Scientific Insights into the Evolution of the Universe and of Life
Pontifical Academy of Sciences, Acta 20, 2009
www.pas.va/content/dam/accademia/pdf/acta20/acta20-nirenberg.pdf

THE GENETIC CODE AND EVOLUTION

MARSHALL NIRENBERG

> "For in the first place, as Augustine says (*Gen. Ad lit*. VI, 10), they [the seminal virtues that determine phenotypic traits] are principally and originally in the Word of God, as 'typal ideas'. Secondly, they are in the elements of the world, where they were produced altogether at the beginning, as in 'universal causes'. Thirdly, they are in those things which, in the succession of time, are produced by universal causes, for instance in this plant, and in that animal, as in 'particular causes'. Fourthly, they are in the 'seeds' produced from animals and plants. And these again are compared to further particular effects, as the primordial universal causes to the first effects produced".
> Thomas Aquinas, *Summa Theologica*, Question 115, Article 2.[1]

The DNA that we inherit from our parents contains the information that is needed to make the thousands of kinds of RNA and proteins that are the molecular machinery of the body. As shown in Fig. 1 (see p. 598), DNA consists of 4 kinds of letters, termed bases, T, C, A, and G, in long sequences. T forms hydrogen bonds with A, and C pairs with G. The backbone of DNA is composed of repeating sugar-phosphate moieties, and two complementary strands of DNA interact via base pairs and form a double helix.

The number of base pairs and genes in the DNA of some organisms is shown in Fig. 2. The sequence of bases in the DNA of each species shown has been determined. *Mycoplasma genitalium* has a very small genome consisting only of 580,000 base pairs and 470 genes. The genome of the bacterium, *E. coli,* consists of 4,600,000 base pairs, which encode 4,288 genes. Rice has a large genome consisting of 466,000,000 base pairs and contains 30,000 genes. The genome of the nematode, *C. elegans,* contains 97,000,000 base pairs and encodes 18,424 genes. The genome of the fruit fly, *Drosophi-*

[1]I thank Professor Mark Sagoff for suggesting this quotation.

NUMBER OF BASE PAIRS AND GENES IN GENOMES OF DIFFERENT ORGANISMS

SPECIES	DNA BASE PAIRS $x10^6$	GENES
Mycoplasma genitalium	0.58	470
E. coli	4.6	4,288
Rice	466	30,000
C. elegans	97	18,424
D. melanogaster	165	~14,500
Man	3,300	~25,000

Figure 2.

la melanogaster, consists of 165,000,000 base pairs and encodes approximately 14,500 genes. The Human genome consists of about 3.2 billion base pairs, which encode 20,000 to 25,000 genes. Only about 1.5 percent of the DNA in man encodes protein; additional DNA regulates gene expression. Some DNA consists of repeated transposable elements. DNA also contains nonfunctional pseudo-genes that may be experiments that failed during evolution. Finally, the function of much DNA is unknown.

There are 20 kinds of common amino acids found in proteins. The average protein consists of about 300 sequential amino acid residues, but some large proteins consist of thousands of amino acid residues. The genetic code refers to the translation of base sequences in DNA, which has a 4 letter alphabet to sequences of amino acids in protein, which has a 20 letter alphabet.

When I started to work on protein synthesis in 1958 the mechanism of protein synthesis was not known. Amino acids were known to be incorporated into protein on organelles termed ribosomes and amino acids had been found to be covalently attached to RNAs termed tRNA. Messenger RNA (mRNA) had not been discovered. The first question I asked using a bacterial cell-free protein synthesizing system was: 'Does DNA directly code for protein synthesis, or does RNA, which is transcribed from DNA, code for protein synthesis?' We found that RNA rather then DNA directs the incorporation of amino acids into protein.

In Fig. 3 is shown a simple outline of protein synthesis. We showed definitively that mRNA exists and directs the synthesis of protein (1). One strand of DNA is transcribed to mRNA, the mRNA then associates with ribosomes, and proteins are synthesized amino acid by amino acid on ribosomes. Enzymes with specificity for each kind of amino acid and the appropriate species of tRNA catalyze the ATP dependent activation of the amino acid and the covalent transfer of the amino acid to the tRNA. We showed that 3 bases in mRNA correspond to 1 amino acid in protein. Each 3 base codon in mRNA is recognized by an appropriate 3 base anticodon in tRNA

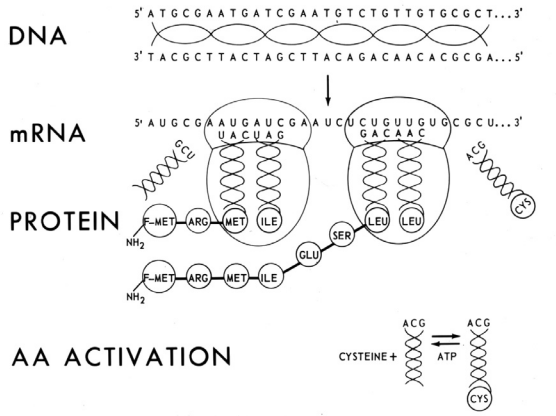

Figure 3.

by the formation of hydrogen bonds. The elongating peptide chain is then transferred to the amino acid attached to tRNA, the free tRNA is released from the ribosome and the tRNA with the attached peptide chain then is transferred to site of the vacated tRNA.

There are 4 kinds of bases in RNA, U, C, A, and G. U in RNA corresponds to T in DNA. U base pairs with A and C base pairs with G. With 4 kinds of bases in RNA there are 64 possible combinations of 3 bases, i.e., triplets. The genetic code, that is the 64 possible triplets which are termed codons and the 3 letter abbreviations of the amino acid that corresponds to each triplet is shown in Fig. 4 (see p. 599). All 64 triplets have meaning. My colleagues and I deciphered the genetic code between 1961 and 1966 (2). We found that the 3rd bases of synonym RNA codons varies systematically. For example UUU and UUC correspond to phenylalanine. Three amino acids, leucine, serine,

and arginine each correspond to 6 synonym codons. For each of 5 amino acids – valine, alanine, glycine, threonine, and proline – there are 4 synonym codons. There are 3 synonym codons for isoleucine. The 3rd base of codons for 6 amino acids can be either U or C. The 3rd base of codons for 3 amino acids can be either A or G. Only 2 amino acids, methionine, shown in green, and tryptophan, each correspond to only a single codon. There are two species of tRNA for methionine, one species initiates protein synthesis (4), the other species corresponds to methionine in internal positions of proteins. Three codons, UAA, UAG, and UGA, shown in red, correspond to the termination of protein synthesis (5-7).

The arrangement of codons for amino acids is not random. For example amino acids with structurally similar side chains, such as aspartic acid and glutamic acid, have similar codons. Asparagine and glutamine also have similar side chains and correspond to similar codons. Most hydrophobic amino acids have U in the central position of the codon; whereas most hydrophilic amino acids have A as the second base in the codon. Thus the effects of mutations due to replacement of one base by another often are minimized.

After we deciphered the code for *E. coli* Richard Marshall, Thomas Caskey, and I (3) asked the question, is the genetic code the same in higher organisms? We determined the genetic code in the amphibian, *Xenopus laevis*, and in a mammalian tissue, guinea pig liver. We found that the genetic code is the same in *E. coli*, the amphibian, and the mammal. We also examined different guinea pig tissues and found that the code is the same in different tissues. We purified tRNA from *E. coli*, yeast, and guinea pig liver, and showed that some species of highly purified tRNA recognize only G in the third position of the codon, others recognize U or C, others recognize A

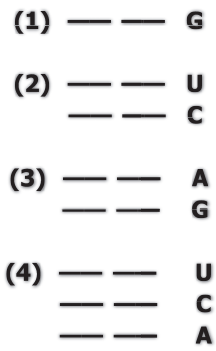

Figure 5.

or G, and still other species of tRNA recognize U, C, or A in the third position of codons (Fig. 5) (8). We showed that yeast alanine tRNA that had been sequenced by Robert Holley recognizes three codons, GCU, GCC, and GCA, and showed that inosine in the tRNA anticodon recognizes either U, C, or A in the 3rd position of the alanine codons (9). Many investigators have shown that there are different modified bases either in the tRNA anticodon or next to the anticodon that result in alternate recognition of 3rd bases in synonym codons.

21st AND 22nd AMINO ACIDS

21st.	SELENOCYSTEINE	UGA
22nd.	PYRROLYSINE	UAG

Figure 6.

In 1986 the 21st amino acid, selenocysteine, was found (Fig. 6) (10,11). Selenocysteine is found in the active centers of some oxidation-reduction enzymes, such as formate dehydrogenase. There is a special tRNA for selenocysteine that accepts serine. There also is an enzyme that catalyzes the acylation of this tRNA with serine and two enzymes that convert the serine attached to the tRNA to selenocysteine. Selenocysteine recognizes the termination codon UGA only if in the downstream region there is a stem-loop secondary structure in the mRNA. The mRNA folds back on itself and base pairs forming a hairpin-like stem-loop structure. There are either 1 or 2 proteins, depending on the species, that recognize both the selenocysteine tRNA and the stem-loop structure, and only then does UGA correspond to selenocysteine.

Pyrrolysine is the 22nd amino acid, which was found in 2002 (12, 13). This is a very rare amino acid, found only in a few species of primitive bacteria. It is found in the active centers of methylamino-, dimethylamino-, and trimethylamino-transferases and in transposase as well. There is a special tRNA for pyrrolysine that recognizes the codon UAG, and an enzyme that catalyzes the acylation of this tRNA with pyrrolysine. Whether this is a conditional recognition in which a protein recognizes pyrrolysine-tRNA and a stem-loop type of mRNA structure is not known.

There is only one genetic code that is used on this planet; hence, the code is a universal code. However, variants of the code have been found in some organisms. For example, in Fig. 7 (see p. 599) are shown some dramatic events that occurred during the evolution of some ciliated protozoa (14). The standard codons recognized by glutamine-tRNA are CAA and CAG. During evolution the gene for glutamine tRNA was duplicated; then a mutation in the anticodon of the second gene for glutamine tRNA replaced G with A; therefore, the tRNA corresponding to the second gene for glutamine tRNA recognized UAA and UAG that are terminator codons in the standard code. Later in evolution the second gene for glutamine tRNA was duplicated and a mutation in the anticodon of the third gene for glutamine tRNA resulted in a replacement of U with C. The tRNA corresponding to the third gene for glutamine tRNA then recognized the codon UAG. So in *Tetrahymena*, CAA, CAG, UAA and UAG correspond to glutamine tRNA. Changes in the meaning of codons are rare events, but there are a number of other organisms that have been found with changes in the translation of some codons.

A number of changes in the genetic code of mitochondria have been found in many organisms. Mitochondria are the organelles that produce energy for cells. Mitochondria have a small amount of DNA that contains about 10 genes and proteins corresponding to these genes are synthesized in mitochondria. Most of the genes for mitochondrial proteins reside in genomic DNA in the nucleus of cells and the proteins that are synthesized in the cytoplasm are imported into mitochondria. Some of the changes in the genetic code in mitochondria are shown in Fig. 8 (see p. 600). In the standard genetic code, UGA corresponds to the termination of synthesis and UGG corresponds to tryptophan. However, in the mitochondria of *Trypanosomes, Neurospora*, yeast, *Drosophila* and mammals both UGG and UGA correspond to tryptophan. In the standard genetic code AUA corresponds to isoleucine and AUG corresponds to methionine; whereas, in the mitochondria of yeast, *Drosophila*, and mammals, both AUA and AUG correspond to methionine. In the standard code CUU, CUC, CUA, and CUG correspond to leucine; whereas, in yeast mitochondria these codons correspond to threonine. In the standard code AGA and AGG correspond to arginine; whereas in the mitochondria of *Drosophila* these codons correspond to serine but in mammalian mitochondria these codons correspond to termination of protein synthesis. Additional changes in the translation of codons in mitochondria have been found in other organisms. The changes that have been found in the translation of codons in mitochondria probably are tolerated because mitochondrial genes only encode about 10 proteins. Similar changes in the translation of

proteins encoded by nuclear genes, which would affect the synthesis of many thousands of proteins, almost surely would be lethal.

A summary of results is shown in Fig. 9. The results strongly suggest that the genetic code appeared very early during biological evolution, that all forms of life on Earth use the same or very similar genetic codes, that all forms of life on Earth descended from a common ancestor and thus, that all forms of life on this planet are related to one another. The messages in DNA that we inherit from our parents contain wisdom gradually accumulated over billions of years. The messages slowly change with time, but the translation of the language remains essentially constant. The molecular language is used to solve the problem of biological time, for it is easier to construct a new organism using the information encoded in DNA then it is to fix an aging, malfunctioning one.

SUMMARY

1. The genetic code appeared very early during biological evolution.
2. All forms of life on Earth use the same or very similar genetic codes.
3. All forms of life on Earth descended from a common ancestor and thus, all forms of life on this planet are related to one another.
4. The messages in DNA that we inherit from our parents contain wisdom gradually accumulated over billions of years. The messages slowly change with time, but the translation of the language remains essentially constant.
5. The molecular language is used to solve the problem of biological time for it is easier to construct a new organism using the information encoded in DNA than it is to repair an aging malfunctioning one.

Figure 9.

REFERENCES

1. Nirenberg, M.W., and Matthaei, J.H.: The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA.* 47, 1588-1602 (1961).
2. Nirenberg, M.W., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. and O'Neal, C.: RNA codewords and protein synthesis. VII. On the general nature of the RNA code. *Proc. Natl. Acad. Sci.,* 53: 1161-1168 (1965).
3. Marshall, R.E., Caskey, C.T. and Nirenberg, M.: Fine structure of RNA codewords recognized by bacterial, amphibian, and mammalian transfer RNA. *Science,* 155: 820-826 (1967).
4. Clark, B.F.C. and Marcker, K.A.: The role of N-formylmethionyl-sRNA in protein biosynthesis. *J. Mol. Biol.,* 17: 394-406 (1966).
5. Brenner, S., Stretton, A.O., Kaplan, S.: Genetic code: the 'nonsense' triplets for chain termination and their suppression. *Nature,* 206: 994-998 (1965).
6. Weigert, M.G. and Garen, A.: Base composition of nonsense codons in *E. coli* evidence from amino-acid substitutions at a tryptophan site in alkaline phosphatase. *Nature,* 206: 992-994 (1965).
7. Brenner, S., Barnett, L., Katz, E.R., Crick, F.H.: UGA: A third nonsense triplet in the genetic code. *Nature,* 213: 449-450 (1967).
8. Caskey, C.T., Beaudet, A. and Nirenberg, M.: RNA codons and protein synthesis. 15. Dissimilar responses of mammalian and bacterial transfer RNA fractions to messenger RNA codons. *J. Mol. Biol.,* 37: 99-118 (1968).
9. Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M. and Anderson, F.: The RNA code and protein synthesis. *Cold Spring Harbor Symp. Quant. Biol.,* 31: 11-24 (1966).
10. Chambers, I., Frampton, J., Goldfarb, P., Affara, N., McBain, W., Harrison, P.R.: The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon, TGA. *EMBO J.,* 5: 1221-1227 (1986).
11. Zinoni, F., Birkmann, A., Stadtman, T.C., Böck, A.: Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli*. *Proc. Natl. Acad. Sci. USA.*, 83: 4650-4654 (1986).
12. Hao, B., Gong, W., Ferguson, T.K., James, C.M., Krzycki, J.A., Chan, M.K.: A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science*, 296: 1462-1466 (2002).

13. Srinivasan, G., James, C.M., Krzycki, J.A.: Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science,* 296: 1459-1462 (2002).
14. Hanyu, N., Kuchino, Y., Nishimura, S., Beier, H.: Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* $tRNAs^{Gln}$. *EMBO J.,* 5: 1307-1311 (1986).
15. Knight, R.D., Landweber, L.F.: Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem. Biol.,* 5: R215-R220 (1998).

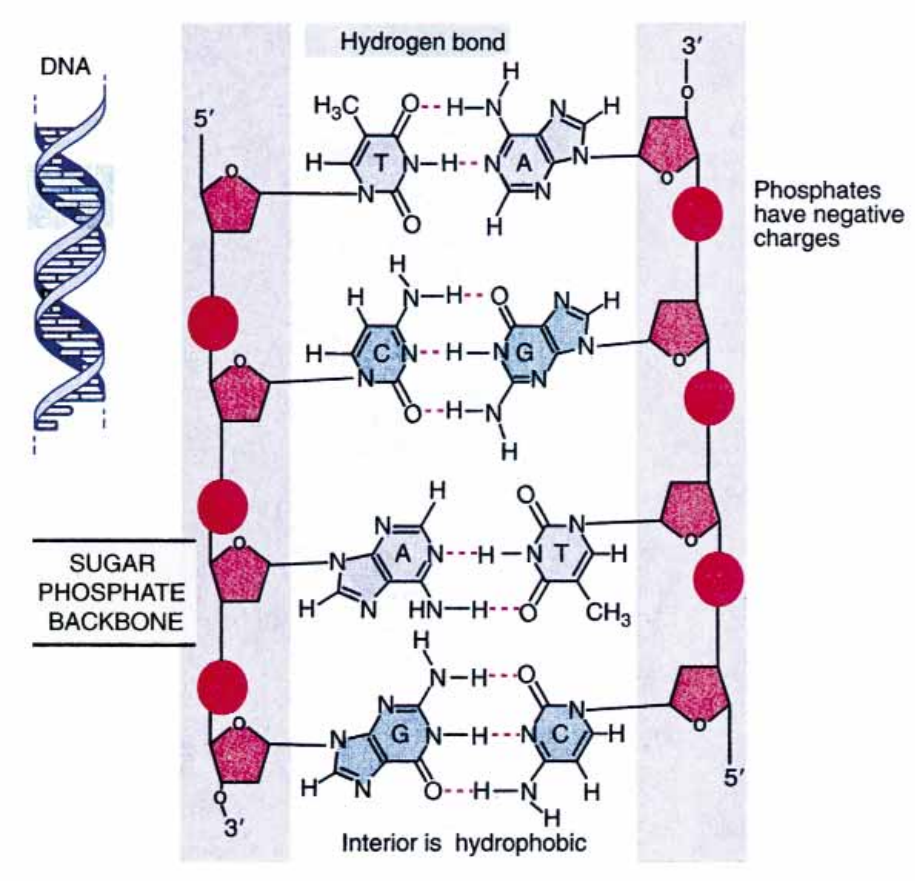

Figure 1. From B. Lewin, *Genes* IX (2008).

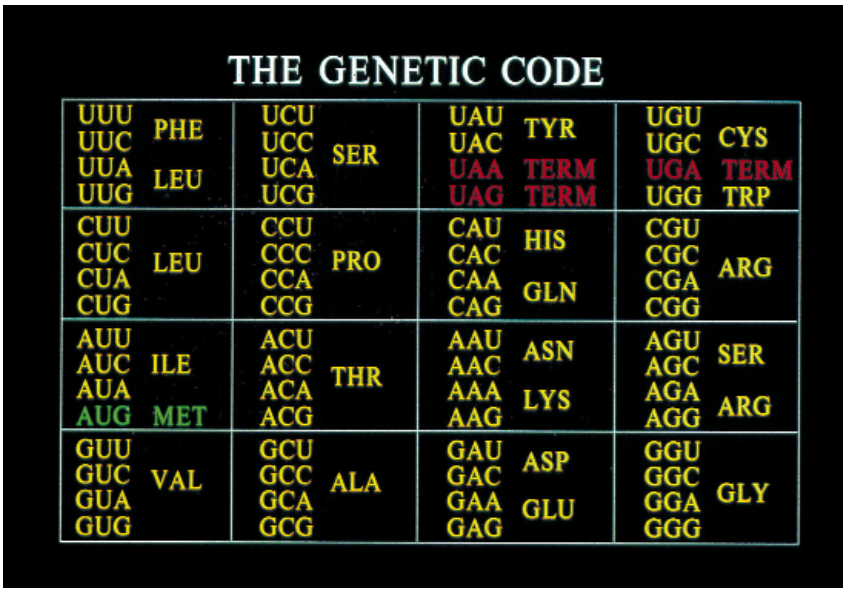

Figure 4.

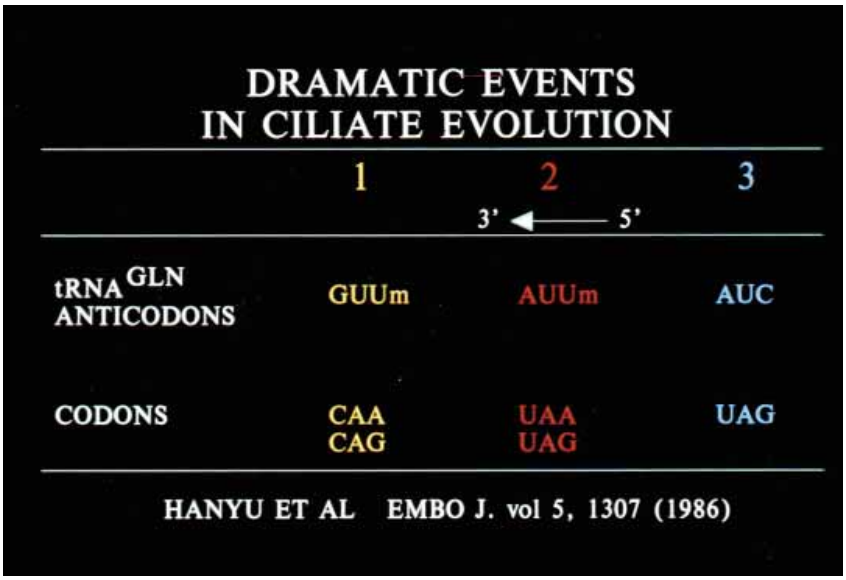

Figure 7.

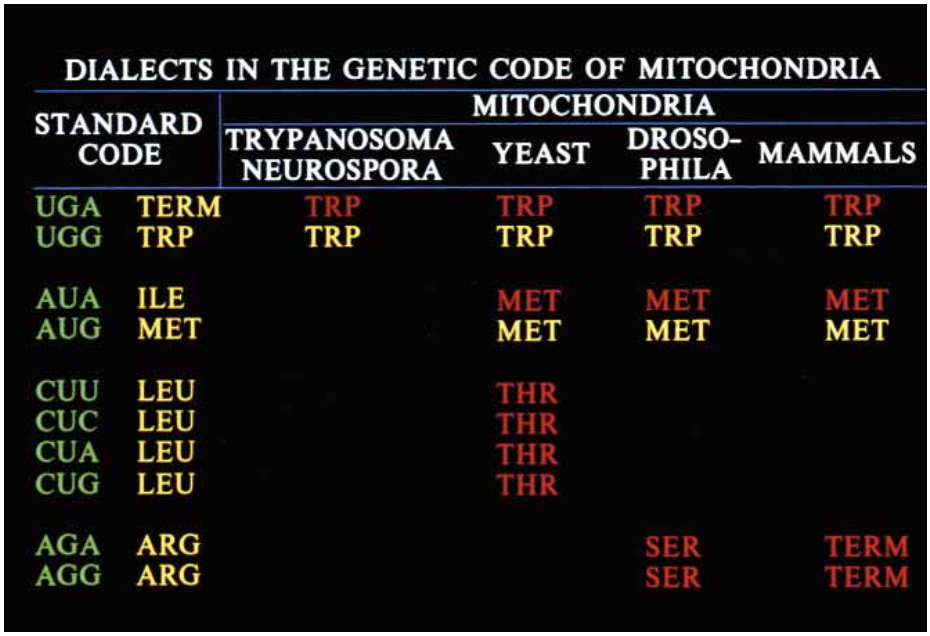
DIALECTS IN THE GENETIC CODE OF MITOCHONDRIA
STANDARD CODE
MITOCHONDRIA
TRYPANOSOMA NEUROSPORA
YEAST
DROSO-PHILA
MAMMALS
UGA TERM TRP TRP TRP TRP
UGG TRP TRP TRP TRP TRP
AUA ILE MET MET MET
AUG MET MET MET MET
CUU LEU THR
CUC LEU THR
CUA LEU THR
CUG LEU THR
AGA ARG SER TERM
AGG ARG SER TERM

Figure 8.