Biol Theory (2016) 11:102–112
DOI 10.1007/s13752-016-0236-4

ORIGINAL ARTICLE

From the Common Ancestor to the First Cells: The Code Theory

Marcello Barbieri[1]

Received: 30 November 2015 / Accepted: 7 March 2016 / Published online: 7 April 2016

Abstract The phylogenetic trees reconstructed from molecular data have led to the discovery that all living creatures belong to three primary kingdoms, or domains, because there are three types of cells in nature. The primary kingdoms are referred to as Archaea, Bacteria, and Eukaya, and their first representatives were the first modern cells that appeared on Earth. All known cells, on the other hand, contain a virtually universal genetic code, and this implies that the code evolved in a population of primitive systems that preceded the first modern cells and is collectively known as the *common ancestor* of all life. This gives us the problem of understanding how the descendants of the common ancestor gave origin to the first modern cells. In this article it is argued that the appearance of the genetic code allowed the ancestral systems to translate genes into specific proteins, but their behavior was still ambiguous because they were unable to produce specific responses to the signals from the environment. To that purpose they needed to evolve *signal processing codes,* and here it is proposed that the development of these codes was a crucial step in the evolution of the first modern cells.

Keywords Cell membrane · Common ancestor · Genetic code · Phylogenetic trees · Primary kingdoms · Signal integration codes · Signal transduction codes

✉ Marcello Barbieri
brr@unife.it

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

Defining the Problem

In *Origin of Species*, Darwin (1859) described evolution as a process that started "*from simple beginnings*" and gave origin to increasingly diverse "*most beautiful forms,*" much like a tree that grows and divides into countless branches. At the base of the tree there are the first "*primordial forms*" and out of them grows a trunk that splits again and again to create an ever-expanding tree. Each branch represents a species, and the branching points are where a species splits in two. Most branches come to a dead end, signifying extinction, but some go all the way up to the top and represent today's organisms. This is the *tree of life*, the graphic description of the relationships that link together all creatures of the present to all those of the past.

The reconstruction of the tree of life has been a holy grail for generations of naturalists, and traditionally it has been conducted with the methods of comparative anatomy combined with the results of paleontology. Ever since the pioneering work of Zuckerkandl and Pauling (1965), however, it has become increasingly clear that the sequences of genes and proteins provide an additional source of information. The protein citochrome-c, for example, has a sequence of amino acids that varies among species in a way that appears to be related to the evolutionary distance that separates them. Between humans and monkeys, for example, the differences are small, and so are those between ducks and pigeons, but between humans and ducks or between monkeys and pigeons they are significantly greater. The numerical values of these differences can be used to build a diagram, and what comes out is remarkably similar to the genealogical tree obtained with traditional methods, thus confirming that molecular sequences contain phylogenetic information. What is particularly important is that the molecular method can be

applied to single cells, and this enormously extends the dimensions of the tree.

It must be underlined, however, that the reconstruction of the tree of life takes place in a theoretical framework that is based on three major biological discoveries.

The greatest generalization of biology is the cell theory, the idea that all living organisms are made of cells and that cells derive from preexisting cells. This implies that all cells of the present are linked to all cells of the past by an uninterrupted chain of descent that goes all the way back to the first cells that appeared on the primitive Earth.

The greatest discovery of paleontology is that our planet has been inhabited exclusively by free-living cells, or microorganisms, for the first three billion years of evolution. For more than 80 % of the history of life, in other words, the microorganisms have been the sole living creatures on Earth and the sole protagonists of evolution.

The greatest discovery of molecular biology is that all known cells contain a virtually universal genetic code, a fact that implies that the code evolved in a population of primitive systems that preceded the first cells and that is collectively known as the *common ancestor* of all life.

We have therefore three distinct problems before us: (1) What do we know about the common ancestor? (2) What were the characteristics of the first cells? And, most importantly, (3) how did the common ancestor give origin to the first cells?

Three Primary Kingdoms

The greatest divide of the living world is not between plants and animals, as has been thought for centuries, but between cells without a nucleus (*prokaryotes*) and nucleated cells (*eukaryotes*).

Prokaryotes, or *bacteria*, have a single DNA molecule, a single cytoplasmic compartment, and the form of the cell is due either to a rigid external wall around the cell membrane or to a rigid cell membrane.

Eukaryotes have various DNA molecules that are repeatedly folded into highly organized chromosomes, a cytoplasm divided in compartments, a variety of organelles (mitochondria, chloroplasts, lysosomes, Golgi, endoplasmic reticulum, etc.), and the form of the cell is not due to a surrounding wall but to an internal cytoskeleton made of three types of filaments (microtubules, microfilaments, and intermediate filaments).

In 1866, Haeckel proposed a phylogenetic tree where the first forms of life were cells without a nucleus (which he called *monera*), which later generated nucleated cells (*protista*) that in turn gave rise to all multicellular organisms. In 1883, Schimper proposed that the chloroplasts in the plant cells had once been free-living bacteria that became incorporated, by a kind of internalization, or *endosymbiosis*, into other cells; later on, Mereschowsky (1910), Portier (1918), and Wallin (1927) also proposed this hypothesis for the origin of mitochondria.

The *endosymbiosis hypothesis* was ignored for decades but in 1970 it was forcefully re-proposed by Lynn Margulis, and within a few years it received the support of an astonishing amount of experimental data. It was found that mitochondria and chloroplasts are still carrying fragments of their ancient bacterial DNA, and have 70S ribosomes which are typical of bacteria, all of which leaves little doubt about their origin.

Today it is universally acknowledged that mitochondria and chloroplasts were acquired by symbiosis, but that tells us nothing about the cells that engulfed them, and on this issue biologists have been divided into opposing camps. Some, like Margulis (1970), maintained that the cells that engulfed bacteria were themselves bacteria, whereas others insisted that they must have been nonbacterial cells.

A major turning point in the reconstruction of cell history came in 1977, when Carl Woese and George Fox discovered that the phylogenetic tree obtained from ribosomal RNAs divides all living creatures not in two but in three groups: two different types of prokaryotes that Woese and Fox (1977) called *archaebacteria* and *eubacteria*, and a third group containing the ancestors of the eukaryotic cytoplasm that they called *urkaryotes*.

This discovery has two outstanding implications:

1. Bacteria do not form a monophyletic group but two distinct kingdoms (archaebacteria and eubacteria).
2. The phylogenetic distance between the two bacterial kingdoms is comparable to the distance that separates any of them from the third kingdom of the urkaryotes, which means that all three groups of cells evolved *independently* from the common ancestor.

Later on, Woese renamed the three groups and proposed that all cells belong to three distinct *primary kingdoms*, or *domains*, that were called *Archaea*, *Bacteria* and *Eucarya* (Woese 1987, 2000; Woese et al. 1990).

This proposal was strongly opposed by Mayr (1998) who argued that from a morphological and physiological point of view archaebacteria and eubacteria are undoubtedly prokaryotes, and their molecular differences cannot be enough to classify them into two distinct kingdoms. Woese (1998) replied that the history of cellular life can only be reconstructed from molecular data, and these data tell us that there have been three distinct types of ancestral cells, not two. The ribosomal RNAs, furthermore, are universal molecules, which means that the tree reconstructed from them embraces all past and present creatures and is therefore a truly *universal* phylogenetic tree (Woese 2000).

Blurring the Tree of Life

The universal phylogenetic tree was first reconstructed from ribosomal RNAs, but in principle it should also be recovered from proteins because they too should carry the signs of what happened in the history of life. When the techniques of molecular phylogeny were applied to proteins, however, the results turned out to be much more complex than expected and provided contrasting phylogenetic information. Some proteins (for example ATPases, RNA polymerases, and ribosomal proteins) confirmed the three domains obtained from the ribosomal RNAs, but other proteins (in particular many enzymes of the metabolic pathways) led to different phylogenetic trees (Brown and Doolittle 1997). The crucial point is that the discordant protein trees do not represent real alternatives because they disagree not only with the RNA tree but also with each other.

The solution to this mystery came from the discovery, in the 1990s, that bacteria routinely swap genetic material in a process called *horizontal gene transfer* (Miller 1998).

The pattern of a tree is realized when genes are transmitted from one generation to the next, i.e., when descent is *vertical*. When genes instead are swapped *horizontally* in every generation they become part of many branches simultaneously, and the resulting pattern is no longer a tree but *a web* (Doolittle 1999; Doolittle and Bapteste 2007).

Most importantly, it turned out that horizontal gene transfer is by no means a secondary process. In prokaryotes it can account for as much as 80 % of the genes (Dagan et al. 2008), and this immediately suggests that in the early stages of evolution most genes were transferred horizontally and the dominant pattern was that of a vastly interconnected web. A treelike pattern probably began to emerge only later, when vertical descent managed to contrast the leveling effects of horizontal gene transfer.

These discoveries have shown that genes are transmitted both vertically and horizontally, but the key point is that they are not equally affected by the two types of transmission. Some are frequently involved in horizontal transfer whereas others are almost exclusively transmitted by vertical descent, and a few of them have been highly conserved in evolution. It is these conserved molecules that maintain a record of what happened in the past and allow us to reconstruct the universal phylogenetic tree.

It is true, therefore, that the tree of life has been heavily *blurred* by horizontal gene transfer, but it is also true that the highly conserved molecules still document its existence and confirm its subdivision into three cellular domains. This conclusion was originally obtained from highly conserved *individual* molecules, but later it was also confirmed by higher-order trees that were reconstructed from whole genomes (Simonson et al. 2005; Snel et al. 2005; Jun et al. 2010). Some differences do remain between the trees of genome phylogeny and those of molecular phylogeny, but both methods converged to the same overall conclusion that the universal phylogenetic tree exists and is split into three primary kingdoms (Harold 2014).

All three kingdoms received the genetic code from the common ancestor, and for this reason that ancestor represents the root of the universal tree of life. But what do we actually know about that distant progenitor?

Ancestral, Ancient, and Modern Genetic Code

The genetic code is an integral part of the apparatus of protein synthesis, and this implies that the common ancestor is the population of primitive systems that evolved not only the genetic code but the entire apparatus of protein synthesis. This extended definition is not only more general but also more useful, because it allows us to divide the evolution of the common ancestors into a sequence of logical steps.

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 when a genetic code already exists, and this implies 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 specific proteins, that did not yet exist, but probably by RNAs (Maizels and Weiner 1987).

The modern genetic code, in other words, was preceded by an *ancient genetic code*, probably based on *RNA synthetases* that later became replaced by *protein synthetases*. The ancient genetic code, in turn, was the result of a previous round of evolution that started when the very first genetic code appeared on the primitive Earth. But what can we say about that first code?

The ribosomal RNAs are among the most conserved molecules in evolution (Woese 1987, 2000), and this means that they appeared very early in the history of life. 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 groups of amino acids to maintain stable conformations and their functions are greatly enhanced by the attachment of peptides and 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 deliver amino acids to the ribosomal RNAs. 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 are 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*.

This amounts to saying that the first 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*.

We come in this way to the conclusion that there have been three distinct genetic codes in the evolution of the common ancestor: the ancestral code, the ancient code, and the modern code (Barbieri 2015).

Woese's Theory on the Origin of the First Cells

The population of primitive systems that last appeared at the root of the universal phylogenetic tree is usually referred to as the *last common ancestor*. After that population, the tree split into three great branches, and the descendants of the last common ancestor independently gave origin to the first modern cells, i.e., to the first Bacteria, the first Archaea, and the first Eukarya. But how did they do it? Carl Woese was the first who addressed this problem and proposed that horizontal gene transfer has been the major driving force in the evolution of the ancestral systems. Let's use Woese's own words to illustrate his theory.

> The universal phylogenetic tree based on ribosomal RNAs is unlike any other phylogenetic tree because it transcends the era of modern cells. Its deepest branches extend back in time to an era when cellular entities were considerably more primitive than cells are today…. When cells are simple enough, horizontal gene transfer is the major, if not the sole, evolutionary source of true innovation and all life becomes a single diverse gene pool…. At such stage evolution was in effect communal: there was a progressive evolution of the whole, not an evolution of individual lineages…. It is only in this way that the radical novelty needed to boot-strap primitive cellular entities into modern cells can occur…. In this process, a stage inevitably will be reached when some cellular entities become complex enough that their cell design starts to become unique. (Woese 2000, p. 8395)

> Biologists have assumed that the "organism" represented by the root of the universal tree was equivalent to a modern cell, in effect it was a modern cell. That is not a scientifically acceptable assumption…. There is evidence, good evidence, to suggest that the basic organization of the cell had not yet completed its evolution at the stage represented by the root of the universal tree…. Because of their loose construction, primitive cells initially did not have stable genealogical records…. Individual lineages (species) emerged from this common ancestral chaos only when cellular organization achieved a certain degree of complexity and connectedness…. As a cell design becomes more complex, a critical point is reached where a more integrated cellular organization emerges…. This critical point is called the "Darwinian Threshold." (Woese 2002, p. 8742)

> Early evolution was dominated by horizontal gene transfer and led to the emergence of modern cell designs from a communal state, not a unique ancestor. Such a communal state existed before the point of emergence of vertical evolution, which has been termed the "Darwinian transition." The defining property of the communal state was that it was capable of tolerating and using ambiguity, as reflected in the pervasive role of horizontal gene transfer. A Darwinian transition corresponds to a state of affairs when sufficient complexity has arisen that the state is incapable of tolerating ambiguity and so there is a distinct change in the nature of the evolutionary dynamics (to vertical descent). We envision that such Darwinian transitions occurred in each of the three major lineages. (Vetsigian et al. 2006, p. 10697)

Let us summarize. The evolution of the genetic code took place in the populations of the common ancestor but did not produce a modern cell design. Carl Woese has convincingly argued that the last common ancestor was still a premodern system and its descendants had to go through other rounds of evolution before they could acquire a modern cell organization. But what did they

actually have to do to become fully modern cells? A possible answer is that *they had to evolve other organic codes,* because the genetic code was not enough to create a modern cell. But is this a realistic idea? Do we have any evidence that the cell contains other codes in addition to the genetic code?

The Presence of Codes in the Cell

A code is *a set of rules that establish a mapping between the objects of two independent worlds*. The Morse code, for example, is a mapping between the letters of the alphabet and groups of dots and dashes. The highway code is a mapping between street signals and driving behaviors, and so on. What is essential in any code is that the coding rules are not dictated by the laws of physics and chemistry. In this sense they are *arbitrary*, and the number of arbitrary relationships between two independent worlds is potentially unlimited. In Morse code, for example, any letter of the alphabet can be associated with countless combinations of dots and dashes, which means that a specific mapping can be realized only by selecting a small number of rules. And this is precisely what a code is*: a small set of arbitrary rules selected from a potentially unlimited number in order to ensure a specific mapping between two independent worlds*.

Organic codes are codes between two worlds of organic molecules and are necessarily implemented by a third type of molecule, called *adaptors*, that builds a bridge between them. The adaptors are required because there is no necessary link between the two worlds, and a fixed set of adaptors is required in order to guarantee the *specificity* of the mapping. The adaptors, in short, are the molecular *fingerprints* of the codes, and we can prove that an organic code exists if we have three things: (1) two independent worlds of molecules connected by adaptors, (2) a potentially unlimited number of arbitrary connections between them, and (3) a selection of the adaptors (a set of coding rules) that ensures a specific mapping (Barbieri 2003).

In the case of the genetic code, the very first hypothesis was the *stereochemical theory*, a model, first proposed by Gamow (1954) and later re-proposed by many other authors, that states that the relationships between codons and amino acids (the coding rules) are determined by stereochemical affinities. This theory automatically implies that the genetic code *is not a real code* because its rules are the inevitable result of chemical processes and do not have the arbitrariness that is essential in any code.

It took a long time and much experimental work to overturn this conclusion. Eventually, however, it was shown that there are no deterministic links between codons and amino acids since any codon can be associated, in principle, to any amino acid (Schimmel 1987; Schimmel et al. 1993). Hou and Schimmel (1988), for example, introduced two extra nucleotides in a tRNA and found that that the resulting tRNA was carrying a different amino acid. This proved that the number of possible connections between codons and amino acids is potentially unlimited, and only the selection of a small set of adaptors can ensure a specific mapping. This is *the genetic code*: a fixed set of rules of correspondence between codons and amino acids that are implemented by adaptors. In protein synthesis, in other words, we find all the three essential components of a code: (1) two independent worlds of molecules (nucleotides and amino acids) connected by adaptors, (2) the proof that the mapping is arbitrary because its rules can be changed virtually at will, and (3) the proof that only a small fixed number of rules has been selected.

The evidence provided by Carl Woese, on the other hand, has shown that the genetic code was not enough to create a modern cell, and we need therefore to find out what else was necessary. To this purpose, let us recall that the cell is a system that is largely composed of subsystems, or *modules*, each of which is specialized in a particular function and has a substantial degree of autonomy (Schlosser and Wagner 2003; Callebaut and Rasskin-Gutman 2005).

A typical example is the apparatus of protein synthesis, but there are at least two other subsystems that play essential roles in all cells, and both of them, as we will see, are based on organic codes. One is the *signal transduction module*, the set of components that receive signals from the environment and transform them into internal signals. The other is the *signal integration module*, the subsystem that combines all internal signals together and delivers the result to the genome.

The Signal Transduction Codes

Living cells react to many physical and chemical stimuli from the environment, and in general their reactions consist in the expression of specific genes. We need therefore to understand how the environment interacts with the genes, and the turning point, in this field, came from the discovery that the external signals (known as *first messengers*) never reach the genes. They are invariably transformed into a different world of internal signals (called *second messengers*) and only these, or their derivatives, reach the genes. In most cases, the molecules of the external signals do not even enter the cell and are captured by specific receptors of the cell membrane, but even those that do enter (some hormones) must interact with intracellular receptors in order to influence the genes (Sutherland 1972).

The transfer of information from environment to genes takes place therefore in two distinct steps: one from first to second messengers, called *signal transduction*; and a

second path from second messengers to genes which is known as *signal integration*. One of the surprising things about signal transduction is that there are literally hundreds of first messengers (ions, nutrients, hormones, growth factors, neurotransmitters, etc.) but only a limited number of second messengers: cyclic AMP or GMP, calcium ions (Ca^{2+}), inositol trisphosphate (IP3), diacylglycerol (DAG), and a few other molecules.

First and second messengers represent two different molecular worlds, and this immediately suggests that signal transduction may be based on organic codes. This is confirmed by the discovery that there is no necessary connection between first and second messengers, because it has been proven that the same first messengers can activate different types of second messengers, and that different first messengers can act on the same type of second messengers (Alberts et al. 2007). The most plausible explanation is that signal transduction is based on organic codes, but of course one would like a direct proof.

The signature of an organic code, as we have seen, is the presence of adaptors and the *transmembrane receptor proteins* of signal transduction do have the defining characteristics of the adaptors. The transduction system consists of at least three types of molecules: a *receptor* for the first messengers, an *amplifier* for the second messengers, and a *mediator* in between (Berridge 1985). This transmembrane system performs two independent recognition processes, one for the first and the other for the second messenger, and the two steps are connected by the bridge of the mediator. This connection, on the other hand, could be implemented in countless different ways since any first messenger can be coupled with any second messenger, and this makes it imperative to have a selection in order to guarantee biological specificity.

In signal transduction, in short, we find the three defining features of a code: (1) two independent worlds of objects (first messengers and second messengers) connected by adaptors, (2) a potentially unlimited number of arbitrary connections between them, and (3) a set of coding rules (a selection of the adaptors) that ensures the specificity of the correspondence. The effects that external signals have on cells, in short, do not depend on the energy or the information that they carry, but on the *meaning* that cells give them with sets of rules that have been referred to as *signal transduction codes* (Barbieri 2003).

The Signal Integration Codes

We have seen that there are only a few families of second messengers in the cell, and yet the reactions that they set in motion can pick up an individual gene among tens of thousands. How this is achieved is still a mystery, but some progress has been made. Perhaps the most illuminating discovery, so far, is that second messengers do not act independently. Calcium ions and cyclic AMPs, for example, have effects that reinforce each other on some occasions but are mutually exclusive at other times (Alberts et al. 2007). The cell, in short, can combine its internal signals in countless different ways, and it is precisely this combinatorial ability that explains why a small number of second messengers can generate an extraordinarily high number of specific genetic responses. The activation of second messengers, in other words, sets in motion a cascade of reactions that normally ends with the expression of a target gene, and again we would like to find out if at least some of them are based on the rules of a code.

One of the most interesting clues, in this field, is the fact that signaling molecules have in general more than one function. Epidermal growth factor, for example, stimulates the proliferation of fibroblasts and keratinocytes, but it has an anti-proliferative effect on hair follicle cells, whereas in the intestine it is a suppressor of gastric acid secretion. Other findings have proved that *all* growth factors can have three distinct functions, with proliferative, anti-proliferative, and proliferation-independent effects. They are, in short, *multifunctional molecules* (Sporn and Roberts 1988).

In addition to growth factors, it has been found that many other molecules have multiple functions. *Adrenaline*, for example, is a neurotransmitter, but it is also a hormone produced by the adrenal glands to spring the body into action by increasing the blood pressure, speeding up the heart, and releasing glucose from the liver. *Acetylcholine* is another common neurotransmitter in the brain, but it also acts on the heart (where it induces relaxation), on skeletal muscles (where the result is contraction), and in the pancreas (which is made to secrete enzymes). *Cholecystokinin* is a peptide that acts as a hormone in the intestine, where it increases the bile flow during digestion, whereas in the nervous system it is a neurotransmitter. *Encephalins* are sedatives in the brain, but in the digestive system are hormones that control the mechanical movements of food. *Insulin* is universally known for lowering the sugar levels in the blood, but it also controls fat metabolism, and in other less-known ways it affects almost every cell of the body.

The discovery of multifunctional molecules means that their function is not decided solely by their structure but also by the *context* in which they find themselves. What matters, in other words, is not their ability to catalyze a specific reaction, but the fact that they are employed as *molecular labels* that can be given one meaning in a certain context and a different meaning in another one.

The signals of the first messengers, in conclusion, undergo two great transformations in the cell. First they are transformed into internal messengers with the rules of the

signal transduction codes, and then these messengers are combined together according to the rules of the *signal integration codes* (Barbieri 2003).

Combinatorial Codes

The discovery of the genetic code has been facilitated by two particularly favorable features: (1) by the fact that the adaptors (the tRNAs) are single molecules and (2) by the fact that the coding elements form a closed set (64 codons and 20 amino acids). In the case of signal transduction and signal integration the situation is different because the adaptors are often combinations of molecules and the domain of the process is open and potentially unlimited. This probably explains why signal transduction and signal integration are not usually referred to as codes, but this terminological habit should not obscure the central issue. A set of relationships is a code when it is made of *arbitrary* rules, even when its domain (or *alphabet*) is open and when its components are combinatorial sets of molecules (*combinatorial codes*). Such cases have already been described in the literature, and it may be worth taking a brief look at some of them.

The Histone Code

In eukaryotes, the DNA filament is wrapped around groups of histone proteins whose tails are subject to a variety of post-translational modifications (in particular acetylation, methylation, and phosphorylation) that have highly dynamic roles and are involved in the activation or repression of gene activity (Kornberg and Lorch 1999; Wu and Grunstein 2000). A crucial breakthrough in this field was the discovery that the post-translational modifications of the histones do not act individually. Most of them are involved in both the activation and the repression of genes (the phosphorylation of histone H3, for example, takes part in the condensation as well as in the decondensation of chromatin), which means that the final result is due to a *combination* of histone marks rather than a single one. This led David Allis and colleagues to propose that the histone marks operate in combinatorial groups, like letters that are put together into the words of a molecular "language" that was referred to as the *histone code* (Strahl and Allis 2000; Jenuwein and Allis 2001). The same concept was independently proposed by Turner (2000, 2002, 2007) who argued that there is an *epigenetic code* at the heart of the regulation mechanisms that are initiated by histone tail modifications. Today, in conclusion, a large number of data support the idea that the regulation of gene activity by histone modifications is based on the rules of a combinatorial code (Berger 2007).

The Splicing Codes

The primary transcripts of the genes are often transformed into messenger RNAs by removing some RNA pieces (called *introns*) and by joining together the remaining pieces (the *exons)*. This cutting-and-sealing operation, known as *splicing*, is carried out by molecular structures that act like adaptors because they perform two independent recognition processes, one for the beginning and one for the end of each exon, thus creating a specific correspondence between primary transcripts and messenger RNAs. Splicing, in other words, is a codified process based on adaptors and takes place with sets of rules that have been referred to as *splicing codes* (Barbieri 2003; Fu 2004; Matlin et al. 2005; Wang and Burge 2008). Splicing, on the other hand, is complicated by the fact that many introns carry sequences that are similar to exons but translate into nonsense and for this reason are called *pseudo exons* or *pseudo genes*. They would create havoc if incorporated into mRNAs and the splicing machinery had to evolve the means to differentiate real exons from pseudo ones. The result is that real exons contain internal identity marks that are known as *exonic splicing enhancers* (ESEs) and *exonic splicing silencers* (ESSs) (Fu 2004; Matlin et al. 2005; Pertea et al. 2007). The presence of these marks, in turn, means that the adaptors of the splicing codes are not single molecules but combinations of molecules, because they must be able to recognize not only the beginning and the end of the real exons, but also their internal identity marks.

It has been shown that a large variety of transcriptional codes are combinatorial codes (Jessell 2000; Marquard and Pfaff 2001; Flames et al. 2007; Osborne et al. 2008; Allan and Thor 2015), and combinatorial rules have also been found in signal processing (Marijuán et al. 2015). We realize in this way that the signal transduction and the signal integration codes belong to a highly heterogeneous family of combinatorial codes with open alphabets.

The Cell Membrane

All living creatures belong to three primary kingdoms, because there are three types of cells in nature that differ in a variety of key components such as cell membranes, cell walls, energy sources, and organelles of movement (Woese and Fox 1977; Harold 2014).

In Bacteria, for example, the cell membrane contains phospholipids, whereas in Archaea it contains isoprenoid lipids. In Bacteria the cell wall is made of peptidoglycans, whereas in Archaea it is made of proteinaceous material. Bacteria move by *flagella,* and these organelles obtain energy from the circulation of protons or sodium ions, whereas Archaea move by totally different organelles that

obtain energy from ATP. Eukarya too have unique features that separate them from the two other types of cells, but on top of that they have some characteristics in common with Bacteria and others in common with Archaea, probably as a result of horizontal gene transfer.

Of all the above cellular components, the cell membrane has attracted special attention because it has the remarkable property of never being constructed de novo. Membranes always grow from preexisting membranes, and this has led to the concept of *membrane heredity*, the idea that membranes are passed down from one generation to the next in an uninterrupted chain of descent (Blobel 1980; Sapp 1987; Cavalier-Smith 2000; Harold 2005).

Genes do not transmit three-dimensional information, and the supramolecular structures of the cell are produced either by self-assembly from their components or by growth from preexisting structures. Chromosomes are produced from preexisting chromosomes and membranes from preexisting membranes, but they carry two very different types of information. Chromosomes transmit genetic instructions, whereas membranes transmit architectural order.

The universal phylogenetic tree tells us that the first modern cells were the first representatives of the three primary kingdoms, and these cells had membranes with modern characteristics that allow us to recognize them as Archaea, Bacteria, and Eukarya. This means that the evolution of the modern cell membranes took place in the descendants of the last common ancestor, but how did that contribute to the origin of the first modern cells? What was the role of the cell membrane in the origin of the modern cell organization?

To this purpose, let us let us keep in mind that the cell membrane is the seat of three distinct processes: (1) it is the site where molecules are transported to and from the environment (*molecular transport*), (2) it is the site where energy is obtained from external sources and converted into internal forms (*energy transduction*), and (3) it is the site where signals are received from the outside world and used to produce internal signals that allow the cell to mount a response reaction.

The first two processes—the exchange of molecules and the access to energy—are so fundamental that we can hardly imagine a common ancestor without them, and this is why it is very likely that the ancestral populations did have some kind of cell membranes for those processes. As for the third process, however, the situation is different. Signal transduction requires transmembrane receptors that create bridges between first and second messengers, and such receptors are proteins that could be produced only after the origin of the genetic code, when the ancestral systems became capable of synthesizing specific proteins.

A signal transduction code, in other words, could hardly be present at the time of the last common ancestor but had to be present in the first modern cells. This suggests a potential solution to our problem: it is possible that the cell membrane had an essential role in the origin of the first modern cells, because it provided the structure where the evolution of the signal transduction code could take place.

The Code Theory

The genetic code is a mapping between 64 codons and 20 amino acids and is therefore a many-to-one code. 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. This is expressed by saying that the genetic code is *degenerate* (or *redundant*) but it is important to underline that it is *not ambiguous* because *any codon codes for one and only one amino acid*.

It has been pointed out, on the other hand, that the first genetic code on Earth was necessarily *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). In that case, a sequence of codons was translated at times into one protein and at other times into a different protein, and the apparatus of protein synthesis was inevitably producing *statistical proteins* (Woese 1965). The evolution of the ancestral genetic code was therefore a process that steadily reduced and finally eliminated the ambiguity of the coding rules (Barbieri 2015). When that happened, it became possible to translate genes into *specific proteins* and life as we know it—life based on *biological specificity*—came into existence.

The origin of a nonambiguous genetic code was a major turning point in the history of life, because it left behind the old world of statistical proteins and set in motion the new world of specific proteins, and yet that major transition was not enough to create a modern cell. The reason is that the descendants of the last common ancestor could produce specific proteins, but without a signal transduction code they could not produce *specific responses to the environment*. They had biological specificity in protein synthesis, but not in their interactions with the world, and it is for this reason that they had not yet become modern cells.

A response to the environment, furthermore, required not only a signal transduction code but also a signal integration code, because it is the combination of these codes, collectively referred to as *signal processing code*, that sets in motion a response reaction to the world.

This is the *code theory of the evolution of the first cells*, the idea that the first living systems with a modern cell organization came into existence when they acquired a

signal processing code that allowed them to mount specific reactions to the signals from the environment.

As in the case of the genetic code, we cannot expect that the rules of the signal processing code appeared all at once; and this inevitably implies that that code was initially ambiguous, and its evolution consisted in a steady reduction of its ambiguity. The appearance of the first modern cells, in other words, can be attributed to the appearance of the first *nonambiguous* signal processing code, because it is only this code that gives *specificity* to the behavior of the cell.

As the genetic code marked the transition from statistical to specific proteins, the signal processing code marked the transition from statistical to specific cell behaviors, and was therefore an equally foundational event in macroevolution. It was the event that allowed the descendants of the last common ancestor to cross what Woese called the *Darwinian threshold*, and give origin to the first modern cells.

Conclusions

The signal transduction system of prokaryotes has long been regarded as a two-components system made of a membrane-bound protein and a soluble cytosolic protein. Each of these proteins, in turn, consists of two domains: the membrane-bound protein contains an "*input domain*" that receives signals from the environment and a "*phosphorylating domain*" that releases a phosphoryl group when a signal hits the protein; the cytosolic protein contains a "*receiver domain*" for the phosphoryl group and a "*response domain*" that sets in motion a specific cell reaction.

In 2005, however, came the discovery that the signal transduction system of many prokaryotes is a one-component system where the input domain and the response domain are no longer located on two different proteins but on a single one (Ulrich et al. 2005). In the same paper, Ulrich and colleagues pointed out that the one-component systems are probably evolutionary precursors of the two-components ones:

> the modular design of one-component systems is obviously simpler than that of two-components systems…. Therefore it is possible that the last common ancestor of archaea and bacteria (i.e., the last common ancestor of all modern life forms) did not have two-component systems, but encoded several one-component regulators. Two-component systems appear to be a subsequent bacterial innovation that emerged as a result of insertion of histidine kinase domains and receiver domains into one-component regulators. (Ulrich et al. p. 55)

This is probably the first model that has been proposed on the evolution of signal transduction, but it must be pointed out that even a one-component transduction system could not have sprung fully formed into existence. It must have been the result of a prior evolutionary process, and we have therefore the problem of finding out, at least in principle, how that ancestral evolution took place. This is the problem addressed by the code theory. The basic goal of the apparatus of protein synthesis was to produce *specific proteins* and was achieved by the evolution of the genetic code. In a similar way, the basic goal of the signal transduction system was to allow the cell to mount *specific reactions* to the environments and was achieved by the evolution of signal processing codes.

For a long time it has been assumed that there are only two types of codes in nature: the genetic code that appeared at the origin of life, and the cultural codes that arrived almost four billion years later, but in recent years many other codes have been discovered (Barbieri 2015). What is slowly coming to light, in other words, is that many organic codes exist in living systems and have appeared throughout the history of life together with the great novelties of macroevolution.

In this larger framework, the existence of signal transduction and signal integration codes becomes an entirely natural phenomenon, because codes are no longer extraordinary exceptions but normal components of life. The great evolutionary potential of the codes is that they can bring absolute novelties into existence because they are not dictated by physical necessity and can establish relationships that have never existed before in the universe.

Acknowledgments I am truly grateful to two reviewers whose comments have greatly improved the first version of this article.

References

Alberts B, Johnson A, Lewis J et al (2007) Molecular biology of the cell, 5th edn. Garland, New York

Allan DW, Thor S (2015) Transcriptional selectors, masters, and combinatorial codes: regulatory principles of neural subtype specification. WIREs Dev Biol 4:505–528. doi:10.1002/wdev.191

Barbieri M (2003) The organic codes: an introduction to semantic biology. Cambridge University Press, Cambridge

Barbieri M (2015) Evolution of the genetic code: the ribosome-oriented model. Biol Theory 10:301–310

Berger SL (2007) The complex language of chromatin regulation during transcription. Nature 447:407–412

Berridge M (1985) The molecular basis of communication within the cell. Sci Am 253:142–152

Blobel G (1980) Intracellular membrane topogenesis. Proc Natl Acad Sci USA 77:1496–1500

Brown JR, Doolittle WF (1997) Archaea and the prokaryote-eukaryote transition. Microbiol Rev 61:456–502

Callebaut W, Rasskin-Gutman D (eds) (2005) Modularity: understanding the development and evolution of natural complex systems. MIT Press, Cambridge
Cavalier-Smith T (2000) Membrane heredity and early chloroplast evolution. Trends Plant Sci 5:174–182
Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci USA 105:10039–10044
Darwin C (1859) On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. Murray, London
Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2129
Doolittle WF, Bapteste E (2007) Pattern pluralism and the tree of life hypothesis. Proc Natl Acad Sci USA 104:2043–2049
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
Flames N, Pla R, Gelman DM et al (2007) Delineation of multiple subpallial progenitor domains by the combinatorial expression of transcriptional codes. J Neurosci 27:9682–9695
Fox GE (2010) Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol 2:a003483
Fu XD (2004) Towards a splicing code. Cell 119:736–738
Gamow G (1954) Possible relation between deoxyribonucleic acid and protein structures. Nature 173:318
Haeckel E (1866) Generalle Morphologie der Organismen. Georg Reimer, Berlin
Harold FM (2005) Molecules into cells: specifying spatial architecture. Microbiol Mol Biol Rev 69:544–564
Harold FM (2014) In search of cell history: the evolution of life's building blocks. University of Chicago Press, Chicago and London
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
Jenuwein T, Allis CD (2001) Translating the histone code. Science 293:1074–1080
Jessell TM (2000) Neuronal specification in the spinal cord: inductive signals and transcriptional codes. Nat Genet 1:20–29
Jun S-R, Sims GE, Wu GA, Kim SH (2010) Whole genome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal resolution. Proc Natl Acad Sci USA 107:133–138
Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell 98:285–294
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
Margulis L (1970) Origin of eucaryotic cells. Yale University Press, New Haven
Marijuán PC, Navarro J, del Moral R (2015) How the living is in the world: an inquiry into the informational choreographies of life. Prog Biophys Mol Biol 119:469–480
Marquard T, Pfaff SL (2001) Cracking the transcriptional code for cell specification in the neural tube. Cell 106:651–654
Matlin A, Clark F, Smith C (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6:386–398
Mayr E (1998) Two empires or three? Proc Natl Acad Sci USA 95:9720–9723
Mereschowsky C (1910) Theorie der Zwei Pflanzenarten als Grundlage der Symbiogenesis, einer neuen Lehre der Entstehung der Organismen. Biologisches Zentralblatt 30:278–303, 321–347, 353–367
Miller RV (1998) Bacterial gene-swapping in nature. Sci Am 278:67–71
Nitta I, Kamada Y, Noda H et al (1998) Reconstitution of peptide bond formation. Science 281:666–669
Orgel LE (1973) The origins of life. Wiley, New York
Osawa S (1995) Evolution of the genetic code. Oxford University Press, New York
Osborne LC, Palmer SE, Lisberger SG, Bialek W (2008) The neural basis for combinatorial coding in a cortical population response. J Neurosci 28:13522–13531
Pertea M, Mount SM, Salzberg SL (2007) A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*. BMC Bioinform 8:159
Portier P (1918) Les symbiotes. Masson et Cie, Paris
Sapp J (1987) Beyond the gene: cytoplasmic inheritance and the struggle for authority in genetics. Oxford University Press, Oxford
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
Schimper AFW (1883) Uber die Entwickelung der Chlorophyllkörner und Farbkorper. Bot Ztg 41:105–114
Schlosser G, Wagner GP (eds) (2003) Modularity in development and evolution. University of Chicago Press, Chicago
Simonson AB, Servin JA, Skophammer RG et al (2005) Decoding the genomic tree of life. Proc Natl Acad Sci USA 102:6608–6613
Snel B, Huynen MA, Dulith BA (2005) Genome trees and the nature of genome evolution. Ann Rev Microbiol 59:191–209
Sporn MB, Roberts AB (1988) Peptide growth factors are multifunctional. Nature 332:217–219
Strahl BD, Allis D (2000) The language of covalent histone modifications. Nature 403:41–45
Sutherland EW (1972) Studies on the mechanism of hormone action. Science 177:401–408
Turner BM (2000) Histone acetylation and an epigenetic code. BioEssays 22:836–845
Turner BM (2002) Cellular memory and the histone code. Cell 111:285–291
Turner BM (2007) Defining an epigenetic code. Nat Cell Biol 9:2–6
Ulrich LE, Koonin EV, Zhulin IB (2005) One-component systems dominate signal transduction in prokaryotes. Trends Microbiol 13:52–56
Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code. Proc Natl Acad Sci USA 103:10696–10701
Wallin JE (1927) Symbionticism and the Origin of Species. Williams and Wilkins, Baltimore
Wang Z, Burge C (2008) Splicing regulation: from a part list of regulatory elements to an integrated splicing code. RNA 14:802–813
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 (1998) Default taxonomy: Ernst Mayr's view of the microbial world. Proc Natl Acad Sci USA 95:11043–11046
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, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eukarya. Proc Natl Acad Sci USA 87:4576–4579

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

Wu J, Grunstein M (2000) 25 years after the nucleosome model: chromatin modifications. Trends Biochem Sci 25:619–623

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366