grafting. This phenomenon has since been recognized as a feature of the plant adaptive immune system, and allows the entire plant to respond to a virus after an initial localized encounter. In response, many plant viruses have evolved elaborate mechanisms that suppress the RNAi response in plant cells. These include viral proteins that bind short double-stranded RNA fragments with single-stranded overhang ends, such as those produced by the action of dicer. Some plant genomes also express endogenous siRNAs in response to infection by specific types of bacteria. These effects may be part of a generalized response to pathogens that downregulates any metabolic processes in the host that aid the infection process.

Although animals generally express fewer variants of the dicer enzyme than plants, RNAi in some animals has also been shown to produce an antiviral response. In both juvenile and adult *Drosophila*, RNA interference is important in antiviral innate immunity and is active against pathogens such as Drosophila X virus. A similar role in immunity may operate in *C. elegans*, as argonaute proteins are upregulated in response to viruses and worms that overexpress components of the RNAi pathway are resistant to viral infection.

The role of RNA interference in mammalian innate immunity is poorly understood, and relatively little data is available. However, the existence of viruses that encode genes able to suppress the RNAi response in mammalian cells may be evidence in favour of an RNAi-dependent mammalian immune response. However, this hypothesis of RNAi-mediated immunity in mammals has been challenged as poorly substantiated. Alternative functions for RNAi in mammalian viruses also exist, such as miRNAs expressed by the herpes virus that may act as heterochromatin organization triggers to mediate viral latency.

Downregulation of genes

Endogenously expressed miRNAs, including both intronic and intergenic miRNAs, are most important in translational repression and in the regulation of development, especially on the timing of morphogenesis and the maintenance of undifferentiated or incompletely differentiated cell types such as stem cells. The role of endogenously expressed miRNA in downregulating gene expression was first described in *C. elegans* in 1993. In plants this function was discovered when the "JAW microRNA" of *Arabidopsis* was shown to be involved in the regulation of several genes that control plant shape. In plants, the majority of genes regulated by miRNAs are transcription factors; thus miRNA activity is particularly wide-ranging and regulates entire gene networks during development by modulating the expression of key regulatory genes, including transcription factors as well as F-box proteins. In many organisms, including humans, miRNAs have also been linked to the formation of tumors and dysregulation of the cell cycle. Here, miRNAs can function as both oncogenes and tumor suppressors.

Upregulation of genes

RNA sequences (siRNA and miRNA) that are complementary to parts of a promoter can increase gene transcription, a phenomenon dubbed RNA activation. Part of the mechanism for how these RNA upregulate genes is known: dicer and argonaute are involved, and there is histone demethylation.

Evolution

Based on parsimony-based phylogenetic analysis, the most recent common ancestor of all eukaryotes most likely already possessed an early RNA interference pathway; the absence of the pathway in certain eukaryotes is thought to be a derived characteristic. This ancestral RNAi system probably contained at least one dicer-like protein, one argonaute, one PIWI protein, and an RNA-dependent RNA polymerase that may have also played other cellular roles. A large-scale comparative genomics study likewise indicates that the eukaryotic crown group already possessed these components, which may then have had closer functional associations with generalized RNA degradation systems such as the exosome. This study also suggests that the RNA-binding argonaute protein family, which is shared among eukaryotes, most archaea, and at least some bacteria (such as *Aquifex aeolicus*), is homologous to and originally evolved from components of the translation initiation system.

The ancestral function of the RNAi system is generally agreed to have been immune defense against exogenous genetic elements such as transposons and viral genomes. Related functions such as histone modification may have already been present in the ancestor of modern eukaryotes, although other functions such as regulation of development by miRNA are thought to have evolved later.

RNA interference genes, as components of the antiviral innate immune system in many eukaryotes, are involved in an evolutionary arms race with viral genes. Some viruses have evolved mechanisms for suppressing the RNAi response in their host cells, an effect that has been noted particularly for plant viruses. Studies of evolutionary rates in *Drosophila* have shown that genes in the RNAi pathway are subject to strong directional selection and are among the fastest-evolving genes in the *Drosophila* genome.

Technological applications

Gene knockdown

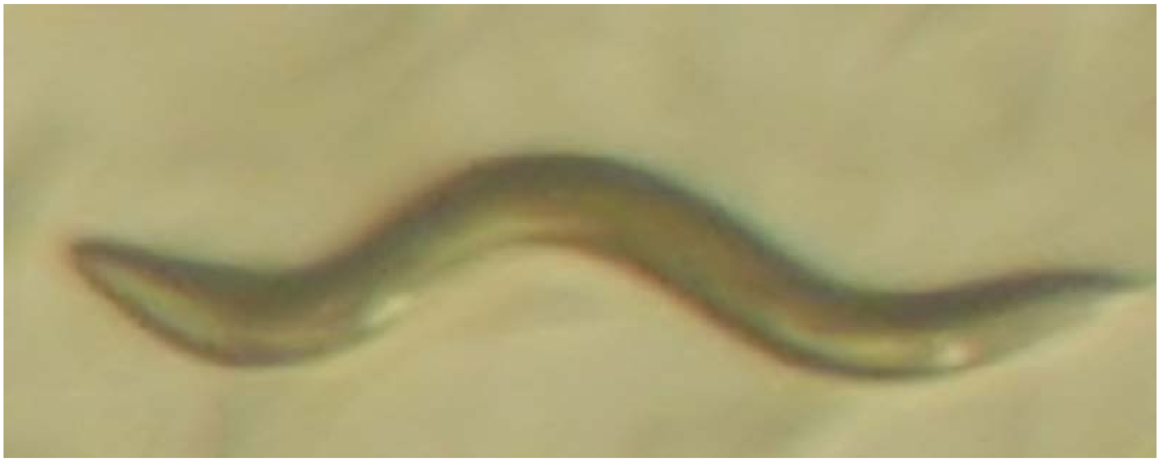
The RNA interference pathway is often exploited in experimental biology to study the function of genes in cell culture and *in vivo* in model organisms. Double-stranded RNA is synthesized with a sequence complementary to a gene of interest and introduced into a cell or organism, where it is recognized as exogenous genetic material and activates the RNAi pathway. Using this mechanism, researchers can cause a drastic decrease in the expression of a targeted gene. Studying the effects of this decrease can show the physiological role of the gene product. Since RNAi may not totally abolish expression of the gene, this technique is sometimes referred as a "knockdown", to distinguish it from "knockout" procedures in which expression of a gene is entirely eliminated.

Extensive efforts in computational biology have been directed toward the design of successful dsRNA reagents that maximize gene knockdown but minimize "off-target" effects. Off-target effects arise when an introduced RNA has a base sequence that can pair with and thus reduce the expression of multiple genes at a time. Such problems occur more frequently when the dsRNA contains repetitive sequences. It has been estimated from studying the genomes of *H. sapiens*, *C. elegans*, and *S. pombe* that about 10% of possible siRNAs will have substantial off-target effects. A multitude of software tools have been developed implementing algorithms for the design of general, mammal-specific, and virus-specific siRNAs that are automatically checked for possible cross-reactivity.

Depending on the organism and experimental system, the exogenous RNA may be a long strand designed to be cleaved by dicer, or short RNAs designed to serve as siRNA substrates. In most mammalian cells, shorter RNAs are used because long double-stranded RNA molecules induce the mammalian interferon response, a form of innate immunity that reacts nonspecifically to foreign genetic material. Mouse oocytes and cells from early mouse embryos lack this reaction to exogenous dsRNA and are therefore a common model system for studying gene-knockdown effects in mammals. Specialized laboratory techniques have also been developed to improve the utility of RNAi in mammalian systems by avoiding the direct introduction of siRNA, for example, by stable transfection with a plasmid encoding the appropriate sequence from which siRNAs can be transcribed, or by more elaborate lentiviral vector systems allowing the inducible activation or deactivation of transcription, known as *conditional RNAi*.

Functional genomics

A normal adult *Drosophila* fly, a common model organism used in RNAi experiments



An adult *C. elegans* worm, grown under RNAi suppression of a nuclear hormone receptor involved in desaturase regulation. These worms have abnormal fatty acid metabolism but are viable and fertile.

Most functional genomics applications of RNAi in animals have used *C. elegans* and *Drosophila*, as these are the common model organisms in which RNAi is most effective. *C. elegans* is particularly useful for RNAi research for two reasons: firstly, the effects of the gene silencing are generally heritable, and secondly because delivery of the dsRNA is extremely simple. Through a mechanism whose details are poorly understood, bacteria such as *E. coli* that carry the desired dsRNA can be fed to the worms and will transfer their RNA payload to the worm via the intestinal tract. This "delivery by feeding" is just as effective at inducing gene silencing as more costly and time-consuming delivery methods, such as soaking the worms in dsRNA solution and injecting dsRNA into the gonads. Although delivery is more difficult in most other organisms, efforts are also underway to undertake large-scale genomic screening applications in cell culture with mammalian cells.

Approaches to the design of genome-wide RNAi libraries can require more sophistication than the design of a single siRNA for a defined set of experimental conditions. Artificial neural networks are frequently used to design siRNA libraries and to predict their likely efficiency at gene knockdown. Mass genomic screening is widely seen as a promising method for genome annotation and has triggered the development of high-throughput screening methods based on microarrays. However, the utility of these screens and the ability of techniques developed on model organisms to generalize to even closely related species has been questioned, for example from *C. elegans* to related parasitic nematodes.

Functional genomics using RNAi is a particularly attractive technique for genomic mapping and annotation in plants because many plants are polyploid, which presents substantial challenges for more traditional genetic engineering methods. For example, RNAi has been successfully used for functional genomics studies in bread wheat (which is hexaploid) as well as more common plant model systems *Arabidopsis* and maize.

Medicine

It may be possible to exploit RNA interference in therapy. Although it is difficult to introduce long dsRNA strands into mammalian cells due to the interferon response, the use of short interfering RNA mimics has been more successful. Among the first applications to reach clinical trials were in the treatment of macular degeneration and respiratory syncytial virus, RNAi has also been shown to be effective in the reversal of induced liver failure in mouse models.

Other proposed clinical uses center on antiviral therapies, including topical microbicide treatments that use RNAi to treat infection (at Harvard University Medical School; in mice, so far) by herpes simplex virus type 2 and the inhibition of viral gene expression in cancerous cells, knockdown of host receptors and coreceptors for HIV, the silencing of hepatitis A and hepatitis B genes, silencing of influenza gene expression, and inhibition of measles viral replication. Potential treatments for neurodegenerative diseases have also been proposed, with particular attention being paid to the polyglutamine diseases such as Huntington's disease. RNA interference is also often seen as a promising way to treat cancer by silencing genes differentially upregulated in tumor cells or genes involved in cell division. A key area of research in the use of RNAi for clinical applications is the development of a safe delivery method, which to date has involved mainly viral vector systems similar to those suggested for gene therapy.

Despite the proliferation of promising cell culture studies for RNAi-based drugs, some concern has been raised regarding the safety of RNA interference, especially the potential for "off-target" effects in which a gene with a coincidentally similar sequence to the targeted gene is also repressed. A computational genomics study estimated that the error rate of off-target interactions is about 10%. One major study of liver disease in mice led to high death rates in the experimental animals, suggested by researchers to be the result of "oversaturation" of the dsRNA pathway, due to the use of shRNAs that have to be processed in the nucleus and exported to the cytoplasm using an active mechanism. All these are considerations that are under active investigation, to reduce their impact in the potential therapeutic applications for RNAi.

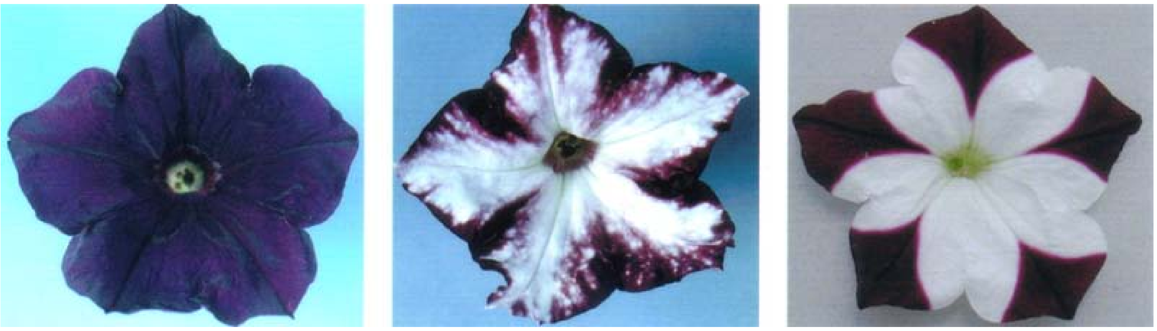
RNA interference-based applications are being developed to target persistent HIV-1 infection. Viruses like HIV-1 are particularly difficult targets for RNAi-attack because they are escape-prone, which requires combinatorial RNAi strategies to prevent viral escape. The future of antiviral RNAi therapeutics is very promising, but it remains of critical importance to include many controls in pre-clinical test models to unequivocally demonstrate sequence-specific action of the RNAi inducers.

Biotechnology

RNA interference has been used for applications in biotechnology, particularly in the engineering of food plants that produce lower levels of natural plant toxins. Such techniques take advantage of the stable and heritable RNAi phenotype in plant stocks. For example, cotton seeds are rich in dietary protein but naturally contain the toxic terpenoid product gossypol, making them unsuitable for human consumption. RNAi has been used to produce cotton stocks whose seeds contain reduced levels of delta-cadinene synthase, a key enzyme in gossypol production, without affecting the enzyme's production in other parts of the plant, where gossypol is important in preventing damage from plant pests. Similar efforts have been directed toward the reduction of the cyanogenic natural product linamarin in cassava plants.

Although no plant products that use RNAi-based genetic engineering have yet passed the experimental stage, development efforts have successfully reduced the levels of allergens in tomato plants and decreased the precursors of likely carcinogens in tobacco plants. Other plant traits that have been engineered in the laboratory include the production of non-narcotic natural products by the opium poppy, resistance to common plant viruses, and fortification of plants such as tomatoes with dietary antioxidants. Previous commercial products, including the Flavr Savr tomato and two cultivars of ringspot-resistant papaya, were originally developed using antisense technology but likely exploited the RNAi pathway.

History and discovery



Example petunia plants in which genes for pigmentation are silenced by RNAi. The left plant is wild-type; the right plants contain transgenes that induce suppression of both transgene and endogenous gene expression, giving rise to the unpigmented white areas of the flower.

The discovery of RNAi was preceded first by observations of transcriptional inhibition by antisense RNA expressed in transgenic plants, and more directly by reports of unexpected outcomes in experiments performed by plant scientists in the United States and the Netherlands in the early 1990s. In an attempt to alter flower colors in petunias, researchers introduced additional copies of a gene encoding chalcone synthase, a key enzyme for flower pigmentation into petunia plants of normally pink or violet flower color. The overexpressed gene was expected to result in darker flowers, but instead produced less pigmented, fully or partially white flowers, indicating that the activity of chalcone synthase had been substantially decreased; in fact, both the endogenous genes and the transgenes were downregulated in the white flowers. Soon after, a related event termed *quelling* was noted in the fungus *Neurospora crassa*, although it was not immediately recognized as related. Further investigation of the phenomenon in plants indicated that the downregulation was due to post-transcriptional inhibition of gene expression via an increased rate of mRNA degradation. This phenomenon was called *co-suppression of gene expression*, but the molecular mechanism remained unknown.

Craig Mello at the 2006 Nobel Prize lecture

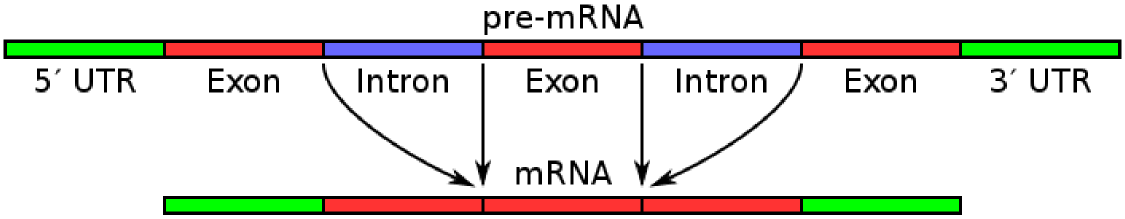
Not long after, plant virologists working on improving plant resistance to viral diseases observed a similar unexpected phenomenon. While it was known that plants expressing virus-specific proteins showed enhanced tolerance or resistance to viral infection, it was not expected that plants carrying only short, non-coding regions of viral RNA sequences would show similar levels of protection. Researchers believed that viral RNA produced by transgenes could also inhibit viral replication. The reverse experiment, in which short sequences of plant genes were introduced into viruses, showed that the targeted gene was suppressed in an infected plant. This phenomenon was labeled "virus-induced gene silencing" (VIGS), and the set of such phenomena were collectively called post transcriptional gene silencing.

After these initial observations in plants, many laboratories around the world searched for the occurrence of this phenomenon in other organisms. Craig C. Mello and Andrew Fire's 1998 *Nature* paper reported a potent gene silencing effect after injecting double stranded RNA into *C. elegans*. In investigating the regulation of muscle protein production, they observed that neither mRNA nor antisense RNA injections had an effect on protein production, but double-stranded RNA successfully silenced the targeted gene. As a result of this work, they coined the term *RNAi*. Fire and Mello's discovery was particularly notable because it represented the first identification of the causative agent for the phenomenon. Fire and Mello were awarded the Nobel Prize in Physiology or Medicine in 2006 for their work.

Chapter- 9

RNA Splicing

In molecular biology, **splicing** is a modification of an RNA after transcription, in which introns are removed and exons are joined. This is needed for the typical eukaryotic messenger RNA before it can be used to produce a correct protein through translation. For many eukaryotic introns, splicing is done in a series of reactions which are catalyzed by the spliceosome, a complex of small nuclear ribonucleoproteins (snRNPs), but there are also self-splicing introns.



Simple illustration of exons and introns in pre-mRNA and the formation of mature mRNA by splicing. The UTRs are non-coding parts of exons at the ends of the mRNA.

Splicing pathways

Several methods of RNA splicing occur in nature: the type of splicing depends on the structure of the spliced intron and the catalysts required for splicing to occur.

Spliceosomal introns

Spliceosomal introns often reside in eukaryotic protein-coding genes. Within the intron, a 3' splice site, 5' splice site, and branch site are required for splicing. The 5' splice site or splice donor site includes an almost invariant sequence GU at the 5' end of the intron, within a larger, less highly conserved consensus region. The 3' splice site or splice acceptor site terminates the intron with an almost invariant AG sequence. Upstream (5'-ward) from the AG there is a region high in pyrimidines (C and U), or polypyrimidine tract. Upstream from the polypyrimidine tract is the branch point, which includes an adenine nucleotide. Point mutations in the underlying DNA or errors during transcription can activate a "cryptic splice site" in part of the transcript that usually is not spliced. This results in a mature messenger RNA with a missing section of an exon. In this way a point mutation, which usually only affects a single amino acid, can manifest as a deletion in the final protein.

Spliceosome formation and activity

Splicing is catalyzed by the spliceosome which is a large RNA-protein complex composed of five small nuclear ribonucleoproteins (snRNPs, pronounced 'snurps'). The RNA components of snRNPs interact with the intron and may be involved in catalysis. Two types of spliceosomes have been identified (the major and minor) which contain different snRNPs.

- Major

  The major spliceosome splices introns containing GU at the 5' splice site and AG at the 3' splice site. It is composed of the U1, U2, U4, U5, and U6 snRNPs and is active in the nucleus. In addition, a number of proteins including U2AF and SF1 are required for the assembly of the spliceosome.

  - E Complex-U1 binds to the GU sequence at the 5' splice site, along with accessory proteins/enzymes ASF/SF2, U2AF (binds at the Py-AG site), SF1/BBP (BBP=Branch Binding Protein);
  - A Complex-U2 binds to the branch site and ATP is hydrolyzed;
  - B1 Complex-U5/U4/U6 trimer binds, and the U5 binds exons at the 5' site, with U6 binding to U2;
  - B2 Complex-U1 is released, U5 shifts from exon to intron and the U6 binds at the 5' splice site;
  - C1 Complex-U4 is released, U6/U2 catalyzes transesterification, that make 5'end of introns ligate to the A on intron and from a lariat,U5 binds exon at 3' splice site, and the 5' site is cleaved, resulting in the formation of the lariat;
  - C2 Complex-U2/U5/U6 remain bound to the lariat, and the 3' site is cleaved and exons are ligated using ATP hydrolysis. The spliced RNA is released and the lariat debranches.

  This type of splicing is termed *canonical splicing* or termed the *lariat pathway*, which accounts for more than 99% of splicing. By contrast, when the intronic flanking sequences do not follow the GU-AG rule, *noncanonical splicing* is said to occur (see "minor spliceosome" below).

- Minor

  The minor spliceosome is very similar to the major spliceosome, however it splices out rare introns with different splice site sequences. While the minor and major spliceosomes contain the same U5 snRNP, the minor spliceosome has different, but functionally analogous snRNPs for U1, U2, U4, and U6, which are respectively called U11, U12, U4atac, and U6atac. Like the major spliceosome, it is only found in the nucleus.

- Trans-splicing

  Trans-splicing is a form of splicing that joins two exons that are not within the same RNA transcript.

Self-splicing

Self-splicing occurs for rare introns that form a ribozyme, performing the functions of the spliceosome by RNA alone. There are three kinds of self-splicing introns, *Group I*, *Group II* and *Group III*. Group I and II introns perform splicing similar to the spliceosome without requiring any protein. This similarity suggests that Group I and II introns may be evolutionarily related to the spliceosome. Self-splicing may also be very ancient, and may have existed in an RNA world present before protein. Although the two splicing mechanisms described below do not require any proteins to occur, 5 additional RNA molecules and over 50 proteins are used and hydrolyzes many ATP molecules. The splicing mechanisms use ATP in order to accurately splice mRNA's. If the cell were to not use any ATP's, the process would be highly inaccurate and many mistakes would occur.

Two transesterifications characterize the mechanism in which group I introns are spliced:

1. 3'OH of a free guanine nucleoside (or one located in the intron) or a nucleotide cofactor (GMP, GDP, GTP) attacks phosphate at the 5' splice site.
2. 3'OH of the 5'exon becomes a nucleophile and the second transesterification results in the joining of the two exons.

The mechanism in which group II introns are spliced (two transesterification reaction like group I introns) is as follows:

1. The 2'OH of a specific adenosine in the intron attacks the 5' splice site, thereby forming the *lariat*
2. The 3'OH of the 5' exon triggers the second transesterification at the 3' splice site thereby joining the exons together.

tRNA splicing

tRNA (also tRNA-like) splicing is another rare form of splicing that usually occurs in tRNA. The splicing reaction involves a different biochemistry than the spliceomsomal and self-splicing pathways. Ribonucleases cleave the RNA and ligases join the exons together.

Evolution

Splicing occurs in all the kingdoms or domains of life, however, the extent and types of splicing can be very different between the major divisions. Eukaryotes splice many protein-coding messenger RNAs and some non-coding RNAs. Prokaryotes, on the other hand, splice rarely and mostly non-coding RNAs. Another important difference between these two groups of organisms is that prokaryotes completely lack the spliceosomal pathway.

Because spliceosomal introns are not conserved in all species, there is debate concerning when spliceosomal splicing evolved. Two models have been proposed: the intron late and intron early models.

Splicing Diversity

	Eukaryotes	**Prokaryotes**
Spliceosomal	+	-
Self-splicing	+	+
tRNA	+	+

Biochemical mechanism

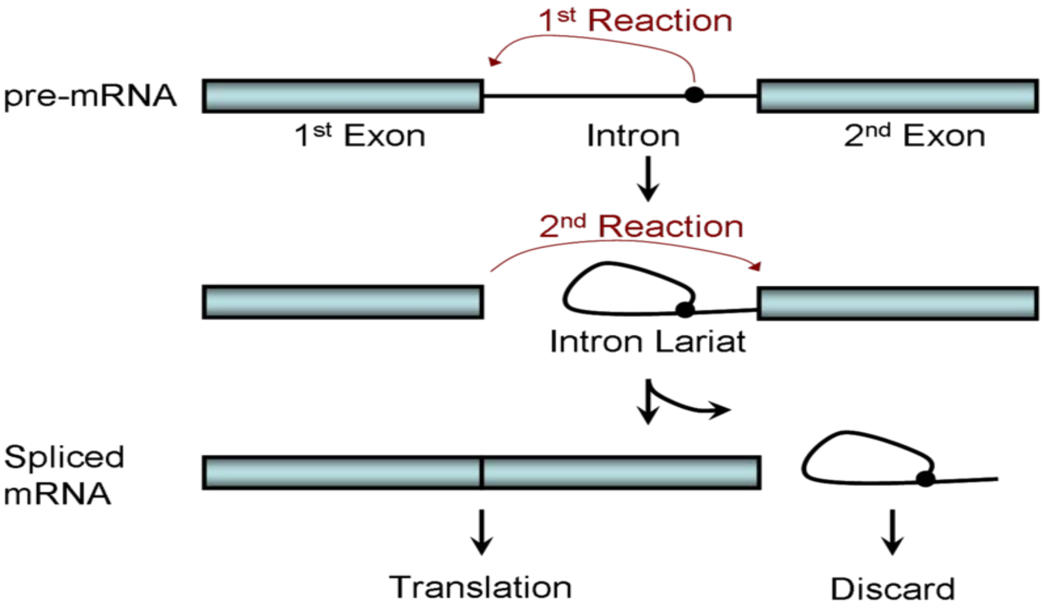


Diagram illustrating the two-step biochemistry of splicing

Spliceosomal splicing and self-splicing involves a two-step biochemical process. Both steps involve transesterification reactions that occur between RNA nucleotides. tRNA splicing, however, is an exception and does not occur by transesterification.

Spliceosomal and self-splicing transesterification reactions occur via two sequential transesterification reactions. First, the 2'OH of a specific *branch-point* nucleotide within the intron that is defined during spliceosome assembly performs a nucleophilic attack on the first nucleotide of the intron at the 5' splice site forming the *lariat intermediate*. Second, the 3'OH of the released 5' exon then performs a nucleophilic attack at the last nucleotide of the intron at the 3' splice site thus joining the exons and releasing the intron lariat.

Alternative splicing

In many cases, the splicing process can create a range of unique proteins by varying the exon composition of the same messenger RNA. This phenomenon is then called alternative splicing. Alternative splicing can occur in many ways. Exons can be extended or skipped, or introns can be retained.

Experimental manipulation of splicing

Splicing events can be experimentally altered by binding steric-blocking antisense oligos such as Morpholinos or Peptide nucleic acids to snRNP binding sites, to the branchpoint nucleotide that closes the lariat, or to splice-regulatory element binding sites.

Splicing errors

Common errors:

- Mutation of a splice site resulting in loss of function of that site. Results in exposure of a premature stop codon, loss of an exon, or inclusion of an intron.
- Mutation of a splice site reducing specificity. May result in variation in the splice location, causing insertion or deletion of amino acids, or most likely, a loss of the reading frame.
- Displacement of a splice site, leading to inclusion or exclusion of more RNA than expected, resulting in longer or shorter exons.

Many splicing errors are safeguarded by a cellular quality control mechanism termed Nonsense-mediated mRNA decay [NMD].

Protein splicing

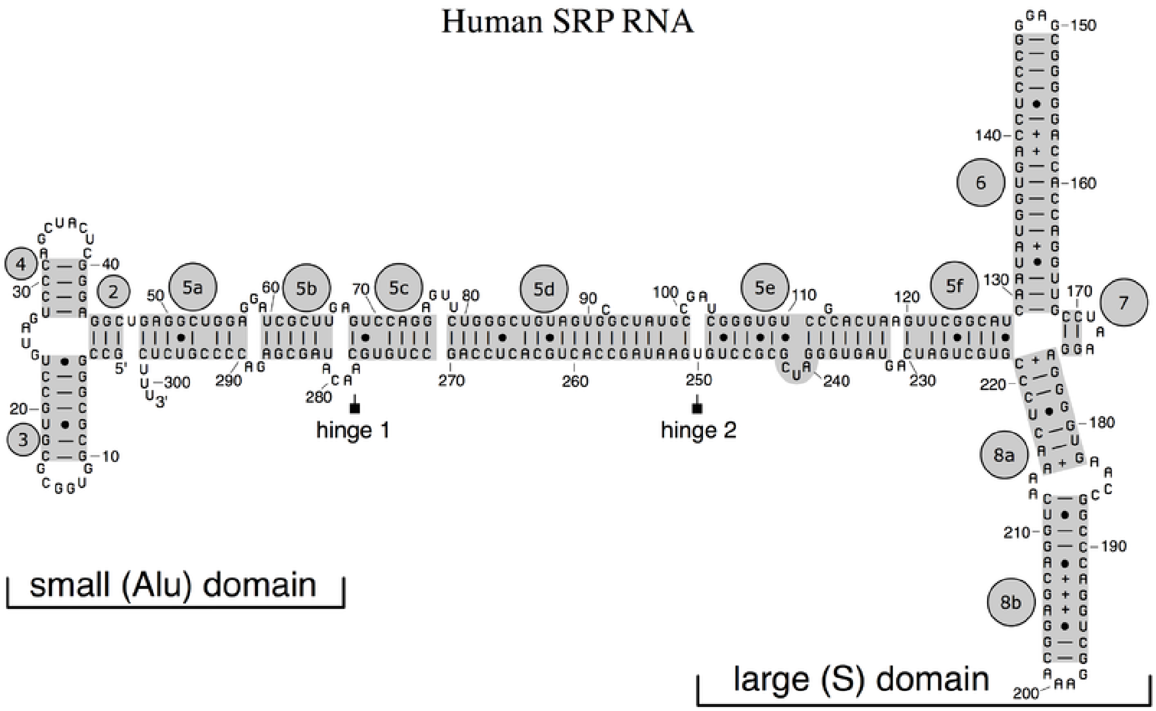
Not only pre-mRNA but also proteins can undergo splicing. Although the biomolecular mechanisms are different, the principle is the same, that parts of the protein, called inteins instead of introns, are removed. The remaining parts, called exteins instead of exons, are fused together. Protein splicing has been observed in all sorts of organisms, including bacteria, archaea, plants, yeast and human.

Chapter- 10

Signal Recognition Particle RNA

RNA, 7SL, cytoplasmic 1

Identifiers		
Symbols	RN7SL1; RN7SL; 7L1a; RNSRP1; 7SL	
External IDs	OMIM: 612177 GeneCards: RN7SL1 Gene	
Orthologs		
Species	Human	Mouse
Entrez	6029	n/a
Ensembl	ENSG00000222619	ENSMUSG00000065062
UniProt	n/a	n/a
RefSeq (mRNA)	NR_002715	n/a
RefSeq (protein)	n/a	n/a
Location (UCSC)	Chr 14: 49.4 - 49.4 Mb	Chr 12: 70.46 - 70.46 Mb



Secondary structure of the human SRP RNA. Helices are numbered from 2 to 8. Helical sections in gray are named with lower case letters. Residues are numbered in increments of ten. The 5'- and 3'-ends are indicated. Highlighted are the two hinges and the small (Alu) and large (S, "specific") domain of the SRP RNA.

The **signal recognition particle RNA**, also known as 7SL, 6S, or 4.5S RNA, is the RNA component of the signal recognition particle (SRP) ribonucleoprotein complex. SRP is a universally conserved ribonucleoprotein that directs the traffic of proteins within the cell and allows them to be secreted. The SRP RNA, together with one or more SRP proteins contributes to the binding and release of the signal peptide. The RNA and protein components of this complex are highly conserved but do vary between the different kingdoms of life.

The common SINE family Alu probably originated from a duplication of a 7SL RNA gene.

The eukaryotic SRP consists of a 300-nucleotide 7S RNA and six proteins: SRPs 72, 68, 54, 19, 14, and 9. Archaeal SRP consists of a 7S RNA and homologues of the eukaryotic SRP19 and SRP54 proteins. Eukaryotic and archaeal 7S RNAs have very similar secondary structures.

In most bacteria, the SRP consists of an RNA molecule (4.5S) and the Ffh protein (a homologue of the eukaryotic SRP54 protein). Some Gram-positive bacteria (e.g. *Bacillus subtilis*) have a longer eukaryote-like SRP RNA that includes an Alu domain.

In eukaryotes and archaea, eight helical elements fold into the Alu and S domains, separated by a long linker region. The Alu domain is thought to mediate the peptide chain elongation retardation function of the SRP. The universally conserved helix which interacts with the SRP54 M domain mediates signal sequence recognition. The SRP19-helix 6 complex is thought to be involved in SRP assembly and stabilises helix 8 for SRP54. binding The human genome in particular is known to contain a large amount of SRP RNA related sequence, including Alu repeats.

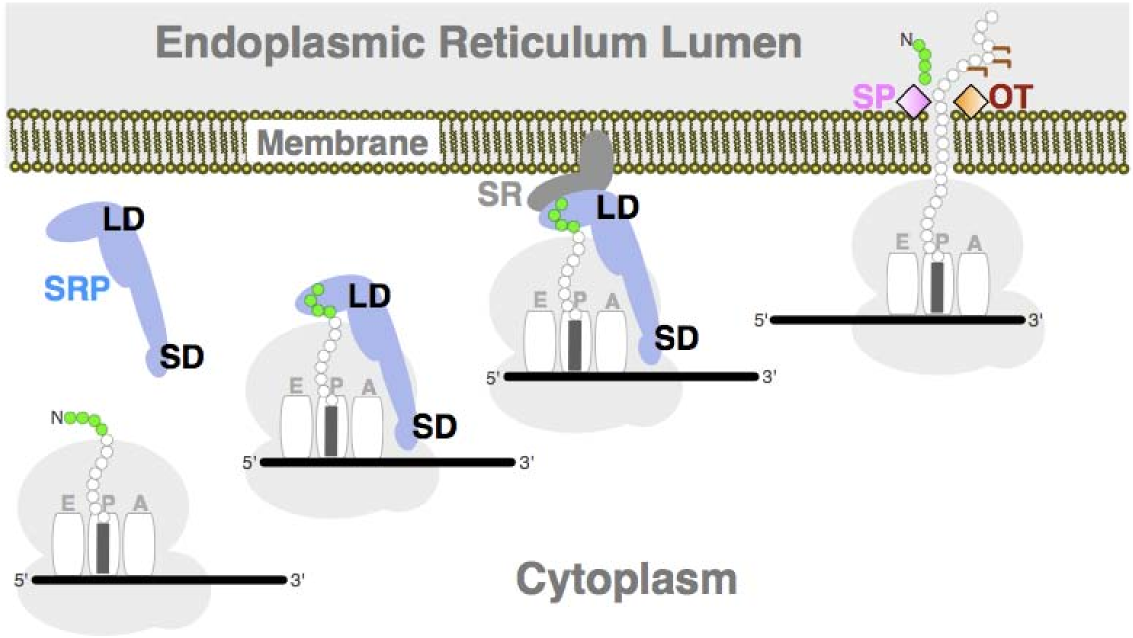
Discovery

SRP RNA was first detected in avian and murine oncogenic RNA (ocorna) virus particles. Subsequently, SRP RNA was found to be a stable component of uninfected HeLa cells where it associated with membrane and polysome fractions. In 1980, cell biologists purified from canine pancreas an 11S "signal recognition protein" (fortuitously also abbreviated "SRP") which promoted the translocation of secretory proteins across the membrane of the endoplasmic reticulum. It was then discovered that SRP contained an RNA component. Comparing the SRP RNA genes from different species revealed helix 8 of the SRP RNA to be highly conserved in all domains of life. The regions near the 5'- and 3'-ends of the mammalian SRP RNA are similar to the dominant Alu family of middle repetitive sequences of the human genome. It is now understood that Alu DNA originated from SRP RNA by excision of the central SRP RNA-specific (S) fragment, followed by reverse transcription and integration into multiple sites of the human chromosomes. SRP RNAs have been identified also in some organelles, for example in the plastid SRPs of many photosynthetic organisms.

Transcription and processing

Eukaryotic SRP RNAs are transcribed from DNA by RNA polymerase III. RNA polymerase III also transcribes the genes for 5S ribosomal RNA, tRNA, 7SK RNA, and U6 spliceosomal RNA. The promoters of the human SRP RNA genes include elements located downstream of the transcriptional start site. Plant SRP RNA promoters contain an upstream stimulatory element (USE) and a TATA box. Yeast SRP RNA genes have a TATA box and intragenic sequences (A- and B-blocks) which likely play a role in transcription. In the bacteria, genes are organized in operons and transcribed by RNA polymerase. The 5'-end of the small (4.5S) SRP RNA of many bacteria is cleaved by RNase P. The ends of the Bacillus subtilis SRP RNA are processed by RNase III. So far, no SRP RNA introns have been observed.

Function



The classical function of SRP in translation-translocation. A membrane separates the cytosol from the endoplasmic reticulum. A ribosome (light gray with A, P, and E sites) synthesizes a protein with a signal peptide (green) encoded by messenger RNA (indicated by a line with 5'- and 3-ends). The elongated SRP (blue), with its large (LD) and small (SD) domains, forms a complex which the membrane-resident SRP receptor (SR). When SRP separates, the protein crosses the membrane through a channel or translocon. The signal peptide may be removed by signal peptide peptidase (SP) and the protein modified by oligosaccharyl transferase (OT).

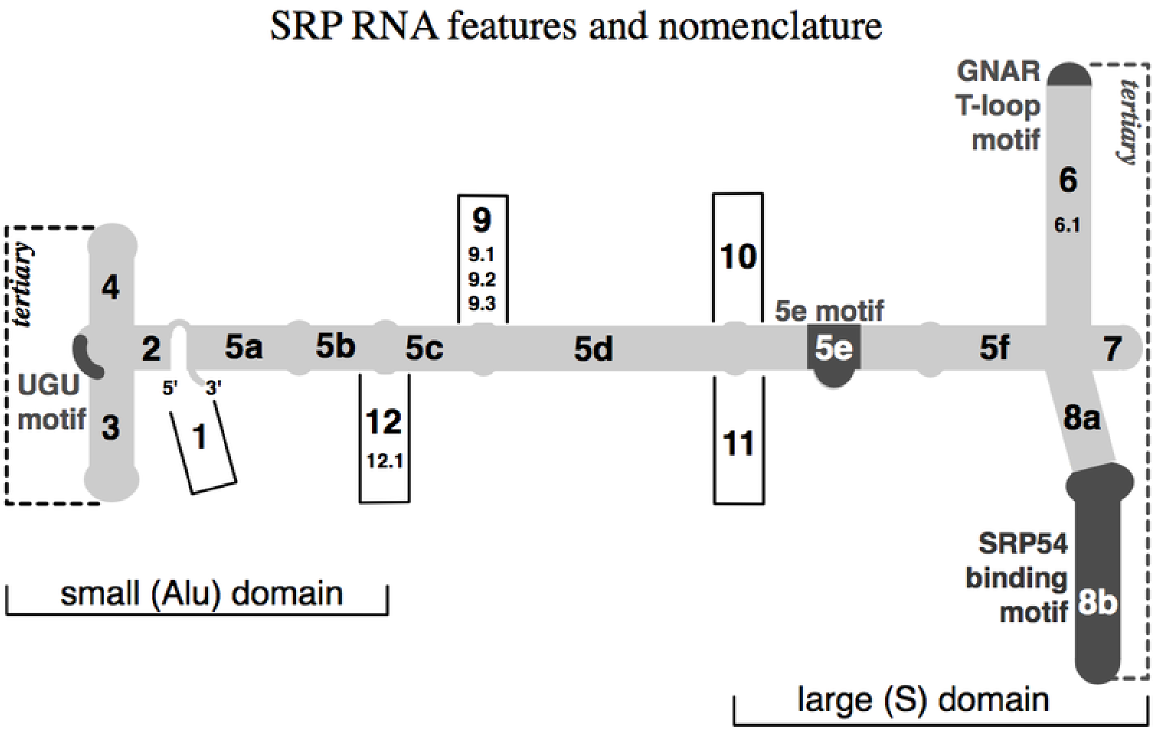
Co-translational translocation

The SRP RNA is an integral part of the small and the large domain of the SRP. The function of the small domain is to delay protein translation until the ribosome-bound SRP has an opportunity to associate with the membrane-resident SRP receptor (SR). Within the large domain, the SRP RNA of the signal peptide-charged SRP promotes the hydrolysis of two guanosine triphosphate (GTP) molecules. This reaction releases the SRP from the SRP receptor and the ribosome, allowing translation to continue and the protein to enter the translocon. The protein transverses the membrane co-translationally (during translation) and enters into another cellular compartment or the extracellular space. In eukaryotes, the target is the membrane of the endoplasmic reticulum (ER). In Archaea, SRP delivers proteins to the plasma membrane. In the bacteria, SRP primarily incorporates proteins into the inner membrane.

Post-translational transport

SRP participates also in the sorting of proteins after their synthesis has been completed (post-translational protein sorting). In eukaryotes, tail-anchored proteins possessing a hydrophobic insertion sequence at their C-terminus are delivered to the endoplasmic reticulum (ER) by the SRP. Similarly, the SRP assists post-translationally in the import of nuclear encoded proteins to the thylakoid membrane of chloroplasts.

Structure



SRP RNA features and nomenclature. The human SRP RNA secondary structure is outlines in light gray and the 5'- and 3'-ends are indicated. Conserved motifs are shown in dark gray. Helices are numbered from 1 to 12, helical sections are designated by lower case letters, and helix insertions by dotted numbers. Tertiary interactions between the apical loops of helices 3 and 4, and between helices 6 and 8 are indicated dotted lines.

In 2005, a nomenclature for all SRP RNAs proposed a numbering system of 12 helices. Helix sections are named with a lower case letter suffix (e.g. 5a). Insertions, or helix "branches" are given dotted numbers (e.g. 9.1 and 12.1).

The SRP RNA spans a wide phylogenetic spectrum with respect to size and the number of its structural features (see the SRP RNA Secondary Structure Examples, below). The smallest functional SRP RNAs have been found in mycoplasma and related species. *Escherichia coli* SRP RNA (also called 4.5S RNA) is composed of 114 nucleotide residues and forms an RNA stem-loop. The gram-positive bacterium *Bacillus subtilis* encodes a larger 6S SRP RNA which resemble the Archaeal homologs but lacks SRP RNA helix 6. Archaeal SRP RNAs possess helices 1 to 8, lack helix 7, and are characterized by a tertiary structure which involves the apical loops of helix 3 and helix 4. The eukaryotic SRP RNAs lack helix 1 and contain a helix 7 of variable size. Some protozoan SRP RNAs have reduced helices 3 and 4. The ascomycota SRP RNAs have an altogether reduced small domain and lack helices 3 and 4. The largest SRP RNAs known to date are found in the yeasts (Saccharomycetes) which acquired helices 9 to 12 as insertions into helix 5, as well as an extended helix 7. Seed plants express numerous highly divergent SRP RNAs.

Motifs

Four conserved features (motifs) have been identified (shown in the Figure in dark gray): the (1) SRP54 binding motif, (2) Helix 6 GNAR tetraloop motif, (3) 5e motif, and (4) UGU(NR) motif.

SRP54 binding

The asymmetric loop between helical sections 8a and 8b and the adjacent base paired 8b section are a prominent property of every SRP RNA. Helical section 8b contains non-Watson-Crick base pairings which contribute to the formation of a flatted minor groove in the RNA suitable for the binding of protein SRP54 (called Ffh in the bacteria). The apical loop of helix 8 contains four, five, or six residues, depending on the species. It has a highly conserved guanosine as the first and an adenosine as the last loop residue. This feature is required for the interaction with the third adenosine residue of the helix 6 GNAR tetraloop motif.

Helix 6 GNAR tetraloop

The SRP RNAs of eukaryotes and Archaea have a GNAR tetraloop (N is for any nucleotide, R is for a purine) in helix 6. Its conserved adenosine residue is important for the binding of protein SRP19. This adenosine makes a tertiary interaction with another adenosine residue located in the apical loop of helix 8.

5e

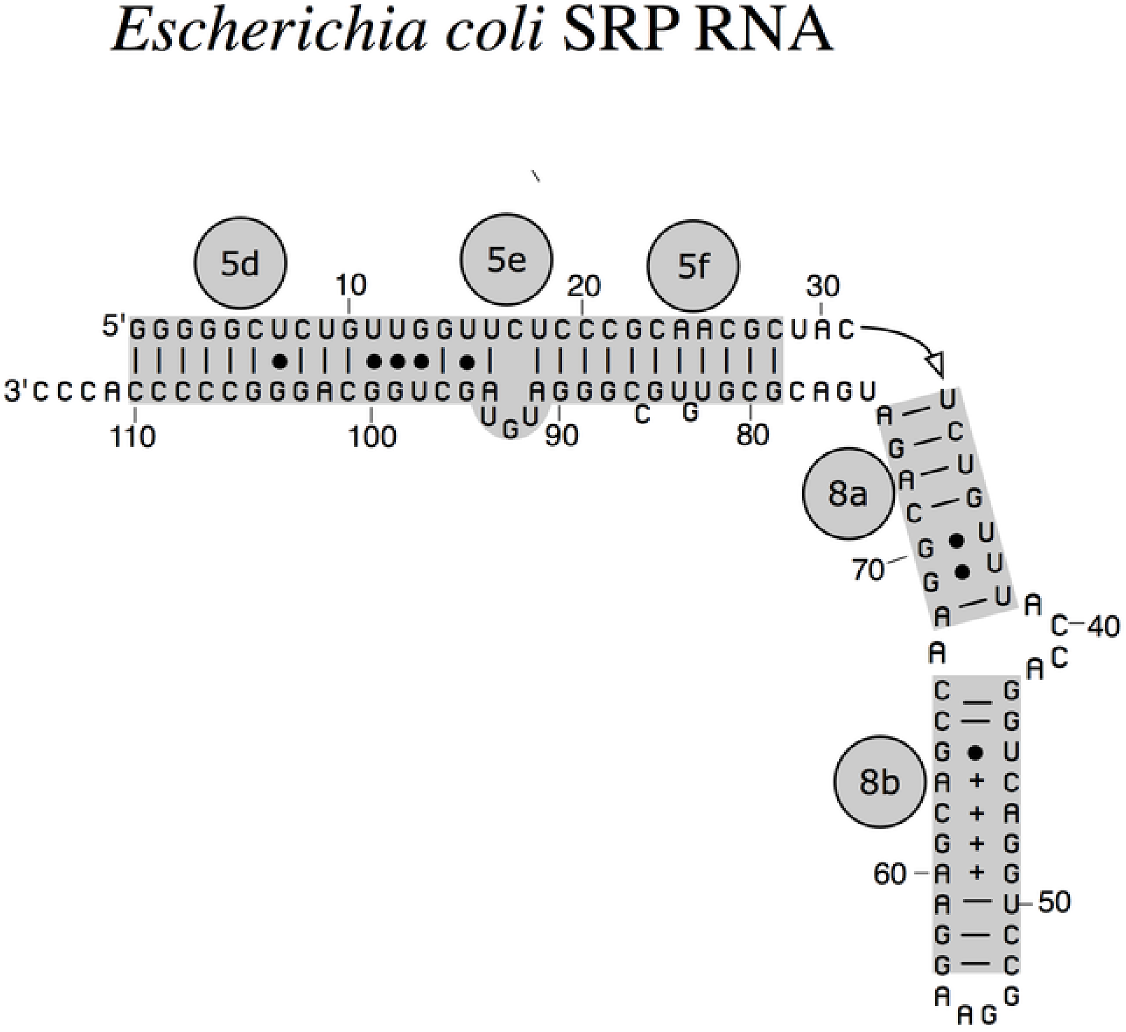
The 11 nucleotides of the 5e motif form four base pairs which are interrupted by a loop of three nucleotides. In the eukaryotes, the first nucleotide of the loop is an adenosine which is needed for the binding of protein SRP72.

UGU(NR)

The UGU(NR) motif connects helices 3 and 4 in the small (Alu) SRP domain. Fungal SRP RNAs lacking helices 3 and 4 contain the motif within the loop of helix 2. It is important in the binding of the SRP9/14 protein heterodimer as part of an RNA U-turn.

Secondary

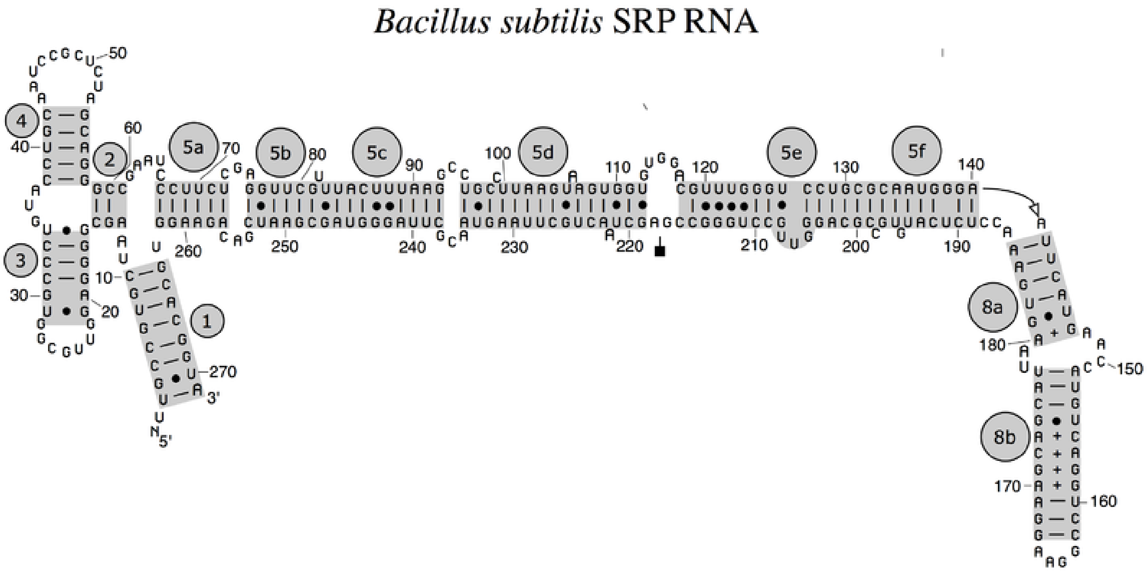
Examples of SRP RNA secondary structures



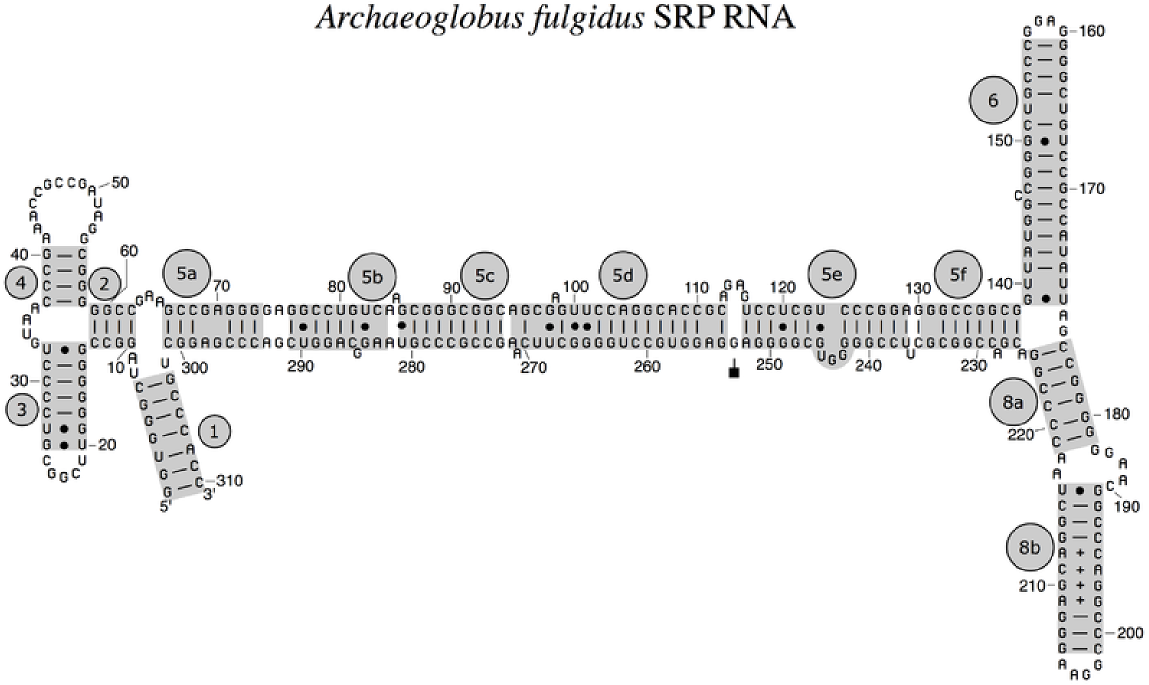
Escherichia coli SRP RNA

5d
5e
5f
8a
8b
10
20
30
40
50
60
70
80
90
100
110
5'GGGGGCUCUGUUGGUUCUCCCGCAACGCUAC
3'CCCACCCCCGGGACGGUCGA AGGGCGUUGCGCAGU

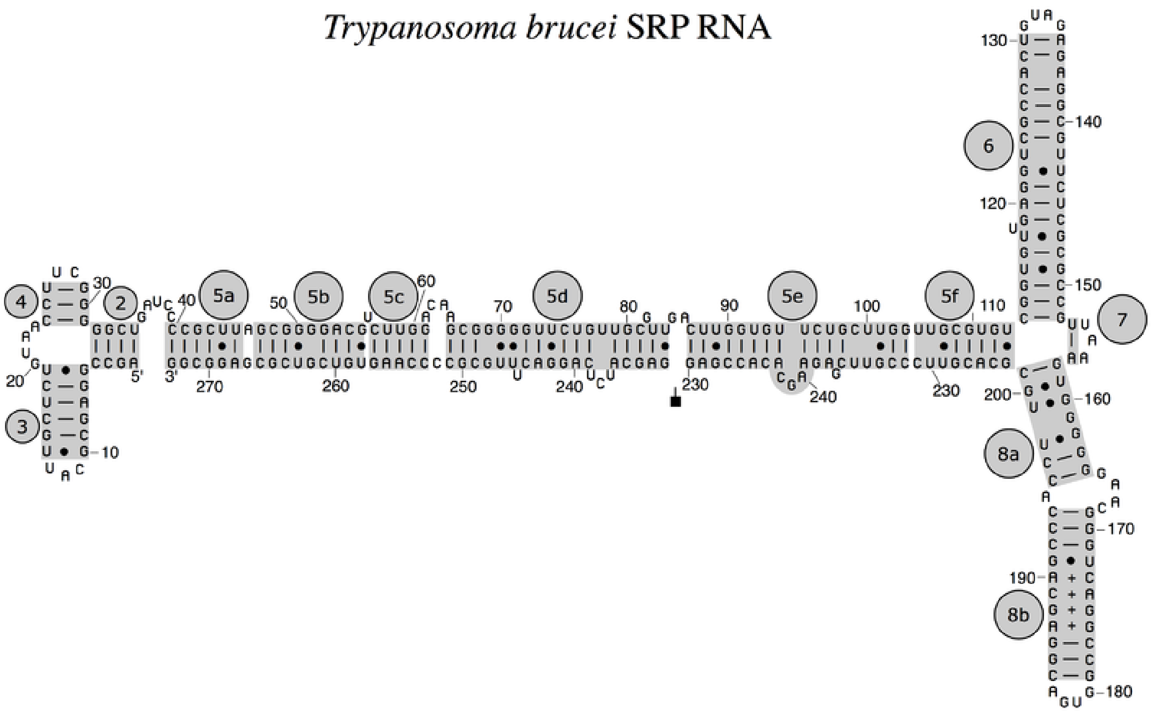
Bacterial SRP RNA (4.5S RNA) from *E. coli*



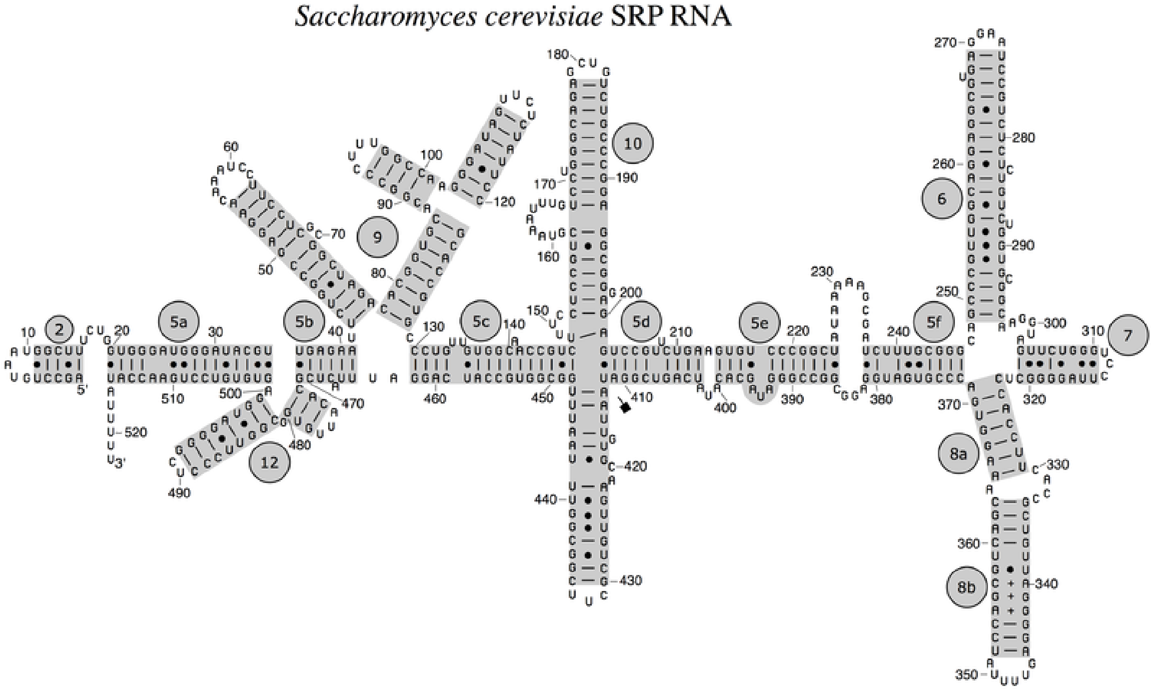
Bacterial SRP RNA (6S RNA) from *Bacillus subtilis*



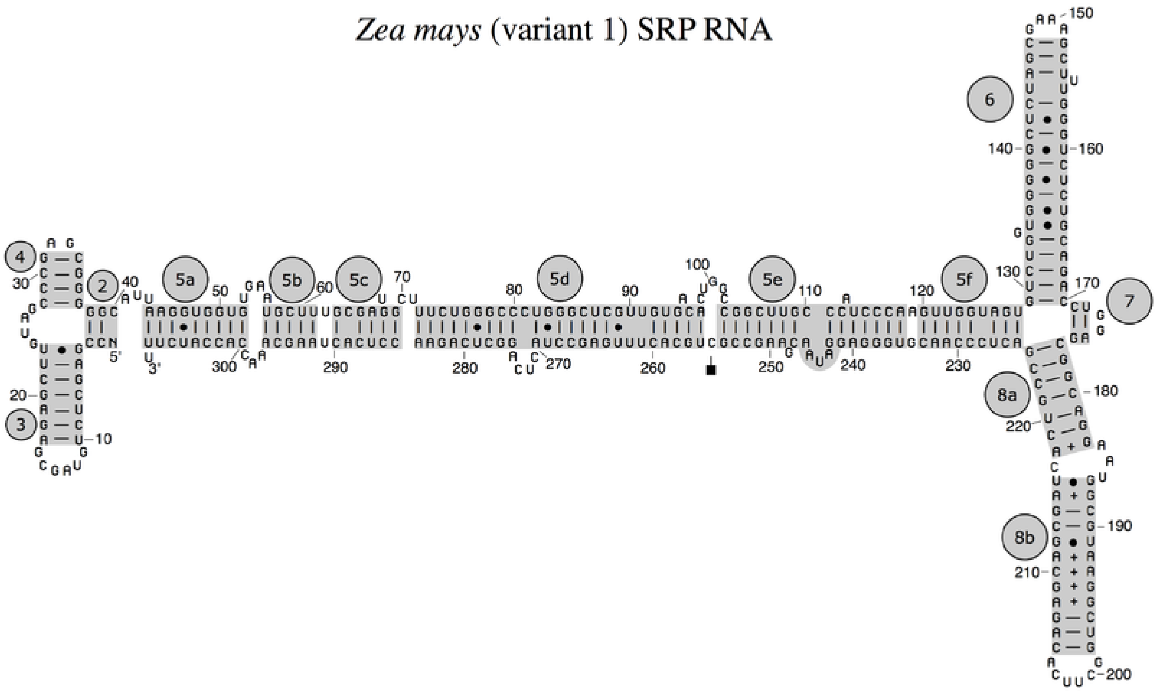
Archaeal SRP RNA *Archaeoglobus fulgidus*



Eukaryotic protist SRP RNA from *Trypanosoma brucei*



Eukaryotic yeast SRP RNA from *Saccharomyces cerevisiae*

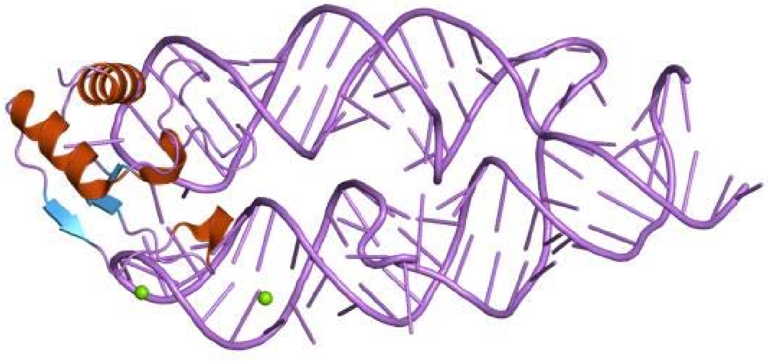


Eukaryotic plant SRP RNA from *Zea mays*

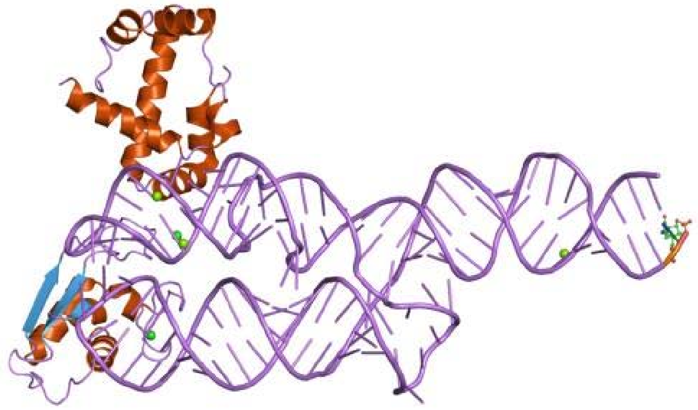
Tertiary

X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM) have been used to determine the molecular structure of portions of the SRP RNAs from various species. The available PDB structures show the RNA molecule either free or when bound to one or more SRP proteins.

Crystallographic structures of representative SRPs



SRP19-7S.S SRP RNA complex from *M. jannaschii*



S-domain of human SRP

Binding proteins

One or more SRP proteins bind to the SRP RNA to assemble the functional SRP. The SRP proteins are named according to their approximate molecular mass measured in kilodalton. Most bacterial SRPs are composed of SRP RNA and SRP54 (also named Ffh for "*F*ifty-*f*our *h*omolog"). The Archaeal SRP contains proteins SRP54 and SRP19. In eukaryotes, the SRP RNA combines with the imported SRP proteins SRP9/14, SRP19, and SRP68/72 in a region of the nucleolus. This pre-SRP is transported to the cytosol where it binds to protein SRP54. The molecular structures of the free or SRP RNA-bound proteins SRP9/14, SRP19, or SRP54 are known at high resolution.

SRP9 and SRP14

SRP9 and SRP14 are structurally related and form the SRP9/14 heterodimer which binds to the SRP RNA of the small (Alu) domain. Yeast SRP lacks SRP9 and contains the structurally related protein SRP21. Yeast SRP14 forms a homodimer. SRP9/14 is absent in the SRP of trypanosoma which instead possess a tRNA-like molecule.

SRP19

SRP19 is found in the SRP of eukaryotes and Archaea. Its primary role is in preparing the SRP RNA for the binding of SRP54, SRP68, and SRP72 by properly arranging SRP RNA helices 6 and 8. Yeast SRP contains Sec65p, a larger homolog of SRP19.

SRP54

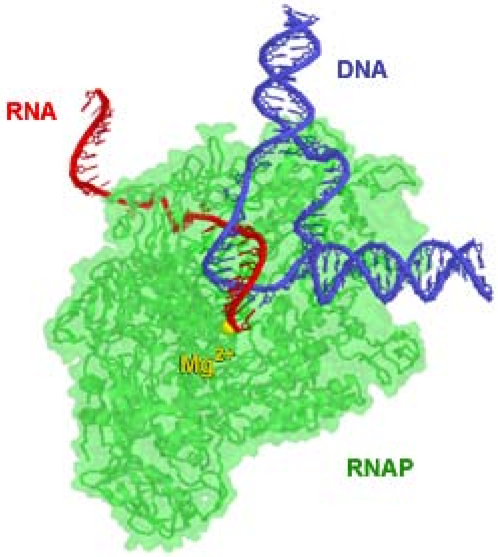
Protein SRP54 (named Ffh in the bacteria) is an essential component of every SRP. It is composed of three functional domains: the N-terminal (N) domain, the GTPase (G) domain, and the methionine-rich (M) domain.

SRP68 and SRP72

Proteins SRP68 and SRP72 are constituents of the large domain of the eukaryotic SRP. They form a stable SRP68/72 heterodimer. About one third of the human SRP68 protein was shown to bind to the SRP RNA. A relatively small region located near the C-terminus of SRP72 binds to the 5e SRP RNA motif.

Chapter- 11

RNA Polymerase



RNAP from *T. aquaticus* pictured during elongation. Portions of the enzyme were made transparent so as to make the path of RNA and DNA more clear. The magnesium ion (yellow) is located at the enzyme active site.

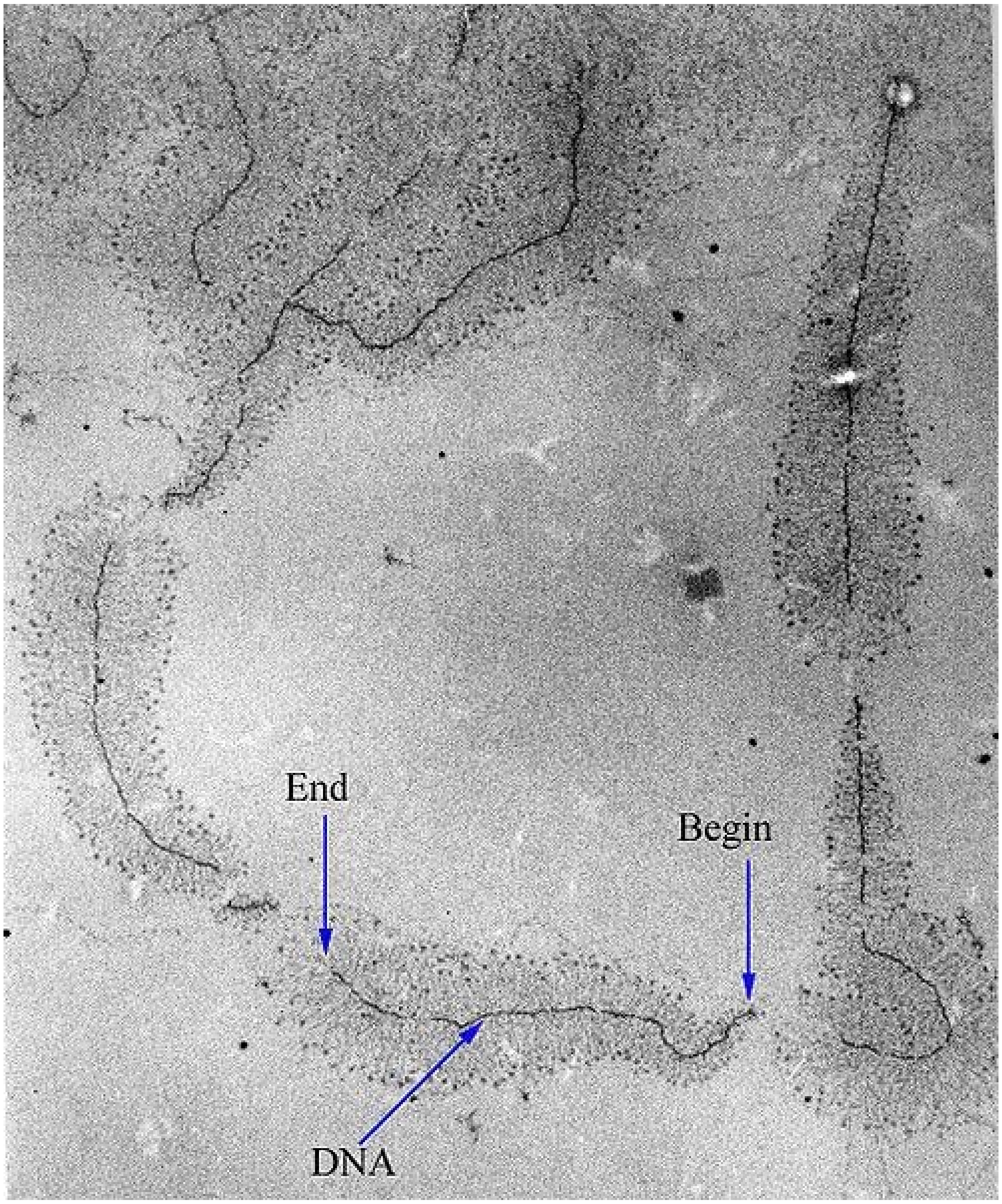
RNA polymerase (**RNAP** or **RNApol**) is an enzyme that produces RNA. In cells, RNAP is needed for constructing RNA chains from DNA genes as templates, a process called transcription. RNA polymerase enzymes are essential to life and are found in all organisms and many viruses. In chemical terms, RNAP is a nucleotidyl transferase that polymerizes ribonucleotides at the 3' end of an RNA transcript.

History

RNAP was discovered independently by Sam Weiss, Audrey Stevens, and Jerard Hurwitz in 1960. By this time the 1959 Nobel Prize in Medicine had been awarded to Severo Ochoa and Arthur Kornberg for the discovery of what was believed to be RNAP, but instead turned out to be polynucleotide phosphorylase.

The 2006 Nobel Prize in Chemistry was awarded to Roger Kornberg for creating detailed molecular images of RNA polymerase during various stages of the transcription process.

Control of transcription



An electron-micrograph of DNA strands decorated by hundreds of RNAP molecules too small to be resolved. Each RNAP is transcribing an RNA strand, which can be seen branching off from the DNA. "Begin" indicates the 3' end of the DNA, where RNAP initiates transcription; "End" indicates the 5' end, where the longer RNA molecules are almost completely transcribed.

Control of the process of gene transcription affects patterns of gene expression and, thereby, allows a cell to adapt to a changing environment, perform specialized roles within an organism, and maintain basic metabolic processes necessary for survival. Therefore, it is hardly surprising that the activity of RNAP is both long and complex and highly regulated. In Escherichia coli *bacteria, more than 100 transcription factors have been identified, which modify the activity of RNAP.*

RNAP can initiate transcription at specific DNA sequences known as promoters. It then produces an RNA chain, which is complementary to the template DNA strand. The process of adding nucleotides to the RNA strand is known as elongation; In eukaryotes, RNAP can build chains as long as 2.4 million nucleosides (the full length of the dystrophin gene). RNAP will preferentially release its RNA transcript at specific DNA sequences encoded at the end of genes known as terminators.

Products of RNAP include:

- Messenger RNA (mRNA)—template for the synthesis of proteins by ribosomes.
- Non-coding RNA or "RNA genes"—a broad class of genes that encode RNA that is not translated into protein. The most prominent examples of RNA genes are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation. However, since the late 1990s, many new RNA genes have been found, and thus RNA genes may play a much more significant role than previously thought.
    - Transfer RNA (tRNA)—transfers specific amino acids to growing polypeptide chains at the ribosomal site of protein synthesis during translation
    - Ribosomal RNA (rRNA)—a component of ribosomes
    - Micro RNA—regulates gene activity
    - Catalytic RNA (Ribozyme)—enzymatically active RNA molecules

RNAP accomplishes *de novo* synthesis. It is able to do this because specific interactions with the initiating nucleotide hold RNAP rigidly in place, facilitating chemical attack on the incoming nucleotide. Such specific interactions explain why RNAP prefers to start transcripts with ATP (followed by GTP, UTP, and then CTP). In contrast to DNA polymerase, RNAP includes helicase activity, therefore no separate enzyme is needed to unwind DNA.

RNA polymerase action

Binding and initiation

RNA Polymerase binding in prokaryotes involves the α subunit recognizing the upstream element (40 to -70 base pairs) in DNA, as well as the σ factor recognizing the -10 to -35 region. There are numerous σ factors that regulate gene expression. For example, σ^{70} is expressed under normal conditions and allows RNAP binding to house-keeping genes, while σ^{32} elicits RNAP binding to heat-shock genes.

After binding to the DNA, the RNA polymerase switches from a closed complex to an open complex. This change involves the separation of the DNA strands to form an unwound section of DNA of approximately 13 bp. Ribonucleotides are base-paired to the template DNA strand, according to Watson-Crick base-pairing interactions. Supercoiling plays an important part in polymerase activity because of the unwinding and rewinding of DNA. Because regions of DNA in front of RNAP are unwound, there is compensatory positive supercoils. Regions behind RNAP are rewound and negative supercoils are present.

Elongation

Transcription elongation involves the further addition of ribonucleotides and the change of the open complex to the transcriptional complex. RNAP cannot start forming full length transcripts because of its strong binding to the promoter. Transcription at this stage primarily results in short RNA fragments of around 9 bp in a process known as abortive transcription. Once the RNAP starts forming longer transcripts it clears the promoter. At this point, the -10 to -35 promoter region is disrupted, and the σ factor falls off RNAP. This allows the rest of the RNAP complex to move forward, as the σ factor held the RNAP complex in place.

The 17-bp transcriptional complex has an 8-bp DNA-RNA hybrid, that is, 8 base-pairs involve the RNA transcript bound to the DNA template strand. As transcription progresses, ribonucleotides are added to the 3' end of the RNA transcript and the RNAP complex moves along the DNA. Although RNAP does not seem to have the 3'exonuclease activity that characterizes the *proofreading* activity found in DNA polymerase, there is evidence of that RNAP will halt at mismatched base-pairs and correct it.

The addition of ribonucleotides to the RNA transcript has a very similar mechanism to DNA polymerization - it is believed that these polymerases are evolutionarily related. Aspartyl (asp) residues in the RNAP will hold onto Mg^{2+} ions, which will, in turn, coordinate the phosphates of the ribonucleotides. The first Mg^{2+} will hold onto the α-phosphate of the NTP to be added. This allows the nucleophilic attack of the 3'OH from the RNA transcript, adding an additional NTP to the chain. The second Mg^{2+} will hold onto the pyrophosphate of the NTP. The overall reaction equation is:

$(NMP)_n + NTP \text{-->} (NMP)_{n+1} + PP_i$

Termination

Termination of RNA transcription can be rho-independent or rho-dependent:

Rho-independent transcription termination is the termination of transcription without the aid of the rho protein. Transcription of a palindromic region of DNA causes the formation of a *hairpin* structure from the RNA transcription looping and binding upon itself. This hairpin structure is often rich in G-C base-pairs, making it more stable than the DNA-RNA hybrid itself. As a result, the 8bp DNA-RNA hybrid in the transcription complex shifts to a 4bp hybrid. These last 4 base-pairs are weak A-U base-pairs, and the entire RNA transcript will fall off of DNA.

RNA polymerase in bacteria

In bacteria, the same enzyme catalyzes the synthesis of mRNA and ncRNA.

RNAP is a relatively large molecule. The core enzyme has 5 subunits (~400 kDa):

- α_2: The two α subunits assemble the enzyme and bind regulatory factors. Each subunit has two domains: αCTD (C-Terminal domain) binds the UP element of the extended promoter, and αNTD (N-terminal domain) binds the rest of the polymerase. This subunit is not used on promoters without an UP element.
- β: this has the polymerase activity (catalyzes the synthesis of RNA), which includes chain initiation and elongation.
- β': binds to DNA (nonspecifically).
- ω: restores denatured RNA polymerase to its functional form in vitro. It has been observed to offer a protective/chaperone function to the β' subunit in *Mycobacterium smegmatis*. Now known to promote assembly.

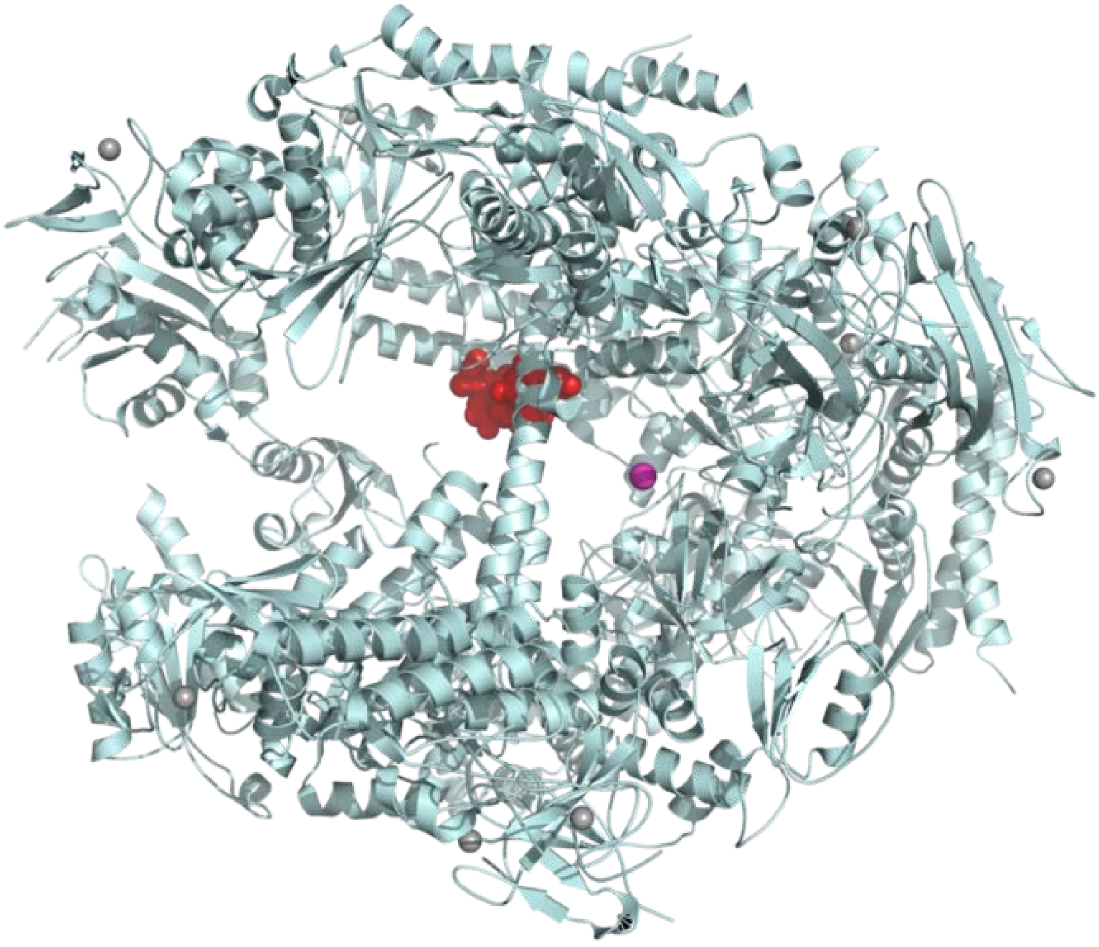
In order to bind promoter-specific regions, the core enzyme requires another subunit, sigma (σ). The sigma factor greatly reduces the affinity of RNAP for nonspecific DNA while increasing specificity for certain promoter regions, depending on the sigma factor. That way, transcription is initiated at the right region. The complete holoenzyme therefore has 6 subunits: $\alpha_2\beta\beta'\sigma\omega$ (~480 kDa). The structure of RNAP exhibits a groove with a length of 55 Å (5.5 nm) and a diameter of 25 Å (2.5 nm). This groove fits well the 20 Å (2 nm) double strand of DNA. The 55 Å (5.5 nm) length can accept 16 nucleotides.

When not in use, RNA polymerase binds to low-affinity sites to allow rapid exchange for an active promoter site when one opens. RNA polymerase holoenzyme, therefore, does not freely float around in the cell when not in use.

Transcriptional cofactors

There are many proteins that can bind to RNAP and modify its behavior. For instance, GreA and GreB from *E. coli* and in most other prokaryotes can enhance the ability of RNAP to cleave the RNA template near the growing end of the chain. This cleavage can rescue a stalled polymerase molecule, and is likely involved in proofreading the occasional mistakes made by RNAP. A separate cofactor, Mfd, is involved in transcription-coupled repair, the process in which RNAP recognizes damaged bases in the DNA template and recruits enzymes to restore the DNA. Other cofactors are known to play regulatory roles; i.e. they help RNAP choose whether or not to express certain genes.

RNA polymerase in eukaryotes



Structure of eukaryotic RNA polymerase II (light blue) in complex with α-amanitin (red), a strong poison found in death cap mushrooms that targets this vital enzyme

Eukaryotes have several types of RNAP, characterized by the type of RNA they synthesize:

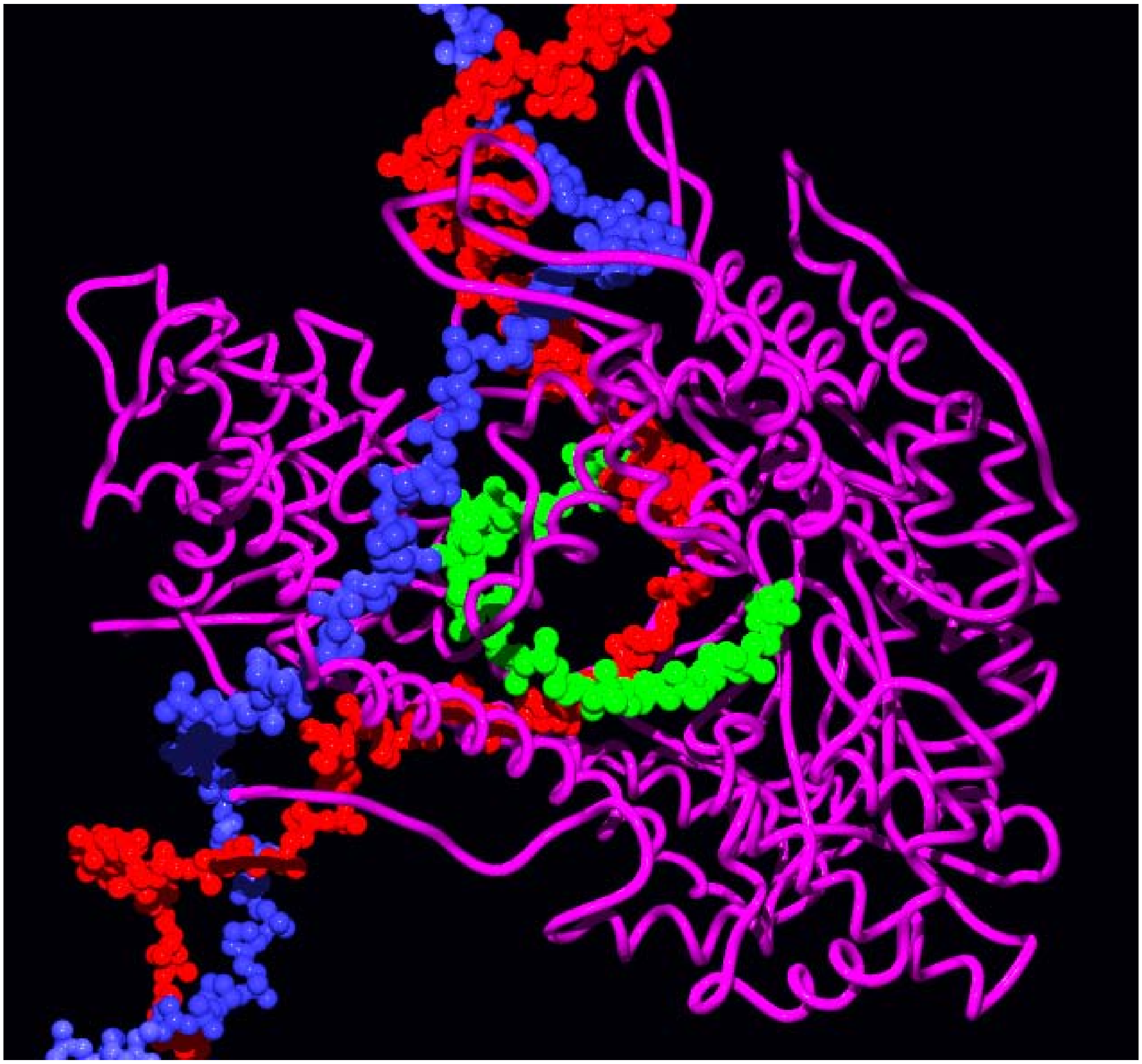
- RNA polymerase I synthesizes a pre-rRNA 45S, which matures into 28S, 18S and 5.8S rRNAs which will form the major RNA sections of the ribosome.
- RNA polymerase II synthesizes precursors of mRNAs and most snRNA and microRNAs. This is the most studied type, and due to the high level of control required over transcription a range of transcription factors are required for its binding to promoters.
- RNA polymerase III synthesizes tRNAs, rRNA 5S and other small RNAs found in the nucleus and cytosol.
- RNA polymerase IV synthesizes siRNA in plants.
- RNA polymerase V synthesizes RNAs involved in siRNA-directed heterochromatin formation in plants.

There are other RNA polymerase types in mitochondria and chloroplasts. And there are RNA-dependent RNA polymerases involved in RNA interference.

RNA polymerase in archaea

Archaea have a single RNAP that is closely related to the three main eukaryotic polymerases (Pol I,II,III). Thus, it has been speculated that the archaeal polymerase resembles the ancestor of the specialized eukaryotic polymerases.

RNA polymerase in viruses



T7 RNA polymerase producing a mRNA (green) from a DNA template. The protein is shown as a purple ribbon. Image derived from PDB 1MSW.

Many viruses also encode for RNAP. Perhaps the most widely studied viral RNAP is found in bacteriophage T7. The single-subunit T7 RNA polymerase is related to that found in mitochondria and chloroplasts, and shares considerable homology to DNA polymerase. It is believed that most viral polymerases therefore evolved from DNA polymerase and are not directly related to the multi-subunit polymerases described above.

The viral polymerases are diverse, and include some forms that can use RNA as a template instead of DNA. This occurs in negative strand RNA viruses and dsRNA viruses, both of which exist for a portion of their life cycle as double-stranded RNA. However, some positive strand RNA viruses, such as polio, also contain these RNA-dependent RNA polymerases.

RNA polymerase purification

RNA polymerase can be isolated in the following ways:

- By a phosphocellulose column.
- By glycerol gradient centrifugation.
- By a DNA column.
- By an Ion exchange column.

And also combinations of the above techniques.

Chapter- 12

RNA-Seq

RNA-Seq, also called "Whole Transcriptome Shotgun Sequencing" ("WTSS") and dubbed "a revolutionary tool for transcriptomics", refers to the use of High-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content, a technique that is quickly becoming invaluable in the study of diseases like cancer. Thanks to the deep coverage and base level resolution provided by next-generation sequencing instruments, RNA-Seq provides researchers with efficient ways to measure transcriptome data experimentally, allowing them to get information such as how different alleles of a gene are expressed, detect post-transcriptional mutations or identify gene fusions.

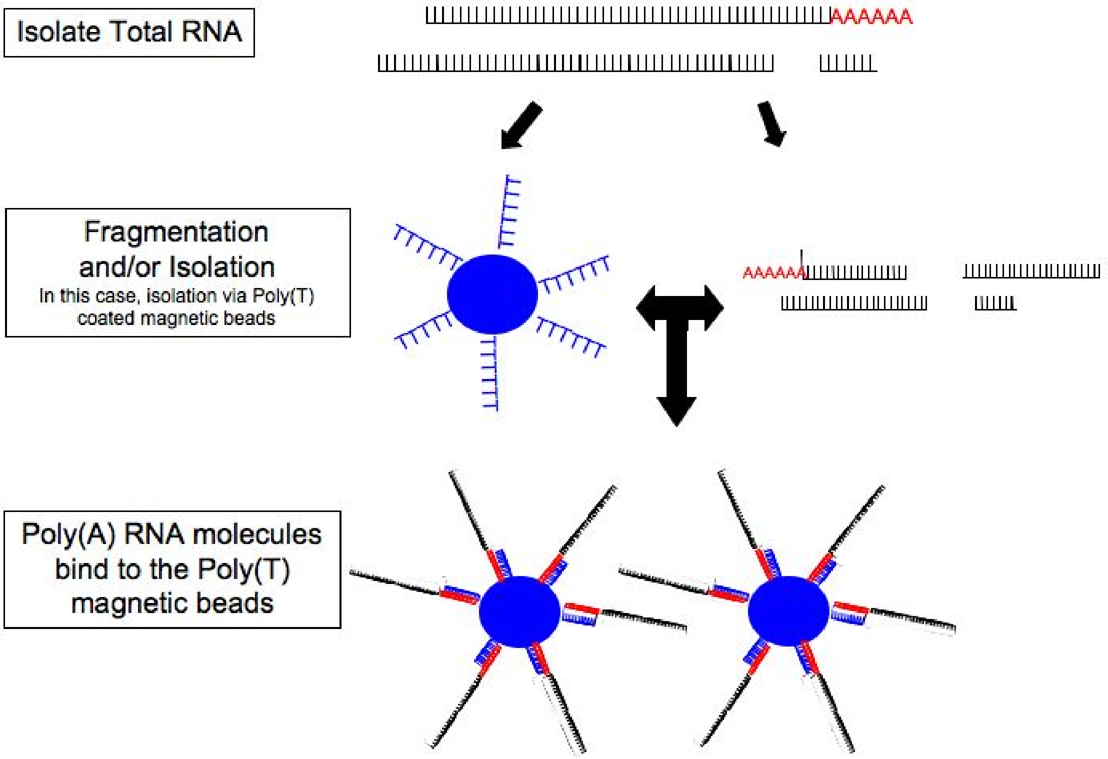
Introduction

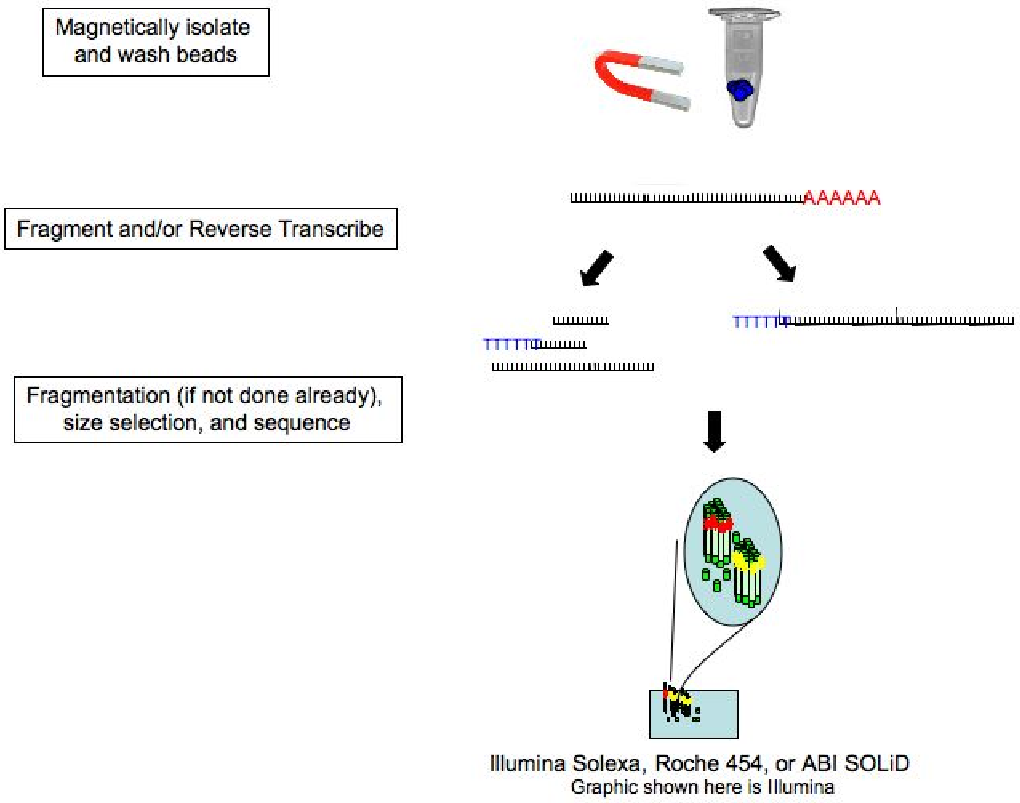
The introduction of Next-generation sequencing or High-throughput sequencing technologies opened new doors into the field of DNA sequencing, however as understanding of these technologies becomes more widespread and new tools are developed, new innovative ways of applying these technologies are being created.

Given High-throughput sequencing technologies' low requirements of nucleotide sequence product, together with its deep coverage and base-scale resolution, its use has expanded to the field of transcriptomics. Transcriptomics is an area of research characterizing the RNA transcribed from a particular genome under investigation. Although transcriptomes are more dynamic than genomic DNA, these molecules provide direct access to gene regulation and protein information. Sequencing transcriptomes is not a new idea. Various methods have been developed previously to directly determine cDNA sequences based mostly around traditional (and more expensive) Sanger sequencing, while others include methodologies such as Serial analysis of gene expression (SAGE), cap analysis gene expression (CAGE) and massively parallel signature sequencing (MPSS).

Transcriptome Sequencing (RNA-seq) can be done with a variety of platforms to test a smorgasbord of ideas and hypotheses. For example, using the Illumina (company) Genome Analyzer platform, recent applications include sequencing mammalian transcriptomes, ABI Solid Sequencing to profile stem cell transcriptomes or Life Science's 454 Sequencing to discover SNPs in maize. Even though each platform has its technical differences, the information gathered from each is of the same nature.

Methods




Magnetically isolate and wash beads
AAAAAA
Fragment and/or Reverse Transcribe
TTTTT
TTTTT
Fragmentation (if not done already), size selection, and sequence
Illumina Solexa, Roche 454, or ABI SOLiD
Graphic shown here is Illumina

RNA *Poly(A)* Library

Creation of a library can change from platform to platform in high throughput sequencing, where each has several kits designed to build different types of libraries and adapting the resulting sequences to the specific requirements of their instruments.

However, due to the nature of the template being analyzed, there are commonalities within each technology. Frequently, in mRNA analysis the 3' polyadenylated (poly(A)) tail is targeted in order to ensure that coding RNA is separated from noncoding RNA. This can be accomplished simply with poly (T) oligos covalently attached to a given substrate. Presently many studies utilize magnetic beads for this step (;) (Invitrogen, MACS mRNA Isolation kit). The Protocol Online website provides a list of several protocols relating to mRNA isolation.

Studies including portions of the transcriptome outside poly(A) RNAs have shown that when using poly(T) magnetic beads, the flow-through RNA (non-poly(A) RNA) can yield important noncoding RNA gene discovery which would have otherwise gone unnoticed

Also, since ribosomal RNA represents over 90% of the RNA within a given cell, studies have shown that its removal via probe hybridization (Invitrogen, RiboMinus

Human/Mouse Transcriptome Isolation kit) increases the capacity to retrieve data from the remaining portion of the transcriptome.

The next step is reverse transcription. Due to the 5' bias of randomly primed-reverse transcription as well as secondary structures influencing primer binding sites, hydrolysis of RNA into 200-300 nucleotides prior to reverse transcription reduces both problems simultaneously. However, there are trade-offs with this method where although the overall body of the transcripts are efficiently converted to DNA, the 5' and 3' ends are less so. Depending on the aim of the study, researchers may choose to apply or ignore this step.

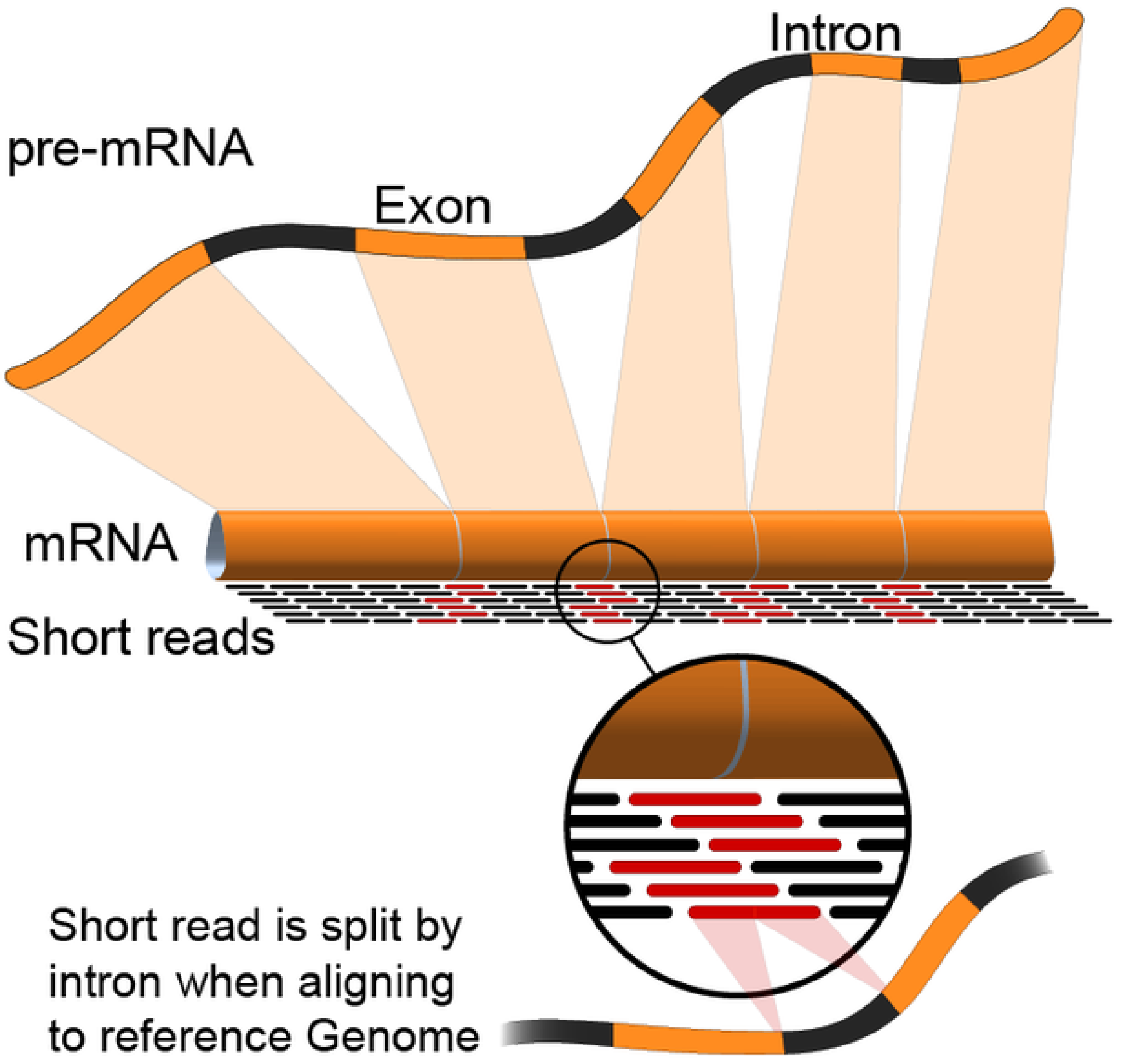
Once the cDNA is synthesized it can be further fragmented to reach the desired fragment length as specified in table 1. The template is now ready to be prepared for the desired sequencing method.

Next-generation sequencing

High-throughput sequencing technologies generate millions of short reads from a library of nucleotide sequences, whether they come from DNA, RNA, or a mixture, the sequencing mechanism of each platform does not vary. The most used technologies and some of their characteristics are shown in the following table

	454 Sequencing	**Illumina**	**SOLiD**
Sequencing Chemistry	Pyrosequencing	Polymerase-based sequence-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired end separation	3 kb	200 bp	3 kb
Mb per run	100 Mb	1300 Mb	3000 Mb
Time per paired end run	7 hours	4 days	5 days
Read length (update)	250 bp (400 bp)	32-40 bp (35-100 bp)	35 bp (35-50 bp)
Cost per run	$ 8,438 USD	$ 8,950 USD	$ 17,447 USD
Cost per Mb	$ 84.39 USD	$ 5.97 USD	$ 5.81 USD

Table 1. Comparing metrics and performance of next-generation DNA sequencers


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

RNA-Seq mapping of short reads in exon-exon junctions

Transcriptome alignment

Due to the small size of the short reads (for Illumina Genome Analyzer this can be around 36 bases) *de novo* assembly may be difficult (though some software does exist: Velvet (algorithm)), as there cannot be large overlaps between each read needed to easily reconstruct the original sequences, and the deep coverage makes the computing power to track all the possible alignments prohibitive. This can be somewhat overcome by having larger sequences obtained from the same sample using other techniques as Sanger Sequencing, and using larger reads as a "skeleton" or a "template" to help assemble reads in difficult regions (e.g. regions with repetitive sequences).

The recommended approach is that of aligning the millions of reads to a "reference genome". There are many tools available for aligning genomic reads to a reference genome (sequence alignment tools), however, special attention is needed when alignment of a transcriptome to a genome, mainly when dealing with genes having intronic regions.

As discussed above, the sequence libraries are created extracting mRNA using its poly(A) tail, which is added to the mRNA molecule post-transcriptionally and thus splicing has taken place. Therefore, the created library and the short reads obtained cannot come from intronic sequences and thus, when trying to align these short reads to a reference genome, only short reads aligning entirely inside exonic regions will be matched while short reads from exon-exon junction regions will not.

A possible method to work around this is to try to align the unaligned short reads using a proxy genome generated with known exonic sequences. This need not cover whole exons, only enough so that the short reads can match on both sides of the exon-exon junction with minimum overlap. The use of paired-end sequencing has been mentioned as a good solution to alignment problems, as besides giving longer length reads, it allows obtaining information in respect to the strand.

Several software packages exist for short read alignment, and recently specialized algorithms for transcriptome alignment have been developed, e.g. *TopHat* and *Cufflinks*.

Analysis

Gene expression

The characterization of gene expression in cells via measurement of mRNA levels has long been of interest to researchers. Even though it has been shown that due to other post transcriptional gene regulation events (such as RNA interference) there is not a strong correlation between the abundance of mRNA and the related proteins, measuring mRNA concentration levels is still a useful tool in determining how the transcriptional machinery of the cell is affected in the presence of external signals (e.g. drug treatment), or how cells differ between a healthy state and a diseased state.

Microarray approach

Prior to RNA-Seq, DNA microarrays were unchallenged as the experiment of choice for transcriptome analysis. Although many experiments are still using microarrays to generate exciting results where the amount of time to retrieve results for a given sample is shorter, intrinsic experimental limitations of microarrays seem to make RNA-Seq the method of choice. One important limitation is a prerequisite for sequence information in order to detect and ultimately evaluate transcripts. As research in the field of RNA-Seq is growing steadily with promising and consistent results, one must now consider, "Is this the beginning of the end for microarrays?". Still, for many applications, microarrays are the method of choice. Not only are they 10-100 times cheaper when compared at the same resolution of accuracy (less than $100 per array for high-throughput applications), but RNA-Seq protocols still suffer from unknown biases such as those implied by the required ligation steps, and known biases where high abundance transcripts (such as from housekeeping genes) make up the majority of sequencing data (e.g. in some tissues 5% of the genes represent up to 75% of the reads sequenced).

Coverage as measure of expression

Expression can be deduced via RNA-Seq to the extent at which a sequence is retrieved. Transcriptome studies in Yeast show that in this experimental setting, a fourfold coverage is required for amplicons to be classified and characterized as an expressed gene. When the transcriptome is fragmented prior to cDNA synthesis, the number of reads corresponding to the particular exon normalized by its length in vivo yields gene expression levels which correlate with those obtained through qPCR.

Single nucleotide variation discovery

Transcriptome single nucleotide variation has been analyzed in maize on the Roche 454 sequencing platform. Directly from the transcriptome anaysis, around 7000 single nucleotide polymorphisms (SNPs) were recognized. Following Sanger sequence validation, the researchers were able to conservatively obtain almost 5000 valid SNPs covering more than 2400 maize genes. This impressive transcriptome analysis is currently being applied to cancer research and microbiology which could quite possibly lead to new forms of medicine.

Coverage/depth

Coverage/depth can affect the mutations seen and given that everything is expression-centric, an allele might not be detected either because it is not in the genome, or because it is not being expressed. At the same time, RNA-seq can yield additional information rather than just the existence of a heterozygous gene as it can also help in estimating the expression of each allele. In association studies, genotypes are associated to disease and expression levels can also be associated with disease. Using RNA-seq, we can measure the relationship between these two associated variables, that is, in what relation are each of the alleles being expressed.

The depth of sequencing required for specific applications can be extrapolated from a pilot experiment.

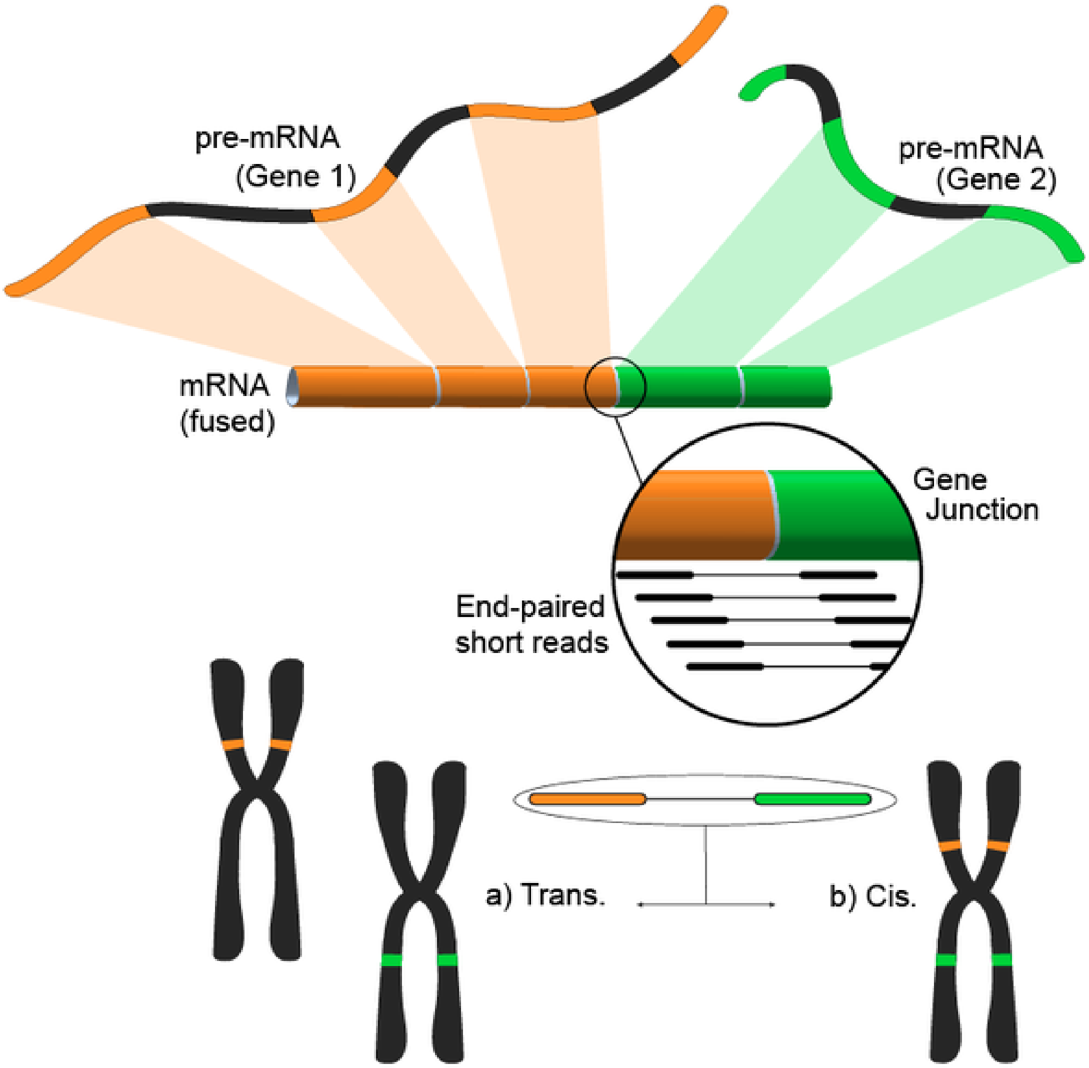
Germline vs expressed alleles

The only way to be absolutely sure of the individual's mutations is to compare the transcriptome sequences to the germline DNA sequence. This enables the distinction of homozygous genes versus skewed expression of one of the alleles and it can also provide information about genes that were not expressed in the transcriptomic experiment.

Post-transcriptional SNVs

Having the matching genomic and transcriptomic sequences of an individual can also help in detecting post-transcriptional edits, where, if the individual is homozygous for a gene, but the gene's transcript has a different allele, then a post-transcriptional modification event is determined.

mRNA centric single nucleotide variants (SNVs) are generally not considered as a representative source of functional variation in cells, mainly due to the fact that these mutations disappear with the mRNA molecule, however the fact that efficient DNA correction mechanisms do not apply to RNA molecules can cause them to appear more often. This has been proposed as the source of certain prion diseases, also known as TSE or transmissible spongiform encephalopathies.



RNA-Seq mapping of short reads over exon-exon junctions, depending on where each end maps to, it could be defined a *Trans* or a *Cis* event.

Fusion gene detection

Caused by different structural modifications in the genome, fusion genes have gained attention because of their relationship with cancer. RNA-Seq's ability to analyze a sample's whole transcriptome in an unbiased fashion makes it an attractive tool to find these kinds of common events in cancer.

The idea follows from the process of aligning the short transcriptomic reads to a reference genome. Most of the short reads will fall within one complete exon, and a smaller but still large set would be expected to map to known exon-exon junctions. The remaining unmapped short reads would then be further analyzed to determine whether they match an exon-exon junction where the exons come from different genes. This would be evidence of a possible fusion event, however, because the length of the reads, this could prove to be very noisy. An alternative approach is using pair-end reads, when potentially a large number of paired reads would map each end to a different exon, giving better coverage of these events. Nonetheless, the end result consists of multiple and potentially novel combinations genes providing an ideal starting point for further validation.

Some considerations

The information gathered when sequencing a sample's transcriptome in this way has many of the same limitations as other RNA expression analysis pipelines. Mainly, the information gathered is:

a) Tissue specific: Gene expression is not uniform throughout an organism's cells, it is strongly dependent on the tissue type being measured;

b) Time dependent: During a cell's lifetime and context, its gene expression levels change.

Because of this, care must be taken when drawing conclusions from the sequencing experiment, as some of the information gathered might not be representative of the individual itself.

An example of this would be during SNV discovery as the mutations discovered are more precisely the mutations being expressed, this is: observing a homozygote location to a non-reference allele in an organism does not necessarily mean that this is the individual's genotype, it could just mean that the gene copy with the reference allele is not being expressed in that tissue and/or at the time snapshot the sample was acquired.

Chapter- 13

RNA Virus

An **RNA virus** is a virus that has RNA (ribonucleic acid) as its genetic material. This nucleic acid is usually single-stranded RNA (ssRNA) but may be double-stranded RNA (dsRNA). The ICTV classifies RNA viruses as those that belong to *Group III*, *Group IV* or *Group V* of the Baltimore classification system of classifying viruses, and does not consider viruses with DNA intermediates as RNA viruses. Notable human diseases caused by RNA viruses include SARS, influenza and hepatitis C.

Another term for RNA viruses that explicitly excludes retroviruses is **ribovirus**.

Characteristics

Single-stranded RNA viruses and RNA Sense

RNA viruses can be further classified according to the sense or polarity of their RNA into negative-sense and positive-sense, or ambisense RNA viruses. Positive-sense viral RNA is similar to mRNA and thus can be immediately translated by the host cell. Negative-sense viral RNA is complementary to mRNA and thus must be converted to positive-sense RNA by an RNA polymerase before translation. As such, purified RNA of a positive-sense virus can directly cause infection though it may be less infectious than the whole virus particle. Purified RNA of a negative-sense virus is not infectious by itself as it needs to be transcribed into positive-sense RNA, however each virion can be transcribed to several positive-sense RNAs. Ambisense RNA viruses resemble negative-sense RNA viruses, except they also translate genes from the positive strand.

Double-stranded RNA viruses

The double-stranded (ds)RNA viruses represent a diverse group of viruses that vary widely in host range (humans, animals, plants, fungi, and bacteria), genome segment number (one to twelve), and virion organization (T-number, capsid layers, or turrets). Members of this group include the rotaviruses, renowned globally as the most common cause of gastroenteritis in young children, and bluetongue virus, an economically important pathogen of cattle and sheep. In recent years, remarkable progress has been made in determining, at atomic and subnanometeric levels, the structures of a number of key viral proteins and of the virion capsids of several dsRNA viruses, highlighting the significant parallels in the structure and replicative processes of many of these viruses.

Mutation rates

RNA viruses generally have very high mutation rates compared to DNA viruses, because viral RNA polymerases lack the proof-reading ability of DNA polymerases. This is one reason why it is difficult to make effective vaccines to prevent diseases caused by RNA viruses. Retroviruses also have a high mutation rate even though their DNA intermediate integrates into the host genome (and is thus subject to host DNA proofreading once integrated), because errors during reverse transcription are embedded into both strands of DNA before integration. Some genes of RNA virus are important to the viral replication cycles and mutations are not tolerated. For example, the region of the hepatitis C virus genome that encodes the core protein is highly conserved, because it contains an RNA structure involved in an internal ribosome entry site.

Replication

Animal RNA viruses are classified into three distinct groups depending on their genome and mode of replication (and the numerical groups based on the older Baltimore classification):

- Double-stranded RNA viruses (Group III) contain from one to a dozen different RNA molecules, each of which codes for one or more viral proteins.
- Positive-sense ssRNA viruses (Group IV) have their genome directly utilized as if it were mRNA, producing a single protein which is modified by host and viral proteins to form the various proteins needed for replication. One of these includes RNA-dependent RNA polymerase, which copies the viral RNA to form a double-stranded replicative form, in turn this directs the formation of new virions.
- Negative-sense ssRNA viruses (Group V) must have their genome copied by an RNA polymerase to form positive-sense RNA. This means that the virus must bring along with it the RNA-dependent RNA polymerase enzyme. The positive-sense RNA molecule then acts as viral mRNA, which is translated into proteins by the host ribosomes. The resultant protein goes on to direct the synthesis of new virions, such as capsid proteins and RNA replicase, which is used to produce new negative-sense RNA molecules.

Retroviruses (Group VI) have a single-stranded RNA genome but are generally not considered RNA viruses because they use DNA intermediates to replicate. Reverse transcriptase, a viral enzyme that comes from the virus itself after it is uncoated, converts the viral RNA into a complementary strand of DNA, which is copied to produce a double stranded molecule of viral DNA. After this DNA is integrated, expression of the encoded genes may lead the formation of new virions.

Group III - dsRNA viruses

- Family Birnaviridae
- Family Chrysoviridae
- Family Cystoviridae
- Family Hypoviridae
- Family Partitiviridae
- Family Reoviridae - includes Rotavirus
- Family Totiviridae
- Unassigned genera
    - *Endornavirus*

Group IV - positive-sense ssRNA viruses

- Order Nidovirales
    - Family Arteriviridae
    - Family Coronaviridae - includes Coronavirus, SARS
    - Family Roniviridae

- Order Picornavirales
    - Family Dicistroviridae
    - Family Iflaviridae
    - Family Marnaviridae
    - Family Picornaviridae - includes Poliovirus, the common cold virus, Hepatitis A virus
    - Family Secoviridae includes subfamily Comovirinae

- Order Tymovirales
    - Family Alphaflexiviridae
    - Family Betaflexiviridae
    - Family Gammaflexiviridae
    - Family Tymoviridae

- Unassigned
    - Family Astroviridae
    - Family Barnaviridae
    - Family Bromoviridae
    - Family Caliciviridae - includes Norwalk virus
    - Family Closteroviridae
    - Family Flaviviridae - includes Yellow fever virus, West Nile virus, Hepatitis C virus, Dengue fever virus
    - Family Leviviridae
    - Family Luteoviridae - includes Barley yellow dwarf virus
    - Family Narnaviridae
    - Family Nodaviridae
    - Family Potyviridae

- Family Tetraviridae
- Family Togaviridae - includes Rubella virus, Ross River virus, Sindbis virus, Chikungunya virus
- Family Tombusviridae
- Unassigned genera
    - Genus *Benyvirus*
    - Genus *Furovirus*
    - Genus *Hepevirus* - includes Hepatitis E virus
    - Genus *Hordeivirus*
    - Genus *Idaeovirus*
    - Genus *Ourmiavirus*
    - Genus *Pecluvirus*
    - Genus *Pomovirus*
    - Genus *Sobemovirus*
    - Genus *Tobamovirus* - includes tobacco mosaic virus
    - Genus *Tobravirus*
    - Genus *Umbravirus*

Group V - negative-sense ssRNA viruses

- Order *Mononegavirales*
    - Family Bornaviridae - Borna disease virus
    - Family Filoviridae - includes Ebola virus, Marburg virus
    - Family Paramyxoviridae - includes Measles virus, Mumps virus, Nipah virus, Hendra virus
    - Family Rhabdoviridae - includes Rabies virus
- Unassigned
    - Family Arenaviridae - includes Lassa virus
    - Family Bunyaviridae - includes Hantavirus, Crimean-Congo hemorrhagic fever
    - Family Orthomyxoviridae - includes Influenza viruses
    - Unassigned genera:
        - Genus *Deltavirus*
        - Genus *Nyavirus* - includes Nyamanini and Midway viruses
        - Genus *Ophiovirus*
        - Genus *Tenuivirus*
        - Genus *Varicosavirus*