Article

Key Steps in the Early Evolution of Life from the Origin of Protein Synthesis to Modern Cellular Life

Sankar Chatterjee *

Department of Geosciences, Museum of Texas Tech University, Box 4319, 3301 4th Street, Lubbock, Texas 79409, USA

* Correspondence: Sankar.chatterjee@ttu.edu; Tel: +1-806-787-4332

Abstract: The emergence of proteins in the prebiotic world was a watershed event at the origin of life. With their astonishing versatility, the protein enzymes catalyzed crucial biochemical reactions within protocells into more complex biomolecules in diverse metabolic pathways, whereas structural proteins provided strength and permeability in the cell membrane. Five major biochemical innovations followed in succession after availability of various kinds of protein molecules during decoding and translation of mRNAs. These are: (1) the modification of the phospholipid membrane into the plasma membrane; (2) the origin of primitive cytoplasm; (3) primitive gene regulation; (4) the beginnings of the virus world; and (5) the advent of DNA. The creative role of viruses during prebiotic synthesis led to the origin of the *DNA world*, when DNA replaced mRNA as the major genome of the protocells. With the advent of DNA, replication of information was entirely dissociated from its expression. Because DNA is much more stable than mRNA with more storage capacity, it is a superb archive for information systems in the form of base sequences. DNA progressively took over the replicative storage function of mRNA, leaving the latter for protein synthesis. Genetic information began to flow from DNA to mRNA to protein in a two-step process involving transcription and translation. In the *biological* stage, DNA replication was central to the binary fission of the first cell, orchestrated by the duplication of genomes and then the division of the parent cell into two identical daughter cells. With the onset of binary fission, the population of primitive cells grew rapidly in the hydrothermal vent environment, undergoing Darwinian evolution and diversification by mutation. The habitat of the earliest fossil record (≥ 3.5 Ga) from the Archean sedimentary rocks of Canada, Greenland, Australia, South Africa, and India offers a new window on the early radiation of microbial life. The development of anoxygenic and then oxygenic photosynthesis from early hyperthermophiles would have allowed life to escape the hydrothermal setting to the mesophilic global ocean.

Keywords: protein/RNA world: plasma membrane; cytoplasm; gene regulation; virus world; pre-retro virus; emergence of DNA; transcription and replication; first cells; hyperthermophiles; LUCA; Bacteria and Archaea; anoxygenic bacteria; oxygenic bacteria; global distribution of cyanobacteria

1. Introduction

The origin of life is one of the great unsolved scientific problems of our age. While the details of this process of origin are still shrouded in mystery, the prevailing scientific hypothesis is that transition from non-living matter, such as simple organic compounds to living cell was not a single event, but a gradual process of increasing complexity. The past decade has seen a resurgence of interest in the origin of life, in part because of new evidence from astrobiology. To understand how life could begin on young Earth, it is essential to know what organic compounds–the essential building blocks of life–were likely to have been available and how they interacted in a suitable environment. Using the early history of the young Earth, I have suggested previously that life probably arose through five hierarchical stages of increasing molecular complexity in a steaming hot environment of hydrothermal crater basins about 4 billion years ago—*cosmic*, *geologic*, *chemical*,

information, and *biological*. Each stage gave birth to a new sequence of biomolecules with new characteristics and increasing complexity that led to the emergence of the first cells. The building blocks of life had their beginnings in the interstellar space during the explosion of a supernova. Most likely, carbonaceous chondrites delivered both building blocks of life and water to early Earth during recurrent impacts on the early Archean crust. Asteroid collisions also created innumerable hydrothermal crater basins that became the perfect cradle for prebiotic chemistry. The hydrothermal crater vents provided both chemicals and energy sources, making them feasible as a niche for the emergence of life [1-3].

The RNA world model has become the main paradigm in the current origin of life research in which RNA assumed informational and functional roles [4-8]. In our previous paper [9], we have argued that the peptide/RNA world is more parsimonious than the RNA world for biogenesis. This view was supported by several authors in recent times [9-15]. We suggested that the demand for a wide range of protein enzymes over peptides in the prebiotic reactions was the main selective pressure for the origin of information-directed protein synthesis [9]. Once the mRNA-directed protein synthesis became established, various enzymes were created to meet the demand of catalysis and metabolism. The next step involved encapsulation of a more complex enzyme system capable of catalyzing fatty acids to form phospholipids. The transition from single-chain lipids to phospholipids was triggered by the availability of a wide range of enzymes in the protein/RNA world that catalyzed the conversion of lipid membranes into phospholipid membranes. In this paper, we discuss the evolution and functional repertoire of translation proteins that led to hierarchical emergences of plasma membranes, cytoplasm, gene regulation, viruses, DNA, and finally the first cells.

2. Protein/RNA World

Proteins were assembled from amino acids using information encoded in mRNA and translated to protein by tRNA. The ensuing peptide elongation was catalyzed by rRNA in the ribosome. Each protein had its own unique amino acid sequence of the mRNA gene. Proteins control most of the functions of a cell, breaking down nutrients, assembling cellular components, copying DNA, and so on. They truly occupy a central position in the organization of the first cell. They act as enzymes that permit only a few of the many possible reactions among cellular components to take place. Enzymes play an essential role in prebiotic synthesis. They are catalysts that greatly accelerate reactions by providing an alternate reaction pathway with a much lower energy barrier. Thus, although they do not create new reactions, they greatly enhance the rate at which a particular substrate is changed into a particular product. Without them, the translation of genetic material into proteins would be impossible. Proteins form channels in plasma membranes, allowing specific substances to enter and leave, while excluding others. They form channels in plasma membranes, allowing specific substances to enter and leave, while excluding others. Protein molecules owe their properties to their three-dimensional shapes, which are themselves determined by their amino acid sequences of their constituent chains. These properties in turn determine how a protein biologically functions: whether it will bind certain organic molecules and catalyze their reactions or form regular structures, such as a helix, and act as a building material.

The first proteins were most likely short, about 25 amino acids long [16]. Protein molecules of this short length displayed enzyme-like activities. In contrast, many modern-day proteins contain several hundred amino acids. Most likely, these long molecules arose by the gradual lengthening of the first products of protein synthesis. Proteins are the primary functional biomolecules of life. Once formed, proteins performed a vast array of functions during biogenesis including catalyzing metabolic reactions and reinforcing cell membranes. The overwhelming number of efficient proteins in the biochemical synthesis occurred in the protein/RNA world. With their astonishing versatility, protein enzymes would have taken ribozyme's role in assisting genetic copying and metabolism. Different kinds of enzymes were in great demand in the protein/RNA world. In order to be useful, an enzyme must necessarily have a specific substrate available in its environment on which to act. The enzyme must also have an outlet for the products it forms. These substrates and outlets must have been provided by the primitive metabolism in the vent environment that supported the

protocells at the time [16]. These newly formed enzymes carried out hundreds of chemical reactions that took place in the protocell. Structural proteins, on the other hand, provided structure and support for the cell membrane. Ancient enzymes, tryptophan synthases, resurrected from the common ancestor of all bacteria indicate that such sophistication of primordial enzyme complexes existed more than 3.4 billion years ago [17].

Proteins mediate most functions of modern cells. Four major events followed after the availability of template-directed proteins but with considerable overlap. These affected, first, the efficiency of the translation machinery, then, resilience of the coding system, and finally, the quality of the synthesized proteins. These protein-mediated events are: (1) the transition from the phospholipid membrane to the plasma membrane; (2) the origin of prebiotic cytoplasm; (3) the beginnings of the virus world; and (4) the advent of DNA. The newly synthesized protein enzymes helped to catalyze and mediate these critical molecular evolutions, favored by strong selective forces.

3. The Evolution of Phospholipid Membranes

Self-assembly of cell membranes that encapsulated the polymers into populations of protocells were essential to the development of life forms in the prebiotic Earth. The primitive cell membranes enclosed the life-building molecules, defined the boundaries from environments, and enhanced chemical reactions inside protocells. Each cell membrane was a very thin film of lipid molecules, about 5 nm thick and was a dynamic fluid structure. The origin of cellular life presumably occurred by self-assembly of cosmic organic compounds into encapsulated systems capable of catalyzed polymer synthesis [18,19]. Thus, the availability of a primitive cell membrane component in a hydrothermal vent environment was a prerequisite for biogenesis. The cell membranes are built of components that have a remarkable capability of spontaneous self-assembly. The chemical requirements for formation of cell membranes from individual molecules are remarkably simple. The lipids that form the lipid bilayer common to all biological membranes are not gene products. Therefore, the development of genetic material need not have to precede the development of the membrane. However, the development of membranes could have facilitated the development of self-replicating genes by providing them a protected space in which to evolve and eventually function.

Most likely, the phospholipids did not become available in the prebiotic environment until various protein enzymes were available and metabolic pathways for their catalyzed synthesis evolved. The phospholipid bilayers were impermeable to most water-soluble molecules. This property made bilayers excellent boundaries that allowed for protocells to maintain an internal composition different from that of the surrounding medium. But protocells could not survive and evolve sealed off from the outside. They must be able to take up nutrients, get rid of waste products, and respond to environmental signals. These functions were carried out by proteins inserted into phospholipid bilayers to form a plasma membrane, improving the bilayer's permeability.

3.1. The Origin of the Phospholipid Membrane

Three critical molecules—fatty acid, glycerol, and phosphate—were available in the hydrothermal vent environment for synthesis of the phosphatidate acid by non-enzymatic synthesis (Fig, 1A,1B) [17]. Abiotic synthesis of aliphatic lipids, fatty acids, and acylglycerols has been reported to occur at elevated temperatures and pressures under simulated hydrothermal conditions [18]. The phospholipid is synthesized by a series of enzyme-catalyzed energy-dependent reactions from the phosphatidate acid (Fig. 1C). This was possible when RNA-directed protein enzymes were available to convert the phosphatidate acid to phospholipid. As such, phospholipid provides a marked contrast with fatty acid as a membrane component. Because of its molecular complexity, it is generally believed that the prebiotic synthesis of the phospholipid was difficult [19-20]. More recently, several studies have achieved the synthesis of the phospholipids and related compounds, such as acylglycerol and glycerol phosphates, suggesting that such molecules may have been present in the prebiotic environment in trace quantities [5, 16, 20-26]. Moreover, the abiotic formation of the ester bond among lipid compounds, including acylglycerides, is possible under simulated hydrothermal conditions, provided the precursors that are present are at sufficient concentrations [27].

The synthesis of the phospholipids requires an activated intermediate, phosphatidate acid (diaglycerol 3-phosphate). The prebiotic pathway from fatty acids to the simplest phospholipid, phosphatidate acid, occurs via successive acyl- and phosphotransfer reactions. The first step in the synthesis of the phospholipids is the synthesis of phosphatidate acid, which is formed by the addition of two fatty acids to glycerol 3-phosphate. Here we explore the non-enzymatic pathways of the emergence of the phosphatidate in the peptide/RNA world. From three building blocks—fatty acids, glycerol, and a phosphate ion, which were available in the prebiotic environment [28], we show the gradual evolution of monoglyceride, diglyceride, and triglyceride by the condensation reaction (Fig. 1B). When glycerol and fatty acid react, a water molecule is expelled, forming an ester linkage. The production of diglyceride is considered first, since it will lead directly to the biosynthesis of the phosphatidate. Diglyceride is formed when the glycerol and fatty acid chains become joined by two ester linkages. Non-enzymatic synthesis of ester bonds to produce the diglycerides might have been the first step toward glycerolipids [16,22,26]. The phosphorylation of diglyceride, in turn, would give rise to the phosphatidate acid, which is essentially a diglyceride in which a phosphate group has been added to a single glycerol molecule (Fig. 1C).

The non-enzymatic synthesis of the activated phosphatidate acid was at a pivotal point in the lipid biosynthetic pathways. It served as the precursor for the formation of the glycerophospholipid (commonly called phospholipid) membrane by enzymatic synthesis. Glycerophospholipids are the main constituents of membrane bilayers. Enzymatic synthesis pathways evolved over time when RNA-directed protein enzymes were available in the protein/RNA world. The phosphate group of the activated phosphatidate acid is esterified to an alcohol to produce a variety of phospholipids including the attachment of choline, ethanolamine, serine, and inositol to the phosphate group of phosphatidic acid. Names of phospholipids then include phosphatidylcholine (phosphate + cholin), phosphatidylethanolamine (phosphate + ethanolamine), phosphatidylserine (phosphate + serine), and phosphatidylinositol (phosphate + inositol). If the alcohol is choline, the product is phosphatidylcholine. Of these, phosphatidylethanolamine is the most common phospholipid in bacterial cell membranes. Different modifiers give the phospholipids different properties and roles in a cell. Three successive enzymatic methylation could convert the phosphatidate to phospholipid.

A phospholipid consists of a polar headgroup on one end of the molecule and fatty acid chains on the end. These chemical structures create an amphipathic liquid. In solution, they instantly form bilayers that are selectively permeable. The phospholipids are composed of a polar head group (usually a negatively charged phosphate group and glycerol); it is hydrophilic. Phosphate is a primary anionic component of most phospholipid membrane lipids (Fig. 1A). The phospholipid tails consist of two long fatty acid chains, which are hydrophobic and avoid interactions with water. Two fatty acids are attached to a glycerol by ester or other bonds. The polar head group and fatty acid chains are attached by a 3-carbon glycerol unit [28].

The phospholipid molecules have a hydrophilic head end and two hydrophobic tails that will not mix with water and will avoid being surrounded by it (Fig. 2A). Because these amphipathic molecules have both a hydrophilic and hydrophobic group on the same molecule, they can undergo self-assembly into a cell. In an oil slick, the hydrophobic tails mix with oil while the heads stay close to the water in a monolayer cell (or micelle). When placed in water, the phospholipids will orient themselves in a bilayer in which non-polar tail regions face the inner layer of the bilayer (Fig. 2B). Being cylindrical, the phospholipid molecules contribute structural stability and create a semipermeable environment. The same forces that drive the phospholipids to form bilayers also provide a self-healing property. Admixture of cholesterol helps to stabilize the bilayer.

Modern phospholipid-based cell membranes are formidable barriers to the uptake of polar and charged molecules ranging from metal ions to complex nutrients and nucleotides. They require special protein transporters to allow their passage through the membrane. The phospholipid membranes are stable under a wide range of temperature, pH, and salt concentrations. A recent experiment suggests that the phospholipid membranes can self-assemble and are surprisingly permeable to transport molecules across its membranes [19].

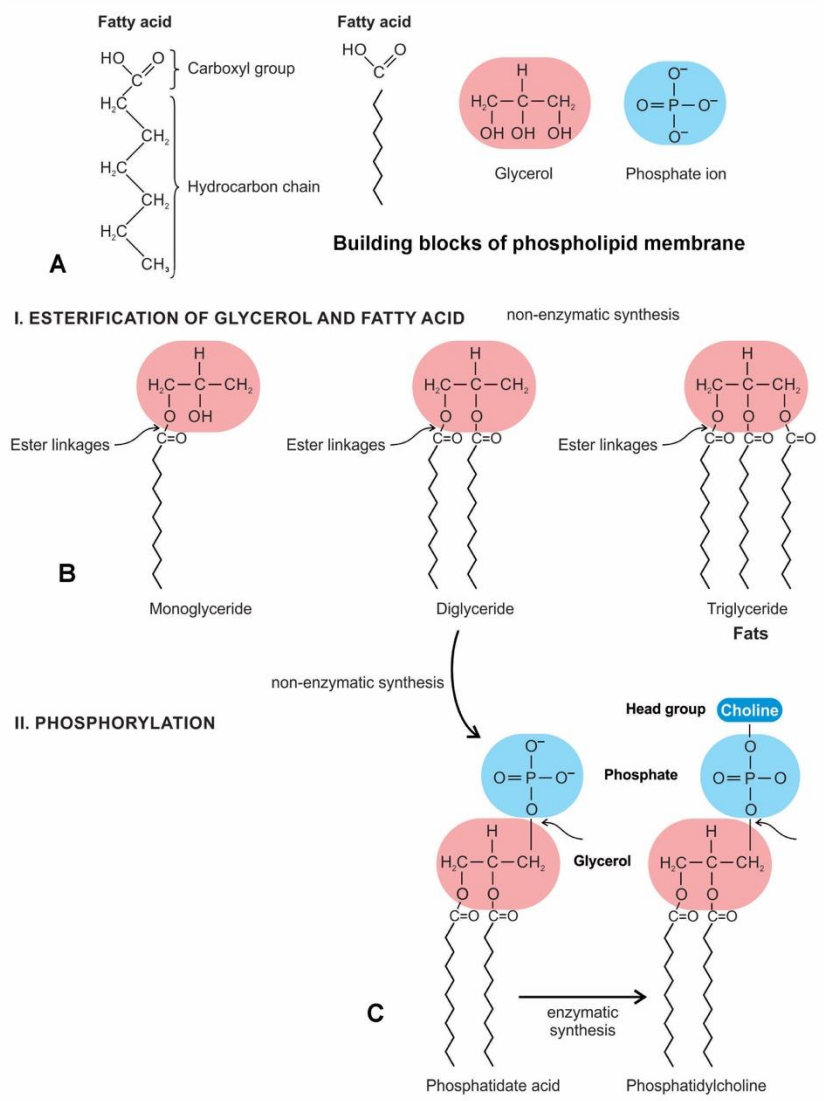

Figure 1. Origin of the phospholipid membrane from simple fatty acids by an intermediate, the phosphatidate acid by a series of non-enzymatic synthesis. A, fatty acid, glycerol, and phosphate ion are the building blocks of phospholipid. B, in the first stage of phospholipid formation, several glycerides such as monoglyceride, diglyceride, and triglyceride (fat) are formed by esterification of glycerol and fatty acid, with the loss of a water molecule; the covalent bond, an ester linkage results from this reaction. C, the next stage of synthesis of phosphatidate acid is by phosphorylation of a diglyceride molecule, when a phosphate ion is joined. Phosphatidate acid, in turn, would give rise to phospholipid by attaching to an alcohol molecule, such as choline, ethanolamine, serine or inositol. Of these various combinations, phosphatidylcholine (shown in the figure) is the most common phospholipid in the cell membrane.

The evolution of phospholipid membranes was a critical and necessary step for the early evolution of cells. The transition from single-chain lipids to double-chain phospholipids had to be gradual. Recent experiments suggest that low levels of phospholipids could drive the growth of fatty acid vesicles by competition of monomers with neighboring vesicles lacking the phospholipids. This competitive growth would have provided strong selective advantage for primitive cells to evolve the catalytic machinery needed to synthesize the phospholipids from their single-chain precursors. Growth results from decreasing fatty efflux from the membrane with increasing phospholipid content, suggesting an evolutionary arms race among primitive protocells [22]. The gradual transition from a fatty acid bilayer membrane to a phospholipid bilayer membrane is shown here in Fig. 3.

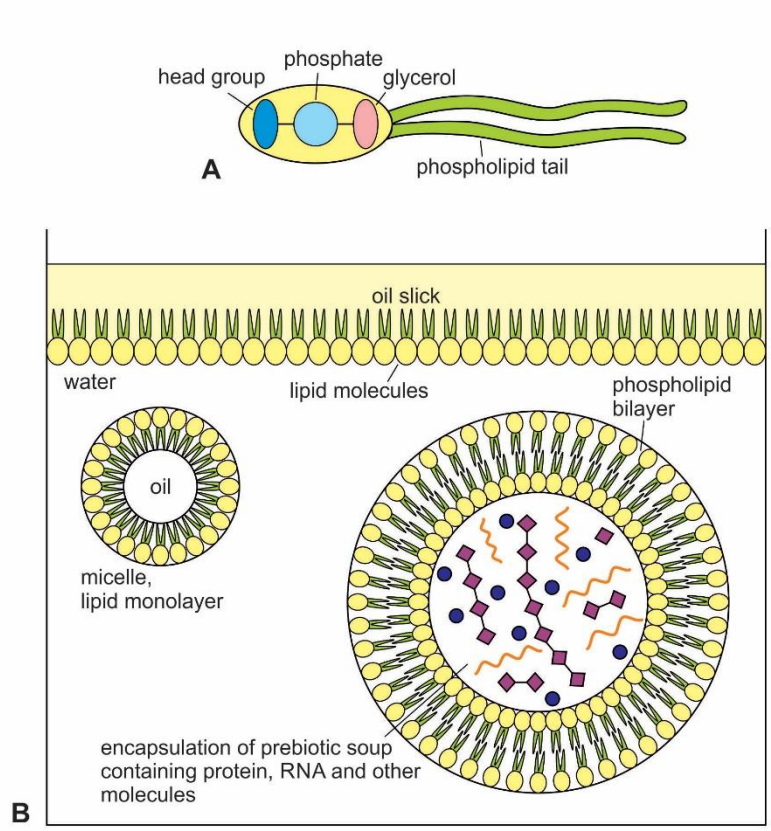

Figure 2. The self-assembly of a phospholipid membrane in a hydrothermal crater lake in the peptide/RNA world. A, a generalized phospholipid molecule has a hydrophilic ('water-loving') head and two hydrophobic ('water-hating') tails that do not mix with water and will avoid being surrounded by it. B, in an oil slick on the surface of a crater lake, the hydrophobic tails mix with oil. whereas the heads stay close to water. During turbulence, phospholipids form two kinds of membranes: a monolayer, which can only capture a drop of oil (left), or a bilayer, which can capture a group of water molecules (right). The bilayer allows the hydrophobic tails to associate with one another, whereas the heads associate with water molecules, on both the inside and outside surfaces of the membrane. A bilayer vesicle is stabilized when it encapsulates protein molecules that interact with the bilayer surface.

Increasing phospholipid content inhibits the permeability of fatty acid membranes through changes in bilayer fluidity. The emergence of the phospholipid membrane would have therefore imposed new selective pressures for the evolution of a more permeable plasma membrane with the protein transport system. Proteins that were amphipathic could be inserted into the phospholipid bilayer to increase permeability.

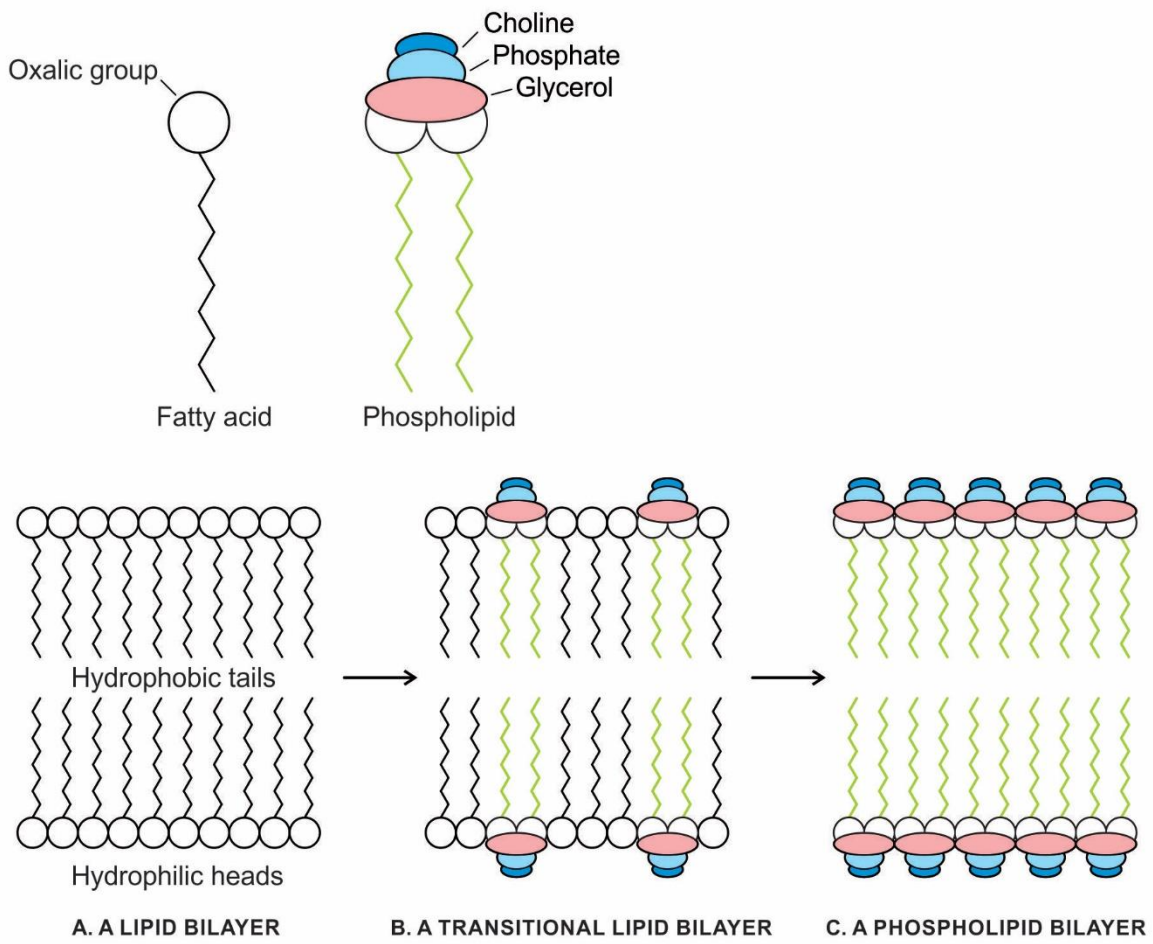

Figure 3. The gradual transition from single-chain, highly permeable fatty acid membrane (A) to selective permeable phospholipid membrane with the increase of phospholipid content (C) via a transitional stage (B). Increasing phospholipid content inhibits the permeability of fatty acid membranes through changes in bilayer fluidity.

3.2. The Origin of the Plasma Membrane

Here, we suggest the likely scenario for the origin of the plasma membrane from the phospholipid membrane (Fig. 4). A new class of proteins emerged in the protein/RNA world that played critical roles for the conversion of the phospholipid membrane to the plasma membrane. Proteins could be amphipathic because they are made of amino acids, and amino acids have R-groups that range from highly nonpolar to highly polar. To incorporate in the phospholipid membrane, the nonpolar amino acids would be selected in the interior of the lipid bilayer, while the polar would be selected alongside the polar heads of the surrounding water.

As the phospholipid membrane began to interact with the new generated amphipathic proteins by translation, endosymbiosis would integrate these proteins to the phospholipid membrane (Fig. 4A). These protein symbionts would stabilize the walls of the membrane, allowing them to resist disruptive forces such as the mechanical shear caused by the convection current of the vent. The resulting increase in osmotic pressure and membrane tension would create a driving force for an increase in the membrane area, therefore encouraging symbiosis between proteins and lipid membranes that triggered the origin of the primitive plasma membrane for stability to prevent a burst of protocells (Fig. 4B). This is how more sophisticated plasma membranes might have formed by endosymbiosis to make them more permeable. The plasma membrane of modern cells is composed of roughly equal parts proteins and lipids by weight. Most likely the primitive protocells had a higher percentage of lipids than proteins from the initial endosymbiosis that favor a high degree of thermostability. The plasma membrane acted as a selectively permeable barrier, preventing some

substances from crossing while permitting other substances to enter and leave the protocell. The selective permeability of the plasma membrane and the specificity of transport proteins made it possible to create an environment inside the protocell that was radically different from the prebiotic soup and amenable to biogenesis.

The first step of survival in confinement was the possibility for the protocells to take in nutrients and energy from the outside environment and get rid of waste material. The simplest way in which fully enveloped protocells could fulfill this condition was by means of pores, mere holes kept open in the phospholipid bilayers by some kind of inserted protein framework. The insertion of proteins in bilayer lipid vesicles was an essential first step that affected the phospholipid bilayer's permeability, facilitating transport and other molecules into the protocells. In primitive membrane transport, passive transport was used when molecules were moved across the plasma membrane of the protocells along, down, or with their concentration gradients. The molecule would flow from where it was at higher concentrations to where it was at a lower concentration.

Next came transport facilitators, which were transmembrane proteins that act as restricted passages for certain specific substances. A more sophisticated kind of molecular passage was the gated channel-like facilitators that let certain substances of a given chemical specificity to move through passively but they were unidirectional and regulated by a gate that needed to be unlocked by some channels or electrical signal.

The next improvement in the building of molecular transport systems was active transport, where molecules were moved against their concentration gradient in energy-requiring processes and the machinery involved was correspondingly more complex. The systems that carry out such active transport are called pumps. The energy used was derived from ATP, the universal currency of energy. There were various ATP-powered pumps that were used to transport ions and molecules against their concentration gradients. The plasma membrane would have played an essential role in the generation of metabolic energy, and transformed it into useful ATP.

The classical fluid mosaic model of the structure of the cell membrane distinguishes between two types of membrane proteins: peripheral proteins and integral proteins; the former occurs only outside the lipid bilayer, while the latter spans the entire membrane for transport of ions and molecules [29]. The phospholipids and the plasma membranes make back and forth movements within the plasma membrane, making the plasma membrane a fluid structure. The key point is that the arrangement of proteins makes the interior and exterior surfaces of the plasma membrane very different (Fig. 4C). Thus, the plasma membranes are a mosaic of the phospholipids and the different types of proteins, and the overall structure is dynamic and fluid. In protocells, the membrane proteins were responsible for the passage of ions, polar molecules, and large molecules that did not readily cross the phospholipid bilayers on their own.

The permeability of the phospholipid bilayer was altered radically by membrane transport proteins, which were scattered throughout the plasma membrane. Selection pressure favored the evolution of the plasma membrane over the phospholipid membrane to overcome the reduced membrane permeability. The integral proteins would allow only certain molecules to enter the protocell. In this way, the protocell could fine-tune the selection of what got in and what did not. The peripheral proteins, on the other hand, acted like sensory antennae, making it possible for protocells to gain information about their immediate environment. In addition to transporting substances into and out of the protocell, the plasma membrane began to function in energy production. It created an internal environment that was more conducive to life synthesis than the external environment. Several layers of the plasma membrane were added with the gradual evolution of the protocells to the first cells.

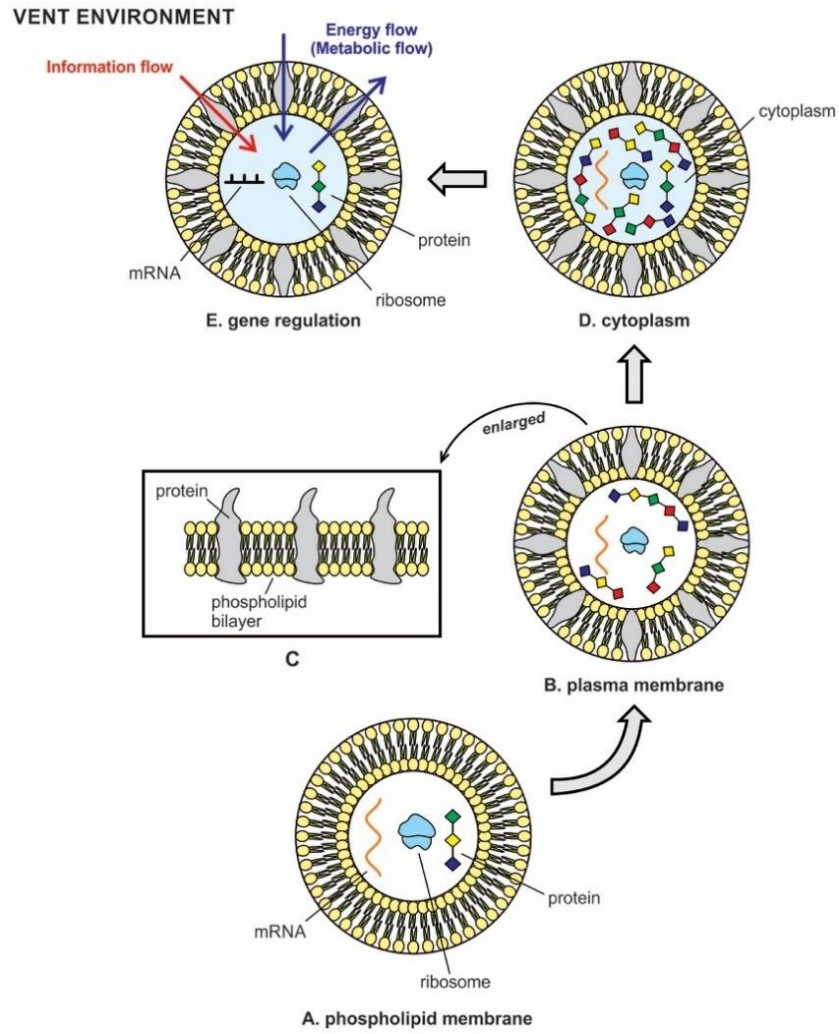

Figure 4. The transition of phospholipid membrane to plasma membrane. A, phospholipid membrane; B, plasma membrane; C, phospholipid membrane evolved into plasma membrane by inserting protein molecules into bilayers that made the cell membrane selective permeable so that certain ions like potassium and sodium can cross the bilayer barrier; D, cytoplasm inside protocell formed when newly synthesized protein molecules by a translation machine began to concentrate inside the protocell, turning internal water into a gel-like substance; E, gene regulation in RNA protocell; energy flows and information flows are represented as the basic exchanges of the protocell with outer/inner environment. RNA protocell began to respond to environmental cues and target mRNA for regulating gene expression.

Proteins are responsible for most of the dynamic processes carried out by the cell membrane, including the transport of molecules into and out of the cell (Fig. 4B,4C). The plasma membrane separates the cell from its environment and is selectively permeable: it chooses what enters and exists in the cell. Receptor proteins are the gatekeepers; they detect signals from the environment of the cells; the transport proteins help some molecules get across the membrane. Certain membrane proteins act as enzymes. The plasma membrane was the ideal microenvironment to experiment with synthesis of more complex nucleic acid such as DNA for the permanent storage of the information system.

4. The Origin of Cytoplasm

A large fraction of cytoplasmic, water-soluble proteins are enzymes that catalyze chemical reactions involved in metabolism and reproduction. Once different kinds of proteins were

synthesized by translation machinery and available for prebiotic synthesis, they began to accumulate in the primitive cytoplasm inside protocells. This primordial cytoplasm became the ready source of a variety of proteins, when needed. The simple aqueous solution (prebiotic soup) inside the protocell was gradually converted to a viscous, gel-like cytoplasm that increased the protocell volume and provided some rigidity of its spherical shape. The first cytoplasm in protocells and the semipermeable plasma membrane from the phospholipid membrane most likely organized at the same time with the availability of proteins (Fig. 4D). The prebiotic cytoplasm provided a stable microenvironment for the organization of all nucleic acids, lipids, enzymes, proteins, other macromolecules, molecules, ions and ribosomes, as well as water and salts that were all encased in the plasma membrane. This primitive cytoplasm became a complex, crowded system containing a wide range of molecules – from ions and small molecules, to macromolecules like proteins, nucleic acids, and ribosomes. Many metabolic reactions, including protein synthesis, and the transition from RNA to DNA began to take place in this primordial cytoplasm. Over half of the molecules were actively involved in the synthesis of proteins. Some of these proteins were used in the synthesis of viruses and DNA molecules, while others were engaged in energy production. The constituents of the cytoplasm were moved across the protocell depending on their requirements. Primitive cytoplasm supported and suspended these molecules in its gel-like substance. This primordial cytoplasm became the site for most of the enzymatic reactions and metabolic activity of the protocell. The primitive cytoplasm was confined to the outside by the plasma membrane, the latter began to regulate the passage of some substances, such as organic molecules, ions and water, preventing the passage of some other substances to maintain the content of the primitive cytoplasm. Other compounds moved passively across the membrane.

Three major innovations took place in succession, exploiting these new protein reserves in the prebiotic cytoplasm: (1) gene regulation and information flow; (2) the origin of the virus world; and (3) the advent of the DNA world. Here we discuss the origin of gene regulation, followed by the primordial viruses in the vent environment that gave rise to the DNA world.

5. mRNA Gene Regulation and Information Flow

The origin of translation—protein synthesis—was the first stage of gene expression. The mRNA protocells were open systems regarding both energy flows and information flows. Another creative outcome of protein synthesis was the development of a rudimentary gene regulatory system. It is about protocells choosing which mRNA gene to use and which not during protein synthesis from environmental signals. In general, a gene was expressed only when its protein product was needed. RNA protocells responded accordingly to environmental fluctuations, transmitted signals across its plasma membrane, and targeted to specific mRNA for protein synthesis (Fig. 4E). With the development of the plasma membrane, protocells might have developed a primitive signal transduction system in order to respond to environmental cues. Plasma membranes contained various receptor proteins to control information flow from the outer environment to the mRNAs in cytoplasm. The signal was then amplified and activated by second messengers to select specific mRNA for gene expression. These mRNA genes were mainly constitutive genes that contained the blueprints for proteins for essential housekeeping functions. Gene expression occurred at translation and post-translational levels. Perhaps a regulatory protein bound a specific mRNA in response to environmental fluctuation and increased the translational rate to build a corresponding protein. These proteins were essential to the survival of the protocells. In general, each mRNA molecule went on to make a specific protein. In some cases, this protein would be an enzyme, in other cases, it would be structural, according to the demand of the protocell. Some proteins were needed for basic metabolism like the enzymes that catalyzed reactions during glycolysis so that protocells could transfer energy from food to ATP. Others would be needed to make the viruses, the protective coat of mRNA, which would in turn, give rise to DNA. Proteins were not fully formed and functional when termination from the ribosome occurred. They were manufactured in an active form and had to be activated by chemical modification, such as the addition of a phosphorous group. This type of modification is post-translational control.

The following steps probably occurred as information flowed from environment to mRNA to proteins represented by arrows in the following expression:
Environmental signal —> receptor proteins —> second messengers —> mRNA —> protein —> activated protein.

6. The Beginnings of the Virus World

Viruses straddle the line between living and nonliving. They are tiny, noncellular, microscopic parasites that infect virtually every type of known cell. Even simpler and smaller than a bacterium, a virus has a diameter of 20–400 nm. Because they are not living organisms in a true sense, they require the biochemical machinery of a cell to reproduce. A virus is nothing more than a few strands of genetic material wrapped in a package of protein—a parasite, unable to function on its own. In order to survive, it must find a cell to infect. Only then can the virus take control of the host's cellular machinery and use it to churn out thousands of copies of itself. These viruses then move from one cell to the next, transforming each new host into a factory that makes even more viruses.

Their hallmark characteristics, namely their small size, tiny genomes, and parasitic dependence on cellular hosts for reproduction, set them apart from all other living things despite their animation. However, the discovery of the giant viruses (> 400 nm), called the Mimivirus, with massive genomes and the most complete resources for building proteins further blurs the established boundaries between viruses and the smallest parasitic cellular organisms. The simple size-based distinction between viruses and cells is no longer tenable. However, its icosahedral ultrastructure of capsid coat, and its typical eclipse phase in its life cycle, support the viral nature of the Mimivirus. Furthermore, the Mimivirus lacks universal bacterial genes, such as encoding ribosomal RNA and proteins [30].

Viruses can be defined as capsid-encoding organisms as opposed to ribosome-encoding cellular organisms [31,32]. Viral particles are by far the most abundant biological entities on our planet, greatly outnumbering all their cellular hosts put together; most of the biomass in the ocean is made up of viruses [33]. The genetic diversity of viruses is enormous as well, in part because they can acquire genomes from their hosts and they can later paste these genes into new hosts. The viruses are agents for gene dissemination, evolution, and biodiversity.

The simplest viruses have just two components: a nucleic acid core and an outer protein capsid shell. The genome, which may be DNA or RNA, contains the instructions for taking over cells, making capsids, and creating more virions, or viral particles. There are many types of viruses, classified by their size and shape, by their genetic material (RNA or DNA), and by their host organisms. The majority of viruses have a genome based on DNA, although a significant minority has RNA genes. Viruses come in three common shapes: helical, polyhedral (such as icosahedral), and complex viruses, the latter often possess a unique structure or an extension on virions. Viruses are highly diverse in their morphology and in the nature of their genetic material. The genomes may be single-stranded RNA (ssRNA), double-stranded RNA (dsRNA), single-stranded DNA (ssDNA), or double-stranded DNA (dsDNA). Encapsidation of viral genomes constitutes a virus particle.

In their overall structure, viruses fall into two categories: enveloped and non-enveloped. Non-enveloped viruses have an extremely simple structure. They consist of genetic material and possibly one or more enzymes that are encased inside the capsid. Enveloped viruses are more derived and complex, where the capsid is surrounded by an envelope. The envelope consists of a phospholipid bilayer with a mixture of viral proteins and proteins derived from the plasma membrane of the host cell [34,35].

Viruses are completely selfish [36]. They cannot reproduce on their own, because they lack ribosomes and the rest of the living cell's protein-making machinery. They can only reproduce by invading host cells and hijacking their ribosomes, enzymes, and energy. When a virus enters a host cell, it sheds its capsid coat, bares its genes, replicates, and induces the cell's own translation machinery to manufacture more viral protein from the viral nucleic acid. The viral genes and capsid protein self-assemble to form virions, turn the host cell into a viral factory, and strain the cell to the bursting point. The host cell bursts and releases hundreds of virus progeny by the lytic cycle. Some viruses are also capable of lysogeny, and their genomes become integrated into the host cell

chromosome. In this condition, the host cell continues to live and reproduce normally. Thus, viruses only need to do two things: they need to have a mechanism for reproduction within host cells, and they need a way to get out of their target cells [37]. Viruses develop a simple way of creating new viruses that require only a minimal investment of molecular machinery.

Viral polymerases play a central role in the viral genome replication and transcription. Several steps in the virus life cycle require the activity of a polymerase. Based on the genome type and specific needs of a particular virus, a variety of enzymes are contributed by viral hallmark genes encoding proteins. These are RNA-dependent RNA polymerase (RdRp), RNA-dependent DNA polymerases (RdDp), DNA-dependent RNA polymerases (DdRp), and DNA-dependent DNA polymerase (DdDp). Viral polymerases slowly transformed RNA viruses to DNA viruses step-by-step during a recurrent infection. The evolutionary networks of primordial viruses, their recurrent infections of protocells, and their polymerases accelerated the origin of the first cells. Two capsid proteins that are most widely distributed among viruses are the jelly-roll capsid proteins (JRC) and the superfamily 3 helicase (S3H) [36].

6.1. Viruses and Evolution

Viruses are not self-sustaining and need to enter a cell in order to complete their life cycle. Viruses emphasize parasitic roles far more than cooperation in the evolutionary process. Therefore, we tend to regard viruses only as pathogens and thereby dismiss their crucial importance for the evolution of life. Viruses were not only the probable precursors of the first cells, but they have helped to shape and build genomes of all species including humans. The impact of viruses on life is dramatic. The symbiotic relationship of viruses and cells is not always restricted to parasitism, but extends to a wide range of mutualism. The majority of known viruses are in fact persistent and inapparent, not pathogenic (toxic). Many viruses are beneficial to their hosts, providing essential functions in others. Viruses are major drivers of evolutionary transitions [34-36].

The history of life is a story of coevolution of viruses and their cellular hosts. All cellular life harbors diverse genetic parasites including transposons, plasmids, viruses and other selfish elements. The parasite-host coevolution is a major aspect of the evolution of life [36,37]. The coevolution is often described as an incessant arms race. The billion-year war between viruses and cells is the major source of evolutionary novelties. Viruses evolve, the host adapts, proteins change, and viruses evade them. It never ends. Many novelties first selected in the viral world might have been transferred to cells as a consequence of the continuous flow of viral genes into cellular genomes. The war has driven a dramatic diversification of viruses and of the host defense system. Viruses have a remarkable capacity to invade, replicate, and evolve within living cells. In response, cells developed an array of defense systems. Viruses and protocells were intertwined since the protein/RNA world. Viral reproduction within a living cell always produces changes in the host cell, sometimes resulting in cell death and sometimes slowly killing the infected cells.

The creative role of viruses in the origin and evolution of life has been known for a century. Viruses are truly nature's genomic laboratory, and they help accelerate evolution of the host in a fast lane. Felix D'Herelle, the discoverer of bacteriophages and one of the founders of virology, proposed as early as 1922 that phages or bacteriophages might be the evolutionary precursors of cells [38]. Similarly, J.B.S. Haldane in his 1928 classic *The Origin of Life* suggested an early 'viral' stage of evolution as an integral part of the proposed scenario for the emergence of life from the prebiotic soup [39]. In our discussion of the origin of life, we have followed Haldane's insight.

Viruses are potentially aggressive, selfish-elements, a property that conferred on a parasitic partnership of viruses and protocells; this association has powerful evolutionary potential. Viruses are adept at transferring genetic information between themselves and hosts. Viruses contributed several key enzymes to the host protocells including RNA-dependent RNA polymerase, DNA-dependent RNA polymerase, and DNA-depended DNA polymerase during RNA-DNA transition. RNA viruses might have coexisted with early protocells that still had RNA genomes. Because viruses have such ancient roots, they preserve a remarkable range of biochemical tricks [40].

The great billion-year war between viruses and cells are the major source of evolutionary novelties [32]. Host cells are under immense evolutionary pressure from their viral invaders. They have evolved numerous immune systems to cope with this pressure. Viruses use an extensive battery of counter-defense strategies to exist in presence of host cells. Without a cellular victim, viruses cannot function. A prominent group of viruses are bacteriophages or phages. The coevolution of viruses and bacteria has a complex and prolonged interaction. CRISPR arrays and associated *cas* genes are a family of DNA sequences found within the genomes of bacterial cells. These sequences are derived from DNA fragments of bacteriophages that have previously infected bacteria and are used to detect and destroy DNA from similar phages during subsequent infections. Hence these sequences play a key role in the anti-phage defense system of bacteria. The discovery and exploitation of CRISPR-*Cas* systems have stimulated a resurgence in the identification and characterization of anti-phage mechanisms and gene editing.

The arms race between protocells and viruses led to the advent of DNA from mRNA precursors. The reverse transcriptase paved the way to generate DNA; they still generate DNA from RNA in retroviruses, cancer cells, and HIV. Viruses donated DNA and their replicating genes to protocells [36,40]; they might have played a central role in the emergence of eukaryotes and their nuclei [36,40,41]; they might have been the cause of the partitioning of biological organisms into three domains of life by horizontal gene transfer [43,44]. Their role in information transfer between extant prokaryotes by horizontal gene transfer complicates efforts to build evolutionary trees depicting early life on Earth and to unravel the origin of particular metabolic pathways.

Many viruses have their own, ancient evolutionary history, dating to the protein/RNA world. They are the relics of the protein/RNA world. Viruses possess genes, replicate, evolve, and adapt to particular hosts, biotic habitats, and ecological niches. From prebiotic protocells to unicellular life to human populations, viruses affect life's outcomes and give an ever-changing shape to the fitness landscape, often determining which organisms will survive [34]. Since the beginnings of life, viruses have been the major drivers of macroevolution in all branches of life by horizontal gene transfer across three cellular domains—Bacteria, Archaea, and Eukaryotes [41,43,44]. They comprise the principal source of novel genes in the biosphere [34,45]. Some viruses have the ability to become dormant inside of a host cell. The genetic material of dormant viruses may remain in the host cell for long periods of time and is copied as the cell reproduces. Other viruses (e.g., retroviruses) integrate their genetic material into the cell they infect, and if this happens to be a germ line, the viral genome (or its relict) can be maintained essentially forever. About 8% of human genetic material originated from RNA viruses rather than from our vertebrate ancestors [46]. Similarly, retroviruses facilitate the rapid evolution of the mammalian placenta [47]. Virus-host interaction is an important evolutionary force and played a crucial role in the origin and evolution of life.

6.2. The Prebiotic Origin of RNA Viruses

The origin of viruses is shrouded in mystery, but recent advances in genomics shed light on their ancient ancestry. Viruses have never been detected in fossils, probably because they are too small and too fragile for fossilization. Therefore, the evolutionary history of viruses is difficult to reconstruct. For many years, the central debating point in discussions of the origin of viruses is whether they are ancient, first appearing before the Last Universal Common Ancestor (LUCA), or evolved more recently, such that their ancestry lies with genes that 'escaped' from the genomes of their cellular host organisms and subsequently evolved through independent reproduction [48].

While the geological record cannot offer any clue of when and how viruses originated, genetics provides increasingly strong support for an ancient primordial origin of viruses. First, the vast majority of viruses do not encode genes for ribosomal proteins or genetic evidence of relics of such genes. Second, this same vast majority of viruses do not contain genetic evidence of ever having encoded enzymes involved in energy metabolism. Third, viral capsid proteins typically do not have obvious homologs among contemporary cellular proteins. Several genes coding for key proteins involved in viral infection as well as major capsid proteins of icosahedral virions are shared by many groups of viruses but are missing in cellular life forms. This is why most viral proteins have no

cellular homologues or only distantly related ones. All these combined evidences argue that viruses did not evolve from free-living cells, but arose independently in the prebiotic world before the first cells. The existence of hallmark genes seems to falsify both the cell degeneration and the escaped genes concepts of viral infection [34,36,49-52]. This primordial origin theory is supported by the strongly inverse relationship between genome size mutation rate across all replications systems, such that pre-LUCA genomes were probably both small and highly error-prone and hence RNA-virus like [48].

In the ancient viral world, the flow of virus-specific genes has gone uninterrupted from the precellular stage of life's evolution to the present day. In our view, three major classes of viruses originated in the prebiotic world: (1) positive-stranded mRNA viruses in the primordial protein/RNA world, (2) retroid viruses in the RNA–DNA transition world, and (3) DNA viruses in the DNA world. These ancient viruses emerged in a hydrothermal vent environment in which the mixing and matching of diverse genetic elements was more extensive than it is in any modern biological community. Phylogenetic analyses have shown that that the RNA polymerase, DNA polymerase, and DNA helicase that transcribe and replicate DNA in modern cells were recruited from the viral world [48,49,54].

The idea that viruses are very ancient and have co-evolved with the protocells has recently led to several hypotheses stating that viruses have played a major role in several critical evolutionary transitions [31-34]. Viruses ubiquitously infect all members of the three cellular domains of life, strongly suggesting that protocells with RNA genomes were already the victims of a viral attack. Moreover, viruses infecting cells from the three domains of life—Bacteria, Archaea, and Eukaryotes—suggests that viruses emerged very early in the prebiotic world before the first cells [37,48,49].

The notion of viral antiquity seems easier to accept for mRNA viruses in the hydrothermal vent environment. High temperatures in the vent environment favored high diversity of virus-like particles [34]. The only organisms with RNA-coded genomes today are RNA-based viruses, which may shed helpful insight into the protein/RNA world. The virus world retained a distinct flow of genes from the repeated infection of protocells containing RNAs, ribosomes, and proteins. Viruses have maintained their identities and unique parasitic lifestyle ever since, notwithstanding the transfer of many genes between viral and cellular genomes. Several genes that are central to viral replication and structures are shared by all viruses but absent from cellular genomes. In this scenario, the principal lineages of viruses emerged from the primordial pool of primitive genetic elements with a distinct suite of viral genes that retained their identity throughout the entire history of life. Viruses enhanced gene mixing in the prebiotic world by infecting protocells containing RNA and protein molecules, and their subsequent endogenization of them. Therefore, the viral evolution in the prebiotic world is closely intertwined with the origin and early evolution of the cell [51,52,55].

6.3. Pre-virus World

Viruses can be viewed as mobile genetic elements (MGE) [36]. The principal lineages of viruses and related selfish agents emerged from the primordial genetic pool of primitive genetic elements in the hydrothermal vent environment, the ancestors of both cellular and genetic elements. The primordial gene pool was crucible for the major virus lineages, where mixing and matching of diverse genetic elements was extensive. Viruses reflect their origin from capsidless selfish replicons, such as plasmids, transposons, and viroids. In this scenario, viruses are direct descendants of primordial genetic elements [32]. These selfish replicons, called viral hallmark genes, encoded capsid proteins with key roles in genome replication, expression, and encapsidation. These selfish replicons are shared by a broad variety of viruses but are missing from cellular genomes suggesting a flow of virus-specific genes that went uninterrupted from the protein/RNA world of biogenesis to this day. These viral genes are genuine viral hallmarks and can originate either through modification of existing genes or *de novo*. Eventually, diverse protein-coding RNA elements would develop a capsid coat to give rise to the first viruses [32,52].

We concur with the ancient virus world hypothesis [3,16,17,39]. In our model, ancestral viral particles, called here 'pre-virus' could have emerged only when RNA-directed proteins were

available in the prebiotic soup in the hydrothermal crater vent environment. As more and more proteins were manufactured inside protocells, these densely packed biomolecules would exert an osmotic pressure on the phospholipid membrane, occasionally resulting in the rupture of the protocells. The biomolecules would then be dispersed in the prebiotic soup, making it an ideal Nature's genomic laboratory. Different kinds of genetic innovations took place in the genetic pool of the hydrothermal vent that mixed, matched, and evolved new, increasingly complex gene ensembles.

The two important capsid proteins are the jelly-roll capsid proteins (JRC) and the superfamily 3 helicase (S3H), the former is more widespread [32]. We speculate that primordial mRNA gene encoding capsid protein JRC was present in the vent environment (Fig. 5A). New JRC protein genes can originate either through modification of existing genes or *de novo*. Viruses contain many *de novo* genes, namely those in which an existing gene has been 'overprinted' by a new open reading frame; mutations of the mRNA gene led to the expression of a second reading frame, overlapping the first. Overlapping genes are very common in viral genomes [56].

The origin and evolution of viruses might have occurred on the mineral substrate of the crater floor, where RNAs and proteins were accumulated side by side. During this prebiotic genomic experiment, naked and fragile mRNA genes might have capped occasionally by proteins that offered protection for stability and durability in the vent environment [1-3]. This protein coat of mRNA molecules was prelude to the evolution of a viral structure. Perhaps, in this milieu of different kinds of mRNAs in the prebiotic soup, JRC capsid genes originated *de novo* within genomes of nonviral mRNAs by overprinting. These capsid genes were capped by proteins on the mineral substrate, transforming them into ancestral viral particles (Fig. 5A). These ancestral pre-viruses had some survival advantages over naked viral genes in the vent environment because of the protective protein coat. Initially, protein coats were random and were not encoded by the enclosed mRNA genes. Perhaps the ancestral pre-viral genes, when engulfed by protocells, could translate it into custom-made capsid protein using ribosomes of the host (Fig. 5B). Primordial viruses could have evolved by encapsidation of these viral genomes.

Some of the ancestral mRNA viruses were accidentally ingested by the protocells by infolding of their membranes while searching for food in the vent environment. The engulfed viruses shed their capsid coat, which might be used by protocells for protein storage. This might have been the beginning of endocytosis. The viral genes, on the other hand began to exploit the translation machinery of the protocell to make the custom-made capsid protein.

6.4. The Reproduction Cycle of mRNA Viruses

The emergence of ancestral virus with capsid-coding sequence of proteins was a big evolutionary step and was mediated by ribosome-coding protocells (Fig. 5B). Here I propose a model for parallel evolution of protocells and the emerging viruses. As more and more proteins were synthesized, various kinds of protocells dominated in the hydrothermal vent environment. Some of these protocells were densely packed with diverse populations of genetic elements, including self-replicating mRNAs, various protein-coding mRNAs, and translation machinery.

A typical viral genome encompasses two core modules that consist, respectively, of genes encoding proteins required for genome replication and proteins involved in capsid formation. In the ancestral virus, the core module might have included all the genes for capsid formation. Initially, protocells with the full set of translation machinery engulfed some of the viral capsid genes from the mineral substrate and inadvertently helped translate their genomes into viral proteins. The engulfed mRNA gene multiplied using replicating enzyme (RNA-dependent RNA polymerase or RdRp) and began to exploit ribosomes of the protocell for synthesis of capsid proteins. Some of the newly created capsid protein strands began to wrap around mRNA as a protective coat where the genome could be maintained as a stable structure. Encapsidation was the hallmark of virus' survival without encapsulation. The association of capsids with capsid genomes was a complex process, but it must result in an energetically stable structure. This is the beginning of the viral world, which evolved in parallel with the protocells.

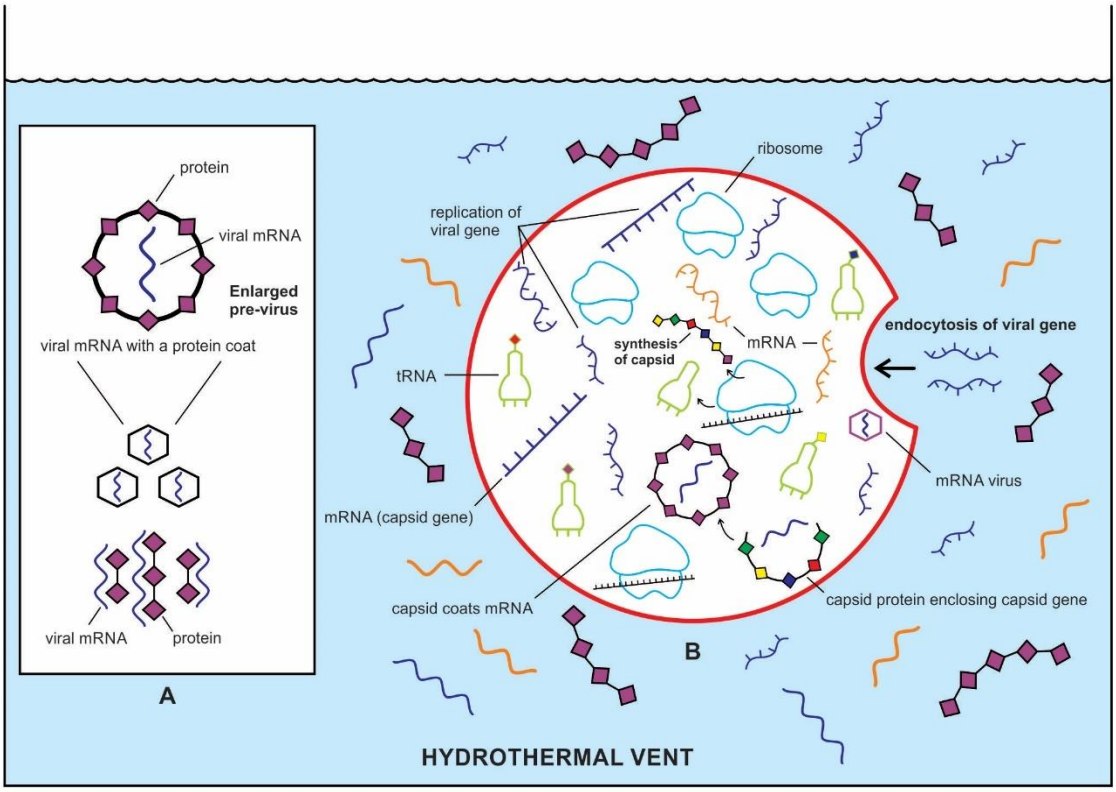

Figure 5. Beginnings of the ancient viral world in the prebiotic soup in the hydrothermal vent environment, which was an ideal Nature's genomic laboratory. A, in my model, viral selfish genes modified from mRNA molecules appeared de novo by overprinting in the prebiotic soup; these capsid genes were encased by a protein shell for stability and durability. However, these proteins were random, easily available in the prebiotic soup, and were not encoded by viral gene. This initial stage of viral structure that afforded protection of fragile mRNA is called 'pre-virus'. B, In the next stage of evolution, some viral genes entered the protocells by endocytosis, utilizing their ribosomes for synthesis of capsid proteins. Once the capsid protein began to coat the viral gene, the first mRNA virus appeared.

Once some mRNA molecules developed the capsid coat, the first mRNA viruses originated. The capsid affords protection of the viral gene and allows viral genes to gain access to appropriate host cells. As more and more mRNA viruses were created, they exerted osmotic pressure on the phospholipid membrane causing a burst of the protocells (Fig. 6A). Slowly, these newly released viruses learned by trial and error how to infect protocells and swap genes with them without inventing their own translation machinery. This innovative short-cut strategy for virus reproduction inside protocells worked efficiently, hijacking the translation machinery of the host. Viruses preferred this parasitic existence from the beginning of biosynthesis, and it has continued to proliferate throughout the geologic ages.

Viruses can reproduce only within host protocells, exploiting their ribosomes. The parental virus (virion) gives numerous progenies, usually genetically and structurally identical to the parent virus. The new generation of viruses began to infect other protocells with an ensemble of translation machinery for their reproduction so that viral mRNAs could be translated into capsid proteins inside the host protocells. Soon, they became obligate intracellular parasites; that is, they developed and reproduced only within the protocells of hosts. They hijacked the host's own machinery to manufacture hundreds of copies of themselves. The reproductive successes of viruses make them archenemies of living cells. Viruses became the capsid-coding particles that began to coevolve with ribosome-encoding protocells. Because of their high mutation rate, viruses were evolution accelerators.

We speculate that primordial viruses were single-stranded mRNA viruses that could function both as a genome and as a messenger RNA. It could be directly translated into capsid protein in the host cell by host ribosomes. Like living counterparts, the genome contains relatively few genes, usually between three and ten, including as an RNA-dependent RNA polymerase (RdRp) or RNA replicase, a viral protein enzyme that synthesizes mRNA from mRNA template. RdRp is an essential protein-encoded in the genomes of all RNA-containing viruses prior to the DNA stage. Pre-viruses would donate an RdRp gene to the host cell facilitating replication of its mRNA. RNA replication before the emergence of RdRp is difficult to comprehend.

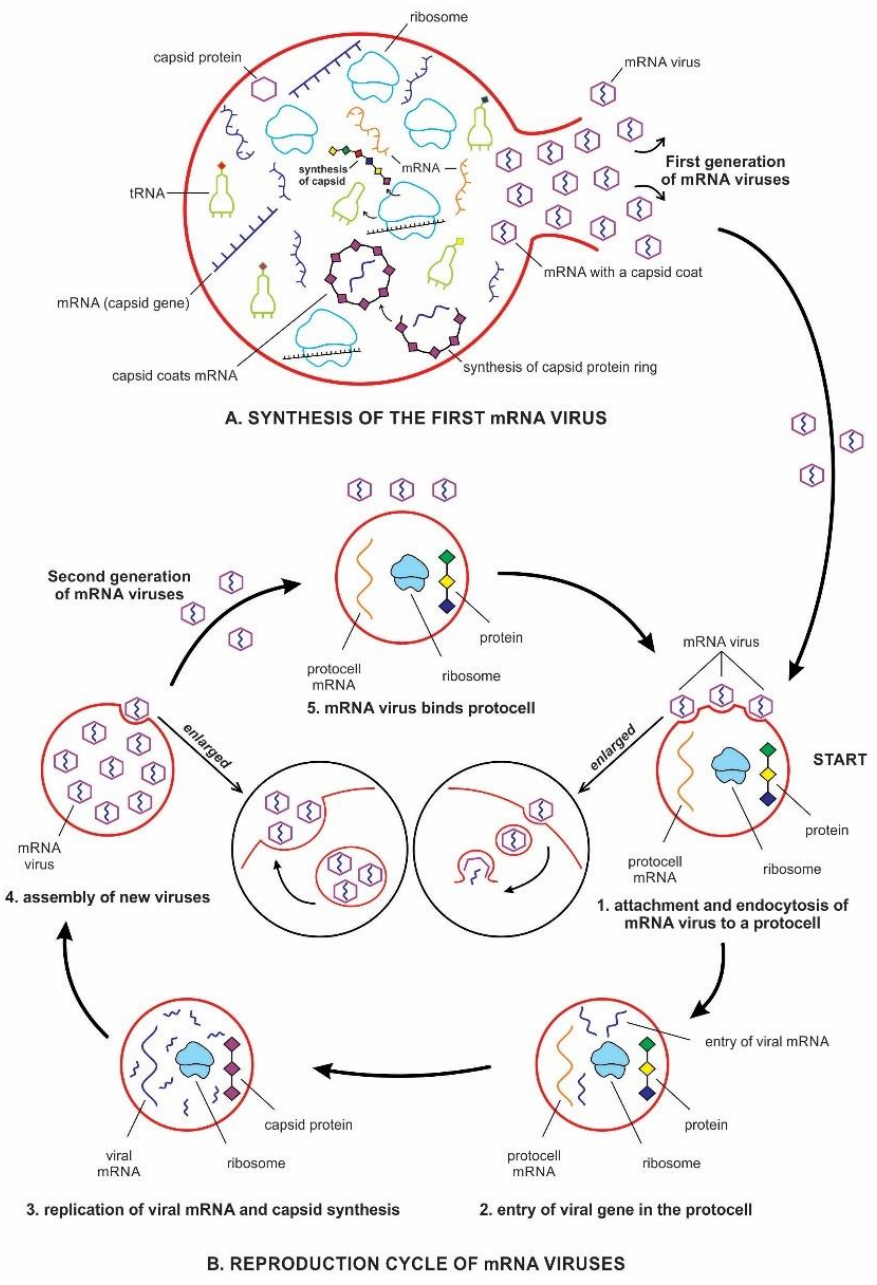

Figure 6. **The** origin of mRNA viruses and their reproduction cycle**.** A, synthesis of first generation of mRNA viruses inside protocells hijacking hosts' translation machinery. mRNA viral genes produced capsid proteins that began to wrap viral genes. As more and more mRNA viruses were synthesized inside protocells, they exerted osmotic pressure on the phospholipid membrane causing a burst of protocells, releasing the first batch of viruses for infection of protocells. B, some likely stages of the reproduction cycle of mRNA virus includes: attachment of a protocell and its entry into protocell via endocytosis (cycle 1), uncoating of capsid shell (cycle 2), mRNA copying and protein synthesis (cycle 2 and 3), self-assembly of viral progeny and release via exocytosis (cycle 4).

We speculate that primordial virions initially didn't kill or loose their host protocells, but utilized their translation machinery for reproduction of genomes. Some of the likely stages of the cycle of infection were (Fig. 6B):

1. Attachment. Capsids on the surface of the mRNA virus attach to the surface of the host protocell (Fig. 6B, cycle 1).

2. Entry via endocytosis. The virus enters the interior of the host protocell through the process of endocytosis.

3. Uncoating of capsid. Inside the protocell, the viral genome emerged from the protein capsid; the capsid in turn destroyed the host mRNA so that viral mRNA occupied its place (Fig. 10B, cycle 2, inset).

4. mRNA copying and protein synthesis. RdRp enzymes copy the viral genome. Energy and ribosome from the host protocell are used to build viral proteins. (Fig. 6B, cycle 2 and cycle 3).

5. Assembly of viral progeny. The viral particles assemble by encapsidation to form progeny virions (Fig. 6B, cycle 4).

6. **Release via exocytosis**. The virus exits in the host protocell by exocytosis (Fig. 6B, cycle 4, inset).

The progeny virions began to infect other protocells to begin the next cycle of infection (Fig. 10B, cycle 5). The life cycle of most viruses is designed to maximize the production of progeny virus particles. Often, the burden of producing a large number of virus particles causes the infected cells to die, the lysis of the host cell. In early stages, primordial viruses probably established a long-term association with the protocell, in which the protocell released a steady stream of viral particles over an extended period of time, benefiting both host and parasite in symbiosis. These ancient RNA viruses had a high mutation rate and underwent evolution and natural selection, just like cellular life, and most of them evolved rapidly. When two viruses infected a protocell at the same time, they might swap genetic material to make new 'mixed' viruses with unique properties. The viral infection of a protocell is a prelude to a modern bacteriophage that infects and replicates within Bacteria and Archaea.

This is the beginning of mRNA viruses and their spread in the vent environment. Today, mRNA viruses amount to a large fraction of known viruses including many pathogens, such as the hepatitis C virus, West Nile Virus, dengue virus, and SARS and MERS coronaviruses, as well as less clinically serious pathogens, such as rhinoviruses that cause the common cold [57].

6.5. Retroviruses

A retrovirus is a highly derived enveloped particle in which the capsid core contains two identical single-stranded RNA molecules, each RNA carries its genetic and structural blueprint. The virion is 80-100 nm in diameter, and its lipid envelope incorporates and displays the viral phospholipids. The hallmark of a retrovirus is its replicative strategy in the sense that it can reverse-transcribe its RNA into DNA using its own reverse transcriptase enzyme. This catalyzed transcription is the reverse flow of information of central dogma, hence the name *reverse transcriptase* and *retrovirus*. The new DNA is then integrated into the host cell genome by an integrase enzyme. The host cell treats the viral DNA as part of its own genome, transcribing and translating the viral genes, producing the proteins required to assemble new copies of a retrovirus [34,35].

The main virion components of a retrovirus are: a phospholipid membrane and two identical double-stranded mRNA encased in a capsid shell. Each mRNA is typically made up of three genes: the group specific antigen gene (*gag*), the polymerase gene (*pol*), and the envelope gene (*env*). The *pol* gene encodes the three viral enzymes: protease, reverse transcriptase, and integrase—that catalyze the steps of retroviral infection. In retrovirus, multiple protein products are synthesized from a single mRNA species by frame shifting. In between 5′ and 3′ ends of RNA is the protein coding domain, which includes *gag*, *pol*, and *env* encoding regions. These three genes are linked to one through recoding by frameshifting. The ability to make two or more proteins from same mRNA is useful, linking structural (e.g., retroviral *gag*) and catalytic polypeptide (retroviral *pol*). This ribosomal

frameshifting mechanism makes the retroviral genome more compact (Fig. 7A). Retroviruses have evolved to exploit this translational plasticity in order to regulate their own expressions [58].

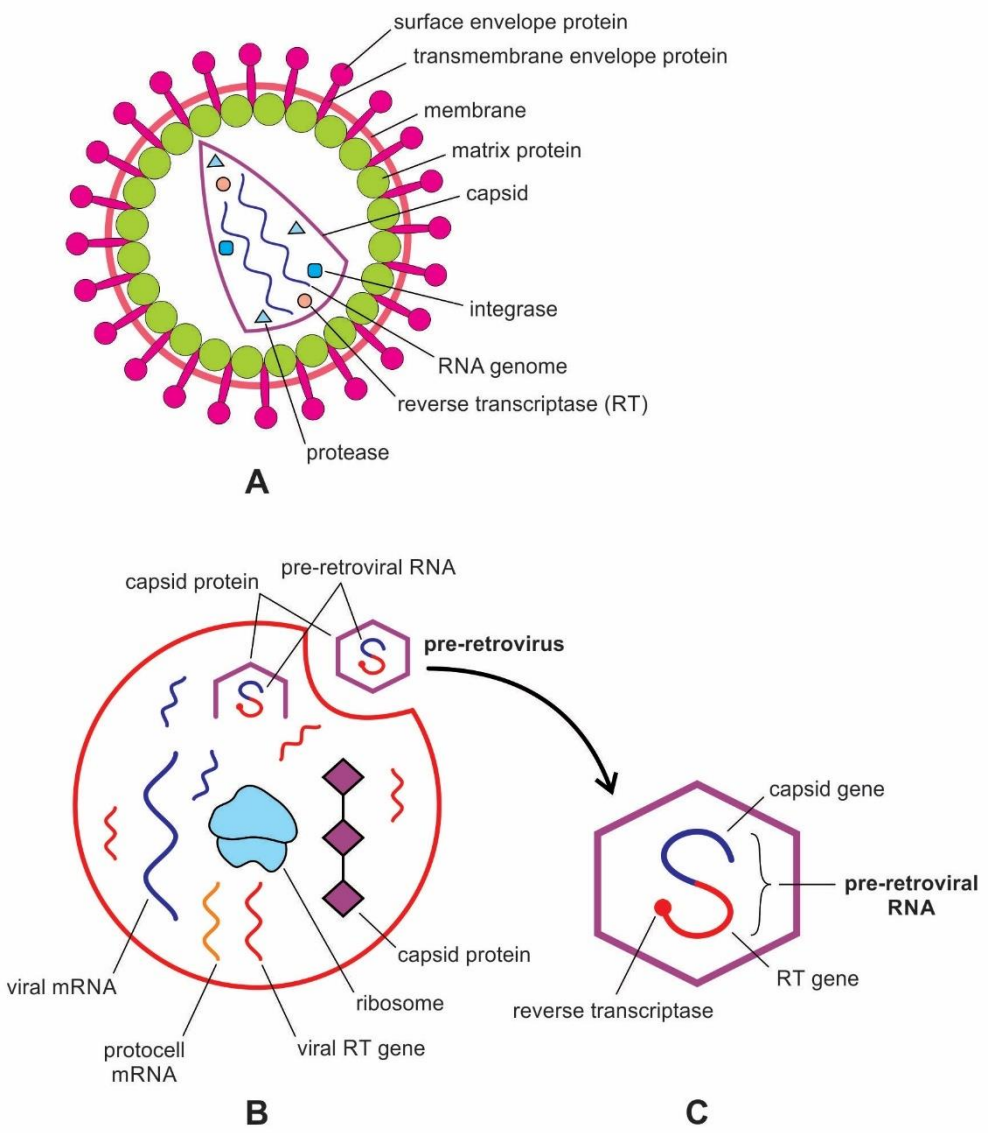

Figure 7. Retroviruses. A, structure of a modern retrovirus, an enveloped particle in which the capsid core contains two identical single-stranded mRNA molecules. Each mRNA is made up of three genes: integrase, reverse transcriptase, and protease. Once inside the host cell cytoplasm, the virus uses its own reverse transcriptase enzyme to produce DNA from its RNA genome. B, likely origin of pre-retrovirus inside infected protocell, where the viral RT gene was linked to the mRNA viral gene, and was encased by a capsid coat. Once two genes were fused into a single gene for close packing, pre-retrovirus released from the protocell via exocytosis. It was a non-enveloped particle; C, structure of a pre-retrovirus showing how two genes—capsid gene and RT gene were fused into a single gene, and were encased by a capsid protein.

Once a retrovirus is inside a protocell (a process mediated by protease), it takes over the host's genetic transcription machinery to construct a DNA *provirus*. Retroviral proteases are key enzymes in viral propagation and are encoded by a part of the *pol* gene. This process, the conversion of retroviral RNA to proviral DNA is catalyzed by reverse transcriptase and is necessary for proviral DNA into host DNA—a step initiated by the integrase enzyme. Retroviruses can be pathogens of many different hosts, including humans. A notable retrovirus is human HIV virus, responsible for AIDS.

6.6. Origin of Retroviruses

The known virosphere consists of three principal viral types: the RNA viruses, retroid viruses, and DNA viruses. Horizontal gene transfer (HGT) is rampant among viruses within each of these principal types, but is generally confined to closely related viruses, or viruses (and plasmids) with similar replication mechanisms [59]. There are many examples of mixing and matching in the virus world, but somehow, they so far have been confined to the same type of nucleic acid. A novel virus genome discovered in an extreme hot spring environment suggests recombination of two unrelated groups— between a ssDNA virus and an RNA virus—a natural chimera not seen before [60]. In this hybrid genome, alongside the RNA-derived gene, it contained a gene for DNA replication typical of a DNA virus. Surprisingly, these hybrid viruses are present not just in the acidic lake, but more widespread in a couple of oceanic samples. This find proves that modern viruses can combine information in the two normally separate genetic molecules. And it lends support to the idea that it was viruses that performed the upgrade from RNA and effectively gave rise to DNA. These authors suggest that the hybrid virus may have formed when an mRNA virus, retrovirus, and DNA virus all infected a cell at the same time. The retrovirus used its reverse transcriptase enzymes to make a DNA copy of an RNA virus gene, which combined with the DNA genome to yield this hybrid. In our view, this hybrid virus provides a first glimpse at the ancient viral birth of DNA in the hydrothermal vent environment by the mixing and matching of mRNA virus, pre-retrovirus, and DNA virus.

Retroviruses infect a wide range of animals from fish to humans, and can occasionally leave genomic fossils within their host genome, known as endogenous retroviruses (ERVs). ERVs consist of the genetic material of extinct, or 'fossil' viruses. Our bodies are littered with the shards of retroviruses. Eight percent of our genome is composed of broken and disabled retroviruses, which, millions of years ago, managed to embed themselves in the DNA of our ancestors. Because they no longer seem to serve a purpose or cause harm, these remains have often been referred as 'junk DNA' [34]. Recent phylogenetic study of ERVs placed the time of their most recent common ancestor in the Early Paleozoic [61]. The origin of retroviruses in the Devonian presents an important framework to investigate evolutionary transitions that led to the emergence of the retroviruses. Since vertebrates originated during the Cambrian evolution, ERVs in vertebrate hosts represent the upper limit of the retroviral origin. Molecular precursors of retrovirus probably began in the prebiotic environment billions of years ago in the RNA-DNA Retro world [36]. Retrovirus-like entities are older than the first cells.

Retroviruses bear much similarity to capsidless selfish genetic elements, such as plasmids and various types of transposons because they have close evolutionary connections, both share hallmark genes. These hallmark genes encode key components of the viral replication apparatus (such as polymerases and helicases). These retroelements—capsidless genetic parasites—are key to understanding the origin of viral genes [52]. We speculate that these retroelements were self-assembled to a viral gene such as a positive-strand mRNA virus step-by-step. Positive-sense RNA are particularly suitable for reverse genetics because their genomes are typically infectious in protocells and can be immediately translated by the host's translation machinery. In our model, ancestral retroviruses could emerge only when various retroelements and protein enzymes were available inside protocells.

One of the critical enzymes synthesized inside protocells was the reverse transcriptase (RT) enzyme. RT is an RNA-dependent DNA polymerase. Due to the limitation of the genome size that can be packaged in the virus shell, viral polymerases are generally active as a single protein capable of carrying out multiple functions related to the viral genome synthesis. Here we propose a simple evolutionary scenario for the origin of retroviruses. The vast class of retroelements is united by a single conserved gene, the RT gene, which also defines the key feature of their reproduction cycle, reverse transcription [52].

The ancestral stage of a retrovirus is called here 'pre-retrovirus,' a retroid, non-enveloped mRNA virus with a minimum functional design. In our model, pre-retrovirus was derived from mRNA virus inside an infected protocell. Perhaps, the viral RT gene evolved *de novo* inside the protocell [56]. Most likely, the mRNA viral gene and the viral RT gene merged by accident into one single-stranded

mRNA and was enclosed in a capsid shell. In this fused viral gene, one gene was used for synthesis of structural capsid protein, the other for the RT gene for synthesis of reverse transcriptase enzyme (Fig. 7B, C). The RdRp enzyme could be used by pre-retroviruses to replicate their genomes.

The ability to make structural protein and enzyme from the same mRNA gene had distinctly selective advantages over two separate genes performing similar functions in a double-stranded RNA virus. Most likely, protocells, by that time, had developed several weapons to ward off viral attacks by destroying viral genomes. In response, the pre-retrovirus altered their RNAs in such a way as to thwart attacks from the protocells. Moreover, this linking of two genes, capsid gene and RT gene, into one mRNA strand allowed a novel compact, space-saving mechanism of the genome. This ability of linking two genes into one was achieved by the ribosomal frameshifting mechanism as discussed earlier [58]. A pre-retrovirus had now become a single-stranded mRNA virus with a capsid coat to play a formidable parasite during protocell infection (Fig. 7B, C). With the emergence of retroviruses, the protein/RNA world transformed itself into the 'retro world' [52].

6.7. The Origin of DNA Viruses from mRNA Viruses

DNA can be considered modified RNA because there are only two chemical differences between RNA and DNA molecules [16]. The first is the removal of a single oxygen atom from RNA (ribonucleic acid) to generate deoxyribonucleic acid, or DNA. The second difference is the addition of a methyl (CH_3) group to the nucleotide base uracil (U) to generate thymine (T).

Forterre [40,49] suggests three stages in the evolution of DNA from RNA through viral infection: the RNA world, the RNA-to-DNA transition, and the DNA world. In the RNA world, RNA viruses emerged from the RNA cell. In this model, LUCA is considered as the primitive RNA cell from which the RNA virus emerged. The RNA virus gave rise to three lineages of the DNA viruses in the RNA-DNA transition. These three lineages of DNA virus evolved in parallel into three domains of life—Bacteria, Archaea, and Eukaryotes.

Forterre [40,49,62] proposes that DNA viruses evolved directly from RNA viruses in two steps during the RNA-to-DNA transition. The critical enzymes for RNA to DNA conversion was supplied by the viral world. In the first step of the RNA-to-DNA transition, the deoxyribose in DNA was generated from the ribose in RNA by enzymes called ribonucleotide reductases, which converted RNA to U-DNA in the genome. In the second step in the evolution of DNA was the conversion of the uracil base to thymine by thymidylate synthases, forming T-DNA (DNA containing thymidine). The emergence of thymidylate synthase activity in some U-DNA virus lineages produced viruses with the modern form of T-DNA. As these new strains of T-DNA viruses infected protocells, the host genomes gradually transformed from U-DNA to T-DNA. Once deoxyribonucleoside triphosphates were available, their assembly into DNA-like chains followed fairly rapidly. The idea that both ribonucleotide reductases and thymidylate synthases were first encoded in viral genomes and were later transferred to protocells is compatible with phylogenetic analyses of these enzyme families. Thus, viruses donated both DNA genomes and their replicating genes to the host protocells [40,49]. The hypothesis of a viral origin for DNA could explain why many DNA viruses encode their own ribonucleotide reductase and/or thymidylate synthase [63]. According to this model, the RNA world and LUCA appeared at the same time with the existence of a few homologous DNA informational proteins in the three cellular domains. Cellular DNA and its replication machineries originated in the DNA world via transfers from DNA viruses to RNA cells. Three such independent transfers led to the origin of Bacteria, Archaea, and Eukaryotes.

Though Koonin et al. [36], accept the viral origin of DNA, they criticize the non-cellular LUCA concept of Forterre in the RNA world. Instead, they favor the traditional view that LUCA emerged after the first cells from which all three domains descended. Moreover, they emphasize the crucial role of retroid RNA in the origin of DNA. DNA could not have emerged from the RNA world without reverse transcription. The retro elements must have been among the first classes of primordial viruses that evolved in the primordial genetic pool in the protein/RNA world after the advent of translation, when several kinds of enzymes were synthesized. Integration of such elements into host genomes must have coevolved with the increase in the size of the DNA genome. I find the retroviral infection

model of the origin of DNA [36,52] very attractive and plausible. In my view, retroid RNA such as pre-retrovirus might have invented DNA step-by-step by converting RNA with the help of the RT enzyme. I suggest three stages in the viral evolution, combining both Forterre's and Koonin's model that gave rise to DNA from RNA: the protein/RNA world, the Retro world, and the DNA world (Fig. 8). In my view, the existence of RNA-only cells (protocells) of Forterre [49] seem to meet formidable difficulties. A more parsimonious scenario includes the protein/RNA world [9].

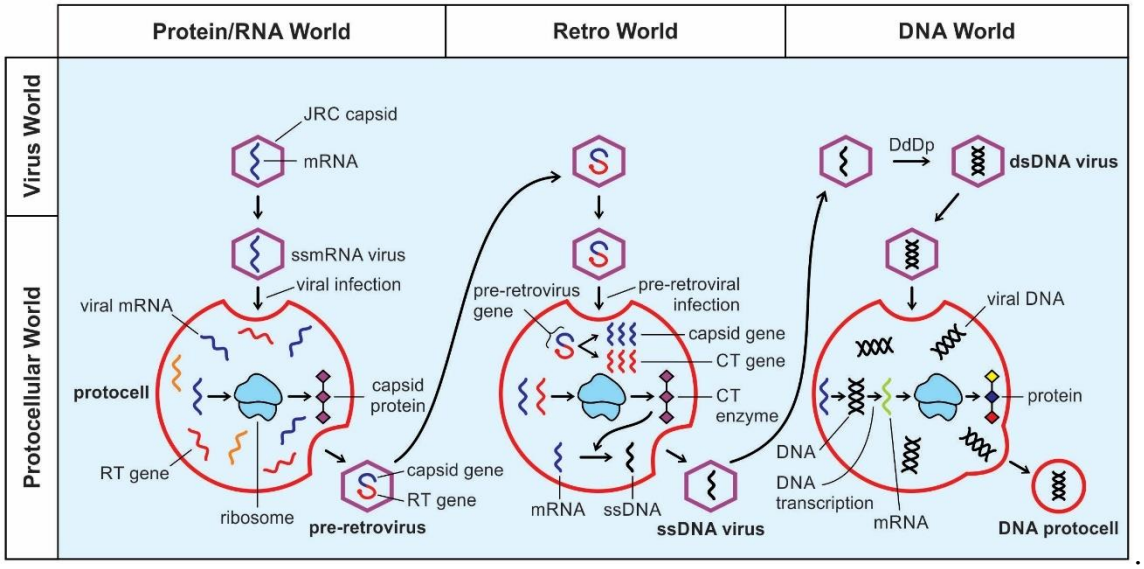

Figure 8. Postulated evolution of the biochemical pathways from the peptide/RNA world to the DNA world and the concurrent building of the information system to contemporary central dogma (DNA-makes-RNA-makes-proteins).

In the protein/RNA world, viral genes might have evolved *de novo* in the vent environment from pre-existing mRNA. These mRNA genes encoded JRC capsid protein for a protective shell for stability and durability to form the first virus inside protocells (Fig. 5). These ancient mRNA viruses began to infect protocells to increase their own populations in the gene pool (Fig. 6).

In the Retro world, pre-retroviruses developed the ability to stitch two genes into one. The first gene encoded the JRC capsid protein, but the second gene encoded the reverse transcriptase (RT) enzyme. The RT enzyme could generate complementary DNA from an mRNA template. The pre-retrovirus would play the crucial role for conversion of RNA to DNA during recurrent protocell infection, and subsequently incorporated into the host genome as a provirus. The reverse transcriptase enzyme, also called RNA-directed DNA polymerase, that catalyzed the conversion of RNA to DNA was not present in the host protocells, but was delivered by the pre-retroviruses that converted RNA of the host genome directly into DNA. The RT enzyme performed three sequential biochemical activities: RNA-dependent DNA polymerase activity, ribonuclease H (RNase H), and DNA-dependent DNA polymerase activity. Collectively these activities enabled the RT enzyme to convert single-stranded RNA (ssRNA) into single-stranded DNA (ssDNA). In our model, pre-retroviruses gradually modified its mRNA to DNA during recurrent retroviral infection. As soon as a pre-retrovirus invaded a protocell by endocytosis, its capsid coat was dissolved. The capsid gene segment would be translated into capsid protein and the RT gene segment into the RT enzyme, thus exploiting the ribosome of the host. The RT enzyme would then transcribe its own single-stranded RNA template to single-stranded DNA (ssDNA); the RT would make another strand of cDNA from another mRNA; two strands of cDNA then combined to make a double-stranded DNA (ssDNA). The conversion of ssDNA to double-stranded DNA was mediated by DNA polymerase (DdDp) (Fig. 9). The information contained in a retroviral gene is thus used to generate the corresponding protein via the sequence:

mRNA —> DNA —> mRNA —> protein.

In the DNA world, once the dsDNA virus appeared, the role of the RNA pre-retrovirus was gradually replaced by DNA viruses. DNA viruses possessed a full set of independent DNA replication enzymes: a helicase to unwind the DNA helix, two DNA polymerases (DNA Pols) to replicate the two strands, and a primase to form RNA primers that DNA Pols extended. As dsDNA viruses infected an RNA protocell, its DNA genome was replicated into multiple copies. Each DNA genome then replaced the mRNA of the protocell, transcribed to the new generation of mRNA that translated into protein. This is the beginning of the DNA world, when DNA replaced mRNA as the major genome of the protocell. The DNA protocell followed a conventional pathway of flow information as in a cell: DNA—> mRNA —> protein.

Time

Prebiotic stages

STAGES OF BIOGENESIS	BIOCHEMICAL PATHWAYS	FLOW OF INFORMATION
VI. Biological stage: Cellular world	First cells → LUCA → 3 domains: Bacteria, Archaea, Eukaryota **cell reproduction**	DNA → mRNA → protein
V. DNA world	DNA virus → DNA protocells **protocell infection**	DNA → mRNA → protein
IV. Retro world Pre-retrovirus	Pre-retrovirus/mRNA protocells → DNA virus **protocell infection**	mRNA $\xrightarrow{RT}$ DNA → mRNA → protein
III. Ancient virus world	Pre-virus/mRNA protocells → mRNA virus **protocell infection**	mRNA → protein
II. Protein/RNA world	Plasma membrane, cytoplasm	mRNA → protein
I. Peptide/RNA world	Protocells with translation machinery	mRNA → protein

Figure 9. Coevolution of viruses and protocells. Viruses became important vectors for donating critical enzymes and modifying genomes of the protocells during recurrent infection. Later, some of those viruses evolved DNA as a way to defend their genomes from attack, and DNA-based viruses became incorporated into hosts. In between are the fundamental catalytic processes that allowed to stepwise generate viral deoxyribonucleotides from ribonucleotides by RNA polymerase (RdRp), reverse transcriptase (RT), and DNA polymerase (DdRp) enzymes.

The abundance of genetic parasites along with the presence of defense systems in all cellular life forms strongly suggest their coevolution might have started in the protocellular stage. The DNA viruses might have emerged as a novel survival strategy. When the RNA protocells were confronted with an invading pre-retrovirus, they might have protected themselves by a number of defense mechanisms and developed an ancestral immune system. Immunity of viral infection allowed the RNA protocells to proliferate. In response, pre-retroviruses might have invented DNA to ward off attacks from the hosts. For pre-retroviruses, DNA might have offered a very powerful, immediate benefit: That pre-retroviruses might have immediate fitness benefit for substituting DNA replacing its RNA genome [64].

It thus appears that the transition from RNA to DNA genomes occurred in the viral world, and that protocellular DNA and its replication machineries originated via transfers from DNA viruses to RNA protocells [62,64]. The DNA virus living in a carrier state in an RNA protocell probably lost the genes for capsid proteins and became established as DNA plasmids. These plasmids were later transferred to the RNA protocell and incrementally transformed into DNA by recurrent infection. In such a scenario, the RNA protocell was transformed, from within, into the DNA protocell cell [49]. With DNA genome and its transcribed mRNA and ribosome, protocells began to function as a DNA protocell that began to synthesize protein. The coded genetic information began to flow from DNA to RNA to protein, beginning the classical central dogma of molecular biology.

There must have been a long transition period during which DNA in a protocell, helped by natural selection, progressively took over the replicative and storage functions of RNA. The DNA genomes are usually far larger than those of the RNA that allowed more storage capacity for

biological information. Once deoxyribonucleotide triphosphates (dNTPs) were available, it is likely that their assembly into DNA-like chains would have followed fairly rapidly. DNA has intrinsically higher replication fidelity, which allows genomes to increase in size and therefore complexity [16].

7. The Advent of DNA

DNA was derived from RNA [16]. In fact, these two molecules are so similar that transformation of RNA to DNA requires two constituents of RNA to be replaced by two close relatives: ribose by deoxyribose, and uracil (U) by thymine (T). As to the information, RNA could be transferred to DNA by reverse transcription, as happens in cells infected with retroviruses. We have already discussed the origin of the DNA virus from the RNA protocell by pre-retroviruses. Although the origin of DNA from RNA via retrovirus is widely accepted [32,49,52], there are some dissenters. The latest twist in the origin of DNA debate is that RNA and DNA might have appeared together in the prebiotic world from building blocks without the assistance of viruses.

Powner et al. [63] proposed a novel pathway for the prebiotic synthesis of several micro components in the assembly of DNA molecules from a mixture of the chemicals thought to have been present in the sulfur-rich prebiotic environment. In this environment both RNAs and DNAs emerged simultaneously, not one after another. The authors argued that the switching of RNA nucleotides into DNA nucleotides needed special enzymes that were costly to produce in terms of energy and material. On the other hand, if DNA molecules were present alongside RNA molecules, this problem of fundamental switching from conversion of RNA to DNA could be solved. The possibility of an abiotic route for the synthesis of deoxyribonucleic acid, according to these authors, provides a new perspective for the origin of DNA.

A similar view has been expressed by other authors. For example, Xu et al. [65] suggested *de novo* assembly of DNA from the building blocks of life in the prebiotic environment. They identified a compound presumably present in the prebiotic Earth called thiouridine that could have linked DNA nucleotides into chain-like DNA. Prebiotic phosphorylation of the 2-thiouridine molecule gave rise to nucleotides of DNA via photoreduction. In their view, both RNA and DNA may have arisen all at once in first life forms.

The possibility that DNA and RNA might have evolved concurrently appears to be the less parsimonious explanation than DNA following RNA. The DNA must have appeared later than the RNA molecule, because RNA degrades and mutates easily [6]. The backbone of a single-stranded RNA molecule is much less stable than the equivalent structure in a double-stranded DNA molecule. Its enhanced stability and longer molecular sequence give DNA greater fidelity and increased memory in its information storage system. RNA replication is intrinsically error-prone compared with DNA replication. DNA was selected over RNA based on its expanded capacity to store information and its dramatically improved error rate during replication [15].

Recent detection of ribose sugar in the Murchison meteorite and other primitive carbonaceous chondrites (NWA 801 and NWA 7020) by an international team of scientists clearly suggests that RNA evolved before DNA, not concurrently [66]. Ribose is a crucial component of RNA, which could have stored information and catalyzed reactions during prebiotic synthesis. The research provides the first direct evidence of ribose in space and the delivery of the sugar to Earth during impacts of carbonaceous meteorites about four billion years ago. It is remarkable that a molecule as fragile as ribose could be detected in such ancient meteorites. In contrast, the sugar in DNA (deoxyribose) was not detected in any of the meteorites analyzed in this study. This finding is important because there could have been a delivery bias of extraterrestrial ribose to the early Earth which is consistent with the hypothesis that RNA evolved first. According to the team, the next logical step would be to investigate the chirality of sugars in more carbonaceous chondrites to check whether the cosmic sugar is right-handed or not. Perhaps, homochirality evolved later in the hydrothermal vent environment [1]. The extraterrestrial ribose might have contributed to the formation of RNA on the prebiotic Earth, which possibly led to the origin of DNA by retroviral infection [36,49].

7.1. The Evolutionary Advantage of DNA over RNA

As the genetic diversity increased and the translation perfected, DNA protocells must have faced dilemmas: how to increase the information storage capacity of mRNA, and how to separate translation from replication? Unlike RNA, DNA is a poor catalyst. RNA acted as both template and catalyst, whereas DNA can function only as a template. There was only one way out: division of labor between the RNA catalyst and the DNA template. Information storage and replication become the prerogative of DNA, while the utilization of this information for protein synthesis and other functions remains the province of RNA. This is when DNA took charge in protocellular function, when replication of information is entirely dissociated from its expression. Because genes no longer had to serve messengers, they could be joined together in strings of increasing length to maximize the storage capacity of information [16,67]. RNA is more versatile, acting in both the storage and replication of genetic information. However, during this transition from RNA to DNA, the DNA lost its catalytic power but transformed into a stable, long-term storage molecule for genetic data. The functional separation between replication and catalysis must have been a tremendous improvement in DNA molecules and favored by strong selective forces. The division of labor between template and catalyst is a fundamental attribute of all living systems. This fundamental property of life is believed to have been absent in the protein/RNA world. Another advantage of DNA is that it allows a selective expression of individual genes by way of transcription, keeping the DNA molecule intact. In contrast, during translation, each mRNA molecule is destroyed.

With the advent of DNA, all the genes could now be kept as stable, double-stranded threads, which are held together by a complementary base pairing and are twisted into a double helix. The extremely stable structure doesn't allow DNA to mutate rapidly, unlike RNA, thus converting it to an efficient information storage structure. Intact stretches of DNA have been recovered from fossil bones that are at least 700,000 years old. The DNA acts as a permanent record—a blueprint containing the information needed to build protein and run the cell. A segment of DNA is transcribed to short-lived mRNA for translation responsibility, but DNA preserved the information storage system intact. Storage in mRNA without the possibility of retrieval would have been useless. Hence the need for transcription. The stored information could thereby be recovered from mRNA for translation. Transcription is entirely dissociated from translation.

The DNA is completely protected, but mRNA strands are continually made, broken down, and recycled. Deoxyribose in its sugar-phosphate backbone makes chains of DNA chemically more stable than chains of RNA, so that much greater lengths of DNA can be maintained containing multiple information of genes without breakage. The DNA double helix replaced mRNA as a more stable molecule for storing the increased amounts of genetic information required by protocells. These distinctions enable the two molecules to work together and fulfill their essential roles during protein synthesis.

The capacity and durability of single-stranded mRNA molecules were severely limited, with a high rate of error during its replication. A backup copy of mRNA was required to protect the original code or restore the prototype if damaged. Because DNA is a double-stranded molecule, it is much more stable than mRNA, and the presence of two strands provides a way of repairing genes. A damaged copy of the gene is restored using the complementary strand, the second copy of the gene, as a guide. This need for a repair mechanism may have been the selection pressure driving the creation of the double strands in DNA, conferring both stability and immortality.

The RNA is much more susceptible to chemical transformation. The DNA backbone is less prone to hydrolysis because it lacks the nucleophilic 2-hydroxyl group, so it represents a more chemically stable material and is less prone to mutation. As a result, DNA replicates with greater intrinsic fidelity than RNA, thus allowing more information storage. RNA polymerase, the enzyme that generates RNA from a DNA template, has no proofreading activity, and whereas DNA has many repair mechanisms, RNA has none. DNA soon superseded RNA as the carrier of genes and became the dominant bearer of genetic information. Being much longer than RNA, the DNA can store information on thousands of genes. Once DNA was established, it allowed genes to become longer and more complex. The fidelity of DNA replication is orders of magnitude greater than that of RNA replication. Unlike the single-stranded RNA molecule, in the DNA molecule, two antiparallel strands

that are complementary in their nucleotide sequences are paired in a right-handed double helix, with about 10 nucleotide pairs per helical turn [67-69].

Once DNA molecules appeared on the scene in the DNA world, they took on the role of the primary genetic molecules, superseding the RNA molecules, which became intermediaries between DNA and proteins. These new DNA-containing protocells rapidly diversified into large populations that easily outcompeted the RNA-based protocells.

7.2. DNA structure

The double helix structure of DNA, proposed by Watson and Crick [69], is an iconic image based on two paired DNA strands that are complementary in their nucleotide sequence. The immortal coil of two intertwined strands creates a molecule with the shape of a twisted ladder. The sides of the ladder are made of the sugar-phosphate backbones of two strands. The four nitrogen bases are adenine (A), thymine (T), guanine (G) and cytosine (C). They project off of the sugar-phosphate backbone and join together by hydrogen bonds, forming the 'rungs' of the ladder (Fig.10A). Each strand of DNA is a long sequence of four bases. The order of these bases is what determines DNA's instructions, or genetic code. Both strands of double-stranded DNA store the same biological information. The two chains are held together by hydrogen bonds. A pairs with T by forming two hydrogen bonds; and G pairs with C by forming three hydrogen bonds. Because the AT and GC pairs are equal length and fit identically into the double helix (like rungs on a ladder), the diameter of the helix is uniform, 2.0 nm (Fig. 10B). These two strands are complementary, but they go in opposite or antiparallel direction. The DNA sequence directly encodes each protein's structure, which determines its activity. The spiral is 'right-handed'—twisting upward as if driven by a right-handed screw, a chiral feature. All these features increase DNA's stability and thus its effectiveness as a reliable information-bearing molecule [68].

Solving the structure of DNA immediately revealed how the two fundamental processes of inheritance and mutation worked at a molecular level [68,69]. That is, a DNA sequence could be faithfully copied and passed on because of the base present at each position on one strand determined its complement on the other strand. Mutations result from errors in copying processes, where the wrong base or extra bases get inserted, or a base may be deleted, generating a change in the DNA sequence.

Two important functions evolved with the advent of DNA: transcription and replication. Transcription is the first step in gene expression. It involves copying a gene of the DNA sequence to make an mRNA molecule for protein synthesis. Replication, on the other hand, is the process by which the double-stranded DNA molecule is copied to produce two identical DNA molecules. As discussed earlier, the critical enzymes needed for transcription and replication were provided by viruses [49].

7.3. DNA Transcription

The storage of information in DNA without the possibility of retrieval would have been useless. The RNA can function as both template-directed polymerase and a template, whereas the DNA can function only as a template. As their genetic diversity and sophistication increased, protocells needed a division of labor between the DNA template and its catalyst mRNA to function properly. In other words, replication had to be separated from translation [67]. Therefore, a transcription mechanism emerged. The information stored in DNA must be recoverable in mRNA for translation. Transcription is the first step of DNA based protein synthesis, in which a particular segment of the DNA template is copied into a complementary mRNA by the DdRp enzyme (RNA polymerase). This is the transfer of genetic information from a segment of the DNA, called a gene, into mRNA. Transcription proceeds in 5′ to 3′ direction. That transcript, an mRNA gene, is then used to produce a protein.

The first step in the process is the unwinding and separation of the two strands of the DNA helix. RNA polymerase then travels along the length of the strand of DNA and binds complementary RNA nucleotides to it, until a complete strand of mRNA is formed, encoding at least one gene (Fig. 10C).

Once a gene is transcribed, the newly made mRNA is dissociated from the DNA template. As the mRNA molecule is formed, the DNA helix zips itself back together. This new generation of mRNA is then translated by ribosome to synthesize a protein chain. The DNA transcription is the beginning of the process that ultimately leads to the translation of the genetic code via mRNA. This is when the central dogma emerged combining the two-step process, transcription and translation, by which the information flows from DNA to mRNA to protein [5] (Fig. 10D).

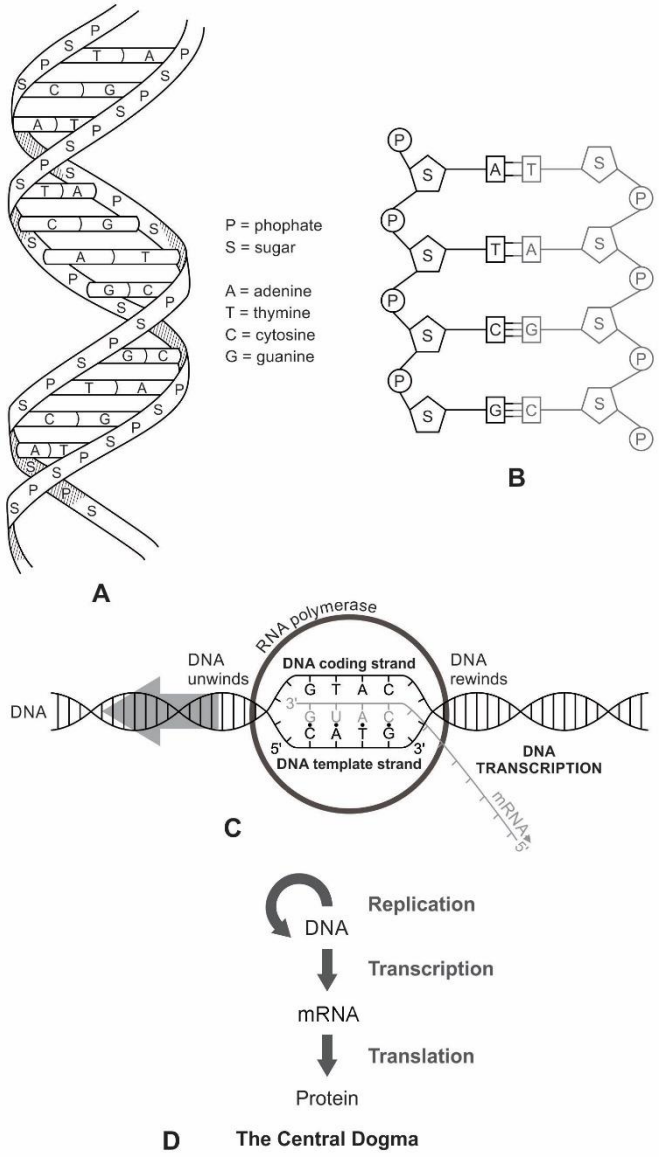

Figure 10. DNA structure and transcription. A, the twisted ladder model of the double helix of DNA showing two antiparallel strands, which are held together by hydrogen bonds between their bases. B, unwounded ladder structure of DNA showing that the sides of the ladder made sugar-phosphate backbones of two strands; the four bases joined by hydrogen bonds forming the rungs of the ladder. C, DNA transcription; C, mRNA protein synthesis begins when a region of DNA is teased apart, and a molecule of mRNA is built along one template strand by an enzyme called 'RNA polymerase'. When the mRNA transcript is formed, it peels away from the DNA, allowing the already transcribed DNA to rewind into a double helix. The mRNA conveys the genetic information from the DNA to the ribosome, where it specifies a sequence of amino acids that will form a particular protein, in a process called translation. The tRNA molecules bring the appropriate amino acids to assemble the protein. D, the central dogma of molecular biology states that 'DNA encodes RNA and RNA encodes proteins. Thus, information flows in one direction when genes are expressed: from DNA to RNA to protein.

Although DNA generated new proteins, this production was balanced by the loss of proteins through their degradation. The interaction between proteins and DNA was of mutual benefit and was important for the survival of the protocell.

7.4. DNA Replication

DNA replication is basically a polymerization process that requires a template and a primer, and the product of the reaction is a new strand of DNA that is complementary to the template strand. DNA replication is central to the reproduction of all cellular life forms and many viruses [69]. Every time a cell divides, the DNA polymerases are required to help duplicate the cell's DNA, so that a copy of the original DNA molecules can be passed to each daughter cell. In this way, genetic information is passed from generation to generation. Cellular DNA replication systems are broadly classified into two types, one in Bacteria and the other in Archaea/Eukaryota. In contrast, double-stranded DNA viruses have a much broader diversity of DNA replication systems. Both Forterre [40] and Koonin [50] suggest that the protocells acquired the DNA replication core enzymes, such as helicase, primase, and DNA polymerases (Pols), from the DNA viruses during their recurrent infection of protocells.

The double helix structure of DNA had striking implications for the processes of DNA replication. There are some similarities between transcription and replication. In both cases, the DNA double helix is untwisted when the hydrogen bonds between the bases are broken. However, there are major differences between these two processes. Transcription copies the DNA into the mRNA, whereas replication makes another copy of the DNA. In transcription, strand separation is mediated by the RNA polymerase, whereas in replication, the DNA polymerase takes this role. Both processes involve the generation of a new molecule of nucleic acid, either mRNA or DNA. However, the function of each process is very different, with one involved in gene expression and the other in cell division. Transcription was the precursor to replication. DNA replication must have evolved as a prelude to cell division.

The DNA replication process is well-known in literature [68] and its basic idea will be discussed briefly to highlight its role in the first cell division and the origin of life. The antiparallel structure of DNA is important in DNA replication because it replicates the leading strand in one way and the lagging strand the other way. The DNA replication is performed by the replisome, a complex molecular machine composed of numerous enzymes. It is a complex process that requires the coordinated effort of a team of enzymes for unwinding, separation, replication, and rewinding the helix. The replisome is composed of two replicative polymerase complexes, one of which synthesizes the leading strand, while the other synthesizes the lagging strand. DNA replication requires at a minimum, a helicase to unwind the DNA duplex, two DNA polymerases (Pols) to replicate the two DNA strands, and primase to form RNA primers that DNA Pols extends. Replication occurs in three major steps: the opening of the double helix and separation of the DNA strands, the priming of the template strand, and the assembly of the new DNA segment.

During separation, the two strands of DNA double helix uncoiled at a specific location called the origin. The initiation of the DNA replication occurs when an initiator enzyme called a helicase unwinds a short stretch of the DNA double helix using the energy of ATP hydrolysis that breaks apart the hydrogen bonds between the bases of the DNA strands and opens up like a zipper in one direction at the Y junction, or replication fork, to form two strands: a continuous leading strand and a small, discontinuous lagging strand. In DNA, these two strands are antiparallel, the 3′ end of the leading strand is paired with the 5′end of the lagging strand. Each strand serves as a template for a new strand. (Fig. 11).

In the next step, several enzymes and proteins work together to prepare, or prime the strands for duplication. A doughnut-shaped DNA polymerase requires a primer—a starter strand of RNA to which they can add new nucleotides. An enzyme called primase synthesizes a short stretch of RNA that acts as a primer for the DNA polymerase; this RNA primer can simply match ribonucleotides directly by complementary base pairing on single-stranded DNA.

Synthesis of the leading strand is straightforward after an RNA primer is in place. A DNA helicase, powered by ATP hydrolysis, propels itself rapidly along the leading template DNA strand forcing the DNA to open the DNA helix ahead of the replication fork. The helicase moves into the replication fork, which unzips ahead of it by another enzyme called topoisomerase that relieves the twisting forces. As the replication fork moves, the DNA polymerase can move continuously along this arm of the Y in 5′ —> 3′ chemical direction, adding complementary nucleotides to the 3′ end of that strand to produce a new daughter DNA molecule. As the replication continues, it creates two double-stranded helices, which are an exact copy of each other.

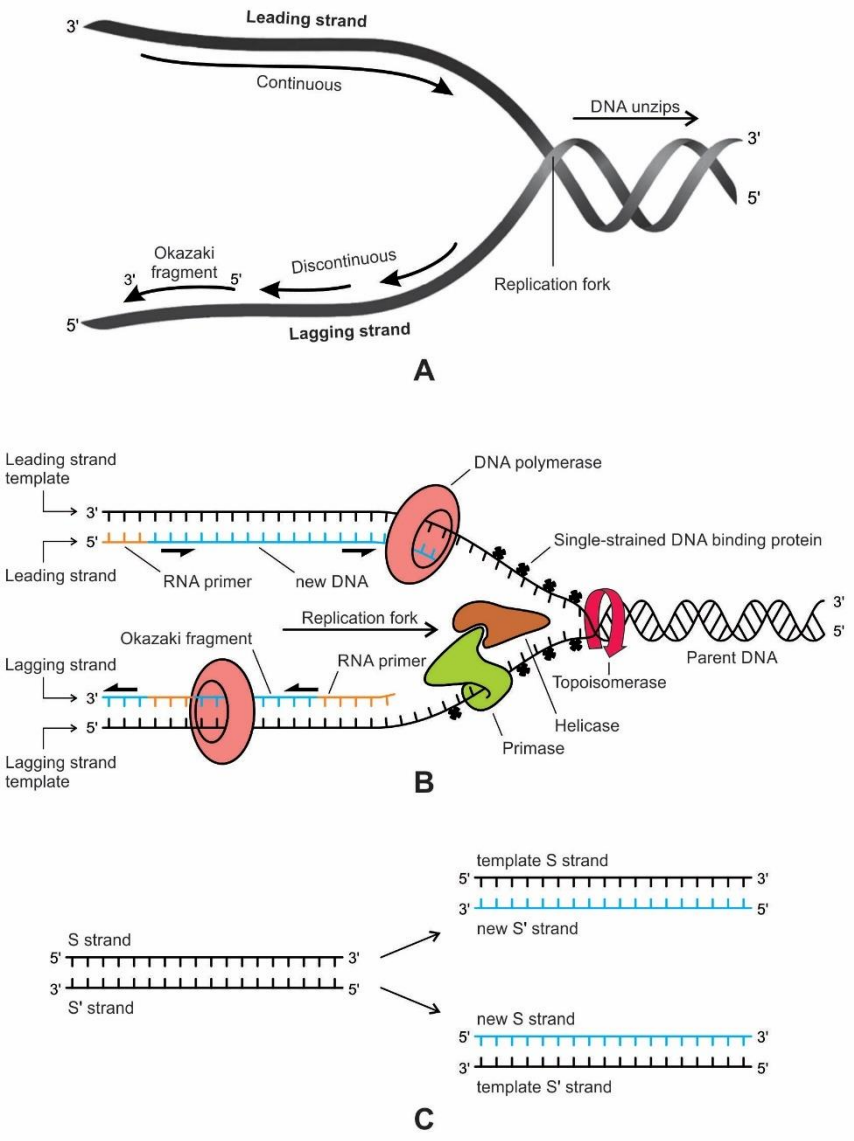

Figure 11. Replication of DNA by DNA polymerase. A, to begin DNA replication, DNA helicase enzyme causes the two parent DNA strands to unwind and separate from one another to form a Y-shaped replication fork. Both new strands are synthesized in the 5′-to-3′ direction. The leading strand grows continuously forward, but the lagging strand grows in short discontinuous stretches called Okazaki fragments. B, many core enzymes—helicase, primase, and DNA polymerases collaborate at the replication fork and are involved in DNA replication. C, primordial DNA replication in protocell; Here, the short nucleobases of DNA were uncoiled and completely separated into two strands as S and S′, S can serve as a template for making a new strand S′, while strand S′ can serve as template for making a new strand S. In this way double-helical DNA can be copied precisely.

The synthesis of the lagging strand is more complicated, because the DNA strand here is antiparallel, so that the DNA polymerase must work in the opposite of the replication fork, 3′—> 5′

chemical direction. The synthesis of the lagging strand starts when a primase synthesizes a short stretch of RNA that acts as a primer. Synthesis requires repeated priming and extension of the lagging strand discontinuously as a series of Okazaki fragments (Fig. 11B). The DNA polymerase then adds bases to the 3′end of the lagging strand. Here the synthesis is discontinuous, where the DNA polymerase can synthesize relatively small, discontinuous stretches in 3′—> 5′ direction. These short stretches of new DNA are called Okazaki fragments, which are then linked by ligase enzyme to a continuous whole. Once all the bases are matched up, an enzyme called Ribonuclease H strips away the RNA primers. The gaps where the primers were then are filled by yet more complementary nucleotides. Finally, an enzyme called DNA ligase seals up the sequences of the DNA into two continuous double strands [68].

The result of the DNA replication is two DNA molecules consisting of one new and one old chain of nucleotides. This is why DNA replication is described as semi-conservative, half the chain is part of the original molecule, half is brand new. Following replication, the new DNA is automatically coiled tightly around an axis to form a double helix for stability and greater compactness. In Bacterium, circular DNA is a higher-order helix-upon-a-helix, known as a superhelix.

The DNA replication apparatus has evolved over billions of years through 'trial and error' since its inception in the protocell in the protein/RNA world. The complex process by which DNA is replicated today could not have been the original version of replication, so a simpler, more primitive mechanism remains to be discovered. One possibility is that in the DNA genome of the protocell was very small to a bare minimum, with few hundreds of nucleotides where replication was less cumbersome, requiring few unwinding and rewinding enzymes and DNA polymerases for replication of each strand (Fig. 11C). It required two replisomes for bidirectional replication. These short nucleobases of two DNA strands were uncoiled and completely separated into two strands, each strand can serve as a template for making a new complementary strand. The net result was formation of two new double-stranded DNA sequences that were exact copies of the original double-stranded DNA. In this way, the double-helical DNA could be copied precisely.

As more and more nucleobases were added to the DNA with time to make longer and longer strands, one enzyme after another was selected and added to the mix of primordial enzymes to the replication fork, presumably for fine tuning to increase the speed, control or overall accuracy of the replication process.

8. Biological Stage

8.1. What is Life?

The most unique feature of Earth is the existence of life. Life's biggest event, of course was its invention, its very beginnings four billion years ago. Life, as we know it today, is an outcome of interactions between genetic opportunities, metabolic capabilities, and environmental changes. How to define 'life' is a sweeping question that affects whole branches of biology, biochemistry, genetics, and ultimately, the search for life elsewhere in the Universe. Yet, there is currently no consensus regarding the definition of life. We need a working definition of life so that we can identify the first cells from their protocellular ancestors. In 1944, the Nobel physicist Erwin Schrödinger, prior to the discovery of DNA, defined life as that which 'avoids the decay into equilibrium.' [70]. This definition refers to the *Second Law of thermodynamics*, which says that entropy always increases. But living things, said Schrödinger, are able to postpone this trend by taking its nutrients by metabolism, an open system, unlike the closed system of the physical world. Life is an open system that exchanges matter and energy with the surrounding environment. However, attempts to define life with thermodynamics have so far failed to distinguish clearly between the living and the non-living. A fire also fits this definition, but it is not considered 'living' [71].

Schrödinger argued correctly that organisms must run a sort of computer program, which is what we now call the genome. He speculated that genes were some kind of 'aperiodic solid' that contained some version of an 'elaborate code-script' that specified all of the future development of the organism. Von Neumann [72] following this argument of genetic software, found striking

similarity between software/hardware of a cell and those of a computer, and created 'Universal Constructor' to mimic the cell division. Today, all living cells utilize DNA as the replicable repository of genetic information, express the information by transcription of the DNA to the mRNA, and translate the mRNA into proteins by the same metabolisms by the same mechanisms, including the same genetic code [16]. Thus, the genetic code was the 'elaborate code-script' conjectured by Schrödinger.

Schrödinger's insight inspired many molecular biologists, including James Watson and Francis Crick, who not only discovered the double helix structure of DNA but also embraced the idea that genes contained a code by which cells turn that information into proteins [69]. Information is the currency of life, where information flows from DNA to RNA to proteins. Living systems store and process the information and transmit it to their offspring. Information is the language of life. Recently, it has been argued that the genetic software provides the singular definition regarding what life is [73]. Biological information separates life from non-life. In this view, life emerged in that instant when information gained control over the biomolecules. It has become clear in recent years that the biological world is computational at its core. Algorithms, or instruction sets to synthesize proteins, are found in every cell. Digital storage of the molecular information system is the key to defining life and understanding its origin. The key mechanism is the origin of the genetic code. The algorithms of life are a much more complex system than today's most sophisticated computer. The information system emerged about four billion years ago during prebiotic synthesis and is still working perfectly [9]. It is universal in all life forms from bacteria to humans. The information-directed protein synthesis is a unique signature of life. Although it is difficult to define what makes life so distinctive, there is a general agreement that its informational aspect is a key property, perhaps the key property.

But there are other attributes of life too. All life has a cell membrane, nucleic acids for replication, and proteins for metabolism. Life maintains itself by making more of itself. Life reproduces life. Life's continuous production of itself, called *autopoiesis*, is a unique hallmark [75]. DNA is an unquestionably important molecule for life, but the molecule itself is not alive. When a DNA molecule produces another DNA molecule, it is replication. When a cell divides into identical daughter cells, we call it reproduction [74]. This is why viruses are not alive. They are not autopoietic. Too small to self-maintain, viruses do not metabolize. Viruses do nothing until they enter an autopoietic entity: a bacterial cell, the cell of an animal or plant. They lack sufficient genes and translation machinery to maintain themselves. Thus, the definition of life is not straightforward.

A living system integrates three critical functionalities in an interactive system: (1) the *cell* membrane maintains an identity over time by localizing all its components and protecting them from the environment; (2) the *protein* uses free energy from its environmental resources in order to maintain itself and grow by metabolism, (3) and the *DNA* carries inheritable information and controls cell division. In order to reproduce successfully, a cell must be able to copy and transmit all of its genetic information to each of its daughter cells. The DNA replication allows cells to do this. During cell division, the replication of the DNA molecule makes occasional errors (mutations). The altered base sequences then produce the variations in a population of cells that are essential for evolution by natural selection. The ability to evolve is the final aspect of any definition of life.

Although we have an intuitive understanding of what it means for something to be alive, it is difficult to come up with a precise definition of life. Mindful of the centrality of evolution, NASA defined life as 'A self-sustaining chemical system capable of Darwinian evolution.' This minimalist definition captures the essence of life. One of the most distinctive characteristics of this definition of life is the concept of Darwinian evolution to establish a new level of a biological system. NASA's definition has a universal application which is not limited to Earth where RNA/DNA/protein-based life emerged [71]. According to this definition, life combines three distinctive characteristics. First, any form of life must be a chemical system. Second, life also grows and sustains itself by metabolism. Finally, living entities display genetic variation—raw material for Darwinian natural selection. This minimalist definition captures the essence of life. While this definition is open enough to include a wide range of potential life forms anywhere in the Universe, it also makes it hard to design a simple

test for life or identifying a fossil. This definition does not require compartmentalization. In the prebiotic system, compartmentalization played the crucial role that led to the origin of life.

There is a practical drawback to the Darwinian definition, when searching for life on other planets: how long do we wait for a system to demonstrate that it is capable of Darwinian evolution, and under what conditions? Or, if we find a microfossil in the Archean hydrothermal chert, or in the crater basin of Mars, can we use this definition to ascertain whether this fossil is the primordial life? Probably not. We need some morphological attributes also. Life is what is common to all living beings, which are made of cells. Here, I have expanded NASA's definition of life based on its attributes: 'Life is a self-sustaining, DNA-based information system enclosed in a plasma cell membrane, which is capable of reproduction and Darwinian evolution.' In case of the microfossil mentioned above, if it retains a distinctive cell membrane and a distinctive shape (spherical, rod-shaped, or spiral), our arguments for identifying primordial life become more robust.

There is no bounded entity smaller than a cell that is capable of independent reproduction. Thus, the RNA cell, often depicted as the first cell [40], does not fit into our definition of life. We regard the RNA cell as a protocell, a long way from the first cell. The lonely RNA cell, without the assistance of peptides, could not create the genetic code, the soul of the biological information system [9,13-15]. Life might have originated from interactions between RNA and peptides. Although replicating molecules, like nucleic acids, and catalytic molecules, like enzymes, are essential for life, by themselves they are not alive. Life is when these components are highly organized and contain specialized coordinated components into an interactive system, then it takes on the properties of life. In our view, the first cell was a highly organized living system using DNA and RNA as genetic material, 20 genetically encoded amino-acids, ribosome for template-directed protein synthesis, and membranes that allowed for chemiosmotic coupling.

Many of the problems in defining life boil down to the fact that we have only one example—life on Earth. The Earth is the only planet currently known to support life. We know what is life from a single sample from our planet, which is surprisingly quite uniform. As revealed by its remarkable biochemical and morphological similarities, life on Earth has a common origin. All Earth-based organisms use cell membranes that separate the interior of a cell from the outside environment, nucleic acids for hereditary information, proteins to control biochemical reaction and catalytic activity, and identical phosphorous-containing molecules (such as ATP) to store energy. It's the same biochemistry from bacterium to human [71]. If life is detected on Mars or other celestial bodies inside our Solar System, perhaps our definition of life would be refined or modified.

8.2. First Cells

The origin of modern cells is arguably the most challenging and important problem in the field of Molecular Biology. Cellular life depends on a number of fundamental properties that must have been achieved step-by-step in protocells. The ability to divide must have been a property of protocells from the moment of their existence. However, the mechanism of cell division with the DNA replication creating identical daughter cells is a breakthrough innovation that defines the first life.

The fossil record suggests that the first cells, a signal of a prebiotic evolutionary event appeared on Earth almost four billion years ago. Morphologically, these Archean microfossils look like bacteria [76]. The central question about the origin of life has been how the first cells arose from the DNA protocells and how these first cells began to reproduce, proliferate, and evolve. The birth of the living cell was the culmination of the prebiotic synthesis. It was the turning point in the history of Earth, transforming a rocky planet into *Gaia*—a living planet. The ability to multiply in exact copies of itself is what makes life so completely different from anything else in the known Universe. As discussed earlier, the prebiotic chemical evolution from cosmic ingredients to a single-celled living organism probably took place in the hydrothermal crater vent environment. These pioneer heat-loving microbes – hyperthermophiles – thrived around the volcanic vent of the crater that belched out several toxic gases such as CH_4, NH_3, SO_2 and H_2S, providing chemicals for nutrients and warmth. These hyperthermophiles in the vent environment were probably the first life forms on this planet [1].

Here I reconstruct the first cells from the combined evidence of the prebiotic evolutionary history, the Archean microfossils, and the morphology and function of modern bacteria. From the fossil record of the oldest microbes, it appears that the first cells were probably simple spheres or rods. They were likely a single-celled microbe with minimalist design, but they acquired the essential components of cellular structure for integration and reproduction. The plasma membrane was a simple, double-layered cell wall, and lacked extensive, complex, external and internal membrane systems. The outer membrane was leaky. It was filled with porin proteins, which formed small holes through the membrane allowing passive diffusion. These holes were big enough for nutrients and ions to pass through by passive diffusion, but small enough to keep the machinery of the cell inside. The membrane encased the cytoplasm in which all soluble components were packed inside. Like bacteria, the first cell might have contained only one circular chromosome in the cytoplasm – a large DNA molecule, packed closely in the nucleoid. Most likely, this chromosome was longer than the diameter of the cell in which it was encapsulated, so the circular configuration of the chromosome was a space-saving device for the crowded mass of DNA. To fit into the nucleoid, the DNA double helix coiled on itself with the aid of enzymes to form the highly compact 'superhelix' circular structure, which was enclosed tightly in a condensed area of nucleoid (Fig. 12). In addition to chromosomal DNA, there might be plasmid, another small, circular, double-stranded DNA molecule, which can replicate independently. Outside the chromosome, the most prominent structure in the soluble portion in the cytoplasm was dispersed ribosomes; other smaller molecules included various enzymes and other translation machinery for protein synthesis. Some molecules were devoted to producing energy from environments. Perhaps the flagella were not yet invented as the fossil record suggests; the first cell lacked this propelling device to power movement.

It was known for many years before the detailed structure of DNA was determined that DNA is replicated during cellular reproduction. It is important here to draw a sharp distinction between replication and reproduction. Molecules can replicate, but only cells can reproduce [48]. The first and most important attribute of life is reproduction. If something is alive (autopoietic), it can make a copy of itself. Reproduction with variation is an essential characteristic that distinguishes life from nonlife. The reproduction of the first cells was the watershed event in the history of life. Without reproduction, life would quickly come to an end, so reproduction is essential for the continuity of life. The biological stage was the tipping point, the boundary between the information age and the biological evolution. How do non-living biopolymers, such as membranes, nucleic acids, and proteins, morph into living organisms—organizing into cells, reproducing, then growing, and evolving? The first life was integrated molecular systems of interconnected parts to produce homeostasis and discovered ways to respond to stresses, both physical and chemical. The major components of the first life were a plasma membrane enclosing the core enzymes and information carrier genomes forming the integrated systems that included metabolizers and the ability to capture energy from vent environments. The first life was autopoietic.

Building new proteins was one of the main tasks of the first cells. Over half of the molecules inside the *Escherichia coli* cell are involved in one way or another with the synthesis of proteins [65]. The making of various types of proteins was one of the most important events for a cell because protein not only forms structural components of the cell; it also composes thousands of enzymes that catalyze the production of the remaining biomolecules necessary for life. In the early cell, different genes were active that produced only those proteins that were needed in the cell. All proteins in the first cell were encoded in one large circular DNA. The cell must control when and where each gene was used. The information held in the DNA genome was highly regulated. Perhaps, a host of repressors and activators interacted with each gene, determining when it would be used to create protein.

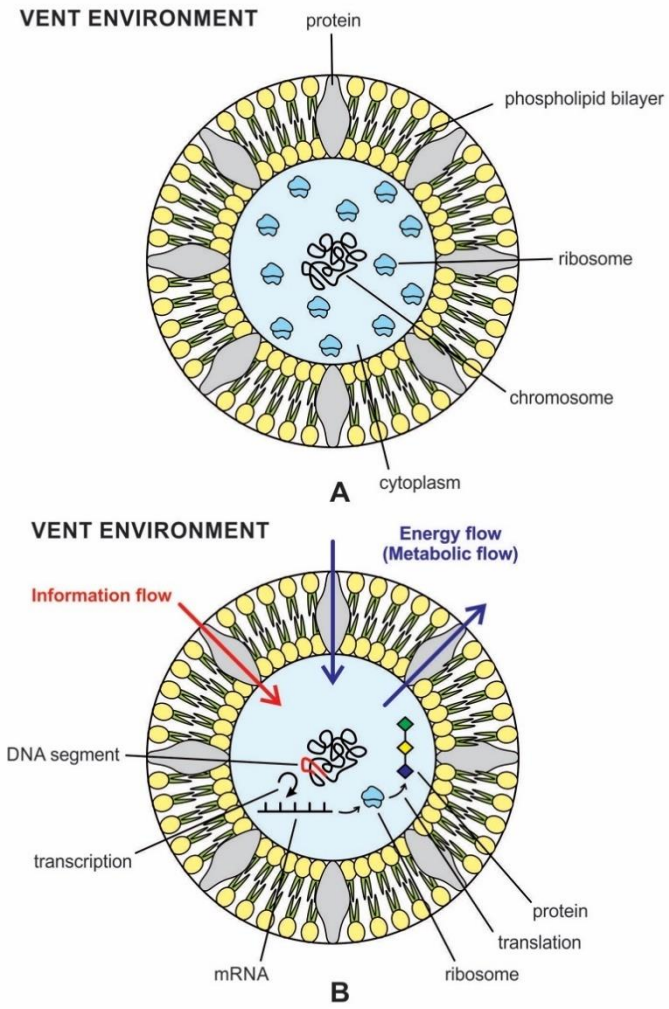

Figure 12. A, reconstruction of the hypothetical first cell. The plasma membrane encloses a primitive cytoplasm that contains a prominent circular chromosome at the center; outside the chromosome, the most prominent structure was numerous ribosomes; other molecules like RNAs, various enzymes and proteins, and other translation machineries were too small to show; B, energy flows and information (or signaling) flows are represented as the basic exchanges of a first cell with the outer/inner environment. The signal from vent environment is transduced when the receptor proteins located in plasma membrane bind to ligands on the outside of the cell and then causes the behavior on the inside of the cell. The signal from the environment controls gene regulation and expression: which genes to use and which not to use during synthesis of protein.

The first life orchestrated how to control gene expression at any step between the synthesis of the mRNA and the activation of the gene product. Gene expression can be controlled at any step between the synthesis of the mRNA and the activation of the first gene product. In the first cell, the prebiotic information systems became increasingly sophisticated to process more and more advanced levels of biological information. In the peptide/RNA world, information flows from the mRNA to the protein [4]. With the emergence of DNA, the central dogma is established; information flows from DNA to mRNA to protein. With the emergence of the first cell, information flows from cell to DNA to mRNA to protein.

8.3. Control of Gene Expression in First Cell

Simple cells were open systems regarding both energy flows and information flows. An information transfer pathway was basically different from the metabolic pathway, the former is unidirectional, the latter is bidirectional (Fig. 12B). In information pathway or signal transduction pathways, only an impulse is related from the environment to the cell membrane. First cells were typically exposed to an ever-changing environment in which nutrient availability might increase or decrease radically. First cells responded to such variations in their environment by altering their gene expression pattern, thus they expressed different enzymes depending on the nutrients available to them. Like the protocell, the first cell responded to the environmental fluctuations of the hydrothermal vent for survival and adaptation. For an early cell to survive and reproduce in this environment, it must use resources efficiently—particularly resources that provided energy and nutrients. It developed a signal transduction pathway to relay messages from the environment to the cytoplasm of the cell for responses. Signal transduction converted an extracellular environmental signal to an intracellular signal. The extracellular signal can be light, temperature, nutrients, chemicals, toxins, osmotic pressure, or cell density, but intracellular signals are chemical messengers. Like modern bacteria, perhaps a two-component signal pathway existed in the first cells. Receptor proteins located in the plasma membrane had binding sites for the signals they recognized. The signal was transduced when the receptor changed shape and became ready to cause a change inside the cell. The signal was then amplified when the receptor activated molecules called second messengers. These second messengers increased the concentration inside the cell, spread the original signal around the cell, and activated a particular gene of a DNA chromosome to be expressed in responses to environmental cues. First cells responded to environmental changes by turning on genes for making proteins that would help them to survive. Genes were turned on and off by DNA-binding proteins that bound to DNA and controlled transcription.

For a cell to function properly, necessary proteins must be synthesized at the proper time. First cells controlled and regulated the synthesis of protein from information encoded in their DNA. Gene regulation is about early cells choosing which genes to use and which not to use for expression. The regulation of gene expression conserves energy and space. It would require a significant amount of energy for an early cell to express genes at all times, so it is more energy efficient to turn on the genes only when they are required. First cells must have developed a complete set of genes in their DNA molecules for blueprints of various kinds of proteins needed for the survival and reproduction, yet these cells used only a fraction of the genes at any given moment. In general, a gene was expressed when its specific protein product was needed. First cells might change what genes they were using depending on signals or changes in the environment. Transcription and translation occurred almost simultaneously in the cytoplasm of first cells. Here the gene expression was regulated at the level of transcription (see Fig. 11 for detailed transcription mechanism). Transcription of a gene by RNA polymerase can be regulated by several mechanisms. Protocells might have already developed constitutive genes for the essential housekeeping proteins that were typically in use all the time. First cells, in addition to constitutive genes, developed regulated genes, which were turned on and off as they were needed by the cells.

Gene expression occurs in two essential steps: transcription and translation. In bacteria, the two processes are tightly coupled in time and space, and highly regulated in a fluctuating environment. In the bacterial cell, genes are organized into operons, or clusters of coregulated genes on the chromosome, where they are transcribed from one promoter (RNA polymerase binding site). In addition to being physically close to the genome, these genes are regulated such that they are all turned off or on together, when proteins are needed for a specific function. Other regulatory proteins are repressors and activators. Grouping related genes under a common control mechanism allows bacteria to rapidly adapt to change in the external environment, optimizing the cell at any given time for survival.

In the first cell, most likely gene expression was much simpler to control and modulate the single gene for one enzyme. In general, a gene was expressed only when its protein product was needed. Most likely, operons were gradually added as the first cell evolved and acquired more functional capability.

Seven steps occurred during the flow of information in the first cell for gene expression, represented by arrows in the following expression:
Environmental signal —> receptor proteins —> second messengers —> DNA—> mRNA—> protein —> activated protein.

The arrow from environmental signal to DNA represents the selection of a gene by the cell to make a particular protein. The arrow from environmental signal to DNA was relayed via receptor proteins and second messengers; the arrow from DNA to mRNA represents the transcription of that gene to mRNA. The arrow from mRNA to protein represents translation, in which ribosomes read information in the mRNA and use that information to synthesize a protein. The arrow from protein to activated protein represents post-translational modifications.

Life came to a climax when algorithms of cells took control. It began in the peptide/RNA world with the origin of the genetic code, then refined with DNA replication, and culminated with the emergence of the first cells. It is the informational role that is the key to transforming nonliving protocells into life. A key feature of biological information in algorithmic life was feedback from the environment. Changes in gene expression allowed first cells to respond to environmental changes. Gene expression could be controlled at three levels: transcription, translation, or post-translation (post-activation).

The emergence of the first cells was the defining moment in the history of our planet that made Earth unique—the *Blue Marble* in the Solar System. Once the first cells formed and began to reproduce, Darwinian evolution began to diversify early life into different species that invaded different environments. The evolution of reproduction in the first cells was the turning point in the early history of life, heralding the biological stage and ultimately leading to biodiversity. Life began to move from the hydrothermal crater basin to the mesophilic global ocean, diversified and proliferated [1].

8.4. Cell Division

Life is organized into cells that grow and divide. The advent of cell division into two identical cells with a complete set of genetic machinery defines the emergence of the first cells from their protocell precursors. Although protocells could divide by physical forces, it was not controlled by DNA replication. The chromosome replication was tightly linked to the first cell division. We can infer the origin of the primordial cell reproduction from the modern bacterial fission and protocell division.

To reproduce, bacterial cells divide into two genetically identical cells to propagate by binary fission. A common bacterium, *Escherichia coli*, which colonizes the mammalian intestine, is an excellent exemplar of the cell cycle and binary fission. The cell cycle of a bacterium is the series of events that takes place in a cell that lead to the duplication of its DNA, the segregation of copied DNA, the splitting of the parent cell's cytoplasm, and its division into two identical daughter cells. The genome consists of a single, circular DNA chromosome that must be replicated and then allocated into the daughter cells; the cytoplasmic contents must also be divided to give both new cells the machinery to sustain life. *Escherichia coli* can divide every 20 minutes [77].

Most bacterial cells spend their time cycling between a state of calm (interphase) and the dividing phase, in what is known as the 'cell cycle'. The starting point of replication, the origin, is close to the binding site of the chromosome to the plasma membrane (Fig. 13). A cell must grow to twice its original size during interphase, and this growth is coordinated with the duplication of the chromosome. In bacterial fission, the circular DNA chromosome is attached to the cell wall, near the midpoint of the cell. DNA replication occurs bidirectionally as the cell grows and elongates. Before binary fission occurs, the cell must replicate its chromosome and segregate these copies to opposite ends of the plasma membrane. The cell division and the DNA replication were coordinated in such a way that the distribution of new DNA copies to each daughter cell was ensured. The seemingly simple process of bacterial growth and division requires an impressive orchestration of functions by several regulatory proteins. These core cycle regulators can cause key events, such as DNA

replication or chromosome separation, to take place. They also make sure that the cell cycle events take place in the right order.

Bacterial cell division depends on the formation of a furrow in the plasma membrane with infolding of membrane and cell wall. A constriction develops by the Z-ring between the duplicated chromosomes so that one is included in each of the two daughter cells. The fact that the bacterial chromosome is anchored to the plasma membrane makes this separation possible without the construction of an elaborate mitotic apparatus. This contractile Z-ring orchestrates the separation of the chromosomes and the cell division in the bacterium. A septum is formed between the nucleoids, extending gradually from the periphery to the center of the cell. When the new cell walls are in place, the daughter cells separate, completing the cell division [68,78].

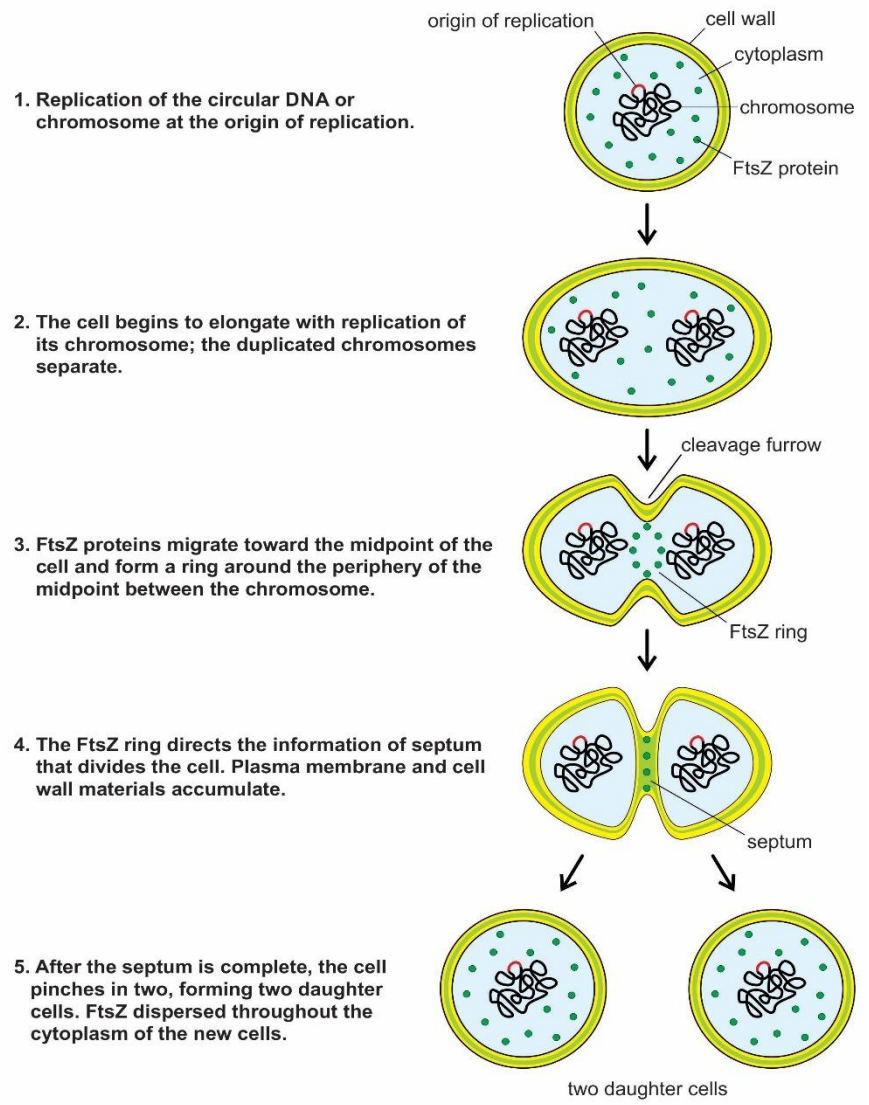

Figure 13. A hypothetical scheme of cell division in primitive first cell using modern bacterium as a guide. The cell-division cycle takes place in a cell leading to duplication of its DNA. The early cell must coordinate its growth, division, cell volume and shape with inheritance of genome. During the process, thousands of FtsZ molecules come together in the middle of the cell and form a circle-like structure known as the Z-Ring. Z-rings are produced at the middle of the cell division leading to constriction. The replicated chromosome and cytoplasm separate into two new identical daughter cells.

Laboratory simulations hint at a solution to the primitive cell division mechanism. Lipid vesicles extracted from the Murchison meteorite undergo spontaneous primitive cell division in the laboratory, with no external forces acting upon them [28]. When a mixture of these cosmic vesicles,

amino acids, and nucleic acids was shaken, the vesicles trapped the organic molecules inside them and began to interact. Therefore, these vesicles can apparently take in substances from outside themselves through their lipid walls, and use them to build new walls and new contents. A large vesicle mimics a primitive kind of cell division.

The physics of 'chemically active' droplets, which cycles chemicals in and out of the surrounding fluid may shed light on the origin of protocell division [80]. The team studied a theoretical model for behavior of a liquid droplet in a chemically disequilibrated system. This 'active droplet' behavior differs from passive and more familiar tendencies of oil droplets, which join together into bigger and bigger droplets without dividing. On the other hand, these chemically active droplets can grow to a stable size by taking resources from the environment. Droplet growth eventually leads to instabilities – linked to the changing shapes of the droplets. The droplet keeps elongating and pinches in at the middle, which has low surface areas. Eventually, surface tension causes it to spilt into a pair of droplets. This process of dividing droplets somewhat mimics the spontaneous vesicle division from the Murchison meteorite [28].

In a laboratory simulation, a genome-rich vesicle increased in size at the expense of an empty vesicle. When its greater size imposed too much osmotic stress, pearling instability developed, and the stretched vesicle divided into two, each daughter vesicle retaining some of the original genomic contents [79]. Recent work on model protocell membranes has demonstrated that vesicles can grow as filamentous structures and divide spontaneously under mild shear forces, and that with photochemical stimulation, a robust 'pearling' mechanism produces many small daughter vesicles [81]. Self-replicating membranes can divide either spontaneously or under the influence of external environmental forces [13], and high environmental shear forces can cause vesicles to divide.

Synthetic biologists use simple 'protocells' to study the origin of cell division, but previous models were not able to reproduce both the genome and the membrane sustainably. Kurihara et al. [81] proposed a recursive self-proliferating model protocell that represents a step towards eventual production of model protocells that are able to mimic cell division. They used a novel system by fusing the self-reproducing vesicles with feeder vesicles, thus allowing the vesicle composition to be sustained over multiple generations. Because of competition, the larger vesicle grows more quickly and fuses with the feeder vesicles. Thus, feeding the protocells by vesicle fusion offers a practical pathway for indefinite self-reproduction [82].

The mechanism of cell division is a complicated multistep process. A new study shows that a mutant bacterium can reproduce without a wall or division machinery, supporting the idea that primordial cells could have divided using physical mechanisms, such as simple shearing alone. Leaver et al. [83] generated a mutant strain of *Bacillus subtillis* that lacked cell walls. Although the Z-ring retains in this mutant form, it does not participate in cell division. They found that these cells divide without the Z-ring but by an extrusion-resolution mechanism. This novel form of cell division provides insights into how early forms of cellular life may have proliferated. This pattern is strikingly similar to 'pearling instability' seen in lipid vesicles. The study supports a model in which the constriction of the Z-rings is dependent on wall synthesis.

Although the structure of the bacterial cell is very different from that of the DNA protocell because of billions of years of evolution, we may assume some basic common mechanism of protocell division from bacterial fission. The protocell might have achieved a rudimentary form of cell division using physical mechanisms alone, and was modified with the availability of proteins. Initially, the DNA protocell division may have been an accidental burst, when the cell size increased as the cell accommodated more and more complex biopolymers, until it reached an unstable size. The surplus molecules generated inside the protocell, causing it to bulge, and the protocell continued to elongate and became pinched in the middle. Eventually, the surface tension caused the cell to split into a pair of daughter cells. As the cell divided, the two new cells were not necessarily identical, because there was no mechanism in place to ensure an equal distribution of the parent cell's contents. The dissimilarity in the daughter cells was an advantage at this stage of evolution because it promoted a diversity upon which natural selection could operate, but true cells required mechanisms that guaranteed identical daughter cells [84].

DNA replication was the major driver of symmetrical binary fission in early cells. However. it took many generations of DNA cells to invent the cell reproduction by countless trial and error. We can speculate how the DNA cell might have learned how to reproduce into an identical daughter cell. Perhaps, initially the cell grew in size, then elongated for cell instability, and pulled apart by physical forces, but divided asymmetrically without DNA replication, so one daughter cell might have intact DNA, while the other was devoid of it (Fig.14). In the second generation of the attempt, the DNA might have replicated, but it was not coordinated with cell division, so one daughter cell might have two chromosomes, the other got only cytoplasm without any DNA. Eventually, the DNA protocell learned how to coordinate DNA replication with cell division, probably aided by a few replicators. As DNA was replicated, the replicated DNA molecules were segregated at the two ends of the cell in an energy-dependent process, forcing the cell to grow, elongate, and pull apart, making the equatorial region highly stretched, weak, and constricted. DNA replication exerted osmotic pressure on the cell wall by increasing the concentrations of entrapped biomolecules, and DNA separation actively dragged the nucleotides apart. At that time cytoplasm was enriched with FtsZ protein that aided to constrict the membrane at the midpoint. The Z-ring established the location of the division site, acted as the scaffold for the division apparatus, and provided the contractile force to precisely orchestrate the binary fission. As the contractile Z-ring closed like a purse string, the protocell cytoplasm was divided into two, completing the cell division (Fig. 14). For the division to produce viable daughter cells, it must be coordinated in time and space with other major events of the cell cycle, such as DNA replication and segregation.

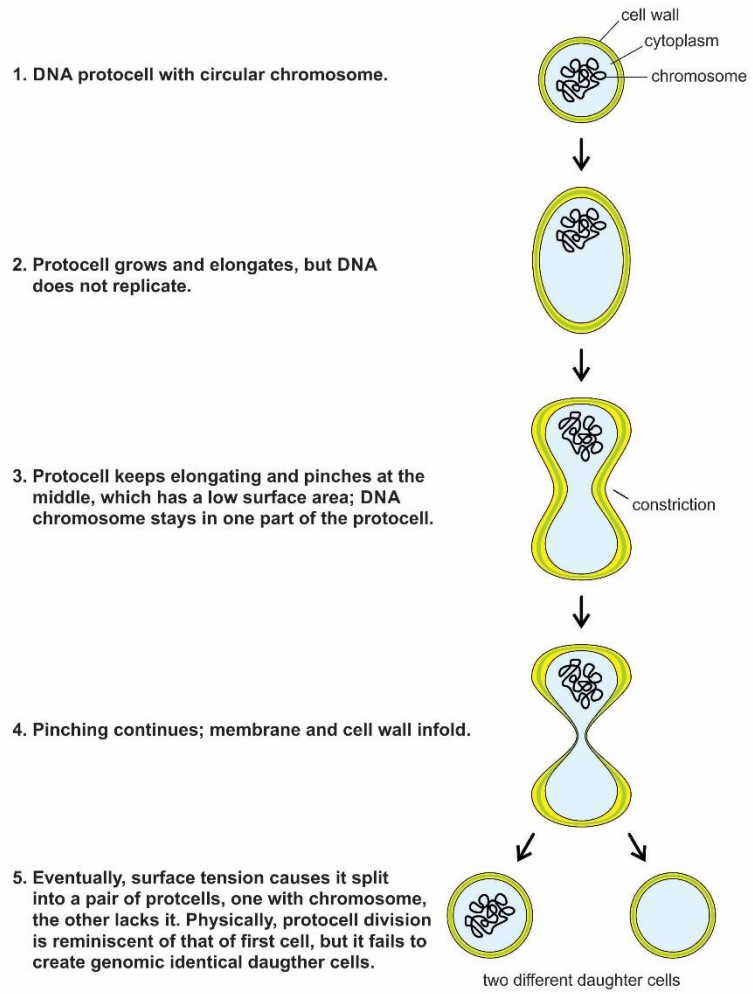

Figure 14. DNA replication was the major driver of symmetrical binary fission in early cells. Most likely there were many trials and errors before the cell division was perfected. Here a hypothetical scheme is shown where chromosome duplication was not perfected, thus giving rise to different kinds of daughter cells.

The first cell was self-sustaining, DNA-based and chemically sophisticated, possessing many of the housekeeping proteins, and capable of mutation and Darwinian evolution. It had developed capacities for harnessing energy from the hydrothermal environment. However, these primordial organisms were continuously infected by viruses and developed some immune systems for their survival. The coevolution of viruses and the first cells were the source of innovation, gene enrichment, and diversity of the early life. The first cells were stabilized, perfected in cell division, multiplied, mutated for innumerable generations, and spread across hydrothermal systems in the young Earth.

The first cells became the earliest self-sustaining organisms and their emergence was the turning point in the early history of life. Suddenly, the first cells began to reproduce and multiply quickly, colliding with each other and crowding the vent environment (Fig. 15). Reproduction was an essential life process for the first cells because it allowed them not only to survive but to continue as a population. Reproduction of the first life was the most momentous event in the early history of our planet, transferring the barren alien rock into a living world. Since then, life itself has in many ways helped to shape our planet. Earth is much more complex than all other Solar System objects that we know of because of life. As soon as life acquired the ability to spread and multiply, it also carried the potential to change the environment chemistry from which it had emerged. The early history of life on Earth may have been characterized by coevolution of microbial metabolism and atmospheric composition. Earth is the only planet we know of that can support life. This is an amazing fact, considering that it is made out of the same matter as other planets in our Solar System, was formed at the same time and through the same processes as every other planet, and gets its energy from the Sun.

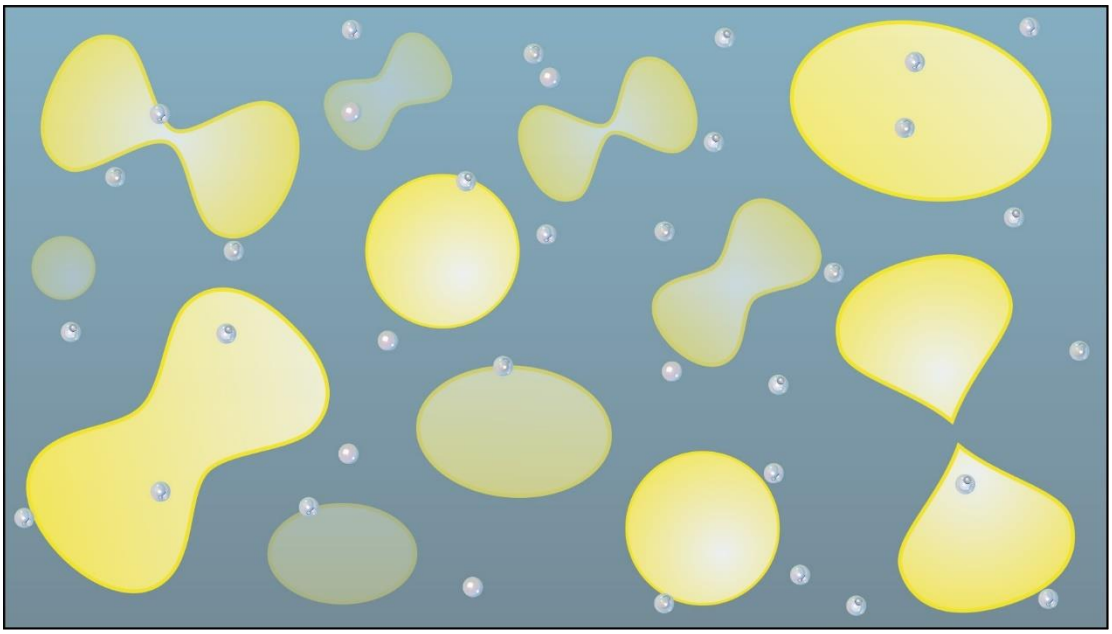

Figure 15. Once the cell division was perfected, the hydrothermal vent environment was crowded with a new generation of daughter cells.

8.5. Hyperthermophiles

The first traces of life on Earth dates back to the Early Archean, about four billion years ago. Although the recognition of ancient microfossils just by morphology turned out to be difficult, the geologic setting of these fossil-bearing strata indicates the hydrothermal environment. Only hyperthermophiles would have been able to thrive and survive in this extremely hot environment [1,85]. They have solved the protein and membrane problems: they possess specially heat-stable proteins and have high-melting fats in their membranes [86,87].

The habitats of hyperthermophiles (superheat-loving microbes), which are the most primitive living organisms, may shed new light on the oldest ecosystems on our planet—the cradle of life. Hyperthermophiles occupy the short deepest branches of the universal phylogenetic tree closest to

the root, making the hydrothermal ecosystem the most ancient continuously inhabited ecosystem on Earth. Hydrothermal systems prevailed throughout Earth's history, the extant organisms in these systems and their genomes are living records of changes over geologic time [86]. Discovered in 1977, submarine hydrothermal vents astounded many scientists when it was discovered that the hyperthermophilic bacteria and archaea thrive in these deep, dark, anaerobic, hostile and volcanic environments through their ability to utilize the chemical nutrients that arise from the hot vent fluids interfacing with cooler sea water as a source of energy [87]. Today, hyperthermophiles are found in geothermally heated subterranean rocks such as the boiling hot springs of Yellowstone National Park. They grow optimally above 80º C and exhibit an upper temperature border of growth up to 113º C. Hyperthermophiles have certain heat-stable enzymes (that are proved to be very important in biotechnology) and unusual rigid membranes that are specially geared to working in high temperatures. The cell membrane contains high levels of saturated fatty acids to retain its shape at high temperatures. Based on their growth requirements, hyperthermophiles were probably the most primitive living organisms that could have existed on early Earth about four billion years ago [85-87].

9. The LUCA: The Genetic Portrait of the Ancestor of All Life

Darwin's theory of common descent with modification is the central pillar of modern evolutionary biology [88]. It states that all life on Earth that has ever lived on Earth has descended from one original primordial form that diverged with time – like tree branches from a single trunk. Darwin recognized that species not only evolve but also divide. As species evolve, they split and diversify through time, increasing morphological divergence like the branching pattern of a tree. He presented a metaphorical tree in his book that showed how species change through time from the common ancestor. Darwin's tree is a visual representation that shows how species are related by a common descendent.

Darwin left to later biologists to figure out what the real evolutionary tree looked like. Before the development of sequence-based molecular methods, it was impossible to know the evolutionary relationships connecting all of life and thereby to draw a universal evolutionary tree. Although the topology of the tree of life has changed over time as more and more genetic and proteinic information of organisms are available, they all confirm Darwin's theory of common descent. A genetic portrait of the ancestor of all living things has been slowly emerging in recent years. This venerable ancestor, the last universal common ancestor or LUCA, was most likely a single-cell, bacterium-like organism that lived in the hydrothermal vent environment [89-94].

But how can we formally test the idea of common descent? The fact that living organisms have a common ancestor explains why they share many traits. There is a compelling list of circumstantial evidence – for example the universal genetic code and the universal use of the same biomolecules – DNA, RNA, and protein, and homochirality of certain key molecular structures from bacterium to human. The organisms that exist today contain numerous proteins that are clearly homologous, most notably ATP synthase. The organisms themselves must therefore be homologous, descended from common ancestors, the most recent of which we understand as LUCA. The proposition that all extant life is genetically related—is perhaps the most fundamental premise of modern evolutionary theory, providing a unifying foundation for all of life science.

Darwin's dream of the tree of life was realized on the grandest scale when Woese [89,91] proposed that all cellular life could be proposed into three separate fundamental groups of domains—the Bacteria, the Archaea, and the Eukarya, based upon sequence comparisons of ribosomal RNA (rRNA) sequences. According to the 'three-domains tree' the Eukarya and Archaea are more closely to each other than they are to Bacteria. The 'three-domains tree' is the most visible image depicting the diversity of cellular life, but it has not gone unchallenged. An alternative 'two-domains tree,' in which Eukaryotes are nested inside Archaea, has gathered support from recent phylogenetic analyses [90].

Molecular phylogenetics have now provided overwhelming evidence that all living organisms on Earth descended from a single ancestral form, the last universal common ancestor (LUCA). The concept of LUCA is central to the study of early evolution of life's origin, yet its nature and

phylogenetic placement in the tree of life are controversial. Woese [91] identified LUCA as 'progenote,' an organizational level similar to an RNA protocell that preceded with the first cell. A similar view of LUCA has been suggested by Forterre [49]. Today, LUCA is considered as a sophisticated organism with a complex structure recognizable as a true cell. LUCA is probably a bacterium-like organism that could help establish how early life on Earth got established. LUCA is not thought to be the *first* cell, but the *last* before the divergence of the two-domain tree – Bacteria and Archaea (Fig. 16B). There was life before LUCA, when the first cell was mutating and evolving into diverse populations [92]. LUCA is the most common ancestor of all living organisms, but only one of many early prokaryote cells, which is still extant, whereas the others became extinct.

The question of whether or not all the life on Earth has an ultimate common ancestor is a subtle one, complicated by the phenomenon of horizontal gene transfer (HGT). Present evidence suggests that this is a widespread phenomenon in the bacterial world. Some scientists believe that LUCA was not an organism, but a collection of diverse organisms exchanging their genes by rampant HGT more or less without constraint; LUCA in this view looks like a tangled tree of life contradicting its monophyletic origin [44]. The same objection applies to the possibility that certain genes that were not present in LUCA arose later, in separate lines, by convergent evolution. With due regard to these uncertainties, we have enough information to sketch a portrait of the universal ancestor.

Theobald [93] analyzed the vast array of molecular sequences now available from the three domains of life using powerful statistical techniques and concluded a monophyletic origin of LUCA regardless of HGT or multiple origins of life. He studied amino acid sequences from a set of 23 universally conserved proteins found in the three domains of life. He then applied standard programs for inferring evolutionary trees. His study was based on several simple assumptions about how diversity of protein arose. He applied a model of selection theory to a molecular phylogeny that favored the existence of a single origin of LUCA over a wide suite of alternative hypotheses. According to Theobald, LUCA was a microbe living in the early Archean world from which all life evolved.

Weiss et al. [94] genetically analyzed 6.1 million protein-coding genes and 286,524 protein-coding clusters from sequenced prokaryotic genomes of various phylogenetic trees, and identified 355 protein families that were probably common to LUCA. The team realized that HGT between Bacteria and Archaea about four billion years ago masked much of LUCA's original genetic signal. Genes found in both Bacteria and Archaea could have been shared through HGT and hence would not necessarily have originated in LUCA. The team searched for 'ancient' genes that have exceptionally long lineages but do not seem to have been shared around by HGT on the assumption that these ancient genes should therefore come from LUCA. Once they finished their analysis, they found only 355 genes that definitely belonged to LUCA and can tell us something about how LUCA lived. LUCA shares these genes with two groups of modern microbes: *Clostridia*, a genus of thermophilic Bacteria, and the methanogens, a group of hydrogen-metabolizing Archaea. Most likely, LUCA lived in an anaerobic hydrothermal vent rich in H_2, CO_2, and iron. LUCA was chemosynthetic and autotrophic, deriving free energy in the form of redox potentials and pH gradients from the vents. Furthermore, hyperthermophiles have adapted to vent conditions, so maybe life began here.

Although the physiology and habitat of LUCA in the hydrothermal vent environment suggested by Weiss et al. [94] has been endorsed by previous researchers, their conclusion about the portrait of LUCA – a progenote, only 'half-alive'– has created a great deal of controversy. Such a small number of genes (~355), of course, would not support life as we know it, and critics immediately jumped onto its apparent gene shortage, pointing out that essential components capable of nucleotide and amino acid biosynthesis, for example, were missing. Gogarten and Deamer [96] have rightly criticized that these authors reintroduced an old misconception view of LUCA, a progenote, which had been rejected by most scientists. LUCA had evolved far beyond the origin of life; it was a full-fledged prokaryote cell that could accomplish the complicated task of synthesizing proteins.

LUCA was a sophisticated organism preceded by a long period of Darwinian evolution. The emergences of the first cells and LUCA are separate events, the former led to the latter. As we discussed earlier, there was life before LUCA. Once LUCA appeared, it quickly gave rise to two

domains – Bacteria and Archaea (Fig. 16B). Although LUCA is long gone, its closest relatives may still be with us.

LUCA itself was a hyperthermophilic chemoautotroph, a view supported by molecular phylogeny [85,91-96]. Allowing for uncertainty, LUCA was not very different from modern hyperthermophilic organisms that appeared 4–3.5 Ga, before the split between Bacteria and Archaea (Fig. 16B). The geological setting of the earliest microbial life is generally considered to have been a hot hydrothermal crater vent environment, thus supporting the physiology and habitat of LUCA [1,86,95].

10. Radiation of Archean Microbial communities

The conditions at the surface of the young Earth were suitable for the emergence and evolution of life. The early atmosphere of Eoarchean Earth was dominated by CO_2 and N_2, not by CH_4 and N_2; the climate was probably similar to that of more recent times [2]. The best sources of information to infer the presence of a stable liquid water veneer on early Earth, which was the prerequisite for the origin, evolution, and propagation of life, are sedimentary rocks. Minerals, such as zircons, and water-lain sediments in the ancient Hadean/Archean crust indicate that liquid water was prevalent as early as 4.3 billion years ago [97].

The fossil record of Early Archean life had been erased mostly by plate tectonics, meteorite impacts, weathering, erosion, and recycling of the crust. The older the rock, the lesser is the chance of its preservation of these delicate fossils. The oldest volcano-sedimentary record of Earth's early Archean history provides tantalizing clues to the emergent biosphere [37,100]. The hydrothermal systems of ancient Greenstone belts including the Nuvvuagittuq Craton of Canada, Akilia-Isua Craton of West Greenland, the Pilbara Craton of western Australia, the Kaapvaal Craton of South Africa, and the Singhbhum Craton of India have yielded the earliest evidence of life. This finding is consistent with molecular biology that the universal ancestor and the ancestor of Bacteria were hyperthermophiles [85,91-101]. The evolution of Bacteria and Archaea from LUCA illustrates the astonishing durability of prokaryotic life. They are built to grow and multiply as fast as possible.

Our discussion of early Archean life is entirely based on prokaryotes—Bacteria and Archaea. Although prokaryotes are microscopic, they outweigh macroscopic creatures at least ten-fold. They are extremely diverse in their metabolism—the range of nutritional and respiratory strategies, which are adapted later by plants, animals, and fungi. They can swap genetic information without the elaborations of sex in Eukaryotes. They are far more variable in appearance than is often imagined: some are spherical, some are rod-shaped, some have whip-like appendages called flagella for swimming. They can communicate by quorum sensing to determine whether other microbes are close to them.

The hydrothermal crater vent environment was highly favorable for early habitation during the last stage of the bombardment period that enabled early life to get a foothold in a bunker, leading to sequestered communities. Their ecosystem was energized not by solar power, but heat from vents and nourished by hydrogen sulfide, methane and other toxic chemicals spewed from vents. Experimental study indicates evidence for the hyperthermophilicity of ancestral life [81]. Recent phylogenetic analysis suggests that the first enzymes were fully adapted in those extreme hot environments [101,102]. Modern hydrothermal crater settings represent a snapshot of the primordial niche for early life. These subaerial hydrothermal lakes are widely colonized by hyperthermophilic Bacteria and Archaea and provide a glimpse of early history of life on our planet. Similarly, the red, orange and blue shades around the geysers and mud holes in Yellowstone National Park disclose hyperthermophiles' ecology on land in extreme harsh environments. The pigments of hyperthermophiles are carotenoids with red, yellow and orange hues that make them colorful.

In a previous paper, Chatterjee [1] reviewed the known Archean biosignatures, microfossils, and stromatolites from hydrothermal deposits during the Eoarchean (4.0–3.6 Ga) and the Paleoarchean (3.6–3.2 Ga) eras in different parts of the world. These microfossils give us valuable clues to the successive evolution of the microbial world (Fig. 16A, B). Three distinct microbial regimes can be reconstructed from the Archean fossil records in stratified sequences of the hydrothermal crater vents

from the bottom to the top (Fig. 16C). They harnessed different sources of environmental energy to grow and nourish. In the earliest stage, hyperthermophilic Bacteria (such as Thermotogales) and Archaea (such as Methanococcales) emerged concurrently with LUCA and adapted independently a benthic hyperthermophilic life style around the vent of the hydrothermal crater basin. These hyperthermophiles were chemosynthetic and harnessed energy stored in chemicals such as iron, hydrogen, sulfur, and methane, from the vent environment. These simple organisms consumed the chemicals from the vent environment and employed enzymes to speed up chemical reactions, which released energy that the organisms harnessed along with ATP for their metabolism. Hyperthermophilic Bacteria employed sulfur reduction metabolism in the vents, whereas Archaea produced methane as a metabolic process.

In the second stage of microbial evolution, there was a gradual change of temperature gradient and niche from extremely hot vent site (hyperthermophilic) to the near surface of the crater basin at moderate temperature (thermophilic) when primitive, anoxygenic, photosynthetic bacteria (such as green and purple sulfur bacteria) began to appear in a hydrogen-sulfur world, but tapped infrared from the Sun. These early photosynthetic life forms were chemoautotrophic, which used bacteriophyll pigment to harvest light, and energy from chemical substances from vent environments to build glucose. These ancestral Bacteria used hydrogen sulfide, not water, to carry out a type of photosynthesis that did not produce oxygen.

In the third stage of microbial evolution, advanced, oxygenic, photosynthetic Bacteria (such as cyanobacteria) emerged at the water surface of the crater-lake that adapted to the normal water temperature niche (mesophilic). As they began to harness solar energy using chlorophyll pigment, they began to spilt water molecules, used hydrogen from water to build glucose and produced oxygen as a byproduct. These phototrophs converted sunlight into biologically useful energy in the form of chemical gradients and reduced molecules. The evolution of aerobic photosynthesis is undoubtedly the single most important metabolic innovation in the history of life. Some anaerobic photoautotrophs, such as halophiles (salt-loving bacteria), are the most closely related to the cyanobacteria. They live in an extremely salty environment and carry out a peculiar type of anaerobic photosynthesis using rhodopsin as a pigment [76].

Though closely related, the anoxygenic photosynthesis and oxygenic photosynthesis differ in their metabolic pathways. Both combine hydrogen and carbon dioxide to build glucose, but the hydrogen comes from different sources. In the primitive anoxygenic process, hydrogen is supplied by hydrogen gas (H_2), small organic compounds, or hydrogen sulfide (H_2S). Light energy is harvested by bacteriochlorophyll pigments in anoxygenic bacteria to build glucose, the universal cellular fuel.

A more complicated photoautotrophy evolved next, based on chlorophyll-containing light-sensitive photosynthesis [76]. Here, hydrogen is always provided by water, and for that reason oxygen is liberated by breakdown of the water molecule, and hydrogen atoms are split away. A key component in oxygenic photosynthesis is the oxygen-evolving complex that is based on manganese compounds exploiting a transition from Mn_4O_4 to Mn_4O_6 that suggest the vicinity of a hydrothermal vent [103]. Cyanobacteria are the inventors of 'green plant photosynthesis' using chlorophyll and the ultimate source of breathable air. Photosynthesis is the end result of a long chain of evolutionary selection at the molecular level that shows gradual evolution of pigment with a changing niche to harness energy.

These vertically stratified temperature regimes and niche partitions in crater basins encouraged different kinds of microbes to adapt thermotolerance and diversity (Fig. 16). As cyanobacterial life invaded the ocean surfaces from a hot to normal temperature regime, the associated enzymes coevolved to keep life's chemical reactions going and shifted their optimal temperature range gradually to a cooler environment [103].

Both kinds of photosynthetic bacteria, anoxygenic and oxygenic, created distinctive stromatolite horizons in the Pilbara and Kaapvaal sequences that indicate their microbial activity and distinctive habitats. Anoxygenic photosynthesis predates the oxygenic photosynthesis in standard phylogenetic models. Possible microfossils with cyanobacteria-like filamentous morphology and stromatolites are present in the Pilbara Craton of Australia around 3.5 Ga [76]. Today, anoxygenic photosynthetic

bacteria and cyanobacteria coexist in layered stromatolitic communities, because they have different light-capturing pigments. Life on present-day Earth is largely dependent on the products of oxygenic photosynthesis. The development of anoxygenic and then oxygenic photosynthesis would have allowed life to escape the hydrothermal setting and invade the global oceans by switching to solar energy from hydrothermal energy.

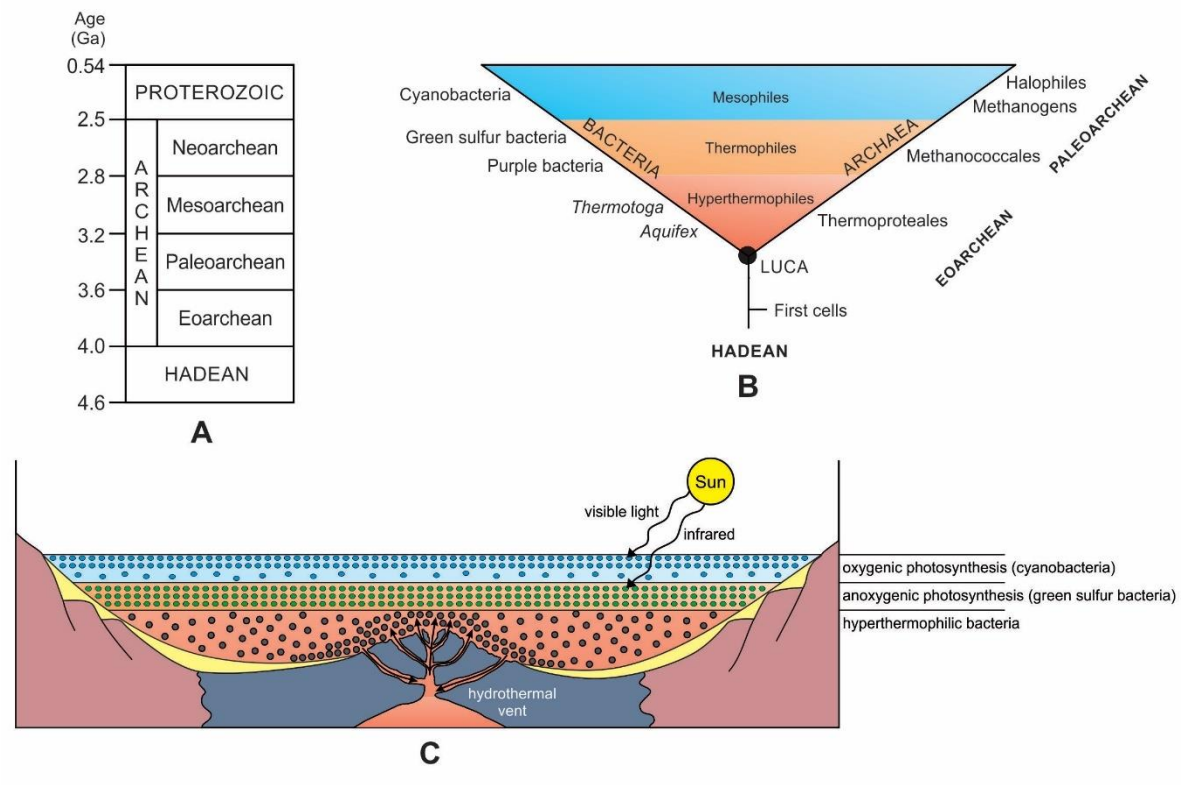

Figure 16. The evolution of the Archean biosphere. A, delineates the geological timescale of the Archean eon. B, illustrates the origin and early evolution of life in an Archean hyperthermophilic, benthic crater-lake. Two distinct domains of life, Bacteria and Archaea, are preserved in the fossil record (~3.5 Ga) of the ancient Vaalbara continent. Both domains show the gradual reduction of thermotolerance over time: from hyperthermophilic to thermophilic to mesophilic. In the first hyperthermophilic habitat both Bacteria and Archaea appear. In the second, thermophilic habitat Bacteria evolve as anoxygenic photosynthesizers. In the final mesophilic habitat, Bacteria congregate at the upper surface of the crater-lake; they begin to tap solar energy, and evolve into oxygenic cyanobacteria. Over time, these cyanobacteria spread globally through the ocean and begin to produce oxygen. Hyperthermophilic archaea, on the other hand, evolved through two stages: as thermophilic Methanococcales, and then as mesophilic Methanogens and Halophiles, the latter thriving in the hypersaline environment of ponds and lakes. C, a cross-section of a hydrothermal crater-lake showing the evolution of photosynthesis. This occurs through three stages of an evolving microbial community: first, on the bottom, hyperthermophilic bacteria emerge; next, in the thermophilic middle stage, the anoxygenic photosynthetic green sulfur bacteria appear; in the final stage, at the upper mesophilic level, oxygenic photosynthesizing cyanobacteria form and begin the production of oxygen.

As anaerobic and aerobic Bacteria invaded the shallow-water of the global ocean, they competed side by side to harvest sunlight. However, cyanobacteria produced molecular oxygen, a gas toxic to their anaerobic neighbor, killing countless species of anaerobic Bacteria and Archaea, because their energy-generating metabolic processes were not coupled with the consumption of oxygen. In fact, the presence of oxygen actually poisoned some of the key enzymes of anaerobes. These anaerobes took refuge in the hydrothermal vents and remained there forever, a survival strategy. The anoxygenic photosynthetic bacteria retreated initially to the deeper part of the water column on the

murky ocean floor, shielded by a cover of water where there was no free oxygen. They survived by fleeing the oxygen warfare. Later, they learned how to coexist with oxygenated bacteria by hiding below the cyanobacterial mat, where the infrared could penetrate. Certainly, anaerobes didn't vanish from the Earth, but they were vanquished to low-oxygen environments. It was perhaps the first of the mass extinctions that life would face on our planet, and the killing agent was deadly oxygen gas produced by the cyanobacteria.

Some of the earliest evidence for Archean microbial ecosystems stems from putative stromatolites 3.5 to 3.2 Ga, which display similar structures to modern microbial mats. The earliest microbial mats consisting of Bacteria and Archaea may have formed as biofilms in hydrothermal vents. As photosynthetic bacteria left hydrothermal vents and invaded the shallow ocean, the microbial mats developed profusely in this environment, allowing widespread colonization of the globe and the creation of further aerobic habitat. A key feature of modern-day microbial mats and stromatolites is the fact that they are oriented and stratified in relation to sunlight. Thus, there is reason to believe that ancient stromatolites may have functioned similarly to microbial mats today. Microbial mats are a unique ecological niche representative of early life on Earth. Rapid nutrient cycling across microgradients coupled with putative niche differentiation within mat layers enables diverse metabolic processes to occur in spatial proximity.

On a broad scale, the origin of oxygenic photosynthesis by cyanobacteria led to the rise of oxygen on Earth during the early Proterozoic (~2.4-2.3 Ga), played a major role in the oxygenation of ocean and atmosphere, and paved the way for oxygen-breathing microbes such as respiratory bacteria. The *Great Oxidation Event* is when the atmosphere first became oxygenated. However, oxygen level only reached somewhere between 0.2 to 2% by volume, not today's 21% [103]. Respiration is opposite to photosynthesis. Whereas photosynthesis employs energy from the Sun, respiration releases energy. In photosynthesis, cyanobacteria combine carbon dioxide and water to form glucose and oxygen. In respiration, aerobic bacteria consume sugar and oxygen to gain energy. One biochemical trick that evolved around 2.5 billion years ago to take advantage of oxygen is still being used for respiration by all animals including humans.

The history of life is largely microbial. Microbes once ruled the whole world. The Earth has been a bacterial world for at least the last 3.5 billion years. The diversity of life today is a result of the dynamic interplay between genetic opportunity, metabolic capability, and environmental changes. For most of their existence, Earth's habitable environments have been dominated by microbes and subjected to their metabolism and evolution. Microbes gave life its initial foothold, and they engineered the planet for our use, taking in carbon dioxide and giving off oxygen, day in and day out for billions of years until there was enough oxygen in the atmosphere to support larger life.

11. Discussion and Conclusion

Life arose through five hierarchical stages of increasing molecular complexity in the hydrothermal vent environment about 4 billion years ago [1-3]. Prebiotic synthesis began amidst a chaotic chemical mixture of cosmic ingredients as that present in the Murchison meteorite. Three processes were required for life's origin: (1) condensation reactions that produce essential biopolymers by non-enzymatic reaction; (2) self-assembly of lipid membranes that encapsulate the polymers into populations of protocells; and (3) Darwinian selection during the biogenesis [9,16,28,74]. Molecular selection at each hierarchical level offered a pathway by which smaller components could organize themselves into more complex molecules such as cell membranes, RNAs, proteins, and DNAs that became highly organized leading to the first cell. To paraphrase François Jacob [104], the origin of life does not produce innovation from scratch, but owing to cumulative effects of hierarchical history on prebiotic synthesis. Nature functions by integration. It works on what already exists like a tinkerer who, during millions of years slowly fine-tuned the products step-by-step by natural selection that culminates in the emergence of the highly organized first cell. But once life had started, further evolution had to proceed mainly through mutation or the slight tinkering of already existing DNAs. But these were merely variations on previous cellular structures.

Macroevolution, on the other hand, can accelerate evolution by gene enrichment in a fast lane [41,43,44].

Life depends on the intricate interplay of myriads of different biomolecules, but how such molecular networks arose at the origin of life remains a mystery. Life is the ultimate example of complexity at work. The first cell developed through an incredibly complex series of interactions involving a vast number of components. These components of subsystems, were made up of smaller molecular components, which independently exhibited their own dynamic behavior, such as the ability to catalyze chemical reactions. Yet when they were combined into some larger functioning unit—such as the first cell or LUCA—utterly new and unpredictable properties emerged, including the ability to move, to change shape, to grow, and to divide identical daughter cells. This phenomenon, in which components join together to form larger stable structures having new properties that could not have been predicted from the characteristics of their individual parts, is known as self-assembly. Molecular forces drive self-assembly. The most obvious way to increase the complexity of an encapsulated system of simple molecules is to link them in polymers held together by covalent bonds [28].

The coevolution of protocells and genetic parasites (such as primitive retrovirus) from the protein/RNA world is supported by experimental study [105]. The emergence of DNA and genes gave rise to a new mechanism for generating structural diversity that accelerated evolution. Yet throughout all this time, the rule guiding the process of hierarchical self-assembly remained essentially unchanged. The ability of amphiphilic compounds in hydrothermal water to spontaneously assemble in the evolutionary is obviously an essential step in the evolution of cellular life [28]. Molecular self-assembly underlies the step-by-step construction of cellular components without guidance from an outside source. Complex-systems can self-assemble from the bottom-up in biological systems. Certain organic compounds have the capacity to react with each other to form more complex molecules such as hydrocarbons, amino acids, and simple sugars. Some of these can spontaneously self-assemble into membrane structures, and others can polymerize into molecules similar to polypeptides and RNAs. With the availability of the enzymes, these in turn form larger self-assembled structures such as the double helix of DNA via retrovirus. Both self-assembly and Darwinian selection played crucial roles in the origin of life.

The invention of cell division marks the transition from the prebiotic world to the biotic world. The first cells underwent a vegetative cell division known as binary fission, where the genetic material was segregated equally into two identical daughter cells. Every time a cell divided, DNA polymerases helped duplicate the cell's gene, so that a copy of the original gene could be passed to each daughter cell. All cell divisions were preceded by a single round of DNA replication. Once life began, it was able to diversify and proliferate until it filled every niche on the planet.

Archean volcano-sedimentary rocks (4.0-2.5 Ga) host all the known types of biosignatures, including morphological fossils of cells, colonies, biofilms, or other biological constructions, such as microbial mats and stromatolites. Various Archean microfossils from Canada, Greenland, Australia, South Africa, and India provide valuable clues to the distribution and nature of ancient microbial life [1]. These Archean biosignatures demonstrate that microbial life was abundant and possibly more diverse than currently believed.

LUCA, the last universal common ancestor of life on Earth, was probably a single-cell hyperthermophilic bacterium-like organism, as suggested by recent molecular phylogenetics analyses [93,94,98]. Most likely, the hyperthermophilic bacteria gave rise to the evolution of anoxygenic bacteria as early as 3.4 Ga, and then oxygenic photosynthetic bacteria around 3.0 Ga [76]. The emergence of oxygenic photosynthetic bacteria or cyanobacteria allowed life to escape the hydrothermal crater surface into the surface of the adjacent shallow ocean basin, which depended entirely on the solar power in the mesophilic environment. It was a major transition in the microbial life where oxygen began being toxic to becoming a vital part of metabolism. Cyanobacteria contributed to the geological processes by producing oxygen in our planet for the first time, and left their signature in shallow marine sediments by vast amounts of carbonate sediments and stromatolitic structure. Once cyanobacteria invaded the sea, they had global distribution. The

invention of oxygenic photosynthesis by cyanobacteria forever transformed Earth. The biogeochemical shift set into motion the evolution of subsequent microbial metabolism and lifestyles. Recent fossil studies have provided evidence for the diversification of microbes in the Paleo- and Mesoarchean. Once life proliferated across the planet, it assumed remarkable forms, and wrought the extraordinary changes that have now inextricably linked to biosphere and geosphere.

Over the past decades, the boundary conditions under which microbial life can survive have been pushed in every possible direction, encompassing broader ranges of temperature, pH, pressure, salinity, energy, and nutrient limitation. Extremophiles are unicellular microbes (Bacteria, Archaea, and Eukaryotes) that can thrive in some of the most extreme environments known today from Antarctic frozen lakes to continental hot springs and crater vents, submarine hydrothermal vents, dry desert soils, salt solutions, toxic wastes, organic solvents, heavy metals, radioactive waste, even 7 km-deep rocks inside earth's crust, as well as acid and alkaline niches that were previously considered inhospitable.

The study of these extremophiles has profound implications for the search of life on other planetary bodies. There is a strong possibility that life also emerged beyond Earth wherever the necessary physical and chemical conditions were met. 'Extra-terrestrial life' refers to a form of life resembling Earth life in all basic properties, including the DNA-RNA-protein triad [71]. Early life on Earth may inform us about what we might find on Mars, how it might be preserved, where to search for it, and how to confirm its biogenic origin. Using this information, we can search for potential life on other planetary bodies such as dwarf planet Ceres in the asteroid belt, Jupiter's moon Europa, and Saturn's moons Titan and Enceladus. Recently, the Kepler space telescope has discovered about 4,000 Earth-like exoplanets in our galaxy, each orbiting a Sun-like star in the 'Goldilocks zone' — that is, at just the right distance for liquid water. Some of these exoplanets may be habitable and may harbor life. NASA has detected water vapor and optimum temperatures in the atmosphere of the exoplanet, known as K2-18b that could potentially support life. The question is no longer, is there any life beyond Earth? The question is how do we detect life by remote sensing in our Solar System and beyond.

Funding: The research received no external funding.

Acknowledgments: I thank Michael San Francisco for helpful discussion, Kippra Hopper for editing the manuscript, and three anonymous reviewers for their constructive suggestions that improved clarity and focus. I acknowledge Shigenori Maruyama for his helpful comments. I thank Volkan Sarigul for artwork and Texas Tech University for funding the research. The paper is dedicated to David Deamer who inspired me in the astrobiology research and provided critical information on the origin of the cell membrane.

Conflicts of interest: The author declares no conflict of interest.

References

1. Chatterjee, S. The hydrothermal impact crater lakes: the crucibles of life's origin. In *Handbook of Astrobiology*; Kolb, V.M. Ed.; CRC Press; Taylor & Francis: Boca Raton, FL, USA, 2018; pp.265-295.
2. Chatterjee, S. The RNA/protein world and the endoprebiotic origin of life In *Earth, Life, and System*; Clarke, B. Ed; Fordham University Press, New York, **2015**, pp. 39-79.
3. Chatterjee, S. A symbiotic view of the origin of life at hydrothermal impact crater lakes. *Phys. Chem. Chem. Phys*. **2016**, *18*, 20033-20046.
4. Cech, T.R. RNA as an enzyme. *Scient. Amer*., **1986**, *255(5)*, 64-75.
5. Crick, F.H.C. The origin of genetic code. *J. Mol. Biol*., *38*, **1968**, 367-379.
6. Gilbert, W. The RNA world. *Nature*, *319*, **1986**, 618.
7. Orgel, L.E. The origin of life. *Scient. Amer*., **1994**, *271(4)*, 77-83.
8. Robertson, M.P.; Joyce, G.F. The origins of the RNA world. *Cold Spring Harb. Perspect. Biol*., **2012**, *4(5)*, a003608.
9. Chatterjee, S.; Yadav, S. The origin of prebiotic information system in the peptide/RNA world: A simulation model of the evolution of translation and the genetic code. *Life* **2019**, *9*, 25; doi:10.3390/life9010025.
10. Hartman, H.; Smith, T.F. Origin of genetic code is found in the transition between and Thioester world of peptides and the phosphoester world of polynucleotides. *Life*, **2019**, *9*, 69; doi: 10.3390/life9030069.

11. Bowman, J.C.; Hud, N.V.; Williams, J.D. The ribosome challenge to the RNA world. *J. Mol. Evol.*, **2015**, *80*, 143-161.
12. Carter, C.W. Jr. What RNA world? Why a peptide/RNA partnership merits renewed experimental attention. *Life* **2015**, *5*, 294-320; doi:10.3390/life5010294.
13. Carter, C.W. Jr. An alternative to the RNA world. *Nat. Hist*. **2016**, *125(1)*, 28-33.
14. Carter, C. W. Jr.; Wills, P. R. Interdependence, reflexivity, impedance matching, and the evolution of genetic coding. *Mol. Biol. Evol*. **2017**,
15. Harish, A.; Caetano-Anolles, G. Ribosomal history reveals origin of modern protein synthesis. *PLoS One*, **2001**, doi: e32776.
16. De Duve, C. *Singularities: Landmarks on the Pathways of Life*: Cambridge University Press, New York, USA, 2005.
17. Busch, F.; Rajendran, C.; Schlee, S.; Merkl, R.; Stemer, R. Ancestral tryptophan synthase reveals functional sophistication of primordial enzyme complexes. *Cell Chem. Biol*., **2018**, *23*, 709-715.
18. Hargraves, W.W.; Mulvihill, S.J.; Deamer, D.W. Synthesis of phospholipids and membranes in prebiotic conditions. *Nature*, **1977**, *266*, 78-80.
19. Deamer, D.W. The role of lipid membrane in life's origin. *Life*, **2017**, *7(1)*, 5; https://doi.org/10.3390/life3701005.
20. Mansy, S.S., Schrum, I.P.; Krishnamurthy, M., Tobe, S., Treco, D.A.; Szostak, J.W. 2008. Template-directed synthesis of a genetic polymer in a model protocell. *Nature*, **2008**, *454*, 122-126.
21. Budin. I.; Prwyes, N.; Zhang, N.; Szostak, J.W. Chain-length heterogeneity allows for assembly of fatty acid vesicles in dilute solutions. *Biophys. Jour*., **2014**,*107*, http://dx.doi.org/10.1016/j.pbj.2014.07.067.
22. Monnard, P.A.; Deamer, D.W. Membrane self-assembly processes: steps toward the first cellular life. *Anat. Rec*., **2002**, *268*, 197-207.
23. Zhu, T.F.; Szostak, J.W. Coupled growth and division of model protocell membranes. *J. Am. Chem. Soc*., **2009**, *131*, 5705-5713.
24. Chen, I.A.; Szostak, J.W. A kinetic study of the growth of fatty acid vesicles. *Biophys. J.* **2004**, *87*, 988-998.
25. Budin, I.; Szostak, J.W. Physical effects underlying the transition from primitive to modern cell membranes. *Proc. Nat. Acad. Sci*., **2011**, *108*, 549-5254. http://dx.doi.org/10.1093/molbev/msx265.
26. Apel, C.L.; Deamer, D.W. The formation of glycerol monodecanoate by a dehydration condensation reaction: The chemical complexity of amphiphiles on the early Earth. *Orig. Life. Evol. Biospheres*, **2005**, *35*, 323-332.
27. Simoneit, B.R.T.; Rushdi, A.I.; Deamer, D.W. Abiotic formation of acylglycerols under simulated hydrothermal conditions and self-assembly properties of such lipid products. *Adv. Space Res*., **2007**, *40*, 1649-1656.
28. Deamer, D.W. *First Life: Discovering the Connections between Stars, Cells, and How Life Began*; University of California Press: Berkeley, CA, USA, 2011.
29. Singer, S.J.; Nicholson, G.L. The fluid mosaic model of the structure of the cell membrane. *Science*, **1972**, *175*, 720-731.
30. La Scola, B.; Audic, S.; Robert, C.; Jungang, L.; Lamballerie. X. D.; Drancourt, M.; Birtles, R.; Claverie, J.M.; Raoult, D. A giant virus in Amoebae. *Science*, **2003**, *299*, 2033.
31. Forterre, P. Defining life: The virus viewpoint. *Orig. Life. Evol. Biosph*., **2010**, *40*, 151-160.
32. Koonin, E.V.; Senkevich, T.G.; Dojla, V.V. The ancient virus world and evolution of cells. *Biol. Dir.* **2006**, *1:29*, doi: 10.1186/1745-6150-1-29.
33. Suttle, C.A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol*., **2007**, *5*, 801-812.
34. Villarreall, L.P. *Viruses and Evolution of Life*; American Society of Microbiology Press, Washington, DC, 2005.
35. Wagner, E.K; Hewlett, M.J. *Basic Virology*; Blackwell Publishing; Malden, MA, USA, 2004.
36. Koonin, E.V. Viruses and mobiles elements as drivers of evolutionary transitions. *Phil. Trans. R. Soc. B*, **2016**, *371*, http://dx.doi.org/10.1098/rstb.2015.0442.
37. Villarreall, L.P. Are viruses alive? *Scient. Amer*., **2004**, *291(1)*, 100-105.
38. D'Herelle, F. *The Bacteriophage: Its Role in Immunity*. Williams and Wilkins, Baltimore, 1922.
39. Haldane, J.B.S. The origin of life. *Ration. Ann.* **1929**, 148-169.
40. Forterre, P. The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochemie*., **2005**, *87*, 793-803.

41. Margulis, L. Symbiosis in Cell Division. Freeman, San Francisco, CA, 2005.
42. Takemura, M. Poxyviruses and the origin of the eukaryotic nucleus. *Mol. Evol. Biol*., **2001**, *52*, 419-425.
43. Chatterjee, S. The river of life: a genetic perspective on macroevolution. Forum on Public Policy *Journal of Oxford Round Table*, **2009**, *5(5)*, 1-43.
44. Doolittle, W.F. (2000) Uprooting the tree of life. *Scient. Amer*., **2000**, *282(2)*, 90-95.
45. Daubin, V; Ochman, H. Start-up entities in the origin of new genes. *Curr. Opin. Genet. Dev.* **2004**, *14*, 616-619.
46. Feschotte, C. Virology: bornavirus enters the genome. *Nature*, **2010**, *463*, 39-40.
47. Chuong E.B. Retroviruses facilitate the rapid evolution of the mammalian placenta. *Bioessays*, **2013**, *35(10)*, doi:10.1002/bies.201300059.
48. Holmes, E.C. What does virus evolution tell us about virus origin? *J. Virol*., **2011**, *85*, 5247-525.
49. Forterre, P. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res*., **2006**, 117, 5-16.
50. Koonin, E.V. On the origin of cells and viruses: Primordial virus world scenario. *Ann. N.Y. Acad. Sci*., **2009**, *1178*, 47-64.
51. Koonin, E.V.; Dojla, V.V. A virocentric perspective on the evolution of life, *Curr. Opion. Virol*., **2013**, 3(5), 546-557.
52. Koonin, E.V.; Dojla, V.V. Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol. Mol. Biol. Rev*., **2014**, *78*, 278-303.
53. Krupovic, M.; Koonin, E.V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Nat. Acad. Sci*., **2011**, doi.10.1073pnas.1621061114.
54. Filee, J.; Forterre, P. Viral proteins functioning in organelles: a cryptic origin? *Trends. Microbiol.* **2005**, *013*, 510-513.
55. Durzynska, J.; Gozdzicka-Jozefiak, A. 2015. Viruses and cells intertwined since the dawn of evolution. *Virol. J.* **2015**, 12:169; doi: 10.1186/s12985-0400-7.
56. Sabbath, N.; Wagner, A.; Karlin, D. Evolution of vital proteins originated de novo by overprinting. *Mol. Biol. Evol*., **2012**, *29*, 3767-3780.
57. Ahlquist, P; Noueiry, A.O.; Lee, W.M.; Kushner, D.B.; Dye, B.T. Host factors in positive-strand RNA virus genome replication. *J. Virol*., **2003**, *77*, 8181-8186.
58. Brierly, I. Ribosomal frameshifting on viral RNAs. *J. Gen. Virol.* **1995**, *76*, 1885-1892.
59. Krupovic, M; Prangishvilli, D.; Hendrix, R.W.; Bamford, D.H.; Genomic of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev*., **2011**, *75*, 610-635.
60. Diemer, G.S.; Stedman, K.M. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol. Direct*, **2012**, *7*,13, doi: 10.1186/1745-6150-7-13.
61. Aiewsakun, P.; Katzourakis, A. Marine origin of retroviruses in the Early Paleozoic Era. *Nature Com.* **2017**, 8:13954, doi:10.1038/ncomms133954.
62. Forterre, P. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc. Nat. Acad. Sci*., **2006**, *103*, 3669-3674.
63. Powner, M.W.; Zheng, S.L.; Szostak, J.W. Multicomponent assembly of proposed DNA precursors in water. *J. Amer. Chem. Soc*., **2012**,*134*,13889-13895.
64. Forterre, P.; Prangishvilli, D. The great billion-year war between ribosome- and capsid- encoding organisms (cells and viruses) as major source of evolutionary novelties. *Ann. N.Y. Acad. Sci*., **2009**, *1178*, 65-77.
65. Xu, J.; Green, N.J.; Gibard, C., Krishnamurthy, R.; Sutherland, J.D. Prebiotic phosphorylation of 2-thiouridine provides either nucleotides or DNA building blocks via photoreduction. *Nature Chem*. **2019**, *11*, 457-462.
66. Furukawa Y.; Chikaraishi Y.; Ohkuchi N.; Ogawa N.O.; Glavin D.P.; Dworkin J.P.; Abe, C.; Nakamura T. Extraterrestrial ribose and other sugars in primitive meteorites. *Pros. Nat. Acad. Sci*., *116*, **2019**, doi: 10.1073/pnas.1907169116.
67. Takeuchi, N.; Hogeweg, P.; Koonin, E.V. 2011. On the origin of DNA genomes: Evolution and division of labor between template and catalyst in model replicator systems. *PLoS Comp. Biol*., **2011**, *7*, 3 e1002024.
68. Freeman, S. *Biological Science*, 2nd edition, Pearson Prentice Hall: Upper Saddle River, New Jersey, USA, 2005.

69. Watson, J.D.; Crick, F.H.C. A structure for deoxyribose nucleic acid. *Nature*, **1953**, *171*, 737-738.
70. Schrödinger, E. *What is Life? The Physical Aspects of Living Cell*. Cambridge University Press: Cambridge, UK,1944.
71. Catling, D.C. Astrobiology: A Very Short History of Introduction. Oxford University, Press, Oxford, 2013.
72. Von Neumann, J. *Theory of Self-Reproducing Automata*. University of Illinois Press: Chicago, IL, USA, 1966. Edited and completed by A.W. Burks.
73. Walker, S.I.; Davies, P.C.W. The algorithmic origins of life. *J. R. Soc. Interface*, **2012**, doi: 10.1098/rsif.2012.0869.
74. Dyson. F. *Origins of Life*. Cambridge University Press: Cambridge, UK, 2004.
75. Maturana, H.R.; Valera, F.J. Autopoiesis and Cognition. The Realization of Living. *Boston Stud. Phil. Sci.*, *42*, D. Reidel Publishing: Boston, 1981.
76. Schopf, J.W. *Cradle of life: The Discovery of Earth's Earliest Fossils*. Princeton University Press: Princeton, New Jersey, 1999.
77. Goodsell, D.S. *The Machinery of Life.* Springer, New York, 2010.
78. Chen, I.A. Cell division: breaking up is easy to do. *Curr. Biol*., **2009**, *19*, R327-R328.
79. Zwicker, D; Seyboldt, R.; Weber, C.A.; Hyman, A.A.; Jülicher, F. Growth and division of active droplets provides a model for protocells. *Nature Phys*., **2017**, *13*, 408-413.
80. Saha, R.; Verbanic, S; Chen, I.A. Lipid vesicles chaperone an encapsulated RNA aptamer. *Nature Comm.* **2018**, 9, 2313, doi: 10.1038/s41467-018-04738-8.
81. Kurihara, K.; Okura, Y.; Matsuo, M.; Toyota, T.; Suzuki, K.; Sugawara, T. A recursive vesicle-based model protocell.
82. Saha, R.; Chen, I.A. Origin of life: protocells red in tooth and claw. *Curr. Biol*., **2015**, *25*, R1166-R1185.
83. Leaver, M.; Domingez-Cuevas, P.; Coxhead, J.M.; Daniel, R.A.; Erington, J. Life without a wall or division machine in *Bacillus subtillis*. *Nature*, **2009**, *457*, 849-853.
84. Panno, J. *The Cell: The Nature's First Cell Life Form.* Facts on File: New York. 2010.
85. Stetter, K.O. Hyperthermophiles in the history of life. *Phil. Trans. R. Soc.*, **2006**, *361*, 1837-1843.
86. Reysenbach, A.L.; Shock, E. Merging genomes with geochemistry in hydrothermal ecosystems. *Science*, **2002**, *296*,1077-1083.
87. McCollom, T.M., Ritter, G.; Simonetta, B.R. Lipid synthesis under hydrothermal conditions by Fischer-Tropsch-type reactions. *Orig. Life Evol. Biosph*., **2003**, 29, *153*-166.
88. Darwin, C. *On the Origin of Species*. John Murray, London, 1859.
89. Woese, C.R. On the evolution of cells. *Proc. Nat. Acad. Sci*., **2002**, *99*, 8742-8747.
90. Williams, T.A.; Foster, P.G.; Cox, C.J.; Embley, T.M. An archaeal origin of eukaryotes supports only two primary domains. *Nature*, **2013**, *504*, 231-236.
91. Woese, C.R. The universal ancestor. *Proc. Nat. Acad. Sci*., **1994**, *99*, 6864-6859.
92. Cornish-Bowden, A.; Cardenas, M.L. Life before LUCA. *Jour. Theor. Biol*., **2017**, *434*, 68-74.
93. Theobald, D. L. A formal test of the theory of universal common ancestry. *Nature*, **2010**, *465*, 219-223.
94. Weiss, M.C.; Sousa, F.L.; Mrnjavic, N.; Neukirchen, S.; Roettger, M.; Nelson-Sathi, S.; Martin, W.F. 2016. The physiology and habitat of the last universal common ancestor. *Nature Microbiol.*, **2016**, doi: 10.1038/NMICROBIOL,2016.116.
95. Gogarten, J.P.; Deamer, D. Is LUCA a thermophilic progenote? *Nature Microbiol.* **2016**, doi: 10.1038/NMICROBIOL2016229.
96. Akanuma, S.; Nakajima, Y.; Yokobori. S.; Kimura, M.; Nemoto, N.; Mase, T.; Miyazono, K. Tanokura M, Yamagishi A (2013) Experimental evidence for the thermophilicity of ancestral Life. *Proc. Nat. Acad. Sci.*, **2013**, *110*, 11067-11072.
97. Mojzsis, S.J.; Harrison, T.M.; Pidgeon, T.T. Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4,300 Myr ago. *Nature* **2001**, *409*, 178-181.
98. Gaucher, E.A.; Govindan, S.; Ganesh, O.K. Paleotemperature trend for Precambrian life inferred from resurrected proteins. *Nature*, 2008, *451*, 704-707.
with a primitive model cycle. *Nature Comm*., **2015**, doi: 10.1038/ncomms9352.
99. Parnell. J.; Lee, P.; Cockell, C.S.; Osinski, G.R. (2004) Microbial colonization in impact- generated hydrothermal sulphate deposits, Haughton impact structure, and implications for sulfate on Mars. *Int. J. Astrobiol.* **2004**, *3*, 247-25.
100. Pace, N.R. A molecular view of microbial diversity and the biosphere. *Science*, **1997**, *276*, 734-740.

101. Reysenbach, A.L.; Cady, S.L. Microbiology of ancient and modern hydrothermal systems. *Trend Microbiol.*, 2001, 9, 79-86.
102. Nguyen, V.; Wilson, C.; Hoemberger, M.; Stiller, J.B.; Agavonof, R.V.; Kutter, S.; English, J.; Theobald, D.L., Kern, D. Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science*, **2017**, *355*, 289-294.
103. Nisbet, E. G.; Sleep, N. H. The habitat and nature of early life. *Nature*, **2001**, *409*, 1083-1091.
104. Jacob, F. Evolution and tinkering. *Science*, **1977**, *196*, 1161-1166.
105. Bansho, Y.; Furubayashi, T.; Ischihasi, N.; Yomo, T. Host-parasite oscillation dynamics and evolution of compartmentalized RNA replication system. *Proc. Nat. Acad. Sci.*,**2016**, 113, 4045-4050.