Biol Theory (2015) 10:301–310
DOI 10.1007/s13752-015-0225-z

ORIGINAL ARTICLE

# Evolution of the Genetic Code: The Ribosome-Oriented Model

**Marcello Barbieri**[1]

Received: 24 August 2015 / Accepted: 2 September 2015 / Published online: 7 October 2015


**Abstract** There are currently three major theories on the origin and evolution of the genetic code: the stereochemical theory, the coevolution theory, and the error-minimization theory. The first two assume that the genetic code originated respectively from chemical affinities and from metabolic relationships between codons and amino acids. The error-minimization theory maintains that in primitive systems the apparatus of protein synthesis was extremely prone to errors, and postulates that the genetic code evolved in order to minimize the deleterious effects of the translation errors. This article describes a fourth theory which starts from the hypothesis that the ancestral genetic code was ambiguous and proposes that its evolution took place with a mechanism that systematically reduced its ambiguity and eventually removed it altogether. This proposal is distinct from the stereochemical and the coevolution theories because they do not contemplate any ambiguity in the genetic code, and it is distinct from the error-minimization theory because ambiguity-reduction is fundamentally different from error-minimization. The concept of ambiguity-reduction has been repeatedly mentioned in the scientific literature, but so far it has remained only an abstract possibility because no model has been proposed for its mechanism. Such a model is described in the present article and may be the first step in a new approach to the study of the evolution of the genetic code.

✉ Marcello Barbieri
brr@unife.it

[1] Department of Morphology and Embryology, University of Ferrara, Ferrara, Italy

**Keywords** Code ambiguity · Code evolution · Codepoiesis · Genetic code · Ribosomal proteins · Ribosomes

## The Apparatus of Protein Synthesis

In 1946, Jean Brachet argued that "DNA is primarily confined to the nucleus, while RNA is mainly found in the cytoplasm, and protein synthesis is associated with RNA." More precisely, Brachet (1944, 1946) argued that protein synthesis takes place on heavy cytoplasmic particles that he called *ribonucleoprotein granules*.

Shortly afterwards, Boivin and Vendrely (1947) carried the scheme of Brachet to its logical conclusion and proposed that "DNA makes RNA makes Proteins." This 1947 version of the central dogma of molecular biology was largely ignored, but the idea was born and the evidence started accumulating.

In 1952, Alexander Dounce made another revolutionary proposal. He suggested that there must be a *code of correspondence* between nucleic acids and proteins, and since there are 20 amino acids but only four nucleotides, he proposed that each amino acid is coded by a group of three nucleotides (a *codon*). On top of that, Dounce (1952, 1953) proposed that the attachment of the 20 canonical amino acids to nucleotides is promoted by 20 specific catalysts that he called *activating enzymes*. Since their discovery, these molecules have been called *aminoacyl-tRNA-synthetases*, or, more briefly, *synthetases*.

Between 1946 and 1952, in short, the basic concepts of molecular biology had already been formulated but were largely ignored. The turning point came in 1953, when James Watson and Francis Crick proposed the model of the double helix, the idea that in one swift stroke illuminated

the structure of DNA and suggested that hereditary information is carried by linear sequences of nucleotides (Watson and Crick 1953).

In 1957 Francis Crick argued there must be intermediary molecules between nucleotides and amino acids, molecules that are necessarily made of RNA because they must be able to recognize a codon by a complementary triplet of nucleotides that he called *anticodon*. Crick (1957) called these molecules *adaptors*, but in that same year they were discovered by Hoagland et al. (1957) and became known as *transfer RNAs*.

A year later, Crick (1958) re-proposed the central dogma, and this time the idea was immediately accepted: *DNA makes RNA* (transcription) *and RNA makes proteins* (translation). In the same year, Roberts (1958) gave the name *ribosomes* to the molecular machines that make proteins (the ribonucleoprotein granules described by Brachet), and the conclusion that *the ribosome is the decoder of genetic information* acquired the status of an experimental fact.

Ribosomes account for more than 80 % of the total RNA of a cell, and in the 1950s it was taken for granted that the information of the genes is transported by ribosomal RNAs, but things turned out very differently. In 1961, François Jacob and Jacques Monod proved that the carriers of genetic information are a completely different family of RNA molecules that they called *messenger RNAs* (Jacob and Monod 1961). Later on, it was discovered that the scanning of the messenger RNAs requires a whole new battery of enzymes that were called *initiation*, *elongation*, and *termination factors* (Nomura et al. 1974).

The apparatus of protein synthesis, in conclusion, consists of ribosomal RNAs, messenger RNAs, transfer RNAs, ribosomal proteins, aminoacyl-synthetases, scanning factors, and amino acids. It is a huge supramolecular system made of more than 120 different types of molecules.

What is most extraordinary, however, is that the rules of the genetic code are virtually identical in all living creatures (Hinegardner and Engelberg 1963; Woese et al. 1964). A few exceptions do exist but they are very minor changes and occur in an infinitesimal number of organisms. The genetic code, in other words, is virtually universal, and this means that it has been transmitted to all forms of life by a population of primitive systems that has become known as the *common ancestor*. Ever since this discovery, the origin of the genetic code has become one of the greatest problems of biology.

## Characteristics of the Code

A messenger RNA is scanned by a ribosome in groups of three nucleotides, called *codons*, and every codon is recognized by the *anticodon* of a transfer RNA (tRNA) that carries one of the 20 canonical amino acids. The total number of codons that can be obtained with four nucleotides (A, U, C, G) is 64 ($4^3$), and the rules of correspondence between 64 codons and 20 amino acids represent, collectively, *the genetic code*.

In 1961, Nirenberg and Matthaei announced that an artificial messenger containing only uracil (U) is translated into a protein that contains only phenylalanine (Phe), which means that the codon UUU codes for phenylalanine (Niremberg and Matthaei 1961). They had deciphered the first code word of the genetic code. Other code words were identified with artificial messengers made of nucleotides arranged in various orders (Speyer et al. 1963). It was found, for example, that a sequence of alternating uracil and cytosine (…UCUCUCUC…) codes for a polypeptide made of alternating serine and leucine (…Ser-Leu-Ser-Leu…), thus proving that UCU is a codon for serine and CUC is a codon for leucine (Niremberg and Leder 1964; Nishimura et al. 1965).

Various other techniques were designed to unlock the meaning of the other codons, and by 1966 the genetic code was completely deciphered (Khorana et al. 1966; Niremberg et al. 1966).

It turned out that 61 codons code for amino acids, and that one of them (usually AUG) is also used as a start signal, whereas the remaining three (UAA, UAG, and UGA) are termination signals. Between 61 codons and 20 amino acids there is necessarily a many-to-one correspondence, and this is expressed by saying that the genetic code is *degenerate* (or *redundant*). More precisely, some amino acids are specified by six codons, some by four, others by two, and only two amino acids are coded by a single codon. The genetic code is therefore redundant but *not ambiguous* because *any one of the 61 codons codes for one and only one amino acid*.

In principle, this implies that every cell contains 61 different types of tRNAs, one for each codon, but in practice the actual number is about 40 per cell. The best explanation for this surprising fact was proposed by Crick (1966) with what has become known as the *wobble hypothesis*. Crick pointed out that the three nucleotides of an anticodon stick out like fingers from the surface of their tRNA and this allows them to oscillate, or *wobble*. The result is that a nucleotide in an anticodon (especially in third position) can form a temporary bond not only with its complementary nucleotide but also with another one with which it has a partial similarity. A uracyl in third position, for example, can form a bond not only with adenine but also with guanine, and in this case its tRNA can associate the same amino acid to two distinct codons. This means that only one tRNA, rather than two, is sufficient when two codons specify the same amino acid, and this is indeed what happens in various cases. It has been found, for

example, that the codons that end with adenine (XYA) and those that end with guanine (XYG) code for the same amino acid, and the same is true for the codons that end with uracyl (XYU) and with cytosine (XYC). Lisine, for example, is codified by AAA and AAG, whereas tyrosine is codified by UCU and UAC.

There are, in short, regularities in the genetic code that allow a cell to carry far less than 61 different types of tRNAs, and that have, as we will see, important biological effects. This implies that the modern code has been the result of an evolutionary process, but first we must address a preliminary question: what was the starting point of that process?

## The Ancestral Genetic Code

The ribosomal RNAs are among the most conserved molecules in evolution (Woese 1987, 2000) and this means that they appeared very early on the primitive Earth. It is also known that they contain regions that have the ability to form peptide bonds (Nitta et al. 1998), and this means that some primitive ribosomal RNAs could stick amino acids together at random and produce statistical proteins. These proteins did not have biological specificity but could still be useful because the RNAs can barely work on their own. They need amino acids and peptides to maintain stable conformations, and their functions are greatly enhanced by the attachment of small proteins (Orgel 1973). This is why an apparatus of protein synthesis started evolving from pieces of ribosomal RNAs, possibly stabilized by random polypeptides.

The next step in the evolution of this apparatus was the acquisition of transfer RNAs, molecules that have the ability to carry amino acids to the site of protein synthesis. All modern transfer RNAs are small molecules (75–90 nucleotides long) with a basic cloverleaf structure that has been highly conserved in evolution, which strongly suggests that they descended from a common ancestor. The contribution of these molecules to protein synthesis, on the other hand, was greatly enhanced by a third type of RNAs, because at the site of synthesis it is necessary that the amino acids be kept in place for a long enough time to allow the formation of a peptide bond (Wolf and Koonin 2007; Fox 2010). This means that the transfer RNAs required temporary *anchoring sites*, and in primitive systems these were provided by *anchoring RNAs*, the ancestors of the *messenger RNAs* (Osawa 1995).

The combination of ribosomal RNAs, transfer RNAs, and anchoring RNAs gave origin to an apparatus of protein synthesis where the transfer RNAs were automatically creating a bridge, or a *mapping*, between codons and amino acids, and any such mapping is, by definition, a *genetic code*.

We realize in this way that the genetic code appeared on Earth when transfer RNAs and anchoring RNAs joined the ribosomal RNAs and became an integral part of the apparatus of protein synthesis. Here, this first code is referred to as the *ancestral genetic code*.

But what type of code was it? We know that the modern code is *nonambiguous* because any one of its codons codes *for one and only one amino acid*, but what about the ancestral code? It has been underlined that that code was almost certainly *ambiguous* because at such an early stage nothing could prevent a codon from coding for two or more amino acids (Fitch and Upper 1987; Osawa 1995). This means that a sequence of codons was translated sometimes into one protein and at other times into a different protein, and the apparatus was inevitably producing *statistical proteins*.

It is a fact, on the other hand, that a fully *nonambiguous* genetic code did appear on the primitive Earth, and this means that the ambiguity of the ancestral code was steadily reduced until it reached a point in which any codon could code for one and only one amino acid. When that happened the first nonambiguous code came into existence, a code that here is referred to as the *ancient genetic code*.

The transition from ancestral to ancient genetic code transformed the early systems based on statistical proteins into the first systems that were producing specific proteins, but how did it happen?

Koonin and Novozhilov (2009) have shown that today there are three major theories on the origin and the evolution of the genetic code—the *error-minimization theory*, the *stereochemical theory*, and the *coevolution theory*—and we need therefore to examine the mechanisms that they propose.

## The Error-Minimization Theory

The translation of a sequence of nucleotides into a sequence of amino acids is subject to a variety of errors that can be studied in vitro. In particular, they have been studied by experiments where protein synthesis takes place in different environmental conditions and with a variety of artificial messengers. The most used messenger is the poly-U molecule where all codons are UUU and all amino acids should be phenylalanine (Phe). In this case, the appearance of any other amino acid can only be the result of a translation error, and it is possible therefore to make a statistical study of these errors. The overwhelming result of these studies is that the error rate in the third position of a codon is about 100 times greater than that in the first position, which in turn is about ten times greater than the error rate in the second position. The third position, in other words, is the most error prone, whereas the second position is the most stable.

On the basis of these results, in 1965 Carl Woese proposed a theory on code evolution that consisted in two main concepts:

(1) The first is the idea that in ancestral systems "the translation mechanism was a far more rudimentary thing than at present, in particular far more prone to make translation errors…. We shall assume that errors in translation were extreme, to such an extent that the probability of translating correctly any given messenger-RNA was essentially zero. From this concept of error-ridden translation in the primitive cell it follows that the proteins produced by any given gene will have to be statistical proteins" (Woese 1965, p. 1548).

(2) The second concept is the idea that the codons of the ancestral code were "readjusted" or "reallocated" in order to minimize the effects of the translation errors. More precisely, the ancestral code evolved in such a way that the codon resulting from a translation mistake would code either for the same amino acid or for an amino acid with very similar chemical properties (Woese 1965).

This has become known as the "error-minimization theory," but it must be underlined that what is minimized are not the translation errors but their biological effects. Woese pointed out that the translation apparatus did improve its performance in the course of evolution, and became less and less prone to errors, but that was an altogether different process, and one that necessarily took place at a later stage.

More precisely, Woese argued that it would be tautological to say that the primitive cells evolved a more efficient translation apparatus by learning to translate more efficiently. "The way out of this paradox is that although unable to reduce the translation error rate, the primitive cell can do something tantamount to this by adjusting the code so that the 'effect' of the translation errors is lessened" (1965, p. 1549).

Hence the idea that the primitive cells had first to evolve a genetic code that could minimize the effects of the translation errors, and only after that could they start improving the translation apparatus. Woese proposed in this way a theory where "… the evolution of the genetic code starts with a primitive cell possessing random, ambiguous codon assignments, and a very error-ridden translation process, and it was the 'necessity' of minimizing the effects of translation errors that led to the highly ordered code that we observe today" (1965, p. 1550).

Woese acknowledged that a similar theory was proposed by Sonneborn (1965) with the idea that the genetic code evolved in order to minimize the lethal effects of ordinary mutations, and the concept of "error minimization" was extended to all errors deriving from translation mistakes and from mutations.

At a later stage, Woese proposed that horizontal gene transfer is another powerful mechanism of code evolution and suggested that it was most probably that mechanism that was responsible for the near universality of the modern genetic code (Woese 2002; Vetsigian et al. 2006).

## The Stereochemical Theory

In 1954, George Gamow proposed the *stereochemical hypothesis*, the idea that the amino acids fit with a lock-and-key mechanism into "holes" formed by four nucleotides, and that it is the three-dimensional shape of each hole that determines which amino acid binds to which quartet of nucleotides. According to this *diamond code* (Gamow 1954) the rules of the genetic code are the result of chemical affinities between codons and amino acids and are therefore determined by chemistry.

Gamow's model was quickly abandoned when it became clear that in protein synthesis there are no direct contacts between codons and amino acids, but the idea of stereochemical interactions between them was not discarded and has been re-proposed ever since in many different forms.

Pelc and Weldon (1966) suggested that there is a stereochemical complementarity between amino acids and their codons; Dunnill (1966) argued that the anticodon loop of the transfer RNA forms a molecular pocket in which the amino acid can be trapped; Melcher (1974) proposed that amino acids have a stereochemical correlation with their anticodons; and Shimizu (1982) maintained that a complementary relationship exists between amino acids and groups of four nucleotides in the transfer RNAs.

It must be underlined that the interactions between nucleotides and amino acids are beyond dispute. The negative charges of the nucleotide phosphates attract the positive charges of the basic amino acids and this gives origin to countless interactions between them. The crucial point is that these interactions take place between any nucleotide and any amino acid and in no way account for the *specificity* of the coding rules. The stereochemical theory is the idea that *in addition* to the standard chemical interactions there are also specific affinities between codons and amino acids, and the history of this theory has been but a long journey in search of such alleged affinities.

A systematic study on chemical affinities was conducted by Saxinger et al. (1971) by filtering nucleotides on gels containing amino acids and by measuring the quantities of amino acids that were selectively retaining triplets of nucleotides. The results were disappointing and indicated that there are at best very weak specific interactions between amino acids and codons.

The stereochemical theory was revived again when Yarus discovered that there is a selective interaction

between the amino acid arginine and a nucleotide that is present in all four codons that code for arginine (Yarus 1988, 1998). Later on Yarus and colleagues extended the research to eight amino acids and reported other positive correlations (Yarus et al. 2005), but in these cases the evidence was much weaker and arginine remained an isolated exception.

Another difficulty for the stereochemical theory is its potential conflict with the error-minimization mechanism. If it is true that codons can be reallocated to different amino acids in order to minimize the effects of translation errors, one is bound to conclude that there are no specific chemical affinities between them. Yarus replied to this objection with the argument that only some coding rules are determined by stereochemistry whereas others are free to change and allow the system to mitigate the effects of mutations and translation errors (Yarus et al. 2005).

After decades of research, in conclusion, there still is no real evidence in favor of the stereochemical theory, and what keeps it alive is the *possibility* that stereochemical interactions between codons and amino acids might have been important at some early stages of evolution (Koonin and Novozhilov 2009).

## The Coevolution Theory

The origin of the genetic code is still a mystery, but we do have some interesting clues. The first is that the number of amino acids *changed* during the early evolution of the code. This is because only *less than half* of the 20 canonical amino acids can be synthesized from inorganic molecules and for this reason are referred to as "primary" (or "precursor") amino acids. The others are always synthesized taking primary amino acids as starting points and are referred to as "secondary" (or "product") amino acids.

The crucial point is that only primary amino acids are produced in laboratory experiments that simulate prebiotic conditions, which strongly suggests that less than ten primary amino acids appeared on the primitive Earth (Wong and Bronskill 1979). This conclusion is supported by the discovery that the amino acids that are missing in laboratory syntheses are also missing from meteorites (Higgs and Pudritz 2007). The implication is that the evolution of the genetic code started with less than ten amino acids and went all the way up until it reached the canonical set of 20 that has been strongly conserved ever since.

The second important clue on the genetic code is that the codons assigned to the secondary amino acids differ very little from the codons of their precursors amino acids. The best explanation of this pattern, so far, is the theory proposed by Jeffrey Wong, the so-called *coevolution theory* of genetic code and amino acid biosynthesis (Wong 1975, 1981).

Wong proposed that the secondary amino acids received their codons from the primary amino acids that served as precursors in their biosynthesis. This theory predicts that "the codons of precursor-product amino acids should be contiguous, i.e., separated by only a single base change" (1975, p. 1909), and in many cases this is what is actually observed.

Wong concluded that the first genetic code that appeared on Earth was codifying only primary amino acids, and all codons were assigned to them. Later on, during the evolution of the genetic code, the secondary amino acids steadily increased in number by new biosynthetic pathways and received their codons from the primary amino acids that served as precursors in their biosynthesis.

The coevolution theory proposed by Wong starts from a genetic code where all codons are assigned to less than ten primary amino acids, but does not say how this first code came into existence. This issue was taken on by Di Giulio (2008) who proposed an "extension" of the original theory that applies the same mechanism also to the primary amino acids of the first genetic code.

The key idea of the coevolution theory, in both the original and the extended form, is the hypothesis that the codon of any secondary amino acid comes from the codons assigned to its precursor amino acid because in the early stages of evolution the synthesis of amino acids was taking place on transfer RNAs, and there was therefore a *metabolic continuity* between RNAs and amino acids. In this case, the relationships between codons and amino acids are no longer due to chemical affinities, as in the stereochemical theory, but continue to be deterministic relationships because they are dictated by metabolic reactions.

The problem with the coevolution theory is that the synthesis of amino acids *could* have taken place on RNAs in the RNA-world but this is certainly no longer true in the protein world. When the amino acids ceased to be synthesized on RNAs, the ancient metabolic connections between codons and amino acids disappeared and with them disappeared the rules of the ancestral genetic code, so how did these rules reemerge in the protein world? The coevolution theory, in other words, *might* account for the origin of the genetic code in the RNA world, but leaves us with the huge problem of understanding how the same coding rules reappeared in the protein world.

## Arbitrariness

The stereochemical theory and the coevolution theory assume that the genetic code originated respectively from chemical affinities and from metabolic relationships between codons and amino acids, and in both cases it

would not be a real code because its rules would not have the arbitrariness that characterizes all true codes.

The crucial point, here, is the relationships that exist between the recognition of the amino acids, performed by the synthetates (*aminoacyl-tRNA-synthetases*) and the recognition of the codons performed by the anticodons of the transfer RNAs. If the synthetases could recognize both the amino acids and the anticodons, they would establish direct connections between them and the rules of the genetic code would be deterministic, but this is precisely what the evidence has ruled out. Experiments have shown that the recognition of the amino acids is *independent* from the recognition of the anticodons because in many cases the synthetases have no access to the anticodons (Schimmel 1987; Schimmel et al. 1993). On top of that, it has been shown that the links between codons and amino acids could have been made in countless different ways. Hou and Schimmel (1988), for example, managed to introduce two extra nucleotides in a tRNA without changing its anticodon, and found that that the resulting tRNA was recognized by a different synthetase and was carrying therefore a different amino acid. They had changed one of the rules of the genetic code, a result that achieved in vitro what a few microorganisms have achieved in vivo in the course of evolution (Jukes and Osawa 1990, 1993).

The lesson that comes from these experiments is that the rules of the genetic code are the result of interactions between synthetases and tRNAs that can be modified *virtually at will* by adding or subtracting a few molecules. This means that the number of adaptors between codons and amino acids is potentially unlimited, and only the selection of a fixed number of them can ensure a specific correspondence. It also means that the rules of the genetic code are not dictated by any form of chemical necessity, and in this sense they are *arbitrary*.

It is worth mentioning, at this point, a fairly widespread argument according to which the rules of the genetic code cannot be arbitrary because they have been optimized in the course of evolution. In reality, the two things are not incompatible. The rules of the Morse code, for example, have been optimized by associating the most frequent letters of the alphabet with the shortest combinations of dots and dashes, and yet they continue to be arbitrary rules. An optimization of the coding rules, in other words, simply means that they are not *random*, and is perfectly compatible with their arbitrariness.

We reach in this way the conclusion, first expressed by Monod (1970), that the genetic code is *chemically arbitrary* because its rules are not dictated by necessity. The genetic code, in short, is a real code and this makes the problem of its origin all the more challenging and interesting.

## The Missing Theory in Code Evolution

If we admit (1) that the ancestral genetic code was ambiguous and (2) that the ancestral apparatus of protein synthesis was error prone, we conclude that the primitive systems contained two different types of statistical proteins: those produced by code ambiguity and those produced by translation errors. This in turn implies that the evolution of the genetic code took place by two distinct mechanisms, one that reduced code ambiguity and one that minimized the effects of translation errors. We should have therefore two distinct theories on code evolution, but what we have had up to now is only the error-minimization framework. So far, the ambiguity-reduction framework has remained an abstract possibility and can rightly be regarded as the *missing theory* in code evolution.

One may be tempted to suggest that ambiguity reduction can be included in the error-minimization category, but this is not the case. Error minimization implies that some associations between codons and amino acids are normal and others are the result of errors, whereas in code ambiguity all associations between transfer RNAs and amino acids are equally "normal." A solution of the error-minimization problem, in other words, is in no way a solution of the ambiguity-reduction problem. A genetic code can continue to be ambiguous even when the translation apparatus has become completely error free.

In his 1965 paper, Woese underlined that the ancestral genetic code was *necessarily* ambiguous, but then he concentrated on error minimization only and did not mention a separate mechanism for the reduction in code ambiguity. Some forty years later, Woese and collaborators underlined again that codon ambiguity has nothing to do with errors, and said so in no uncertain terms: "Ambiguity is therefore not the same thing as error" (Vetsigian et al. 2006, p. 10696). That seminal paper, however, was dedicated to the role of horizontal gene transfer in the evolution of the genetic code, and the reduction in ambiguity was mentioned only as something that necessarily took place but whose mechanism is still unknown.

The ambiguity-reduction concept has remained in this way a missing theory, a project for the future, and it is high time therefore that we try to address it. More precisely, that we try to figure out the mechanism that brought it about. To this purpose, it may be useful to recall the "little parable" of the hotel keys proposed by Nino (1982).

In any hotel there are two types of keys: the familiar keys that open individual doors and the passkey that opens all doors. At first, one may be forgiven for thinking that the passkey is the most complex of all, whereas the truth is precisely the other way round. The passkey is the simplest one because what is complex in a key is not the ability to

open a door (that can easily be done with a screwdriver) but the ability to open one door *and not all the others*. This suggests an interesting parallel with the evolution of the genetic code.

An ancestral ribozyme might have been able to attach amino acids to all transfer RNAs—like a passkey that opens all doors—thus giving origin to a completely ambiguous genetic code. A reduction in the ambiguity of this ancestral code would have been achieved by evolving ribozymes that became capable of attaching fewer and fewer amino acids to any transfer RNA, until the point was reached when a ribozyme could deliver to any transfer RNA one and only one amino acid.

The problem is: why did this happen? What was the driving force that fueled a systematic reduction of ambiguity in the ancestral genetic code?

## The Ribosome-Oriented Model

Ribosomes consist of a small and a large subunit that are made of ribosomal.

RNAs and ribosomal proteins. The ribosomal-RNAs account for more than half of the huge molecular weight of the ribosomes (over 2 million), and the rest is accounted for by more than 50 different ribosomal proteins of low molecular weight, each present in one or a few copies (Nomura et al. 1974). These proteins are the descendants of the statistical proteins of the ancestral ribosomes, but how did they evolve? Today we have at least two important clues about this major transition.

The first comes from the fact that there was an evolutionary advantage in increasing the total number of ribosomal proteins irrespective of their individual characteristics. The reason comes from a general principle in engineering that Burks (1970) expressed in this way: "there exists a direct correlation between the size of an automaton—as measured roughly by number of components—and the accuracy of its function." In our case, this principle means that increasing the number of ribosomal proteins was making the ribosomes more heavy, more resistant to thermal noise, and therefore more reliable in protein synthesis.

The second clue comes from the fact that ribosomes are formed by self-assembly from their components, and it has been possible to discover the contribution of individual ribosomal proteins by studying what happens when ribosomes are reassembled without any one of them in turn. These experiments have shown that the ribosomal proteins fall into three major groups: some are necessary for function, others are required for self-assembly, and those of the third group have a stimulating effect but are fundamentally disposable (Kurland 1970; Fox 2010).

The evolution of the ribosomal proteins was therefore a process that gave origin first to three great families and then to an increasing number of subfamilies. These subfamilies, on the other hand, could be transmitted to future generations only if the ambiguity of the genetic code was lower than the statistical differences between them. The ambiguity of the code, in other words, was the limiting factor that determined how many subfamilies of statistical ribosomal proteins could reappear in the descendants. Which means that only by decreasing the ambiguity of the ancestral genetic code was it possible to promote the evolution of the ribosomal proteins—an evolution that increased their number, that diversified their functions, and that favored the reappearance of increasingly similar types of ribosomal proteins in the descendants because this made it easier for ribosomes to self-assemble in every new generation.

It will be noticed that the evolution of the ribosomal proteins was not about this or that protein or this or that protein function. It involved all statistical proteins at the same time. It was not about *individual* features but about *collective* relationships. It was an evolution that went on until the ambiguity of the genetic code was completely removed and *biological specificity* came into existence.

It is possible, in other words, that the evolution of the ancestral genetic code was *ribosome oriented*. But do we have any evidence of this? Luckily we do. If the evolution of the ancestral genetic codes was in function of the ribosomal proteins, we should find that these proteins were *the first specific proteins* that appeared on the primitive Earth, and the evidence does seem to support this conclusion.

The last common ancestor gave origin to the cells of the three primary kingdoms (Archaea, Bacteria, and Eukarya), and it has been shown that most ribosomal proteins are present in all kingdoms (Woese 2002; Fox 2010). This means that the primary kingdoms received from the last common ancestor not only a universal genetic code, but also a set of universal ribosomal proteins. Which in turn means that the evolution of these proteins had already taken place when the last common ancestor came into being. The molecular trees, on the other hand, do not reveal the existence of older proteins, and this strongly suggests that the ribosomal proteins were indeed the first *specific* proteins that appeared on the primitive Earth.

## The Modern Genetic Code

The modern genetic code is a mapping between 64 codons carried by transfer RNAs and 20 amino acids carried by 20 aminoacyl-tRNA-synthetases, each of which attaches one amino acid to one or more tRNAs. The synthetases are specific proteins that can be produced only by an apparatus

that already has a genetic code, and this gives us a classic *chicken-and-egg* paradox: how could the genetic code come into existence if its rules are implemented by proteins that can be made only when the code already exists?

A possible solution to this paradox is that the *modern* apparatus of protein synthesis was preceded by an *ancient* apparatus where the amino acids were attached to the transfer RNAs not by proteins but by RNAs (Maizels and Weiner 1987). The *modern genetic code*, in other words, was preceded by an *ancient genetic code* based on *RNA synthetases* that were later replaced by *protein synthetases*. Such a replacement, on the other hand, was bound to have major biological consequences. Proteins can mimic RNAs but only up to a point, and replacing the RNA synthetases with protein synthetases could well have modified the rules of the ancient genetic code. But did that actually happen?

Evidence in support of the idea that the ancient genetic code was repeatedly modified has come from computer studies that suggest that the modern genetic code performs better than most of its many potential alternatives (Haig and Hurst 1991; Freeland and Hurst 1998; Bollenbach et al. 2007). Gilis et al. (2001) have shown that the modern code is optimal with respect to the stabilization of protein structure; Itzkovitz and Alon (2007) have argued that the modern code is nearly optimal for the acquisition of additional information into genetic sequences, whereas Drummond and Wilke (2008) have suggested that the modern code is ideally suited to favor the process of protein folding. It must also be reported, however, that some authors have warned against reaching overoptimistic conclusions about this issue. Novozhilov et al. (2007) have pointed out that there are $10^{84}$ possible codes and that a huge number of them are more robust than the modern code. Their computer simulations revealed that the genetic code did go through processes of optimization but apparently went only halfway up the optimality ladder.

Altogether, the computer simulation data leave little doubt that the genetic code was, at least partially, optimized, but it is unlikely that the optimization process was conducted on countless proteins. A more realistic scenario is that it was optimized in a group of 50 or so ribosomal proteins, and *once it was optimized for them it was adopted without further changes for all other proteins*.

It is possible, in conclusion, that the step-by-step introduction of protein synthetases in the apparatus of protein synthesis provided the means for optimizing its performance until the point was reached when the accuracy of protein synthesis became so high as to be virtually error free.

This suggests that the two historical evolutions of the genetic code were both *ribosome oriented*: the evolution from ancestral to ancient genetic code was driven by the ribosomal proteins, whereas the evolution from ancient to modern genetic code was driven by the synthetase proteins (Barbieri 2015).

## The Conservation of the Genetic Code

The modern genetic code appeared on Earth before the first cells of the three primary kingdoms and has been highly conserved ever since (Woese 2000, 2002). Today, this extraordinary process of conservation is usually explained by saying that the genetic code is a set of *constraints* (Pattee 2001) and that physical constraints cannot be changed, an idea that appears to explain why the genetic code has been "frozen" since the origin of life.

The conclusion that the genetic code is a set of constraints is formally correct because a code is indeed a set of rules that impose limitations on a virtually unlimited number of possibilities. It must be underlined, however, that the rules of the genetic code are *biological* constraints, not *physical* ones.

They are *biologically generated* rules and in no way can be assimilated to physical constraints. This is because the genes of the genetic code are constantly subject, like all other genes, to mutation and neutral drift. They are in a continuous state of flux and the fact that they have been highly conserved in evolution means that there is a biological mechanism that *actively and continuously* restores their original structure. The conservation of the genetic code, in other words, is not the passive result of physical constraints. It can only be the result of an active biological mechanism that is continuously at work, a mechanism that has been referred to as *codepoiesis* (Barbieri 2012).

The concept of *autopoiesis*, or self-production, describes the ability of living systems to produce their own components and eventually to generate copies of themselves. Before the genetic code, however, specific proteins did not exist, and the ancestral systems were producing descendants that were inevitably *different* from themselves. Autopoiesis, in short, did not exist before the first cells, so it was not the mechanism that gave origin to them.

The ancestral apparatus of protein synthesis was engaged in the process of evolving coding rules and was therefore a *code-generating system*. After the origin of the genetic code the situation completely changed, and the system in question became a *code-conservation system*. Another part of the living systems, however, maintained the potential to evolve other coding rules and behaved as a new *code-generating,* or *code-exploring, system*. In the early Eukarya, for example, the cells had *a code-conservation part* for the genetic code, but also a *code-exploring part* for the splicing code. The evolution of the first cells, in other words, was based on two complementary processes: one was the *generation* of new organic codes and the other

was the *conservation* of the existing ones. Taken together, these two processes are the two sides of *codepoiesis*.

The ancestral systems, in conclusion, were not autopoietic systems but they had to be codepoietic systems. And all cells that came after them were not always engaged in autopoiesis but were inevitably engaged in codepoiesis. This is the great message of the conservation of the genetic code: what is always and necessarily present in all living systems is codepoiesis, not autopoiesis.

## Conclusions

The history of the genetic code can be divided into four great phases: (1) the origin of the ancestral code, (2) the evolution from ancestral to ancient code, (3) the evolution from ancient to modern code, and (4) the conservation of the modern code. The present article takes as a starting point the hypothesis that the ancestral code was ambiguous and proposes that the second of those four major transitions—the evolution from ancestral to ancient genetic code—took place with a mechanism that gradually reduced the ambiguity of the ancestral code and eventually removed it completely, a mechanism that here is described, at least in first approximation, by the ribosome-oriented model.

The ambiguity-reduction theory and the ribosome-oriented model have implications also for the other evolutionary phases of the genetic code. On the third phase—the evolution from ancient to modern code—the implication is that the synthetase proteins had a driving role similar to that of the ribosomal proteins in the second phase. About the fourth phase—the conservation of the modern genetic code—it is argued that the underlying mechanism is probably far more complex than that of a *frozen accident*. As for the first phase—the origin of the ancestral code—the implication is that the result of that process was an *ambiguous* code.

The ribosome-oriented model describes a mechanism that accounts for a steady reduction of ambiguity in the evolution of the genetic code, a process that ended with the origin of *biological specificity*, the very hallmark of life as we know it. The model is likely to require further developments, but the important point is that ambiguity reduction is no longer an abstract possibility. Perhaps the most significant implication of this article is the fact that the ribosome-oriented model describes a *mechanistic* approach to the problem of the origin of coding (and therefore of *meaning*) in a population of primitive systems based on statistical proteins.

**Acknowledgments** I am indebted to two anonymous referees whose comments greatly improved the initial version of this manuscript.

## References

Barbieri M (2012) Codepoiesis: the deep logic of life. Biosemiotics 5:297–299

Barbieri M (2015) Code biology: a new science of life. Springer, Dordrecht

Boivin A, Vendrely R (1947) Sur le rôle possible deux acides nucleic dans la cellule vivant. Experientia 3:32–34

Bollenbach T, Vetsigian K, Kishony R (2007) Evolution and multilevel optimization of the genetic code. Genome Res 17:401–404

Brachet J (1944) Embriologie chimique. Masson et Cie, Paris

Brachet J (1946) Nucleic acids in the cell and the embryo. Symp Soc Exp Biol 1(213–215):222

Burks AW (1970) Essays on cellular automata. University of Illinois Press, Urbana

Crick FHC (1957) The structure of nucleic acids and their role in protein synthesis. Biochem Soc Symp 14:25–26

Crick FHC (1958) On protein synthesis. Symp Soc Exp Biol 12:138–163

Crick FHC (1966) Codon-anticodon pairing: the wobble hypothesis. J Mol Biol 19:548–555

Di Giulio M (2008) An extension of the coevolution theory of the origin of the genetic code. Biol Direct 3:1–37

Dounce AL (1952) Duplicating mechanism for peptide chain and nucleic acid synthesis. Enzymologia 15:251–258

Dounce AL (1953) Nucleic acid template hypothesis. Nature 172:541

Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352

Dunnill P (1966) Triplet nucleotide-amino-acid pairing; a stereochemical basis for the division between protein and non-protein amino-acids. Nature 210:1267–1268

Fitch WM, Upper K (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harb Symp Quant Biol 52:759–767

Fox GE (2010) Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol 2:a003483

Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248

Gamow G (1954) Possible relation between deoxyribonucleic acid and protein structures. Nature 173:318

Gilis D, Massar S, Cerf NJ, Rooman M (2001) Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol 2:41–49

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Mol Evol 33:412–417

Higgs PG, Pudritz RE (2007) From protoplanetary disks to prebiotic amino acids and the origin of the genetic code. In: Pudritz RE, Higgs PG, Stone J (eds) Planetary systems and the origins of life, vol 3., Cambridge series in astrobiology. Cambridge University Press, Cambridge, pp 62–88

Hinegardner RT, Engelberg J (1963) Rationale for a universal genetic code. Science 142:1083–1085

Hoagland MB, Zamecnik PC, Stephenson ML (1957) Intermediate reactions in protein biosynthesis. Biochem Biophys Acta 24:215–216

Hou Y-M, Schimmel P (1988) A simple structural feature is a major determinant of the identity of a transfer RNA. Nature 333:140–145

Itzkovitz S, Alon U (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res 17:405–412

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356

Jukes TH, Osawa S (1990) The genetic code in mitochondria and chloroplasts. Experientia 46:1149–1157
Jukes TH, Osawa S (1993) Evolutionary changes in the genetic code. Comp Biochem Physiol 106 B:489–494
Khorana HG, Büchi H, Ghosh H et al (1966) Polynucleotide synthesis and the genetic code. Cold Spring Harb Symp Quant Biol 31:39–49
Koonin EV, Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61(2):99–111
Kurland CG (1970) Ribosome structure and function emergent. Science 169:1171–1177
Maizels N, Weiner AM (1987) Peptide-specific ribosomes, genomic tags and the origin of the genetic code. Cold Spring Harb Symp Quant Biol 52:743–757
Melcher G (1974) Stereospecificity and the genetic code. J Mol Evol 3:121–141
Monod J (1970) Le Hasard et la Necéssité. Editions du Seuil, Paris. English edition: Monod J (1971) Chance and necessity (trans: Wainhouse A). Knopf, New York
Ninio J (1982) Molecular approaches to evolution. Pitman Books, London
Niremberg M, Leder P (1964) RNA codewords and protein synthesis. Science 145:1399–1407
Niremberg M, Matthaei H (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci USA 47:1588–1602
Niremberg M, Caskey T, Marshal R et al (1966) The RNA code and protein synthesis. Cold Spring Harb Symp Quant Biol 31:11–24
Nishimura S, Jones DS, Khorana HG (1965) The in vitro synthesis of a co-polypeptide containing two amino acids in alternating sequence dependent upon a DNA-like polymer containing two nucleotides in alternating sequence. J Mol Biol 13:302–324
Nitta I, Kamada Y, Noda H et al (1998) Reconstitution of peptide bond formation. Science 281:666–669
Nomura M, Tissières A, Lengyel P (1974) Ribosomes, Cold Spring Harbor monograph series. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
Novozhilov AS, Wolf YI, Koonin EV (2007) Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biol Direct 2:24
Orgel LE (1973) The origins of life. Wiley, New York
Osawa S (1995) Evolution of the genetic code. Oxford University Press, New York
Pattee HH (2001) The physics of symbols: bridging the epistemic cut. BioSystems 60:5–21
Pelc SR, Weldon MGE (1966) Stereochemical relationship between coding triplets and amino-acids. Nature 209:868–870
Roberts RB (1958) Microsomal particles and protein synthesis. Pergamon Press, Washington
Saxinger WC, Ponnamperuma C, Woese CR (1971) Evidence for the interaction of nucleotides with immobilized amino-acids and its significance for the origin of the genetic code. Nat New Biol 234:172–174
Schimmel P (1987) Aminoacyl tRNA synthetases: general scheme of structure-function relationship in the polypeptides and recognition of tRNAs. Ann Rev Biochem 56:125–158
Schimmel P, Giegé R, Moras D, Yokoyama S (1993) An operational RNA code for amino acids and possible relationship to genetic code. Proc Natl Acad Sci USA 90:8763–8768
Shimizu M (1982) Molecular basis for the genetic code. J Mol Evol 18:297–303
Sonneborn TM (1965) Degeneracy of the genetic code: extent, nature, and genetic implications. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 377–397
Speyer J, Lengyel P, Basilio C et al (1963) Synthetic polynucleotides and the amino acid code. Cold Spring Harb Symp Quant Biol 28:559–567
Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code. Proc Natl Acad Sci USA 103:10696–10701
Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. Nature 171:737–738. Genetical implications of the structure of deoxyribose nucleic acid. Nature 171:964-967
Woese CR (1965) Order in the genetic code. Proc Natl Acad Sci USA 54:71–75
Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271
Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97:8392–8396
Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci USA 99:8742–8747
Woese CR, Hinegardner RT, Engelberg J (1964) Universality in the genetic code. Science 144:1030–1031
Wolf YI, Koonin EV (2007) On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. Biol Direct 2:14
Wong JT (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912
Wong JT (1981) Coevolution of genetic code and amino acid biosynthesis. Trends Biochem Sci 6:33–36
Wong JT, Bronskill PM (1979) Inadequacy of prebiotic synthesis as origin of proteinous amino acids. J Mol Evol 13:115–125
Yarus M (1988) A specific amino acid binding site composed of RNA. Science 240:1751–1758
Yarus M (1998) Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. J Mol Evol 47:109–117
Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: the escaped triplet theory. Ann Rev Biochem 74:179–198