DRAFT - 8/15/88
For Conference Distribution Only

THE EVOLUTION OF THE GENETIC CODE

by

Guido Pincheira
Professor of Genetics
University of Chile
Santiago, CHILE

The Seventeenth International Conference on the Unity of the Sciences
Los Angeles, California November 24-27, 1988

INTRODUCTION

Life, as a result of evolution, is characterized by a very rich variety of molecules and chemical processes. Nevertheless, there are basic features that seem to be common or similar to all organisms. One of these aspects is the chemical nature and physiology of the genetic material, or genome, that contains the coded information for the transient or permanent characteristics of the organisms.

It is really surprising that, in spite of the tremendous variations produced by evolution in living forms, speciation has been very conservative with the nature and processing of the genetic information. As a consequence, it is generally said that the genetic code is "universal" or identical to all living things.

Although the progress achieved by molecular biology in the last decades has contributed a great deal to clarify the characteristics of the genetic code; its origin and evolutionary path are still very puzzling problems for science.

WHAT IS THE GENETIC CODE?

The molecular carriers of genetic information are nucleic acids, desoxiribonucleic acid (DNA) or ribonucleic acid (RNA), formed by long and linear sequences of nucleotides that differ in a nitrogen base. As such, we have nucleotides formed by Guanine (G) Adenine (A) Cytosine (C) Thymine (T) or Uracil (U).

Desoxiribonucleic acid (DNA) is the genetic material in the majority of the organisms. The complementarity between two base pairs, G with C, and A with T, provides DNA with the unique property of making exact replicas of itself during reproduction.

Many viruses have ribonucleic acid (RNA) as the genetic material. In this molecule, Thymine is replaced by Uracil.

We usually refer to the genetic code as the correspondence of sequences of three nucleotides (codons) in a polymer, called messenger RNA, and the position of aminoacids in a protein. The code was deciphered in the early 60's with the assigment of 61 codons to 20 protein amino-acids and 3 codons responsible for the termination of the translation of the genetic message. As such, the genetic code may be called an aminoacid code. See Fig. 1.

In essence, the genetic code for aminoacids is the correspondence between a four letter alphabet, organized in triplets of nucleotides, and a twenty letter alphabet that represents the protein aminoacids. Genetic codons may be synonyms for different aminoacids, a condition called redundancy or degeneracy of the code. Only the aminoacids methionine and tryptophan are specified by one codon. See Fig. 2.

The redundancy of the code provides many possibilities to specify a sequence of aminoacids in a protein. For example, a polypeptide made of methionine - leucine - serine - arginine - alanine-glutamic acid may be genetically coded in many different messages as a result of the possibilities of aminoacids to be specified by different codons.

Ex: met - leu - ser - arg - ala - glu

1 x 6 x 6 x 6 x 4 x 2 = 1728

Theoretically, DNA may have 1728 ways to code this sequence of aminoacids in a protein.

Another example: met - asp - lys - ala - trp

1 x 2 x 2 x 4 x 1 = 16 ways.

The processing of the genetic information to synthesize a given protein includes two main steps: DNA makes (a) RNA makes (b) Protein

a) The first step is the process of genetic transcription, in which an RNA message (mRNA) is synthesized by copying DNA strands.

b) The second step is the process of translation, in which the information coded in the RNA message is interpreted through the intervention of decoding molecules (tRNAs) porting the

aminoacids for the peptide formation. These tRNA molecules have other triplets of nucleotides, called anticodons, to recognize the codons in the mRNA according the complementarity of nucleotides.

In fact, both steps are based in the complementarity of nucleotides.G pairs with C and A pairs with U, forming hydrogen bonds.

Nevertheless, the interactions between codons and anticodons takes place under rules that partially relax the restriction of G paired with C and A paired with U and also under the influence of other factors.

The actual genetic code for aminoacids is structurally made of three letter words, but operationally it may be considered a two letter code, because the precision with which the letters are read reside in the first two nucleotides of the codon. The possibility of reading a codon by two out of three nucleotides is know as the "wobble" characteristic of the code. This is illustrated by the examples of glycine, alanine, valine, threonine and proline, in which the 3rd. nucleotide of the codon is completely redundant, because it can be read by any of the four nucleotides.

In the case of isoleucine, phenilalanine, tyrosine, aspartic, asparragine, glutamic acid, glutamine, cysteine and histidine; the 3rd. nucleotide of the codon is only partially redundant, because it can be read by U or C; but not by A or G.

In only 3 codons, the reading of the nucleotide third is specific. These codons are:

AUG ——————— methionine
UGG ——————— tryptophan
UGA ——————— stop signal

In spite of what we already mentioned, our understanding of the coding role of the genome is still very limited and it is possible that more complex coding principles may be involved in mechanisms of action of the genetic material. Coding signals may not reside exclusively in specific DNA sequences. They may also depend of other molecular events. For example: minor modifications,like methylation of the

nitrogen bases of the nucleotides may alter the meaning of the genetic language. We already know more than 50 modified nucleotides in transfer RNAs. These modifications involve the alteration of or the addition to existing bases in tRNAs. It is known that adenine and uracil are almost not employed in the first position of the anticodon.

Adenine is usually converted by deamination to Inosine (I)which can pair with U, C or A.

Uracil is also modified and appears as pseudouridine, dihydrouridine, 4-thiouridine; with variations in the property to recognize other nucleotides by complementation. In one case, it is known that the enzymatic exchange of a nitrogen base, called queuosine, for guanosine in tRNA affects the specificity of codon recognition.

Variations in the genetic code for aminoacids have been found in the mitochondrial genome, as compared which the genetic code processed in the cytoplasm.

In human mitochondria some of these differences are:

Codon	In the mitochondria	In the cytoplasm
AG $\begin{matrix}A\\G\end{matrix}$	Termination	arginine
AUA	methionine	isoleucine
UGA	tryptophan	termination

As a consequence, in human mitochondria all aminoacids have, at least, two codons; and there are four codons for ending the message, because AGA and AGG, which correspond to arginine in the "standard" genetic code, in the mitochondria do not code for any aminoacid.

Two codons, AUG and AUA are codons for methionine and indicate the start of a message.

In some protozoans, such as Paramecium, the mitochondrial genetic code has only one codon for termination. The other "standard" or cytoplasmic stop codons code for aminoacids in the protozoan mitochondria.Under these circunstances it is rather improbable that a mitochondrial mRNA may be correctly translated by the cytoplasm decoding system.

Code interpretation may also be related to aspects of structure of nucleic acids. Unusual genetic codes have been found in mitochondria of different organims.

This condition determines that codons AGA and AGG, responsible for arginine in the "standard" code, mean serine in the genome of starfish mitochondria.

The explanation given by Himeno et al (1) is the possibility of a different decoding capacity of transfer RNA in relation to variations in size of a region known as the dihydrouridine (D) arm of this molecule. See Fig. 3.

The number of nucleotides in D arms of tRNAs may vary in different species of organisms, For example: 5 nucleotides in mammals

10 " " Xenopus

11 " " Drosophila

12 " " starfish

As a consequence, the manner of base pairing in the D arm may influence the decoding capacity of tRNAs.

Another aspect of nucleic acid structure very relevant to interpret a genetic message is the formation of a complex between the anticodon nucleotides and the nucleotide located at the 4th position from the 3' end in the tRNA molecule. According Shimizu (2), the mechanism of tRNAs to recognize the aminoacids to be transported to the site of protein synthesis depends to a great deal of the nucleotide located in the 4th position indicated above, which would play a discriminatory role in conjuntion with the anticodon.

We can agree that from a structural point of view the genetic code for aminoacids is universal, because it is always made of 3 letter codons. Nevertheless, the results of research in genome physiology in cellular organelles of different organisms, indicate that the code may be operationally different; because codons may be read in different ways.

The complexity and variations that are begining to be detected in the interpretation of the genetic code are not sufficient yet to advance explanations for the lack of recognition or exclusion of 100

other aminoacids in the process of protein synthesis.

OTHER CODING FUNCTIONS OF THE GENOME

Besides the formation of proteins, the genome codes also the information for other important life processes.

1) As already mentioned, one is the formation of new genetic material (DNA) by means of replication. Another one is the synthesis of structural RNA (transfer RNA and ribosomal RNA) through the process of transcription.

2) The genetic code also contains the information for the recognition and interaction of the genome with many other molecules, including proteins. This sort of "recognition code" is crucial for the initiation or termination of many genetic activities, such as the specific recognition of enzymes for transcription or regulatory and structural proteins.

   The positions and orientations of nucleotides in DNA or of aminoacids in a protein and the overall fit of the protein and DNA surfaces play significant roles in the decodification of the genetic information. Hydrogen bonds between side chains of aminoacids and the edges of base pairs in the nucleotides are the main instruments for the interactions between DNA or RNA with proteins.

   The "recognition code" also seems to be "degenerate". because each nitrogen base in a nucleotide may be recognized by different aminoacids, and each aminoacid may bind to different nitrogen bases. Ex: Adenine is recognized by glutamine and serine. This last aminoacid may also bind to guanine.

ORIGIN OF THE GENETIC CODE

In modern Biology, the origin of the genetic code is still a very puzzling problem due to the difficulties to investigate or to create experimentally the conditions that led to its appearance.

A good hypothesis for the origin and evolution of the genetic code must fullfil several requirements. Besides of being supported by experimental evidence, the hypothesis must consider:

a) the almost "Universal" condition of the code,
b) its specificity to operate and to exclude a large number of aminoacids in the formation of actual proteins,
c) it must also take into consideration the "degeneracy" and partial flexibility (wobble) in the codon-anticodon pairing, and
d) the gradual organization of a very complex translation mechanism to assure an exact interpretation of the coded information; specially the rol accomplished by tRNA molecules, not only in the operation of the system, but in the evolution of the genetic code.

Several thesis or theoretical models have been proposed. Authors, like F. Crick, (3) consider that the actual genetic code is arbitrary and a survivor of many other codes, also arbitrary, as a result of natural selection. This theory proposes that nucleic acids and proteins appeared independently and their colinear corrspondence in the actual genetic code was a matter of chance in a series of "random" accidents. The idea is difficult to prove and does not satisfy very well the concept of a "universal" code, because, it would be possible to expect the appearance and prevalence of several versions of the code.

Other theories postulate that the actual genetic code is the consequence or result of physicochemical events. In other words, this hypothesis considers that the correspondence between sequences of aminoacids in a protein and nucleotides in a nucleic acid

are derived from a series of structural characteristics that may have favored stereochemical reactions between them. Such relationships may be traced back to common precursors for both molecules.

Fujio Egami (4) considers that aminoacids and nitrogen bases appeared simultaneously during chemical evolution and that their correspondence in the genetic code is derived from a common origin

According Egami, the actual 20 protein aminoacids derive from each other according 4 different pathways for their synthesis.

Each one of these pathways would also lead to the synthesis of a nitrogen base. So, we have a pathway for the synthesis of aminoacids linked to the formation of guanine; another pathway is linked to the synthesis of adenine, other is linked to the synthesis of cytosine and a fourth one is connected with the formation of uracil.

These relationships would be reflected in the composition of codons for the different aminoacids; in such a way that the nitrogen base connected with their synthesis is a constituent of the first two nucleotides of the corresponding codons.

According Egami, the first aminoacids to have codons were the more simple ones. Glycine, with 2 C, was synthetically related to guanine, and its codon is GGG. Alanine and serine, with 3 C atoms, would derive from glycine and would be coded by codons GC and AG. Then aspartic acid, with 4 carbon atoms, would be coded by GA, and so on. This group of aminoacids are related to purines. Another group of aminoacids would be related, in a similar pattern to the synthesis of pyrimidines, (cytosine and uracil).

Egami's propositions have been tested in the laboratory and they seem to be feasible. He considers that the lack of correspondence of some codons with some of the basic assumptions of his theory may be related to the ambiguity of the code or can be explained by the similar physicochemical properties of certain aminoacids, such as aspartic acid and glutamic acid, which would be coded by GA in the first two positions of the codon.

According Eigen (5) the primordial soup contained sugars, aminoacids and many other substances that today might be considered laboratory curiosities. These molecules, as products of chemical evolution, were not biologically important. In this background, the genetic code could have emerged, extracting the molecules that would better serve the organization of biological polymers, like nucleic acids and proteins. With reference to nucleic acids, it is possible to imagine that primordial routes of synthesis led to the formation of short polymers that ressembled today's RNA.

These primordial RNA strands, soon developed the capacity of selfreplication, according rules of base pairing favored by energy reassons, (G with C, A with U).

In this respect it is interesting that short polymers of the adenine nucleotide (poly-A) may form spontaneously, without the intervention of enzymes or other type of catalysts, when adenine monomers are exposed to templates formed by polymers of Uracil nucleotides (poly-U).

It has already been suggested that RNA could have been the original genetic material. The RNA polymers, depending on the sequence of nucleotides, could have folded up and became a replicating structure.

In a similar way, when poly-C is the template, the presence of lead makes possible the incorporation of G and a minor amount of adenine.

The continuos chemical production of RNA molecules provided increasing numbers of RNA strands, with the possibility of ocurrence of errors (mutations) and a greater variety of nucleotide sequences.

Later, DNA would appear by the replacement of uracil by thymine. Its higher stability in aqueous media and its advantages for the preservation of the genetic information made him a better molecule for the rol of information storage.

The self-replication property of nucleic acids gave them a great advantage for preservation of the information.

CODING FUNCTIONS

How RNA became the code for aminoacids in a polypeptide is still a mystery.

It has already been postulated that the first codons were formed by only one type of nucleotide. Perhaps an elementary code was necessary to produce primitive proteins, made from fewer amino-acids. The four homocodons would correspond to 4 aminoacids, which, according Miller (6), were the more prevalent in the primordial soup.

Glycine	GGG
Lysine	AAA
Phenil alanine	UUU
Proline	CCC

Later, the availability and necessity to incorporate other aminoacids required combinations of two nucleotides and more late, of 3 nucleotides. The mechanism could also have evolved to a higher precision in reading the 3rd. nucleotide in the codon, making possible the incorporation of new amino-acids.

Although, a one letter or a two letter code could have been enough to start with, it is also possible that the code from the begining was a three letter code, with a more specific meaning for the first two nucleotides and a more flexible rol for the third.

The development of the capability for the detection and correction of errors may have appeared later in evolution by means of mechanisms effective for proofreading and suppresion of errors through the intervention of other molecules, such as enzymes.

This step required, of course, the translation of RNA information into proteins and the operation of the genetic code for aminoacids.

Root - Bernstein (7) has proposed ideas about the origin of the code that take into consideration the ocurrence of chemical operations present in modern biochemical processes, and which also may have taken place in the primordial soup. His hypothesis considers the possibility of pairing between aminoacids, between nucleotides and aminoacids or between short nucleotide polymers.

Today, it is known that pairing between aminoacids or polypeptides occur in a parallel conformation and play a significant rol in the tertiary structure of proteins. It is also possible that a primitive relationship between aminoacids and nucleic acids may also have been a parallel disposition, facilitating the organization of a colinear genetic code.

Pelc and Welton (8) have proposed a model for codon - aminoacid pairing. Their ideas consider the existence of selective affinities between aminoacids and nucleotides that may have been very relevant for an early coding system.

Particular sequences of nucleotides may offer patterns for specific contacts with aminoacids; which then may be stabilized by weak chemical interactions with other molecules.

Saxinger and Ponnamperuma (9) provided support for this idea when they demonstrated a high degree of specificity in the binding of tryptophan with the triplet AAU, which is in some way related to the structure of the codon for this aminoacid.

In a similar context, authors, like Mayer et al. (10), have demonstrated a great affinity between the polynucleotide poly-A and lysine-lysine peptide.

Fox and Dose (11) reported data about interactions between homocodons GGG; AAA; UUU and CCC and the aminoacids they respectively code (glysine, lysine, phenilalanine and proline). These situations provide good possibilities to explain the origin of the genetic code based in stereochemical operations.

Root - Bernstein (7) has also suggested that codons for glycine (GGG) and proline (CCC) may pair. The same can occur between codons of lysine (AAA) and phenil alanine (UUU).

The "wobble" phenomenon may be explained according the following binding pattern:

-The first nucleotide generally binds only to the COOH group of any aminoacid and, as such, is unspecific.

-The second nucleotide is always bound to a COOH or NH2 group, and also to the side chain of the aminoacid. This condition makes the middle nucleotide more important than its neighbors for amino acid recognition and binding.

-The third nucleotide interacts with the aminoacids side chain only.

The rol of tRNA molecules would be to facilitate the transition from a parallel aminoacid-nucleotides interaction to a nucleotide-nucleotide antiparallel recognition, inverting the directionality of the code interpretation.

It is also possible that the rather complex translation mechanism or decoding device was developed as a mean to separate the storage of genetic information from the system to process such information.

The basic condition required by this hypothesis is the simultaneous appearance of aminoacids and nucleic acids. Nevertheless, the idea of a primitive genetic code, based in codon aminoacid pairing, needs further testing and experimentation.

A similar proposition has been provided by Eigen (5), who has demostrated that polypeptides have a high affinity for the RNA chain in which they are encoded. This type of interaction ressembles what we observ in the "recognition" code and may have been the chemical precursor for the sophisticated decoding system we have today.

To conclude we may say that it may be possible that an important part of the genetic language has been derived following a chemical logic, but it is also possible that chance may have been important in the origin and evolution of some codons.

Exact genetic codes have been studied in very few organisms and very few issues related to its origin are really clear. As such, it is evident the relativity of the concept of universality of the code, making necessary additional work with the new technologies available for studying the genome. These efforts may uncover variations of the genetic code that may illustrate better its original conception.

R E F E R E N C E S

1. Himeno, H.; Haruhiko Masaki; Tatsuski Kawai; Takahisa Ohta; Izumi Kamagai; Kin-ichiro Miura and Kimitsuna Watanabe, 1987. Unusual genetic codes and a novel gene structure for tRNA Ser AGY in starfish mitochondrial DNA. Gene 56: 219 - 230.

2. Shimizu, M.; 1982. Molecular basis for the genetic code. J. Molec. Evol. 18: 297 - 303.

3. Crick, F.H.C.; 1968. J. Mol. Biol. 38: 367.

4. Egami, F.; 1979. J. Agric. Chem. Soc. (Japan) 53.

5. Eigen, M.; Gardiner, W.; Schuster, P. and R. Winkler - Oswatitsh. 1981. The origin of genetic information. Scient. Amer. 244: 78 - 95.

6. Miller, S.L. and Orgel, L.E. 1974. The origin of life on the Earth, Englewood Cliffs. New Jersey. Prentice Hall.

7. Root - Bernstein, R.S.; 1982. On the origin of the genetic code. J. Theor. Biol. 94: 895 - 904.

8. Pelc, S.R. and Welton M.G.E. 1966. Nature (London) 209:868

9. Saxinger, D. and C.A. Ponnamperuma. 1974. Origin life 5: 180. Reproduced in Calvin, M. 1975. Am. Sci 63: 169.

10. Mayer, R.; Toulme, F.; Montenay - Garestier, R. and Helene C. 1979. J. Biol. Chem. 254: 75.

11. Fox S.W. and K. Dose. 1977. Molecular evolution and the origin of life. Marcel Dekker. New York. pp. 235 - 238.

THE GENETIC CODE

1st ↓ 2nd →	U	C	A	G	↓ 3rd
U	PHE	SER	TYR	CYS	U
	PHE	SER	TYR	CYS	C
	LEU	SER	STOP	STOP	A
	LEU	SER	STOP	TRP	G
C	LEU	PRO	HIS	ARG	U
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLN	ARG	A
	LEU	PRO	GLN	ARG	G
A	ILEU	THR	ASN	SER	U
	ILEU	THR	ASN	SER	C
	ILEU	THR	LYS	ARG	A
	MET	THR	LYS	ARG	G
G	VAL	ALA	ASP	GLY	U
	VAL	ALA	ASP	GLY	C
	VAL	ALA	GLU	GLY	A
	VAL	ALA	GLU	GLY	G

The names of the twenty amino acids and their abbreviations are:

ALA	–	Alanine	LEU	–	Leucine
ARG	–	Arginine	LYS	–	Lysine
ASN	–	Asparagine	MET	–	Methionine
ASP	–	Aspartic acid	PHE	–	Phenylalanine
CYS	–	Cysteine	PRO	–	Proline
GLN	–	Glutamine	SER	–	Serine
GLU	–	Glutamic acid	THR	–	Threonine
GLY	–	Glycine	TRY	–	Tryptophan
HIS	–	Histidine	TYR	–	Tyrosine
ILEU	–	Isoleucine	VAL	–	Valine

The abbreviation STOP shows the three triplets which can terminate the polypeptide chain.

Fig. 1

Amino acid	Base triplets
alanine	GCU, GCC, GCA, GCG
arginine	CGU, CGC, CGA, CGG, AGA, AGG
asparagine	AAU, AAC
aspartate	GAU, GAC
cysteine	UGU, UGC
glutamate	GAA, GAG
glutamine	CAA, CAG
glycine	GGU, GGC, GGA, GGG
histidine	CAU, CAC
isoleucine	AUU, AUC, AUA
leucine	UUA, UUG, CUU, CUC, CUA, CUG
lysine	AAA, AAG
methionine	AUG
phenylalanine	UUU, UUC
proline	CCU, CCC, CCA, CCG
serine	UCU, UCC, UCA, UCG, AGU, AGC
threonine	ACU, ACC, ACA, ACG
tryptophan	UGG
tyrosine	UAU, UAC
valine	GUU, GUC, GUA, GUG
start	AUG (methionine)
stop	UAA, UAG, UGA ('nonsense')

Fig. 2