CHAPTER 1

RNA MOLECULES: THE NEED AND THE APPROACH TO STUDY THEM
1.1 RNA MOLECULES AS A CLASS

1.1.1 Functional Characteristics of RNA Molecules

1.1.2 Structural Aspects of RNA Molecules

1.2 RATIONALE FOR THE CHOICE OF TRANSLATION PROCESS TO STUDY RNA MOLECULES

1.2.1 Translation Process Exercises Controls on Gene Expression

1.2.2 Manifestation of RNA Structures in the Translation Process

1.3 RNA MOLECULES CAN BE STUDIED THROUGH THEORETICAL APPROACH

1.3.1 Theoretical Studies Complement Experimental Studies

1.3.2 Nucleic Acid Sequence Data is a Source of Information for the Biological Processes

1.4 AIDS AND TOOLS FOR THEORETICAL STUDIES ON NUCLEIC ACID SEQUENCE DATA

1.4.1 Data Bases for Nucleic Acid Sequences

1.4.1.1 General data bases

1.4.1.2 Specialised data bases

1.4.2 Tools for Computer Analysis of Nucleic Acid Sequence Data

1.4.2.1 Types of analysis on nucleic acid sequence data

1.4.2.2 Approaches for sequence data analysis

1.5 SCOPE OF THE PRESENT STUDY
1.1 RNA MOLECULES AS A CLASS

1.1.1 Functional Characteristics of RNA Molecules

DNA acts as the carrier of genetic information. The information contained in DNA determines the phenotypic characteristics of the organism. The DNA sequences can be categorized as (a) coding sequences that code either for RNA or for protein, and (b) noncoding sequences which contain signals for various processes, such as, transcription-initiation, recombination, etc. The coding DNA can further be classified as (a) informational sequences that are transcribed as messenger RNA (mRNA) molecules for translation into protein, and (b) noninformational sequences that are transcribed as structural RNA molecules. Though the structural RNA molecules are not translated into proteins, they function as important regulatory or auxiliary entities, such as, transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA) etc. Thus, within the cell, it is the RNA and not the DNA that acts as the information molecule (Siolin, 1972).

RNA molecules perform various biological functions as indicated in Table 1.1. RNA molecules work as informational, catalytic, regulatory, replicase molecules, etc, (Lamond and Gibson, 1990) as elaborated below.

(a) Altman and coworkers (Guerrier-Takada et al, 1983) demonstrated that the ribonucleolytic activity of the enzyme RNAase P is contained in its RNA component. Cech and coworkers (Kruger et al, 1982) showed that the rRNA transcript of the protozoan, Tetrahymena contains an intervening sequence that is
<table>
<thead>
<tr>
<th>RNA</th>
<th>Function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>message specifying the type &amp; the order of the amino acid</td>
<td>Crick (1968)</td>
</tr>
<tr>
<td>rRNA</td>
<td>1. decoding of mRNA</td>
<td>Dahlberg (1989)</td>
</tr>
<tr>
<td></td>
<td>3. translocation of peptidyl-tRNA</td>
<td>Moazed and Noller (1989)</td>
</tr>
<tr>
<td>tRNA</td>
<td>adaptor carrying amino acids used in protein synthesis</td>
<td>Crick (1968)</td>
</tr>
<tr>
<td>U1, U2, U4 /U6 &amp; U5 snRNAs</td>
<td>pre-mRNA splicing</td>
<td>Reviewed in Padgett et al (1986)</td>
</tr>
<tr>
<td>U3 snRNA</td>
<td>rRNA processing</td>
<td>-do-</td>
</tr>
<tr>
<td>U7 snRNA</td>
<td>processing the histone-mRNA</td>
<td>-do-</td>
</tr>
<tr>
<td>U11 snRNA</td>
<td>cleavage/polyadenylation of pre-mRNA 3' termini</td>
<td>-do-</td>
</tr>
<tr>
<td>2.5S RNA</td>
<td>glucano transferase (catalytic moiety of the RNP)</td>
<td>Shvedova (1987)</td>
</tr>
<tr>
<td>M1 RNA</td>
<td>tRNA processing (catalytic moiety of the RNP)</td>
<td>Guerrier-Takada et al (1983)</td>
</tr>
<tr>
<td>Telomerase</td>
<td>template for telomere synthesis</td>
<td>Greider and Blackburn (1989)</td>
</tr>
<tr>
<td>7S RNA</td>
<td>protein translocation (Component of SRP)</td>
<td>Siegel and Walter (1986)</td>
</tr>
<tr>
<td>micF</td>
<td>Modulation of translation efficiency (Regulator RNA)</td>
<td>Inouye (1988)</td>
</tr>
<tr>
<td>RNA I &amp; II</td>
<td>Regulation of replication of the E. coli plasmid colE1</td>
<td>Eguchi and Tomizawa (1991)</td>
</tr>
</tbody>
</table>

(List is not exhaustive and contains only major functional classes of known RNAs).
removed by a series of self-catalysing trans-esterification reactions. The above mentioned self-splicing intron not only excises itself but also undergoes two additional self-cleavage reactions giving rise to a simple form of RNAase, called the 'ribozyme' (Zaug and Cech, 1986). The authors further showed that the above RNA could synthesise polycytidylic acid using pentacytidylic acid as a substrate. Thus, the RNA molecule showed, in addition to the ribonucleolytic activity, a polymerase activity.

(b) The RNA component of the enzyme 'telomerase' from *Tetrahymena thermophila* has been shown to act as an internal template for the addition of telomeric repeats to the ends of chromosomes during replication.

(c) The signal recognition particle (SRP) that plays a key role in targeting secretory proteins to the membrane of the endoplasmic reticulum contains a 7S RNA molecule. This RNA moiety is required not only for the structural integrity of the SRP, but also for the SRP activity (Siegel and Walter, 1986).

(d) 'Guide RNA' works as the template for the correct processing of the RNA transcripts from *Tetrahymena* in a process called 'RNA editing' (Blum et al, 1990).

Through such studies, functional capabilities of RNA molecules are revealed emphatically. As a result, more studies are being directed towards the structural and functional aspects of RNA rather than towards ribosomal proteins. The studies on the ribosomal RNA molecules, with respect to their involvement in the translation process (Moore, 1988), best illustrates this departure.
1.1.2 Structural Aspects of RNA Molecules

The formation of base pairs between complementary bases in the RNA molecule gives rise to double helical (stem) regions interspersed with single-stranded regions. These single-stranded regions are of various types, such as, hairpin, bulge, multibranched and interior loops (Figure 1.1). The ensemble of these structural elements in a planar form represents the secondary structure of RNA. RNA attains a three-dimensional folded structure by means of long-range base pairing interactions between the various regions on the secondary structure.

The predominant element of RNA secondary structure is the hairpin stem-loop motif. Hairpins are formed when a local region of the RNA chain folds back on itself to form a short, intramolecular base paired helix (Figure 1.1-a). Hairpin loops provide potential nucleation sites for RNA folding and also act as motifs for interaction with other macromolecules. There seems to be a preferential selection of a particular nucleotide sequence in the hairpin loops. 70% of all the tetranucleotide loops in rRNAs are of the type UNCG or GNRA (Woese et al, 1983; Gutell and Fox, 1988). Such loops have been found to provide more stability to the hairpin stems.

The helical regions are seldom regular as the segments on the chain brought into opposition do not have entirely complementary sequences. The non-bonded bases 'loop out' of the structure giving rise to bulge, interior, and multibranched loops (Figure 1.1-b,c,d). Most of the ribosomal proteins and translation
Figure 1. Structural motifs on RNA molecules.
repressors bind to the rRNAs by recognising specific types of bulges and stems (Draper, 1989).

Another important structural element on RNA is a pseudoknot (Pleij et al, 1985). Pseudoknot is formed from stem-loop structure, in which the bases outside the stem-loop are paired with those in the loop so as to create a second stem (Figure 1.1-e). The second stem stacks upon the first to form a continuous coaxial helix. This stacking requires the single-stranded connecting nucleotides to cross the grooves of the helix. The pseudoknots are specifically recognized by components of the translation apparatus (Schimmel, 1989; Tang and Draper, 1989; Brierley et al, 1989).

Predictions of the secondary structure of RNA, whether based on free energy calculations or inferred from comparative sequence analysis, are more reliable than those for protein from amino acid sequence (Turner et al, 1988). The double-stranded RNA assumes the A-form DNA structure. Because of the additional structural contributions from the single-stranded regions on RNA (which are not common in the case of DNA), the RNA molecules can assume more varied overall 3D shapes than the duplex DNA. As a result, discrimination among the various functionalities can be very easily achieved in RNA (Steitz, 1990).

The versatility in both the functional and the structural aspects of RNA molecules as enumerated above initiated us to undertake a study on RNA molecules.
1.2 RATIONALE FOR THE CHOICE OF TRANSLATION PROCESS TO STUDY RNA MOLECULES

1.2.1 Translation Process Exercises Controls on Gene Expression

Certain genes are expressed only at specific time frames of the cell cycle; while some are expressed in response to the 'stress trauma' (heat shock, osmotic pressure, etc). Cellular proteins can be categorised into 'abundant' and 'rare' proteins. Thus, there is a differential expression of genes. The importance of transcription as a control step in gene expression was recognised relatively early in the history of molecular biology. However, there is a growing evidence that the translation process provides a number of control points contributing to the differential expression of gene (Bergmann and Lodish, 1979; McCarthy and Gualerzi, 1990; Gold, 1988). Some examples of such controls are given below.

(a) The interactions between the components of translation machinery and mRNA (in particular, the translation initiation region on mRNA) define the molecular basis for translational control on gene expression (McCarthy and Gualerzi, 1990). mRNAs with less secondary structure in the translation initiation region have a higher rate of translation-initiation (Sorensen et al, 1989). mRNAs of highly expressed genes are found to possess 'enhancer sequences' (in addition to ribosome binding sites) that would form additional contacts with the ribosome (McCarthy and Gualerzi, 1990).

(b) A bias in the codon context exerts control on gene expression via its effects on the kinetics of tRNA translocation
on the ribosome. mRNAs of high expression use only a subset of codons for which the levels of tRNAs are very high (Grosjean and Fiers, 1982).

(c) The expression of genes in poly-cistronic operons is often not fully independent, but is coupled to the expression of neighbouring genes. The downstream cistrons are prevented from being translated by inhibitory structures on mRNA. These structures are 'opened' by the ribosomes that are translating the end of the upstream cistron. As a result, fresh ribosomes access the second cistron.

(d) A number of regulatory circuits have been identified at the level of translation. The types of regulation are (i) protein-mRNA repressor systems in which the protein represses the translation-initiation by attaching itself to the ribosome binding site on mRNA; (ii) protein-mRNA activation systems in which the protein activates the initiation by melting the inhibitory secondary structure; (iii) ribosome-mediated control as reflected in translational coupling.

1.2.2 Manifestation of RNA Structures in the Translation Process

Translation can be visualised as a programmed process with the ribosome as the apparatus, and the RNA components (tRNAs, rRNAs) as the mechanical tools to build the polymer of amino acids by reading the triplet code from the mRNA. The translation process involves 3 rRNAs, a minimum of 40 tRNAs, and the mRNA that is being encoded. Thus, there is a large congregation of RNA molecules participating in the translation process. This
process encompasses a varied manifestation of RNA structures as enumerated below.

(a) The process of 'tRNA identity' (assigning correct amino acid residue to the tRNA) brings out very fine features of RNA structures (Schimmel, 1989; Hou et al, 1989). tRNAs typically comprise 74-93 nucleotides. A set of 15 conserved nucleotides facilitates the formation of an L-shaped tertiary structure, which is believed to be similar for all tRNAs (Rich and RajBhandary, 1976). Some of these nucleotides provide common bases that are important for interactions with the translation machinery. The rest of the nucleotides are varied within the constraints of the cloverleaf form of secondary structure and this variation enables the aminoacyl tRNA-synthetases to distinguish one tRNA from another. It has been shown that the G3.U70 base pair in the acceptor arm of E. coli alanine-tRNA is enough for the specificity of alanyl tRNA-synthetase (Hou and Schimmel, 1988); alteration of this G.U pair stopped any aminoacylation. Changing the G3.C70 base pair of the acceptor arm from the phenylalanine- and cysteine-tRNA to G.U base pair makes these tRNA molecules to recognise the alanyl tRNA-synthetase (Hou and Schimmel, 1988; McClain and Foss, 1988).

(b) Specific structures on mRNAs change the coding frame. In the case of the gene 60 of the bacteriophage T4, the E. coli ribosome 'hops' by 50 nucleotides from a glycine codon (that specifies the amino acid 46 in the mature protein) to a leucine codon, specifying the amino acid 47 (Huang et al, 1988). This hopping of the ribosome is achieved by recognising a stem-loop structure on the mRNA.
(c) An inframe terminator codon UGA in the gene for RF2 protein of E. coli is avoided by a +1 frameshift event (Craigon and Caskey, 1986). Such a ribosome 'slippage' is aided by the presence of the Shine-Dalgarno sequence in the region 5' to the frameshift site. The Shine-Dalgarno sequence forms base pairs with 16S rRNA so as to anchor the mRNA with the ribosome to facilitate the ribosome slippage (Weiss et al, 1988).

(d) Specific sequence structures on the mRNAs make the ribosome pause, thereby aiding the sequential folding of protein domains during translation. Such a mechanism has been shown in the case of yeast pykl gene where a string of five consecutive rare codons occurs at a position corresponding exactly to the boundary of two protein domains in the encoded protein (Purvis et al, 1987).

The above observations indicate that the translation process is a fertile ground for the study of RNA molecules.

1.3 RNA MOLECULES CAN BE STUDIED THROUGH THEORETICAL APPROACH

1.3.1 Theoretical Studies Complement Experimental Studies

The structural and functional aspects of RNA molecules can be studied both theoretically and experimentally. As shown below, there are instances in the past which show that both the approaches have complemented each other.

(a) The X-ray crystallography data on tRNA molecules not only confirmed the theoretically predicted structures but also brought out the other types of tertiary interactions that are
possible in RNA. The secondary structures of rRNAs as proposed
through the approach of 'comparative sequence data analysis' [as
advocated by Fox and Woese (1975)] have been confirmed subse-
quently by chemical modification and crosslinking studies.

(b) On sequencing the 3' end of 16S rRNA from *E. coli* and
with a very small number of available *E. coli* mRNA sequences,
Shine and Dalgarno (1974) could propose a base pairing interac-
tion between mRNA initiation region and 3' end of 16S rRNA. They
further proposed that such a base pairing aids in the binding of
mRNA to ribosome for translation. This was subsequently demon-
strated experimentally (reviewed by Steitz, 1980).

(c) Lerner et al (1980) and, Rogers and Wall (1980) pro-
posed that the RNA component of U1 snRNP recognises the splicing
region and forms base pairs with both the ends of an intron.
This base pairing scheme aligns the intron for cutting and
splicing. This was subsequently proved (Kramer et al, 1984;
Padgett et al, 1986) and has now become the accepted mode of
operation for the snRNPs.

(d) The theoretically proposed possibility, that RNA (and
not DNA) could have been the original genetic material (Crick,
1968; Orgel, 1968), has gained widespread acceptance following
the more recent discoveries of the catalytic and enzymatic
properties associated with RNA molecules (Lamond and Gibson,
1990; Kruger et al, 1982; Guerrier-Takada et al, 1983) and has
opened up a new approach of 'ribozyme-mediated anti-viral thera-
py' (Rossi and Sarver, 1990).
1.3.2 Nucleic Acid Sequence Data is a Source of Information for the Biological Processes

The genome is organised in terms of genes interspersed with non-transcribed portions. The building blocks of the genome are the four nucleotides A, C, G, T in the case of DNA genome and A, C, G, U in the case of genomic as well as non-genomic RNAs. The primary function of the specific order of the nucleotides in the coding RNA is to determine the amino acid sequence of the encoded protein. However, the redundancy in the genetic code has given the flexibility to express the same amino acid order in various possible ways of arranging the four nucleotides in the gene. As a result, the same region on the gene is enabled to have other functions. The region may possess sites for regulatory mechanisms or it may be involved in various processes, such as, positioning histones on the DNA, folding the mRNA, etc. The ordering of nucleotides is further subjected to other constraints, such as, the GC/AT content of the genome, the tRNA pool size, the level of expressivity of the gene (reflected by the codon bias/context), etc. It has been shown by earlier workers that in mRNAs, the constraints on the choice of a nucleotide at a particular position seem to arise from the nucleotides present at as far as 5 bases away (Pandit et al, 1986). This is supported by the codon context effect on the choice of synonymous codons (Shpaer, 1986).

Thus, the occurrence of the nucleotides in the genes is quite non-random and has evolved with respect to the functional requirements. A given sequence can be classified as of prokary-
otic or eukaryotic origin by the simple calculation of the preferences/discriminations of some dinucleotides versus others (Nussinov, 1991). We can guess whether it is a coding or a non-coding sequence by scrutinising the base sequence and the reading frame for minimal deviation from the primitive message of the form RNY coding triplets (Shepherd, 1981). One can determine whether it is an intron or an exon region (Konopka, 1989); simple statistics of the occurrence of the nucleotide oligomers is able to uniquely characterize functional domains like peptide coding, intervening segments, etc, on the genome (Smith et al, 1983).

The sequences possess specific patterns that differentiate them from random permutations (Nussinov, 1991). Some examples of specific patterns are (a) the RNY triplet pattern (Shepherd, 1981) in the coding region, (b) the frame-monitoring code - in the form of preference for G-nonG-N triplets in the coding region (Trifonov, 1987), and (c) the flanking of homooligomers by the complementary nucleotide/complementary doublet in the DNA sequences (Nussinov, 1991). In addition, the DNA and RNA molecules contain signals that are recognised by regulatory proteins, enzymes or other RNA molecules. Signals for various biological processes, such as, splicing, transcription-initiation and -termination, translation, recombination, etc, occur either in the form of mandatory context signals (that are found in the vicinity of the biological feature site) or as 'enhancer elements' (that are found in the 'upstream' or 'downstream' regions of the biological feature sites). A collection of various signal
sequences is given by Trifonov and Brendel (1986).

Thus, the order of nucleotides in biological sequences follows some basic rules. Nucleic acid sequence has been modelled as a chain generated by 'sequence generator' (Almagor, 1983). The generator is guided by a set of rules or constraints that are supposed to act on the sequence. Sequence data analysis has emerged as a powerful tool to decipher the information encoded in the nucleic acid sequence to understand the various biological processes.

The aim of the work reported in this thesis is to understand, through sequence data analysis, the functional and structural organisation of RNA components involved in the translation machinery.

1.4. AIDS AND TOOLS FOR THEORETICAL STUDIES ON THE NUCLEIC ACID SEQUENCE DATA

1.4.1 Data Bases for Nucleic Acid Sequences

1.4.1.1 General data bases Ever since it was realised that the nucleic acid sequences provide the fundamental starting point for describing and understanding the biological processes, molecular biologists were trying to sequence the genes of their interest. Developments in sequencing techniques were primarily brought about by Sanger and coworkers (1977), and by Maxam and Gilbert (1977). When the methods developed by these workers were subsequently improved, a large number of genes were sequenced and reported in the literature. As a result, the need to collect,
collate and store the sequence data was felt. Due to the joint efforts of the European Molecular Biology Laboratory (EMBL), Germany and the National Institute of Medical Sciences of the U.S. National Institute of Health, the centralised collection of all published sequences began in 1980. In 1982, these laboratories came out with their first release of the data bases, namely, EMBL data library and GenBank, respectively. In 1987, DNA Data Bank of Japan (DDBJ) began operations. However, the EMBL and GenBank data bases are more popular than DDBJ. Both these data bases have chosen flat file organisation so that they can be accessed even by non-computer professionals. They are distributed to users in various types of media, like, magnetic tape spool, tape cartridge, floppy diskette, CD-ROM diskette, etc, for machines ranging from personal computers to super computers.

The primary data in these data bases is the nucleotide sequence. The data base provides the following information about the individual sequence.

(a) Identification - a unique name associated with each entry along with the gene name, and the encoded protein name.

(b) Bibliographic context - journal citation in which the sequence is reported.

(c) Physical context - organism, chromosome, map position (taxonomic classification) pertaining to the sequence.

(d) Functional context - secondary structural elements, and features, such as, protein coding regions, regulatory sites, signal sites, etc.

(e) Administrative context - date of entry, review of the
sequence.

(f) Sequence - as 60 bases in each line with a subdivision of groups of 10 bases.

1.4.1.2 Specialised data bases

Specialised data bases are created in the following situations:

(a) sequences for homologous genes from various species are available (e.g. data bases on rRNAs, tRNAs, etc.);

(b) a major number of genes of a particular species have been sequenced (e.g. E. coli data base); and,

(c) regions responsible for particular biological function are available from various genes (e.g. eukaryotic promoter sequence data base).

These data bases, unlike the EMBL or the GenBank, contain only those sequences that deal with a specific biological process. Thus the specialised data bases serve as the platform to further characterise specific processes, such as, translation-initiation in E. coli or transcription-initiation in eukaryotes etc. The following are some of the commonly used specialised data bases.

(a) Compilation of small subunit ribosomal RNA sequences

This is a collection of all known (complete and partial) small subunit ribosomal RNA sequences (Neefs et al, 1990). The sequences are listed in 5 groups, eukaryotic-, archaeabacterial-, eubacterial-, plastidial- and mitochondrial- small subunit ribosomal RNAs. The sequences are presented in the aligned form. The secondary structural specifications, such as, the double-stranded
and/or the single-stranded regions, are also given. This data base is available both on hard copy as well as computer readable format.

(b) **Compilation of large subunit ribosomal RNA sequences**
This provides the collection of all the published large subunit rRNA sequences (Gutell and Fox, 1988). The sequences are presented in the form of secondary structure diagrams in which the individual nucleotides are indicated. This data base is presented in the form of hard copy wherein the structures are drawn to a single scale so that they may be superimposed or intercompared.

(c) **Compilation of DNA sequences of *E. coli***
This is a compilation of DNA sequence data of the genes from *E. coli* K12 (Kroger, 1989). The genetic map data is also given and the sequences are arranged accordingly.

(d) **Compilation of sequences of tRNAs and their genes**
This data base contains all the published tRNA sequences and presented in the aligned form (Sprinzl et al, 1989). The numbering system derived for yeast tRNA-phe, as adopted in the Cold Spring Harbour Symposium on tRNA (1979), is used. The secondary structures of tRNAs are indicated. The data base is available on diskettes as well as in the printed form.

In addition, data bases of cloned human DNA, histone genes, restriction enzyme recognition sequences, and eukaryotic promoters are also available.

1.4.2 **Tools for Computer Analysis of Nucleic Acid Sequence Data**

The need to search, manipulate, and analyse nucleic acid/
protein sequences has attracted the attention of several investigators from various fields, like, mathematics and statistics, computer science and graph theory, protein and nucleic acid chemistry, etc. As a result, several methods and computer programs have become available to analyse sequences (Methods in Enzymology, volume 183).

1.4.2.1 Types of analysis on nucleic acid sequence data

The common analysis on nucleic acid sequences can be broadly classified as follows.

(a) Searching data bases for similarity

Often a data base search provides the first insight into the mechanism of action of a newly synthesised gene. The sequence determination itself is often guided by comparison with homologous sequence (Doolittle, 1990). 'Homology domains' between two or more sequences can be found out. Such domains will define the common structural, functional, or genetic regions among the genes.

(b) Searching for patterns

This will include finding the functional feature sites, such as, protein coding regions, splice junctions, promoter sequences, tRNA-like sequences, regulatory sites, repeat sequences, etc, on the genes. The approach commonly used is to assess the statistically significant features or to look for signatures that characterise the known signal sequences.

(c) Predicting RNA secondary structures

Biological functions, such as, regulation of transcription and translation, catalysis, and transfer of proteins across membranes are carried out by specific secondary structures of RNA molecules. Predic-
tion of the secondary structures of RNA is essential to study such biological functions. There are two methods of predicting the secondary structure - phylogeny and energy minimisation. Phylogeny relies on alignment and subsequent folding of functionally analogous RNA sequences (from several species) into similar structures. Energy minimisation relies on thermodynamic parameters to determine the minimum free energy foldings of RNA sequences. In addition to these methods, an algorithm based on graph theoretical approach has been developed to predict secondary structure of large RNA molecules (Thanaraj et al, 1989). The method takes into account available experimental data on the conformation of bases and also the phylogenetic information.

(d) Aligning nucleic acid sequences  
Alignment defines the relationship between sequences on a base-to-base basis. Aligned bases are presumed to be related in an evolutionary and/or functional sense. Alignment studies aid in deciding whether two or more sequences exhibit similar structure/function or share a common ancestor. They also help to establish the genomic divergence through gene rearrangements (Sankoff, 1990).

The widely used packages that facilitate the above mentioned analyses are PC-Gene (Intelligenetics, Inc., USA) for IBM/PCs, and UWGCG (University of Wisconsin, USA) for VAX computers. However, there are several packages which are developed by individual workers to suit their own requirements and are available through private circulation (Staden, 1989; Zuker, 1989).

1.4.2.2 Approaches for sequence data analysis  
Sequence data
analysis mainly addresses the question of what information is encoded in the sequences, i.e. what the sequences tell us about their biological functions. The analysis of sequence data is carried out using two complementary approaches, namely, the structural approach and the functional approach (Nussinov, 1987; Trifonov, 1989). These are described below.

(a) The structural approach deals with the statistical analysis of the sequences. It detects and substantiates the non-randomness of nucleotide occurrences (Blaisdell, 1983; Karlin and Ghandour, 1985). It also involves search for strings and patterns unusual by their occurrence or structure (Konopka et al, 1986; Arques and Michel, 1987; Nussinov, 1991). After the features, such as, dyad symmetry elements, runs of purines, conspicuous repeats, sequence homologies, etc, are found, the question arises as to what might be their functions.

(b) The functional approach, in contrast, starts from the other end and asks question such as what would be the characteristic sequence/structure required to serve a given biological function (Galas et al, 1985; Kozak, 1987). Every biological function is realised through some molecular structure, which manifests most of the time through a sequence of nucleic acid. This justifies the search for the functional codes in the sequences, rather than the search for the non-randomness of the nucleotides. These searches for functional code can be classified into 2 types - the first involves searches on the sequence in the vicinity of the biological feature site and the second type involves searches within a given specified distance from the
biological feature site.

The work reported in this thesis mainly utilises the functional approach of the sequence data analysis.

1.5 SCOPE OF THE PRESENT STUDY

The discussion in the previous sections brings out the following features.

(a) RNA molecules perform a broad spectrum of functions. They assume more varied structures than DNA molecules. The versatility in both the structural and functional aspects of RNA molecules provides a good scope to undertake studies on RNA molecules.

(b) The translation process involves a large interplay between a variety of RNA molecules. Of all the cellular processes, the translation process seems to be a fertile ground for the study of RNA molecules.

(c) Theoretical approach of sequence data analysis is a valuable tool to extract the information about the functional and structural features of RNA molecules.

(d) 'Sequence data analysis' of nucleic acid sequences has emerged as a separate branch of Life Science.

The features such as mentioned above initiated us to undertake a study to understand the involvement of RNA molecules in the translation process through the approach of sequence data analysis.