The Genetic Code

Patricio Jeraldo

May 5, 2006

Abstract

In this paper I will review the genetic code. An overview of its functions and inner workings will be given. Also the current theories on the origin and emergence of the canonical genetic code from early codes will be explored.

1 Introduction

Some introduction I'll also fill out later. Say what you will say. Of course, write it after you write the main document. Even before the genetic code was cracked, the question of the origin of life has fascinated scientists, physicists in particular. Erwin Schrödinger summoned people to tackle the problem when he saw that the possibility of peering into life's innermost mechanisms, even at the molecular level, was open. Ukrainian born physicist George Gamow was one of the first to realize that the nucleotide sequences were in fact a code. He tried many, yet mostly unsucessful, explanations of the the properties of the code.

Once you cut your way into the veil of technical jargon, it is easy to understand why people get fascinated with such beautifully complex machinery hidden inside cells. It has been optimized to attain maximum efficiency; it is fine tuned to minimize errors, yet allowing some mutations to occur, so that the organism as a whole can evolve and adapt.

In this paper I will try to explain in simple terms the importance of the genetic code. The paper is divided into two parts. The first part explains what the genetic code is and where does it fit inside protein biosynthesis. And the second part will survey some theories on how the code came into being given its characteristics.

2 The Genetic Code

Before defining what the genetic code is, first it it necessary to have a basic understanding the process of protein biosynthesis, and the actors involved.

2.1 Protein biosynthesis

First, we have the *deoxyribonucleic acid (DNA)* molecule, arguably the most important molecule for life. In DNA all the information to create a living organism is encoded in its composition. DNA is a double-stranded aperiodic polymer of a base with a deoxyribose (a sugar) backbone. The base + sugar molecule is called *nucleotide*. There are four DNA bases, divided in two groups: *purines* and *pyrimidines*. The purines are *Adenine* (A) and *Guanine* (G) and the pyrimidines are *Cytosine* (C) and *Thymine* (T). In DNA, a strand is a series of these nucleotides, and the other strand is complementary in the sense that an A in a strand always is chemically bound to a T in the other strand, and a G in one strand is bound to a C in the other. In this way the information encoded in DNA is redundant, so one strand could be reconstructed from the other. This particular pairing scheme is known as the *Watson-Crick pairing*.

Information in DNA is subdivided into subsets called *genes*. Each gene has the information that can be used to build a specific *protein*. A protein is an organic molecule that consists of amino acids joined by peptide bonds, and perform different functions in an organism. The process by which a gene is read and its corresponding protein is synthetized is called, appropriately, protein biosynthesis.

Before I go into the description of the process, I must introduce RNA. The *ribonucleic acid* is a single stranded, aperiodic linear molecule. Like DNA, it has 4 nucleotide bases, namely A, G ,C and *Uracil* (U). U, like T, is a pyrimidine and is also complementary to A, but is only present in RNA. The difference between RNA and DNA, besides being single stranded, is that RNA has a ribose backbone instead of a deoxyribose one.

Protein biosynthesis has two major steps, transcription and translation. *Transcription* is the process where a gene is read off DNA and its information is literally transcribed onto a RNA by a specialized enzyme. This enzyme, called *RNA polymerase*, “reads” a strand of the gene nucleotide by nucleotide and synthetizes RNA with nucleotides complementary to the ones found in the gene. For example, if the gene read ATCTGACCG, the corresponding RNA would read UAGACUGGC. This RNA will then carry this information to the ribosome to be translated. Hence this particular RNA is termed *messenger RNA* (mRNA). *Trans-*

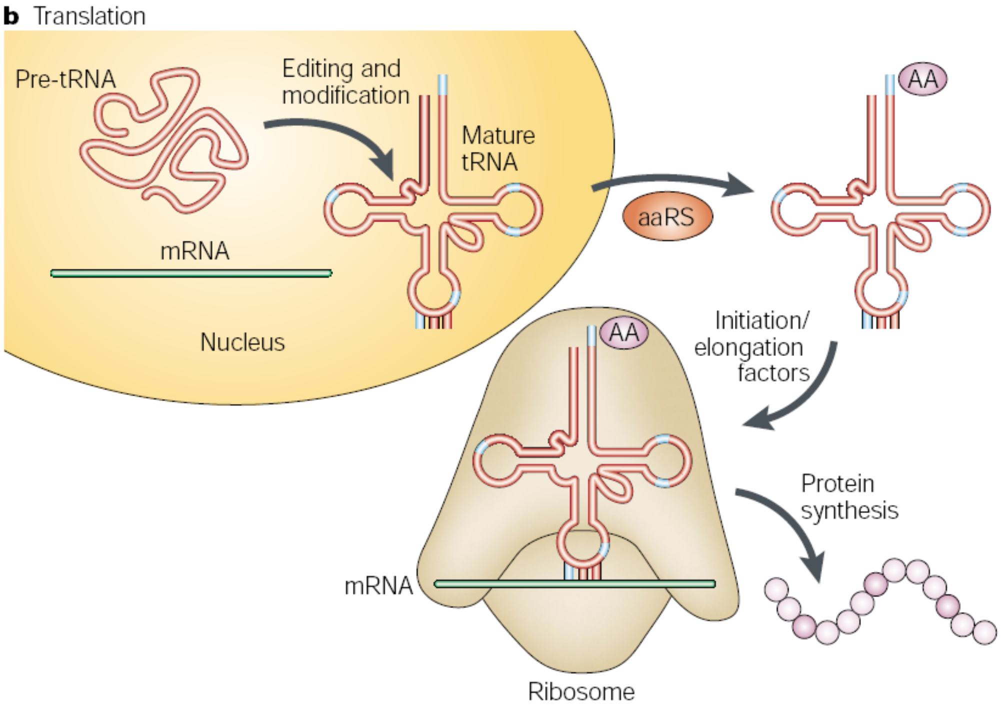
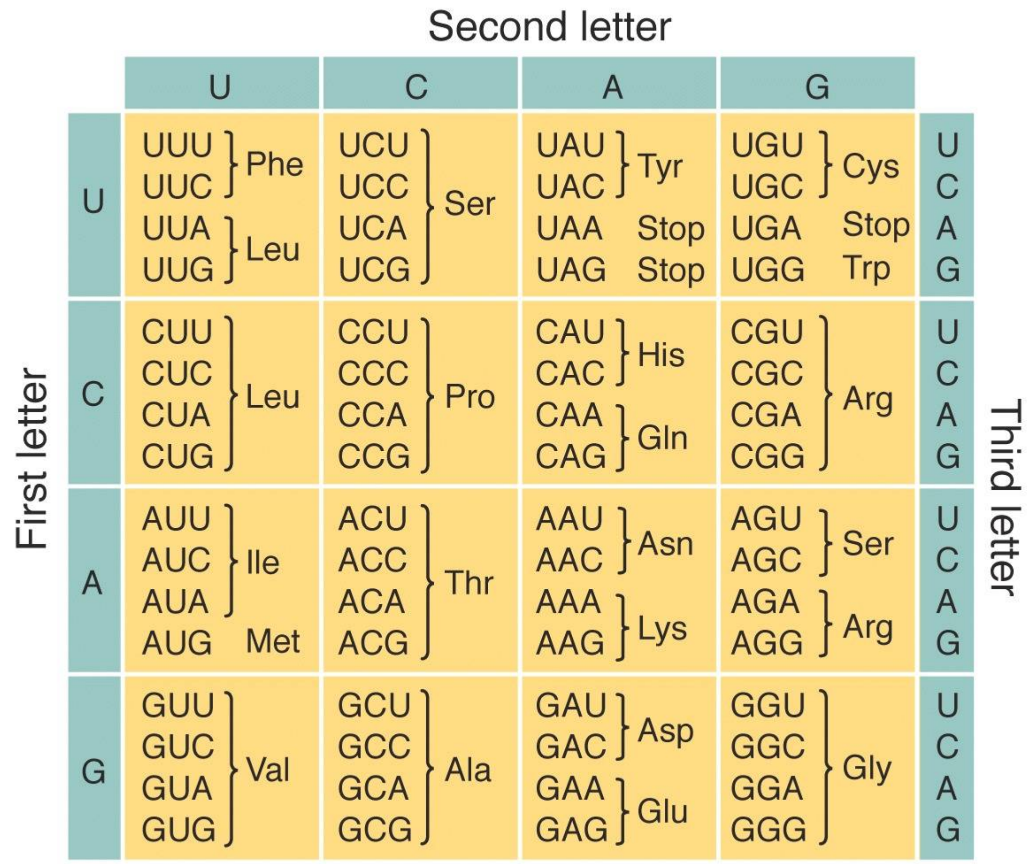


Figure 1: A graphical overview of the translation process. From [5]

lation is the process where information in mRNA is read off by the ribosome and then a protein is assembled. Information is stored in mRNA in nucleotide triplets known as *codons*. Each codon stands for an aminoacid (note that, as we will see later, an aminoacid may be coded by more than one codon). Codons are non-overlapping, in the sense that once a codon is read, the ribosome then jumps to the next codon. For example, UAGCCG has two codons, UAG and CCG. The assembly process involves yet another RNA molecule, *translation RNA* (tRNA). tRNA is a short, non-coding RNA whose purpose is to bring amino acid molecules to the ribosome. Each tRNA has two very special sites. There is a binding site where an amino acid is “loaded” onto the tRNA via an enzyme called aminoacyl-tRNA synthetase (aaRS). The other special site contains a nucleotide triplet known as *anticodon*. When building a protein, the ribosome matches a codon in mRNA with its corresponding anticodon, and then it unloads the amino acid from tRNA. Said amino acid is then added to the growing protein. Then the ribosome repeats the process with the next codon until a Stop codon is found.

This was a very simplistic overview of the protein biosynthetic process, but it’s necessary to understand the context where the code itself appears, and hence we can appreciate its importance.

2.2 The Genetic Code



	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 2: The Canonical Genetic Code. Figure taken from the Genetics, Evolution, and Molecular Systematics Laboratory at the Department of Biology of the Memorial University of Newfoundland.

The genetic code is set of rules that maps codons to amino acids (see figure

2). We can think of it as the alphabet by which you can encode the information needed to build a protein. The code has many properties[1], namely:

- There are 64 codons, each of which is a triplet of nuceotide bases.
- Only 20 amino acids are used. These are called the *standard amino acids* (see figure 3).
- There appear to be some rules, together with some exceptions:
  - XYU and XYC always code the same amino acid.
  - XYA and XYG often code the same aminoacid.
  - In 8 out of the 16 possible cases, XY· encodes a single amino acid, where · represents any of the four bases.
- The code is nearly universal. That is, it seems that the vast majority of living organisms on Earth use this code. This particular code is known as the *Canonical Genetic Code*.
- Besides simple grouping, it seems that the code is not just a random association of codons and amino acids. There seems to be an intriguing underlying order. For instance, all codons with U in the second place code for hydrophobic amino acids.

Many questions have arisen about how the code as we see it right now came into being, as well as what is so special about this particular code -there could be other codes. I'll go into some details in the next section.

3 Origin of the Code

It seems rather obvious that the beautifully complex biosynthetic machinery and the code it uses must have evolved from simpler systems. The problem lies in the fact that when this processes were developed, any primitive synthetic system was wiped off existence due to evolutionary pressures. This code was a enormous advantage at the time. Nevertheless, some information can be hypothetized and inferred about how a code was necessary.

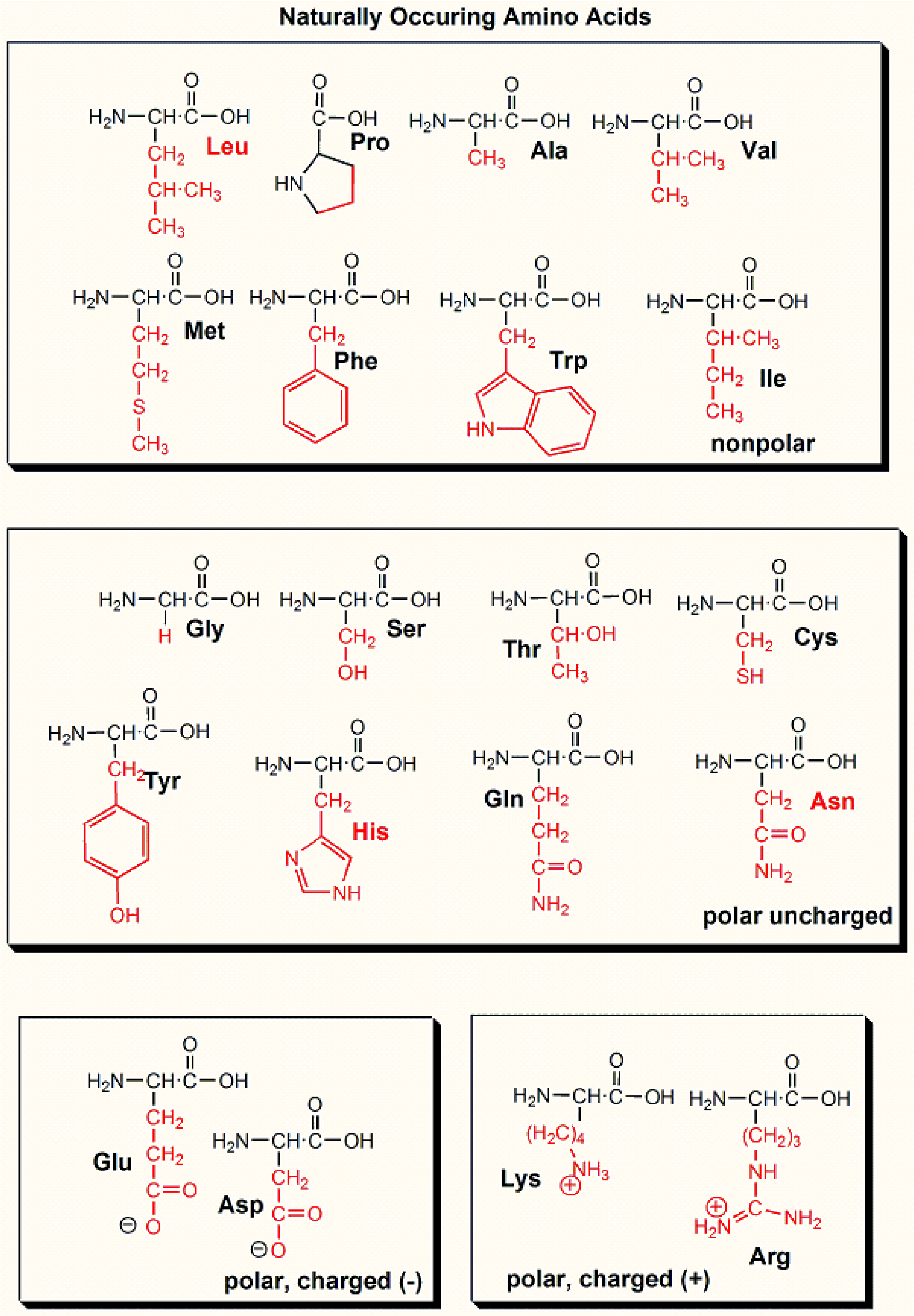


Figure 3: The 20 standard amino acids, and their chemical structures. Figuen from `http: www.csbsj.edu`

3.1 A coded state

Some people have questioned why there is a need for a coded system. By "coded" it is meant that there is a univocal relationship between codon and amino acid (the converse is not true). So the following scenario was suggested [4]. Assume there exists a primitive self-replicating protocell in a RNA World[1] with enough abundance of nucleotides and amino acids for the biosynthetic apparatus to create a single protein, a RNA replicase. The rest of the system can be composed of RNA molecules used for coding and as catalysts (ribozymes). In this scenario, an initial condition would be that each codon can account for many different aminoacids. Hence, given the gene that encodes the replicase, there is a probability that a replicase will actually be synthetized. A protocell with a higher yield would predate protocells with lower yields, since it can replicate itself faster and more efficiently. A way to maximize yield is to maximize the specificity of the codon to amino acid relationship. In particular, an optimum, stable yield occurs in the coded state, where each codon accounts only for one aminoacid. Still, some errors in translation are required, since in this way the system will be error resistant (that is, despite errors, a functional replicase is succesfully made) via suppressor mutations (mutations that cancel out the deleterious effects of other mutations).

This by no means attempts to be the real explanation of how the actual coded state came to be, but it shows a plausible scenario where a coded state could emerge naturally.

3.2 One in a million

Once a coded state is reached, the primitive organism could evolve more specific proteins and the coding system could be refined by including additional amino acids or an extra pair of nucleotide bases (you can have a very simple functional code with only 2 nucleotides).

Eventually a steady state would be reached. Francis Crick proposed[4] that this state may occur because once the code reached its present, complex form, any further alterations will be lethal due to the functional change in proteins because of the change in amino acid assignment for a certain codon. He said that the canonical code was a "Frozen Accident". Evidence for the contrary has been found, though, as organisms with variant codes have been discovered, and that these codes have evolved in a comparatively recent time.

[1] The *RNA World* is a hypothesis which claims that prebiotic self-replicators and simple primitive life could have existed in Earth, and that DNA was a later addition.

There another way in which the code could reach a steady state. The code could be optimal. The problem arises in finding out which property the code was optimized for. A way out is looking at the “frozen accident” hypothesis. The code is probably optimized for minimizing the effects of translation errors. If a translation error is made, the resultant amino acid should have a similar function as the original amino acid.

If we look at the code carefully at the code, we see that acidic amino acids are grouped together, as well as the hydrophobic grouping I pointed out earlier. Hence the error minimization theory is consistent with this observation.

A way to test this assertion is by trying out many codes and minimize a certain quantity. Carl Woese proposed that the key quantity is a measure of amino acid hydrophobicity. He dubbed this quantity *polar requirement*. Together with other technical constraints, it was required that the optimal code should minimize the change in polar requirement when mistranslating a codon by misreading one of the nucleotides of the codon. By trying about a million different assignments it was found that the canonical code excels at satisfying this requirement. It can be said, literally, that the genetic code is “one in a million”.

4 Conclusion

In this paper a quick review of the genetic code and its properties was done. There are many questions left that are still the subject of current research, so we will be hearing news about this topic soon.

References

[1] F.H.C. Crick, *J. Mol. Biol.* **38** (1968), 367

[2] C. Woese, *PNAS* **54** (1965), 1546

[3] R.D. Knoght and L.F. Landweber, *Cell* **101** (2000), 569

[4] V. Bedian, *Biosystems* **60** (2001), 39

[5] R.D. Knight, S.J. Freeland and L.F. Landweber, *Nature Rev. Gen.* **2** (2001), 49

[6] G. Weberndorfer, et al, *SFI Working paper 02-08-034*