Dedicated to my parents and those selfless people who helped in my development

CHAPTER ONE

GENERAL INTRODUCTION

I.1 The Genetic Code – an Overview

I.2 Protein Biosynthesis
I.1 The Genetic Code – An Overview

Where did the building blocks for life come from? How were amino acids and nucleotide bases first formed in the early earth, and how did they come together in such a way as to encode and transmit information? Organic molecules are easy to come by (and have been isolated from extraterrestrial sources), but how they have managed to arrange
themselves to form macromolecules and, ultimately, organisms, is far from clear even till today. In 1952, it was Dounce [1] who first proposed a model for protein synthesis based on a relationship between RNA and amino acids. He postulated that "the specific arrangement of amino acid residues in a given peptide chain" is derived from the "specific arrangement of nucleotide residues in a corresponding specific nucleic acid molecule" and that amino acids were specified by the "immediate surroundings" of the nucleic acid bases. In 1953, a year later, Watson and Crick [2] revealed that it was the pairing pattern of bases in DNA that suggests a possible copying mechanism for the genetic material. This immediately brought the genetic code into the limelight as the most important unsolved question till then in biology and the quest to solve the mystery began. The cosmologist Gamow in 1954, proposed a "diamond code", with each amino acid specified by a quartet of bases in the DNA double helix [3]. This proposal was doomed from the start, because proteins are synthesized in the cytoplasm and not in the nucleus of eukaryotic cells, as had been shown by Borsook et al. in 1949 [4]. This proposal was also unsupportable because it allowed only certain amino acids to adjoin each other in proteins, and no such restrictions on amino acid sequences actually exist.

The period between 1961 and 1962 is often referred to as the “coding race” because of the competition between Ochoa’s and Nirenberg’s labs. Indeed, the two laboratories completed the base composition part of the code almost simultaneously. However, it was Har Gobind Khorana and Robert Holley who shared the 1968 Nobel Prize for Physiology and Medicine along with Marshall Nirenberg. In 1961, Nirenberg and Matthaei showed that a synthetic messenger RNA made up of only uracils can direct protein synthesis. Their experiment showed that an RNA polymeric chain consisting of only the repeating base uracil forced a protein chain made of one repeating amino acid, phenylalanine. The polyU messenger RNA (mRNA) resulted in a polyphenylalanine protein [5]. By 1965, Nirenberg, with help from his NIH colleagues, had become the first to complete the sequencing of the code. In subsequent years, Nirenberg and his group deciphered the entire genetic code by matching amino acids to synthetic triplet nucleotides. Nirenberg and his group also showed that, with some few exceptions, the genetic code was universal to all life on earth. Khorana was also responsible for producing the first man-made gene in his laboratory [6]. His team had established that the mother of all life codes, the biological language common to all living organisms, is
spelled out in three-letter words: each set of three nucleotides codes for a specific amino acid. Holley (1964) determined the sequence and structure of alanine transfer RNA or tRNA [7,8] - the first nucleotide sequence of an RNA fully determined - which was a key discovery in explaining the synthesis of proteins from messenger RNA.

The following features together make the genetic code an important field of research today:

1. Creation of the code: How the genetic code was set up is still not known. Most features of the code appear as largely universal and independent of species. It has also survived in this same form for at least one billion years. Research needs to probe why the genetic code has had such an invariant form for so long. Some recent findings as enumerated above seem to indicate that it is slowly evolving to this day, and the reasons for this need to be discovered.

2. Universality of the code: Apart from a few primitive forms of life, the genetic code is universal for all species of life, pointing to a common origin. Even the few exceptions known differ from the universal code only in some relatively minor features.

3. Molecular basis of the code: Each word (or codon) of the genetic code is a triplet (a trinucleotide sequence) and the words are arranged continuously without punctuation and without any overlapping. The triplet form of the code thus calls for a total of $4^3$ or 64 codons to code for the 20 amino acids of life, which implies degeneracy of the genetic code.

I.2 Protein Biosynthesis

The cells in our body synthesize proteins by a complex, multi-step and highly controlled process under the direction of DNA, which carries all the genetic information for the primary structure of all proteins. This information from DNA is transferred (copied) to the messenger RNA (mRNA) during the process called transcription. The mRNA is then utilized for the actual assembly of amino acids to form polypeptides/ proteins through peptide linkages during the process called translation. In some organisms where DNA is
absent, protein biosynthesis occurs under direct control of genetic RNA. In a living system the process of protein synthesis requires a number of enzymes, transfer RNAs and ribosomes. The two processes of transcription and translation that convert the base sequence of a particular segment of DNA to the amino acid sequence of synthesized poly-peptide chains are governed by the central dogma. The flow of information is one-way - that is, from DNA, information is transferred to RNA and from RNA to protein. However, the reverse flow of transcriptional information can take place from RNA to DNA, as in retroviruses, but never from protein to RNA. **Fig.1.1 (next page)** gives a schematic depiction of the process of biosynthesis of proteins, including the steps of transcription of the gene from DNA (mRNA synthesis) and translation (synthesis of the protein chain itself).

The process of transcription of genetic information results in the synthesis of an mRNA corresponding to the gene which encodes for the
Fig. I.1 Schematic depiction of the process of protein biosynthesis

protein concerned. This process in prokaryotes is different from that in eukaryotes. In prokaryotes, for example in bacterial systems like *E. coli*, a single RNA polymerase species is responsible for synthesis of all kinds of RNAs (rRNA, mRNA and tRNAs). However, in the case of eukaryotes there are at least three different nuclear RNA polymerases for synthesis of various RNAs. The transcriptional processes in mitochondria and chloroplasts are also different from the normal nuclear transcriptional processes in many respects. The primary transcript (nascent RNA) displays a paucity of
primary sequence information which is remedied by various post-transcriptional processes. One such process is that of nucleotide modification. Modified nucleotides perform a large number of functions [9,10] in life.

The process of translation is more complex than transcription and involves initiation, elongation and termination of polypeptide chains. For initiation and elongation of a polypeptide chain, the formation of aminoacyl transfer RNAs (AA-tRNAs) is a prerequisite. Formation of an aminoacyl-tRNA (see below) requires amino acids, tRNAs, aminoacyl-tRNA synthetases, ATP and Mg\(^{2+}\).

![Aminoacylation Reaction](image)

The overall process of translation (Fig. 1.2, next page) is summarised as follows:

1. Protein synthesis is initiated by formylmethionine attached to an initiator tRNA\(_{\text{fMet}}\) which recognizes the start codon (AUG) by anticodon-codon pairing.

2. The next (incoming) aminoacylated tRNA recognizes the next codon in the mRNA, and so on till the entire mRNA for a given protein is read.

3. A peptide bond is formed between the terminus of the synthesized chain and the incoming AA-tRNA charged with the amino acid next in sequence.

4. Translocation (movement of the ribosome relative to the mRNA) allows the cycle of protein synthesis to continue.

5. Protein synthesis ends when the stop codons (UAA, UAG or UGA) are read along the mRNA sequence.
In the eukaryotic translation process, the initiating amino acid (methionine) needs to be formylated, so there are two separate tRNAs for methionine, viz., tRNA_fMet and tRNA_Met. In prokaryotes the initiating amino acid methionine is not formylated so there is only one tRNA for methionine, viz., tRNA_Met. Another important difference between the eukaryotic and prokaryotic translation processes is that the ribosomes enter the mRNA at the AUG codon or the nearby Shine-Delgarno site in eukaryotes, while in
prokaryotes, the ribosomes enter at the capped 5' end of mRNA and then advance to the AUG codon by linear scanning.

The translation apparatus in mitochondria and chloroplasts differs from that in the cytoplasm of eukaryotes. Ribosomes in these organelles are smaller in size (70S) than those in eukaryotic cytoplasm (80S). The number of tRNAs in mitochondria is 22, while in the cellular cytoplasm the number is 55. The antibiotic effects on translation in organelles also differ from that in cytoplasm. For example, the 70S ribosomes (found in organelles) are sensitive to chloramphenicol but not to cycloheximide, while the 80S ribosomes (found in cytoplasm) are sensitive to cycloheximide but not to chloramphenicol.

**The adaptor hypothesis.** Among Francis Crick's great contributions in the area of the genetic code and of the expression of the genetic information was the enunciation of the "adaptor hypothesis" and, later on, the "wobble hypothesis". The "adaptor hypothesis" arose when Crick analyzed some new ideas of George Gamow. He generalized them and explained them in a paper (never published in a "proper" journal). The main idea was that it is very difficult to consider how DNA or RNA, in any conceivable form, could provide a direct template for the side-chains of the twenty standard amino acids. He therefore proposed a theory in which there were twenty adaptors (one for each amino acid), together with twenty special enzymes. Each enzyme would join one particular amino acid to its own special adaptor. This combination would then diffuse to the RNA template. An adaptor molecule could fit in only those places on the nucleic acid template where it could form the necessary hydrogen bonds to hold it in place. Sitting there, it may carry its amino acid to just the right place where it is needed.

The processes that ensure the fidelity of DNA replication have been studied extensively, most probably because of the vast importance of gene mutations in evolution, hereditary diseases and cancer [11]. On the contrary, till date, less is known about the fidelity of transcription (the process where genetic information is read from the DNA) and translation (the process where genetic information is translated into the amino acid sequence), which indisputably are very crucial for the functional maintenance of cells and organisms. Functional degeneration of non-dividing cells limits
the life-span of organs such as the heart or brain. Therefore, decreasing fidelity of gene expression may be an underlying cause of cellular aging, degenerative diseases, and ultimately termination of life. Accurate gene expression is dependent on the fidelity of both transcription of DNA and translation of mRNA. Transcription and translation together constitute the process of protein biosynthesis in all kinds of organisms. In the course of protein expression, the mRNA sequence is faithfully and accurately read by the ribosome according to the sequence of codon triplets that starts with the methionine codon AUG and constitutes the reading frame through to a termination (stop) codon.

The processes of the genetic code (transcription and translation) can operate only within the same generation of cells or individual life form, whereas DNA replication spans the generations. DNA replication is responsible for passing on the parental characters from one generation to its next by maintaining supreme fidelity. Changes in the genetic material is believed to have helped the organisms to adapt themselves to the gradually changing geographical and environmental conditions of the earth by slowly but certainly changing them genotypically as well as phenotypically, or, in other words, making them the fittest to survive (evolution). Unfortunately, fidelity of DNA replication is also responsible for the inheritance of various hereditary diseases like colour blindness, haemophilia, sickle cell anaemia, phenylketonuria etc.

Fidelity of the translation process is maintained by allowing only certain codon-anticodon pairs to occur (called 'cognate') and prohibiting all others (called non-cognate). This ensures that only the correct protein is synthesized, having an amino acid sequence corresponding directly to the base sequence of the mRNA and the DNA gene.

1.3 Features of the Genetic Code

A coding system as exact and non-ambiguous as the genetic code of life requires various features to ensure its fidelity and efficiency in the transmission of genetic information from DNA to mRNA to protein.
The codon dictionary. The complete genetic code given below in Table I.1 was established around 1966 by using mainly *Escherichia coli* systems [12]. It consists of 64 triplet codons built up from the four major RNA nucleosides uridine (U), cytidine (C), adenosine (A) and guanosine (G). 61 of these triplets code for the 20 primary amino acids of proteins, namely, phenylalanine (Phe), serine (Ser), tyrosine (Tyr), cysteine (Cys), leucine (Leu), tryptophan (Trp), proline (Pro), histidine (His), arginine (Arg), glutamine (Gln), isoleucine (Ile), asparagine (Asn), threonine (Thr), lysine (Lys), valine (Val), alanine (Ala), aspartic acid (Asp), glutamic acid (Glu) and glycine (Gly).

<table>
<thead>
<tr>
<th>First Position (5'-end)</th>
<th>Second Position</th>
<th>Third Position (3'-end)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U</td>
<td>C</td>
</tr>
<tr>
<td>U</td>
<td>UUU Phe</td>
<td>UCU Ser</td>
</tr>
<tr>
<td></td>
<td>UUC Phe</td>
<td>UCC Ser</td>
</tr>
<tr>
<td></td>
<td>UUA Leu</td>
<td>UCA Ser</td>
</tr>
<tr>
<td></td>
<td>UUG Leu</td>
<td>UCG Ser</td>
</tr>
<tr>
<td>C</td>
<td>CUU Leu</td>
<td>CCU Pro</td>
</tr>
<tr>
<td></td>
<td>CUC Leu</td>
<td>CCG Pro</td>
</tr>
<tr>
<td></td>
<td>CUA Leu</td>
<td>CGA Pro</td>
</tr>
<tr>
<td></td>
<td>CUG Leu</td>
<td>CGG Pro</td>
</tr>
<tr>
<td>A</td>
<td>AUU Ile</td>
<td>ACU Thr</td>
</tr>
<tr>
<td></td>
<td>AUG Met*</td>
<td>AGC Thr</td>
</tr>
<tr>
<td></td>
<td>AUU Ile</td>
<td>ACA Thr</td>
</tr>
<tr>
<td></td>
<td>GUA Val</td>
<td>GCC Ala</td>
</tr>
<tr>
<td></td>
<td>GUG Val</td>
<td>GCA Ala</td>
</tr>
</tbody>
</table>

Table I.1 The universal codon dictionary

Features of the universal genetic code. The code is firstly non-overlapping, which means that a base in an mRNA is not used for two different codons. In Fig. I.3a below, it is shown that if overlapping of codes is allowed then four amino acids will result from only six adjoining bases. Thus CAG, AGU, GUU and UUC will code for four amino acids Gln, Ser, Val and Ile respectively. But in actual practice these six bases code for only two
amino acids, Gln and Ile with codons CAG and UUC respectively, showing the code is non-overlapping. From **Fig. 1.3b** below, we can also explain the fact that the genetic code is comma-free and continuous, *i.e.* no punctuations are needed between two code words.

![Fig. 1.3a Overlapping of codons due to two bases](image)

![Fig. 1.3b Non-overlapping and comma-free codon sequences](image)

The criterion that each codon can specify only a single amino acid is called the *specificity* of the genetic code. There are 64 such code words in the mRNA codon dictionary, of which 61 words code for the 20 essential amino acids, while 3 code words, namely, UAG, UAA and UGA code for chain termination, called the *stop* or termination or nonsense codons. For initiating the polypeptide chain, two codons, namely, AUG (most often) and GUG (rarely), are used, called *initiation* or start codons.

The code words present in DNA and mRNA occur as trinucleotide sequences or triplets, whose identity is determined by the three nucleobase residues present. Why a triplet code is favored over a doublet or a quartet code is because there are 4 distinct bases in mRNA, and the total number of code words in the mRNA codon dictionary is 64. For doublet (2 letter) codons, the total number of code words would be $4^2 = 16$ which is inadequate to code for all the 20 amino acids. For triplet (3 letter) codons, the total number of code words would be $4^3 = 64$ which is adequate to code for all 20 amino acids, but more than one codon will code for most amino acids. This implies that the genetic code is *degenerate*, which perhaps is the most intriguing feature of the genetic code [13, 14].
Table 1.2 Degeneracy of the genetic code

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>No. of codons</th>
<th>Amino acid</th>
<th>No. of codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met</td>
<td>1</td>
<td>Tyr</td>
<td>2</td>
</tr>
<tr>
<td>Trp</td>
<td>1</td>
<td>Ile</td>
<td>3</td>
</tr>
<tr>
<td>Asn</td>
<td>2</td>
<td>Ala</td>
<td>4</td>
</tr>
<tr>
<td>Asp</td>
<td>2</td>
<td>Gly</td>
<td>4</td>
</tr>
<tr>
<td>Cys</td>
<td>2</td>
<td>Pro</td>
<td>4</td>
</tr>
<tr>
<td>Gln</td>
<td>2</td>
<td>Thr</td>
<td>4</td>
</tr>
<tr>
<td>Glu</td>
<td>2</td>
<td>Val</td>
<td>4</td>
</tr>
<tr>
<td>His</td>
<td>2</td>
<td>Arg</td>
<td>6</td>
</tr>
<tr>
<td>Lys</td>
<td>2</td>
<td>Leu</td>
<td>6</td>
</tr>
<tr>
<td>Phe</td>
<td>2</td>
<td>Ser</td>
<td>6</td>
</tr>
</tbody>
</table>

**Degeneracy of the genetic code.** Degeneracy of the genetic code means that, often, many codons are synonymous for the same amino acid. Experiment shows that six-fold degeneracy exists for the amino acids arginine, leucine and serine, four-fold degeneracy for alanine, glycine, proline, threonine and valine, three-fold degeneracy for only isoleucine, two-fold degeneracy for asparagine, aspartic acid, cysteine, glutamine, glutamic acid, histidine, lysine, phenylalanine and tyrosine, and no degeneracy for methionine and tryptophan (Table 1.2 above).
Codon degeneracy may be attributed to either or both of two phases in the translation process where scope exists for accommodating it. The first phase consists of enzyme-assisted aminoacylation of tRNA with the cognate amino acid. The second phase is that of allowed hydrogen-bonded pairing between the codon and the anticodon. In 1966, Francis Crick in his famous wobble hypothesis proposed that the lack of preciseness in pairing between wobble bases during codon-anticodon pairing results in degeneracy of the genetic code. The work done for this Thesis has been devoted to investigate the role of only the second phase – that of codon-anticodon pairing. This is because the Crick wobble hypothesis focuses only on this phase, and not the other. One's chief aim is to elucidate through computational calculations what the precise factors are that accurately explain and predict both the high specificity and the degeneracy of the genetic code – all on the basis of codon-anticodon pairing.

1.4 Codon-Anticodon Pairing

Adopting the assumption that the specificity and degeneracy of the genetic code are dependent only upon the phase or stage of codon-anticodon pairing, the whole question boils down to which codon-anticodon pairs may be allowed, and which are disallowed in the context of the translation process. Furthermore, one seeks to also delineate the physicochemical reasons why some pairs are allowed while others are not. The patterns emerging out of the complete sets of allowed and forbidden pairs should, as per this hypothesis, fully explain both the high specificity and the degeneracy of the genetic code.

**Codon boxes and half-boxes.** Out of the 64 triplet codons, three triplets are recognized as termination or stop codons. This leaves 61 codons as sense codons, all of which represents an amino acid in the final protein. A *codon box* is a set of four different codons that have the first two bases in common, so that there are altogether 16 codon boxes. Thus a set of triplet codons with the same first two letters represent a codon box. If a full box specifies a single amino acid in the genetic code, it is called a *family* box. Eight codon boxes exist which go to form family boxes, each coding for a single amino acid, and may be referred to as four-fold degenerate codon boxes. The remaining eight
codon boxes are called *mixed* codon boxes (each being split into a pair of codon *half-boxes*).

From Tables I.1 and I.2, it can be seen that the cases of codons cognate for a single amino acid are often clustered together, rather than randomly distributed throughout the code. In cases where an amino acid has two codons (forming a codon *half-family*), the codons are the same in the first two positions and differ only by a transition (a change from one purine to another purine or from one pyrimidine to another pyrimidine at the third position). In cases where an amino acid has four codons, those codons vary only in the third position. In cases where an amino acid has six codons, these arise from one four-codon box and one two-codon half-box. However, the existence of non-degenerate codons for Met and Trp, and the three-fold degenerate case of isoleucine, as well as the three stop codons, present anomalies deviating from these simple rules. It is also not immediately obvious why some amino acids should be encoded by a half-box, others by a full box and yet others by a full box plus a half-box, and two others yet by only one codon. Nor is it immediately evident why the termination codons do not code for any amino acid whatsoever.

**Importance of H-bonding.** In chemistry, a hydrogen bond (H-bond) is a type of attractive *intermolecular force* that exists between two *partial electric charges* of opposite polarity. Although stronger than most other intermolecular forces, the hydrogen bond is much weaker than both the *ionic bond* and the *covalent bond*. Hydrogen bonds do vary in strength, though, ranging from the strong H-bond present in KHF$_2$ to the weak interactions present in carbon H-bonds of the type C-H---X. A widely publicized article [15] proved from interpretations of the *anisotropies* in the Compton profile of ordinary ice, that the hydrogen bond is partly covalent. Other important components of the hydrogen bond include the electrostatic, polarization and dispersion interaction terms.

With the discovery of DNA structure in 1953 [2], Watson and Crick realized that the high fidelity observed in the pairing of complementary DNA sequences is largely due to the unique molecular recognition capability present in naturally occurring nucleic acid bases (A, T, G and C) via Watson–Crick pairing and hydrogen-bonding interactions. In addition to these inter-strand H-bonds, intra-strand base-stacking and inter-strand
cross-stacking interactions involving H-bonds are also equally important in maintaining the bases in a stacked structure along the length of the DNA backbone. Similarly, in other bio-macromolecules, H-bonds determine the three-dimensional structures and are therefore one of the key interactions in molecular biology. H-bond interactions also play a critical role in stabilizing higher-order RNA structures, such as hairpin loops, and thus in the broadest sense, hydrogen-bonding interactions involving nucleic acid base pairs must be considered as playing a salient role in critical areas like genetic coding, biological information storage, and protein synthesis.

Most H-bonds are of the $X{-}H\cdots Y$ types, where $X$ is an electronegative atom and $Y$ is either an electronegative atom with lone electron pair(s) or a region of excess of electron density. H-bonds where $X$ and $Y$ are F, O, or N atoms are well-known, [16,17] and the same is now true also for H-bonds with $Y$ being represented by aromatic π-electrons [18-20]. In 1982, a survey of 113 accurate neutron diffraction crystal structures clearly indicated the widespread occurrence of $C{-}H\cdots O$ hydrogen bonds [21]. The $C{-}H\cdots O$ type H-bonds between water molecules and purines, pyrimidines, amino acids, alkaloids, and others were found in the crystal structures of hydrates of these molecules [22] and are thought to stabilize the conformations of the anticodon loop in tRNA [23] as well as other systems like the β-sheets in proteins [24].

**Anticodon structure**. The anticodon arm is a 5'-base pair stem whose loop contains the **anticodon** [25]. 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. The anticodon base sequence is read from the 5' to the 3' end, and pairs in antiparallel fashion with the codon. For example, the codon for **lysine** is AGU; the anticodon of a lysine tRNA might be ACU. Some anticodons can pair with more than one codon due to **wobble base pairing**. Frequently, the first nucleotide of the anticodon is a modified or minor RNA base, like **inosine** and **pseudouridine**, which can **hydrogen bond** to more than one base in the corresponding codon position. In order 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 anticodon 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 tRNAs are needed to translate unambiguously all 61 sense codons of the standard genetic code [26].

**Fig. I.4** below shows two representations of the structure of tRNA. Part (A) is a cartoon showing the cloverleaf-like secondary structure of tRNA, composed of 4 stems and three loops. Part (B) shows a three dimensional model of tRNA(Phe) from yeast, showing how the secondary structure folds into a tertiary structure. The color coding in Part (B) is the same as for Part (A).

**Fig. I.4 Structure of transfer RNA**

*The codon-anticodon helix.* The heart of decoding involves recognition of hydrogen-bonded Watson-Crick base pairing between the codon (a trinucleotide sequence on mRNA) and the anticodon (a trinucleotide sequence on tRNA found in positions 34 to 36 of the anticodon stem-loop (ASL) of tRNA). This base-pairing results in a short double helix that we call the codon-anticodon helix. A specific codon on mRNA (read from the 5' to the 3' end) always pairs with a specific anticodon of the tRNA in an anti-parallel
fashion. Protein biosynthesis is a complex multistep process that requires, in addition to the ribosome, several extrinsic GTP-hydro-lyzing protein factors during each of the three main stages of initiation, elongation and termination.

The selection of cognate tRNA is believed to occur in two stages— an initial recognition step and a proofreading step—that are separated by the irreversible hydrolysis of GTP by EF-Tu [27-30]. In this scheme, the discrimination energy inherent in H-bonded codon-anticodon base pairing is exploited twice to achieve the necessary accuracy. Recent experiments suggest that the binding of cognate rather than near-cognate tRNA results in higher rates of both GTP hydrolysis by EF-Tu as well as accommodation, a process in which the acceptor arm of the aa-tRNA swings into the peptidyl transferase site after its release from EF-Tu [31]. In both these steps, the higher rate is proposed to be the result of structural changes in the ribosome induced by cognate tRNA.

In the context of proofreading mechanisms alone, it is not yet clear whether additional structural discrimination by the ribosome (over that inherent in the energetics of codon-anticodon base pairing) is actually required for decoding. In another view, it was proposed that the ribosome contributes to the specificity of decoding by recognizing the geometry of codon-anticodon base pairing [32,33]. This is consistent with evidence that the ribosome can discriminate between cognate and near-cognate tRNA with high accuracy in a single step under near-equilibrium conditions [34], so that the recognition and proofreading steps may be needed not only for accuracy as proposed earlier, but for the combination of both high speed and accuracy required in protein synthesis.

Data from chemical protection [35], cross-linking [36], and genetic analysis studies [37] suggest that the decoding site is a region around the ribosomal A site that includes helix 44, the 530 loop, and helix 34 of 16S RNA, which are now known from the high-resolution structure of the 30S subunit to be close to one another [38,39]. As the structure of codon-anticodon pairing in the 30S ribosome subunit has been crystallographically determined [40], it is now possible to obtain direct information about the structural basis of the wobble rules. Recently, in 2003, it has been established that the A-minor motif (the most abundant non-Watson-Crick tertiary interaction in the large ribosomal subunit) is efficiently utilized by the small ribosomal subunit to discriminate
between cognate and near-cognate tRNA, since the universally conserved adenine residues A1492 and A1493 of the 16S rRNA read the shape of the first two base pairs of the codon-anticodon helix [41].

![Codon-Anticodon Pairing Pattern](image)

**Fig. 1.5** Scheme showing of the codon-anticodon pairing pattern

**Cognate and non-cognate pairs.** Any two aminoacyl-tRNAs that participate in standard Watson-Crick interactions with the first two bases in a codon and can form suitable pairs (canonical or non-Watson-Crick) at the third or “wobble” position are designated as *cognate* tRNAs [12,42]. In contrast, tRNAs that do not meet these requirements are commonly referred to as either near- or non-cognate tRNAs. Thus codon-anticodon pairs where the codon and the anticodon both encode for the same amino acid and also have the first two bases of the codon involved in Watson-Crick type of H-bonding are called *cognate pairs*.

This does not mean that all such pairs actually occur in nature or in all species. Some cognate pairs are observed and others not. Near-cognate pairs deviate from fully cognate pairs in having only one position which departs from cognate pairs. Near-
cognate pairs have the codon and the anticodon corresponding to different amino acids, and do not occur during the translation process. They are a special case of non-cognate pairs, where "non-cognate" strictly refers to any disallowed codon-anticodon pair where the codon and anticodon either do not code for the same amino acid, or are unable to pair during translation under normal conditions even if they code for the same amino acid. It is recognized that an amino acid may have more than one anticodon, each pairing with its own set of cognate codons or even with only one codon. Thus more than one set of codon-anticodon pairs may exist for a single amino acid.

1.5 Wobble Base Pairing

In this Section, and indeed in the whole of this Dissertation, the term "wobble pair" is used in a broad sense to describe any H-bond mediated RNA base pair which actually or putatively occurs during codon-anticodon between the third base of the codon and the first of the anticodon, which position is referred to as the "wobble position". This is regardless of whether the base pair is canonical (Watson-Crick type) or not. Earlier definitions of this term have been restricted to certain non-canonical H-bonded RNA base pairs which are able to fit into this position during the translation process, incorporating purines on one side and pyrimidines on the other. However, since this Dissertation focuses on a whole variety of base pairing possibilities at the wobble position, regard-less of whether they actually occur or not, these terms "wobble pair" and "wobble position" are used in the broader sense just described above.

The wobble hypothesis. In 1965, Robert Holley [7] determined the sequence of yeast tRNA\textsubscript{Ala}. The sequence of the anticodon bases in this tRNA contains the minor nucleoside inosine at the 5' position in the anticodon. In 1966, Francis Crick proposed the “Wobble Hypothesis” to explain the discrepancy between the number of codons (64) and the number of amino acids (20). He suggested that while the interaction between the codon in the mRNA and the anticodon in the tRNA needed to be exact (canonical) in two of the three nucleotide positions, this did not have to be so at the third position of the codon. He proposed that non-standard base-pairing might occur between the first base of the anticodon and the third position of the codon. Furthermore, he also
suggested that the unmodified uridine at position 34 of an anticodon (wobble position) would recognize A and wobble to G, but will not recognize U or C (small sized pyrimidine-pyrimidine pair). Crick's original wobble concept is described in Table I.3 below, now recognized to be limited.

**Table I.3** Crick’s original wobble concept

<table>
<thead>
<tr>
<th>5' Anticodon base</th>
<th>3' Codon base</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>C or U</td>
</tr>
<tr>
<td>U</td>
<td>A or G</td>
</tr>
<tr>
<td>I</td>
<td>A or C or U</td>
</tr>
</tbody>
</table>

The Fig. I.6 below depicts how degeneracy arises for the case of alanine, where the three codons GCU, GCC and GCA may all pair with the single alanine anticodon IGC due to the wobble properties of the anticodon wobble base I (inosine).
Base pairing between the anticodon of alanyl-tRNA^{Ala1} and the mRNA codons GCU, GCC and GCA according to the wobble hypothesis.

**Extensions of original wobble concept.** Crick's wobble hypothesis stated that a nearly constant geometry (or configuration) of base-pairing is maintained in forming all wobble base-pairs between the tRNA anticodon and mRNA codon on the ribosome. When accepting an aminoacyl-tRNA, the ribosome requires the maintenance of a specific geometry for the anticodon-codon base-pairing. By the early 90s, with the discovery of a large number of modified nucleosides at tRNA position 34, the original Crick’s wobble hypothesis was amended to explain the decoding properties of the newly discovered nucleosides. The current status [43] of the wobble rules is tabulated in Table I.4 (next page). The newly found decoding properties of unmodified uridine and its derivatives (especially
**Table I.4** Current status of wobble rules for codon-anticodon pairing

<table>
<thead>
<tr>
<th>AWB</th>
<th>CWB</th>
<th>Usage</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>hō₅U</td>
<td>U, A, G</td>
<td>Family boxes</td>
<td>Eubacteria</td>
</tr>
<tr>
<td>cm₅₅U</td>
<td>(Ser UCN, Val, Thr, Ala)</td>
<td>(Ser UCN, Val, Thr, Ala)</td>
<td></td>
</tr>
<tr>
<td>mcm₅U</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Um</td>
<td>A, G</td>
<td>Two-codon sets</td>
<td>Mitochondria, bacteria, eukaryotes</td>
</tr>
<tr>
<td>xm₅s₃U, (G)</td>
<td>Two-codon sets</td>
<td>Eubacteria, eukaryotes</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>U, C</td>
<td>Two-codon sets</td>
<td>All</td>
</tr>
<tr>
<td>G</td>
<td>U, C</td>
<td>Family boxes</td>
<td>Bacteria</td>
</tr>
<tr>
<td>Q</td>
<td>U, C</td>
<td>Two-codon sets</td>
<td>Eubacteria, eukaryotes</td>
</tr>
<tr>
<td>I</td>
<td>U, C, A</td>
<td>Arg CGN</td>
<td>Eubacteria</td>
</tr>
<tr>
<td></td>
<td>U, C, A</td>
<td>All family boxes</td>
<td>Eukaryotes except Gly GGN</td>
</tr>
<tr>
<td>---</td>
<td>--------</td>
<td>------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>A (rare)</td>
<td>U, C, G&gt;A</td>
<td>Thr ACU, Arg CGN</td>
<td><em>Mycoplasma</em> spp., yeast mitochondria</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>L (K²C)</td>
<td>A</td>
<td>Ile AUA</td>
<td>Eubacteria, plant mitochondria</td>
</tr>
</tbody>
</table>
the 5-substituted uridines 5-methyl-2-thiouridine, xms²U and methoxy-5-uridine, xmo⁵U) have yielded new insights into the genetic coding mechanisms in the three domains of life. Unmodified uridine binds to A, and wobbles to all four bases; xmo⁵U binds to A and wobbles to G and U; and xms²U binds to A and wobbles to G.

**Wobble base pairs.** As per the original definition, a wobble base pair is a non-[Watson-Crick base pair](#) that often occurs between first base of an anticodon and third base of a codon during codon-anticodon interactions. Due to wobbling, unorthodox base-pairing may occur between the base at the 5' end of anticodon 3'-UCU-5' with the base at the 3' end of codon 5'-AGA-3' or 5'-AGG-3'. The genetic code makes up for disparities in the numbers of amino acids (20) and of codons (64), by using wobble base pairs. Besides the major RNA bases, various modified or minor RNA bases may exist at the first base of the anticodon for wobble base pairing. One of such important modified base is **inosine** which can pair with three bases: uracil, adenine, and cytosine.

![Diagram of mRNA codon and tRNA anticodon](#)

**Fig. I.7** The wobble position in codon-anticodon pairing

The four main wobble base pairs are inosine-cytosine, inosine-uracil, inosine-adenine and guanine-uracil (Fig. I.8, next page). Thermodynamic stability of a wobble base pair is comparable to that of a Watson-Crick base pair. Wobble base pairs are fundamental in [RNA secondary structure](#) and are critical for the proper translation of the genetic code. In the context of this Dissertation, though, the broad term **wobble base pair** also refers
to any real or hypothetical base pair which occurs at the wobble position between the third codon base and the first anticodon base. This is regardless of whether this pair is allowed or disallowed during actual codon-anticodon pairing. The terms "wobble pair" and "wobble position" are, for the purpose of this Thesis, extended to include any real or putative RNA base pair at the third codon position in codon-anticodon pairing.

Fig. I.8 The four main wobble base pairs

**Observed and unobserved wobble pairs.** Unlike DNA, RNA molecules contain, apart from the standard Watson-Crick base pairs, highly versatile non-Watson-Crick types of base pairs. These non-Watson-Crick base pairs, often called non-canonical base pairs, are important factors governing the evolution and folding of RNA structures [44]. They are also important in forming tertiary interactions between remote portions of RNA structures, and sometimes they participate in the formation of structurally specific and evolutionary conserved regions of RNA structures, called RNA motifs [45-48].

Leontis *et al.* studied the possible base pairing patterns in RNA structures in detail and classified the known RNA base pairing types in 12
geometric families [49] on the basis of type of interacting edge and the orientation of the glycosidic bonds. A matrix formulation has also been elaborated for these base pairing types to signify isosteric relationships between these bases [50].

Not all combinations of bases can occur as pairs at the wobble position. Some wobble pairs are “allowed”, being observed in nature in the context of the translation process, while others are "disallowed". Theoretical modeling studies of solitary wobble pairs, incorporating allowed and disallowed situations, may help to decide which are the configurational criteria at the wobble position that decide the issue, and can shed light on the patterns of degeneracy which are observed in the genetic code as well.

Criteria [49] for determining the canonical Watson–Crick base pairs are (a) the angle between the two local x-axes which must lie in the range 0–17°; (b) the angle between the two local y-axes which must be in the range 157–180°; and (c) the angle between the two local z-axes which must be in the range 0–30°. This defines the standard Watson–Crick geometry where the glycosidic bond configuration is automatically cis. The distance separating the two base pair origins must be < 2.5 Å, and the vertical distance between the two base planes < 1.5 Å.

If the base pair does not belong to the canonical Watson–Crick base pair family as defined above, it is examined to see whether it belongs to one of the other 12 non-Watson–Crick base pair families. The criteria for determining the type of base pair family are the following: (a) the angle $\alpha_b$ between two base planes (or equivalently, between the two local z-axes) must be < 65°; and (b) the vertical distance $d$, between the two bases planes must be < 2.5 Å. At least two hydrogen bonds must exist, one of which should be an H-bond that occurs between donor and acceptor atoms, both of which belong to a base, and with the distance between the two atoms (N or O) < 3.4 Å. The second H-bond is determined with extended distances. The acceptable distance between donor and acceptor for (a) base (N, O)...base (N, O) is < 3.75 Å, (b) base (N, O)...base (C–H) is < 3.9 Å, and (c) base (N, O)...ribose (O2') is < 3.75 Å.

**Modifications at tRNA position 34.** It is now well known that fidelity of codon recognition by an anticodon involves small ribosomal active participation. Functional
contributions of tRNA’s anticodon domain modifications to the small subunit recognition of a stereo-chemically correct anticodon-codon interaction could be crucial to the accuracy and stability of the interactions, and the rate of protein synthesis. A physicochemical understanding [51] of modified nucleoside contributions to the tRNA anticodon domain architecture and its decoding of the genome has advanced RNA world evolutionary theory, as well as the principles of RNA chemistry, and the application of this knowledge to the introduction of new amino acids to proteins.

Most classes of cellular RNA (tRNA, mRNA, rRNA and small nuclear RNA) from all organisms contain post-transcriptionally modified nucleosides, which are derivatives of the four major RNA nucleosides adenosine (A), guanosine (G), cytosine (C), and uridine (U). Most modifications involve simple alkylation, hydrogenation, thiolation or isomerization of the four common ribonucleosides, occurring in the base and the 2' hydroxyl group of the ribose. Some of the modifications involve complex chemical transformations, in which case they are referred to as "hypermodified" nucleosides. Till date, 107 different modified nucleosides have been characterized [52] in various RNAs, out of which 90 are found in tRNAs alone. These post-transcriptional modifications are mostly found to occur at two positions in the anticodon region – position 34 (wobble position) and position 37. Modifications in these two positions of tRNAs appear to be critically important for improving the fidelity and efficiency of tRNA when decoding the genetic message in the correct frame on the ribosome. Anticodon sequence alterations to a particular tRNA that permit decoding of multiple codons are part of a growing number of post-transcriptional changes collectively known as tRNA editing [53]. The decoding changes imparted by tRNA editing provide a mechanism to effectively accommodate genetic code degeneracy. Although a number of editing events have direct effects in expanding a tRNA’s decoding capacity, some editing events indirectly affect tRNA function by repairing otherwise non-functional tRNAs.

I.6 Creation and Evolution of the Genetic Code

The excitement in genetic code research was much stifled in 1968 when Francis Crick put forwarded his "frozen accident" theory of textbook fame [54]. The code of Table I.1
was established around 1966 by using mainly *Escherichia coli* systems. This code was thought to be common to all organisms, i.e., universal, because it applied to certain other organisms such as tobacco mosaic virus and vertebrates. Such apparent universality of the code led Crick to propose the "frozen accident" hypothesis. It states that since the proteins had become so sophisticated in a single pool of progenote cells, any changes in codon meaning would disrupt proteins by incorporating unacceptable amino acid substitutions throughout their sequences, so that evolution of the code is stopped. Crick's "frozen accident" theory was a fair and definitive evaluation of the evidence then available, and called for more experimental evidence bearing on the origin and evolution of the genetic code.

The 1980s marked a rebirth of interest in genetic code evolution. In less than 15 years since the code was first deciphered, it was found that the code in vertebrate mitochondria differed from the universal code by using codons AUA for Met and UGA for Trp [55]. To account for these changes, it was proposed that because the genome of the mitochondria is small, the mitochondria can probably tolerate changes in the code that would not be acceptable to a larger and more complex nuclear genome [56]. This proposal, although it is still being reiterated, became obsolete and unlikely in 1985, when it was discovered that the code had changed in intact organisms. In *Mycoplasma capricolum*, UGA codes for Trp. In certain ciliated protozoans, UAR (where R is A or G) codes for Gln. Recently some code changes in other nuclear genomes have also been reported.

The original notion that the code was universal and "frozen" was dependent on the precept that any mutational change in the code would be lethal, because it would produce widespread alterations in the amino acid sequences of proteins. Such changes would destroy protein function, and hence would be intolerable. However, this now seems to be "bypassed" by nature. It is possible for a codon to disappear from an mRNA molecule, often as a result of directional mutation pressure [57,58] in DNA. Thus, all UGA stop codons can be replaced by UAA. The missing UGA codon can then reappear when some UGG tryptophan codons mutate to UGA. The new UGA codons will be translated as tryptophan, as is the case in non-plant mitochondria and *Mycoplasma*.
Therefore, no changes have taken place in the amino acid sequences of proteins. It is now realized that the code is still evolving [59] in both mitochondrial and nuclear genomes, presumably by essentially the same mechanisms which can take place without deleterious changes in the amino acid sequences of proteins.

### 1.7 Methods Useful for Structural Studies

Structural investigation is one way through which the relation between structure and function of large bio-molecules is elucidated. This study may occur through experimental means like X-ray crystallography and NMR spectroscopic methods, or it may be carried out by laying recourse to reliable computational methods based upon theoretical models.

**(a) X-ray crystallography and NMR**

X-ray crystallography and NMR spectroscopy are the principal methods used to study RNA fragments at high resolution. In x-ray crystallography [60,61], the diffracted waves from periodically arranged atoms in crystals can add up in phase according to Bragg’s law (*i.e.*, the interplanar or interatomic distance should be comparable to the wavelength of incident light to give interference), giving a diffraction pattern. On the other hand, Nuclear Magnetic Resonance (NMR) [62,64] is based on the quantum mechanical property of nuclear spin. It determines information about atoms from the fact that their local environment influences their response to applied magnetic field. The kind of information obtained from NMR includes the measurement of interatomic distances, and coupling constants that can be interpreted in terms of torsion angles.

The number of solved X-ray and NMR RNA structures is now rapidly increasing due to the recent advances in chemical and enzymatic oligo-ribonucleotide synthesis. Unfortunately some of the RNA structures depend strongly on the environmental conditions chosen, either to measure the sample (NMR) or used to promote crystal growth (XRD). This is especially pronounced for certain oligoribonucleotide chains that
are capable of salt dependent hairpin/duplex equilibria [64]. Moreover, the presence of small ligands used in X-ray crystallography as precipitation and cryo-protective agents [65,66] can influence the RNA structure. RNA crystallization frequently requires high salt conditions far exceeding those typical for a living cell. This often makes comparison of crystallographic structures with those delivered by high resolution NMR, taken under low salt conditions, complicated and puzzling. Thus, although experimental techniques like X-ray crystallography and NMR give valuable information about various compounds, the preparation and availability of suitable samples continues to hamper experimental investigations of biomolecules and limits the available information. Theoretical or computational probing for such nanoscale investigations is therefore encouraged to overcome such shortcomings.

(b) Computational methods

Computational methods like molecular mechanics, semi-empirical and ab initio molecular orbital approaches have been used for the structural investigations done for this thesis. Computational techniques are used for molecular modeling and simulation (i.e. numerical experiments based on theoretical models of real molecular systems). High speed/large memory computer systems and the ability of various software programs to handle realistic biological/chemical problems have now opened new possibilities for computational probing. The choice of method to be used for carrying out calculations is largely dependent on the size of the molecular system being studied, varying from use of molecular mechanics methods for large nucleic acid or protein systems to the highly sophisticated and accurate quantum mechanical methods for simple base pairs and dipeptides.

Some of the work done in this line include: quantum chemical analysis of the cis Watson-Crick/sugar edge base pair family [67], characterization of the trans WC/SE RNA base pair family using quantum chemical and molecular mechanics calculations [68], molecular dynamics (MD) analysis of the G:U wobble base pairs in the ribosome based on high resolution crystal structures, including the recent E. coli structure [69]. More
detailed information about the computational methods used for this Dissertation is provided in the next Chapter.

I.8 General Scope of this Investigation

This Dissertation employs the computational methods of molecular quantum mechanics along with parametrised classical molecular mechanics models to study hydrogen-bonded pairing between judiciously chosen RNA bases (forming base pairs) and also between triplet RNA nucleotide strands (codons and anticodons) in order to investigate the structural and energetic factors responsible for the specificity and degeneracy of the genetic code. Semi-empirical SCF-MO theory (the PM3 method of the MOPAC package) is used to conduct a preliminary investigation on the H-bonded base-pairing properties of nine different select major and minor RNA bases present at position 34 of the anticodon loop. This study is enhanced and updated by the use of density functional methods (the B3LYP DFT method of the Gaussian package) along with single-point MP2 calculations to investigate the facility of RNA base pairing and its role for the genetic code. Finally, the parametrized force fields of the classical potential molecular mechanics AMBER package are used to study the H-bonded pairing between complete codon and anticodon triplet nucleotide sequences in order to derive further insight into the physicochemical factors dictating the specificity and degeneracy of the genetic code. The main focus of these studies is to define physicochemical criteria (energetic and spatial) which successfully differentiate between allowed and disallowed H-bonded pairing schemes at the level of the solitary wobble pairs as well as at the level of the complete trinucleotide pairs. The results are then expected to yield vital clues which can explain the key features of specificity and degeneracy of the universal genetic code of life.

The major RNA bases/nucleosides on the anticodon side taken up for this study are guanosine, uracil and cytosine. The minor RNA bases included here for study are: inosine aglycoside or xanthine (Ino), N4-acetylcytosine (Cac),
5-carboxy-methoxyuracil (V), 5-methoxycarbonylmethyl-2-thiouracil (Smc), 5-methyl-aminomethyl-2-thiouracil (Smn) and queuosine aglycoside (Que). All these six minor RNA bases have been found to occur at tRNA’s 34 positon (anticodon wobble position) and their wobbling properties are reported in the literature [70]. The following paragraphs briefly discuss these minor RNA bases:

5-methoxycarbonylmethyl-2-thiouridine (Smc or mcm\(5\)s\(2\)U): It belongs to the group \(xm\(5\)s\(2\)U\) of modified uridines as classified in the literature [71], where the uracil ring has a methylene carbon directly bonded to the C5 atom as well as thiolated at position 2. It can be found in yeast tRNA\(^{\text{Lys}}\) and tRNA\(^{\text{Glu}}\). It can also be found in Sacchaomyces cerevisiae tRNA\(^{\text{Glu}}\)\(_{\text{U}^*\text{UC}}\), tRNA\(^{\text{Lys}}\)\(_{\text{U}^*\text{UU}}\), and in the tRNA\(^{\text{Gln}}\)\(_{\text{U}^*\text{UG}}\) (where U* is Smc). It has been reported that the \(xm\(5\)s\(2\)U\) type of modified uridines can recognize only the purine-ending codons and not the pyrimidine-ending codons [72-74]. The presence of Smc at position 34 was originally proposed to allow the tRNA to efficiently read the cognate A-ending codon and simultaneously reduce the ability to pair with the G-ending codon [74,75]. However, more recent data have suggested that a Smc-containing tRNA may read both A- and G-ending codons [76,77], which challenges the notion that eukaryotes do not use U-G wobbling [78].

5-methylaminomethyl-2-thiouridine (Smn or mnm\(5\)s\(2\)U): This also belongs to the group \(xm\(5\)s\(2\)U\) and can recognise A and G. In E.coli, Smn is found in tRNA\(^{\text{Lys}}\), tRNA\(^{\text{Glu}}\) and one of the two tRNAs\(^{\text{Gln}}\) [79].

5-carboxymethoxyuridine or uridine-5-oxyacetic acid (V or cmo\(5\)U): This nucleoside belongs to the other group of the modified uridines that contains an oxygen atom directly bonded to the C5 atom of the uracil ring (xo\(5\)U) and the members of this group are often found to recognize the U-, A- and G-ending codons. Thus V is present in tRNA of Bacillus subtilis, is predicted to to read A, G and U, but not C [74]. Most in vitro experiments with Escherichia coli tRNAs or anticodon stem–loops (ASLs) support the theoretical considerations that a U reads A and G and that V enhances the wobble to include U, but not C [80-86]. The V\(_{34}\)-G\(_{\text{III}}\) base pair seen in the crystal structure solved by Weixlbaumer et al. [87] (Protein Databank accession code 2UU9) was extracted to
compare our DFT B3LYP results. The geometry of the V-G pair is the same as for a Watson–Crick (A-U or G-C) base pair.

**N4-acetylcytosine (Cac or ac4C):** N4-acetylcytidine is a modified nucleoside that was identified at position 34 (the wobble position) of *Escherichia coli* elongator tRNA^Met^ in 1972 [88].

**Inosine aglycoside or xanthine (Ino or I):** Holley [7] determined the sequence of yeast tRNA^Ala^ in 1965. The sequence of the anticodon bases in this tRNA contained the nucleoside inosine at the 5' position of the anticodon. Inosine is created post-transcriptionally (after synthesis of tRNA by RNA polymerase) through oxidative deamination of adenine.

**Queuosine aglycoside or queuine (Que or Q):** The hypermodified nucleoside queuosine (7-[4,5-cis,dihydroxy-1-cyclo-pentene-3-amino-methyl]-7-deazaguanosine) is found in certain tRNAs in bacteria and eukaryotes [89,90]. Originally identified in *E. coli*, Que occupies the first anticodon position of tRNAs for histidine, aspartic acid, asparagine and tyrosine [91]. The presence of Que at the first anticodon position of tRNAs improves accuracy of translation [92-94]. Que is further modified to galQue and manQue in tRNATyr and tRNAAsp respectively in some higher eukaryotes. Formation of the hypermodified queuosine at position 34 in eukaryotic tRNAs requires only one enzyme, the tRNA:guanine transglycosylase, which exchanges the guanosine 34 base with the free precursor queuine. However, in prokaryotes, at least 3 enzymes are needed to accomplish the stepwise formation of the same queuosine containing tRNAs [95].

For the purpose of this Dissertation, the above nine major and minor RNA bases are treated as simple unglycosylated species with no sugar moiety attached. This is because the base-pairing properties of these systems are largely dependent on the donor and acceptor atoms around the H-bonding zone, being usually little affected by the presence or absence of the more distant sugar moiety.

In this Thesis, study of the entire triplet nucleotide sequences involved in codon-anticodon pairing takes into account only the major RNA bases on both the codon as well as the anticodon side, as described in Chapter Six. This phase of one's work will
soon be extended to cover the minor anticodon wobble bases as well. Pairing between codon and anticodon occurs in antiparallel fashion, where the codon bases are read from the 5' to the 3' end, while the reverse occurs for the anticodon bases. The codon and the anticodon are wound round each other to create a minihelix, regarded as the optimal arrangement for the situation \textit{in vivo}.

References


CHAPTER THREE

INITIAL SEMI-EMPIRICAL STUDIES ON

SOLITARY WOBBLE BASE PAIRS

III.1 Introduction

III.2 Allowed and Disallowed Wobble Pairs

III.3 Role of Other Factors

III.4 Previous Theoretical Studies

III.5 Scope of Systems Studied Here

III.6 Theoretical Methodology

III.7 Results and Discussion