1

Before RNA and After: Geophysical and Geochemical Constraints on Molecular Evolution

Stephen J. Mojsis
Department of Earth and Space Sciences
University of California Los Angeles
Los Angeles, California 90024-1567

Ramanarayanan Krishnamurthy
Skaggs Institute for Chemical Biology
The Scripps Research Institute
La Jolla, California 92037

Gustaf Arrhenius
Scripps Institution of Oceanography
University of California San Diego
La Jolla, California 92093-0220

1. INTRODUCTION

This chapter offers a description of some of the physical and chemical settings for the origin of life on Earth. Considering the topic of this book, particular attention is given to the conditions for a precursor RNA World; an ab initio system based on phosphate-sugar backbone structures in linear polymers and currently with nitrogen bases as recognition molecules. Both the appeal and the uncertainties in the assumption of an RNA World are obvious. The biochemical advantage of this model has geochemical and cosmochemical complements such as the abundance in the universe of simple aldehydes, “the sugar of space.” With their relatively high oxidation state, the aldehydes are compatible with plausible models of early terrestrial atmospheres dominated by CO₂, H₂O, N₂, and small mixing fractions of CO, CH₄ and other reductants. Further suggestions of an RNA World come from observed and inferred sources of active oligophosphates and from concentration mechanisms based on the molecular charge conferred by phosphate esters. Finally, the facile formation of ribose phosphate has been successfully modeled under mild aqueous con-
ditions in the laboratory. Obstacles to progress in prebiotic synthesis leading to the inception of an RNA World yet remain. The concentration of neutral molecules such as formaldehyde and glycolaldehyde required to permit the phosphorylation and sugar phosphate formation found in the laboratory remains an impediment to modeling molecular evolution. Furthermore, the oligomerization of nucleosides or nucleotides without the aid of artificial activating groups, and the synthesis and attachment of nitrogen bases in a formaldehyde environment hostile to the stability and uninhibited reactions of hydrogen cyanide, is another complicated problem remaining to be resolved by further laboratory investigations. At least for the time being, these limitations retain the concept of an RNA World as a metaphor in the company of all current theories of the origin of life. This metaphor has, however, several redeeming geophysical and geochemical qualities that may be regarded as strengthening the case for an RNA World as an early phase in the emergence of life on Earth in the earliest Archean (i.e., before 3800 Ma; Ma = 1 × 10^6 years).

2. EARLY ATMOSPHERE AND HYDROSPHERE

2.1 General Aspects

Because there is no known preserved record on Earth for the state of the atmosphere and hydrosphere during the first half billion years, there remains considerable room for speculation about the nature of this system wherein life began (Fig. 1).

It is currently popular to assign early molecular evolution to hot springs or adjacent cooler regions at the surface as in the case of Yellowstone, Wyoming, or at great depth in the ocean as in the case of the submarine hydrothermal or low-temperature vents. The attractiveness of such schemes comes from the conducive environment rich in energy resources that is provided for chemoautotrophs, and from the fact that a few thousand meters of ocean water affords some protection against thermal shocks by meteorites and comets early in Earth’s history (Sect. 4). The appeal also includes the assumption of a hydrothermal supply of organic compounds, suggested by theoretical considerations and laboratory experiments (Shock 1990, 1996; Shock et al. 1997). However, as of today, the analysis of effluents from bioorganically uncontaminated hot or cold springs has failed to show any organic molecules besides methane, perhaps because of insufficient analytical attention paid to this aspect. In contrast, hot springs debouching through sediments transport a variety of organic decomposition and polymerization products, including petroleum,
that are the direct result of thermal breakdown of residual bioorganic matter from past life, now trapped in rocks.

A particularly demanding but neglected link in such schemes for biopoesis in hydrothermal vents is the need for a crustal source of reduced nitrogen available for prebiotic chemical reactions, made acute by the low solubility of molecular nitrogen in the water transfer medium. A solution may be found in the mechanism demonstrated by Summers and Chang (1993), showing that photochemically produced nitrite ion is converted to

\[ \text{Figure 1} \quad \text{Time scale for events with relevance to the origin and early evolution of life.} \]
ammonia by reduction with Fe$^{++}$ in solution. The reaction is enhanced in the anion-concentrating reactive mineral green rust (ferroferric hydroxide). In contrast, inorganic carbon and hydrogen in the form of carbon dioxide and water are readily transported into oceanic crust and potentially (but for CO$_2$ not yet demonstrably) reduced there during hydrothermal alteration.

The fact that thermophilic Archea are found at the bottom of the phylogenetic tree (Woese 1987) has been considered an argument for the origin of life in a hydrothermal environment. However, present-day Archea are already highly evolved organisms with an elaborate polypeptide-based enzymatic machinery, protecting the otherwise unstable nucleic acid structures against decomposition at high temperature in the immediate vent environment. Though possessing deep phylogenetic roots, these organisms are still very far away from the origin of life. These circumstances tend to induce those advocating these deeply rooted hyperthermophiles as a remnant population from early molecular evolution to retreat from conduits of hot fluid at the vent proper. In the surroundings of hydrothermal vents, temperatures range all the way down to near freezing on the present sea floor, and as low or lower around some continental hot springs in cold climates. Hydrothermal advocates would also be drawn into an accretion camp that places a substantial reservoir of extraterrestrial organic solids, like cometary or carbonaceous meteorite material, in the early crust, available for hydrothermal digestion. If not depending on a source organic compounds from space, much will depend on the outcome of currently intensified analysis of virginal hot spring effluents. Expanded measurements of organic production at hot springs could testify to the efficiency of postulated inorganic–organic reactions. With these provisions and reservations, the hypothesis of hydrothermally assisted molecular evolution would seem to fall within the limits of geophysical and geochemical plausibility. The examples presented here point at the interdependence between different disciplines, and at avenues for verification.

2.2 The “Fertile Atmosphere” and the Concept of a Prebiotic Soup

Already early in the century, considerations of the possible nature of Earth’s primordial atmosphere created compelling arguments for a neutral or weakly reducing composition, dominated by carbon dioxide, nitrogen, and water vapor. In his pioneering electrochemical experiments, Walther Löb (1906, 1914) actually used carbon dioxide and carbon monoxide as major components of the cold plasma discharge system (Fig. 2), which
was found to produce aldehydes, including glycolaldehyde, and the amino acid glycine. Löb’s experiments were intended as metaphors of the conversion of inorganic carbon and nitrogen to organic compounds by autotrophic life without apparent concern for their prebiotic significance.

In contrast, the later and more widely known experiments by Oparin (1924; Oparin and Clark 1959) deliberately aimed at modeling the origin of life. Since the formation of “proteinoids” (postulated then to have been important predecessors of life) demanded a reducing environment, Oparin proposed, without scrutiny of physical plausibilities, that a primordial atmosphere of the Earth must have consisted primarily of a mixture of hydrogen, methane, ammonia, and water vapor. A decisive event in the U.S.
efforts at scientifically addressing the problem of the origin of life was Harold Urey’s entry into this field after World War II. Although with a better awareness of geophysical limitations (Urey 1952) and unaware of Löb’s biomimetic experiments, Urey nonetheless championed Oparin’s concept of a strongly reducing atmosphere despite objections from contemporary geochemists such as V.M. Goldschmidt (1952) and W.W. Rubey (1951, 1955) as well as organic chemists (Abelson 1966; Ferris and Nicodem 1972). The objections of these and other workers were put into the background by the impressive results of experiments by Miller (1953, 1987; Miller and Van Trump 1981) that expanded Löb’s findings to a plethora of amino acids and other organic compounds forming under fertile Oparin-like conditions. The fact that a variety of amino acids, considered by science at the time to be the initial building blocks in biomolecular evolution, could be formed from a gas mixture with claims for modeling the primordial atmosphere made a lasting impression on organic chemists as well as the general public, and stimulated prebiotic chemistry as a worthy field of study. However, it is now held to be highly unlikely that the conditions used in these experiments could represent those in the Archean atmosphere. Even so, scientific articles still occasionally appear that report experiments modeled on these conditions and explicitly or tacitly claim the presence of resulting products in reactive concentrations “on the primordial Earth” or in a “prebiotic soup.” The idea of such a “soup” containing all desired organic molecules in concentrated form in the ocean has been a misleading concept against which objections were raised early (see, e.g., Sillén 1965). Nonetheless, it still appears in popular presentations perhaps partly because of its gustatory associations.

The interest in planetary atmospheric physics and chemistry, rekindled in the era of space exploration, brought physical considerations back into focus in addressing the nature of the early atmosphere. Major contributions to the application of atmospheric dynamics to models for primordial planetary atmospheres were made by Walker (1980, 1982), Kasting and Walker (1981) and later by Kasting and collaborators (Kasting and Walker 1981; Kasting 1982, 1988, 1993; Kasting et al. 1983, 1984; Kasting and Ackerman 1986). These studies again indicated carbon dioxide, nitrogen, and water vapor as dominant components for Earth’s primordial atmosphere and of materials such as H₂S, CO, CH₄ and H₂, ranging in concentration from ppm to a few hundred ppm, as minor components.

Melton and Ropp (1958) demonstrated the formation of cyanide ion by the reaction of atomic nitrogen on HCO₂⁻, reactions that may have taken place in cold plasmas in the Archean atmosphere and ionosphere.
Atomic nitrogen reacting with small amounts of methane in the upper atmosphere was also proposed by Zahnle (1986) as a potential source for hydrogen cyanide. Subsequent to Løb’s (1906) work, several authors have continued to investigate the formation of aldehydes from CO₂, crucial for an RNA World and in what appears to be a realistic vision of the early atmosphere.

Modern views have thus returned the atmosphere to only a limited source of strongly reduced compounds for prebiotic synthesis, but still support the production of organic species at the oxidation level of aldehydes. Most importantly, the resulting low concentration in the hydrosphere of organic compounds made available for reaction in a weakly reducing atmosphere has focused the attention on the crucial need for selective concentration mechanisms in models for biopoesis.

A related casualty of the organic aridity of a near-neutral atmosphere is the concept of solution in the ocean, taken for granted almost automatically in much of the literature as the site of early chemical evolution toward complex biomolecules. The dilution in the ocean of soluble compounds from any weak source is forbidding; already sparse, unstable molecules introduced in a volume of $1.3 \times 10^9$ km$^3$ of seawater are mutually unreactive and hardly retrievable by evaporation or other means. Oceanic boundary reactions at the air–sea or crust–sea interface are entropically more advantageous than anything in the free water column and could have involved, for instance, large-scale processing in the hydrothermal plumbing system of the accumulated polymeric tar that constitutes the major fraction of organics from space. Such matter may have been continuously deposited on the ocean floor in the earliest Archean at a rate some $10^2$–$10^4$ times higher than at present (see, e.g., Chyba et al. 1990; Chyba and Sagan 1992). It is conceivable that reactive concentrations of such pyrolysis and hydrolysis products could be maintained in low flow-rate regions of the oceanic crust. Semipermeable membranes of precipitating metal sulfides, observed in laboratory experiments, have been suggested by Russell et al. (1994) as containment for such solutions.

Another way to achieve both preconcentration and processing in schemes for biopoesis was quoted by Charles Darwin: a small lake or “pond” with soluble compounds accumulating from a surrounding drainage area. Solar evaporation in a Darwinian pond would be a questionable concentration process in view of the destructive effect of UV radiation on organic compounds. Selective sorption on surface-active sedimentary minerals may be more effective based on experimental results. In such an environment, wide local ranges of pH and temperature could be achieved (Arrhenius et al. 1997).
A conceptual representation of a number of processes proposed in the literature that could contribute to selective synthesis in the Darwinian environment is schematically illustrated in Figure 3. Extraterrestrial input by proposed higher rates of impact to the early Earth (Sect. 4) is represented by atmospheric ablation of meteors and a cometary impact pond (Clark 1988). Volcanic exhalation products issuing in a glacial environment (Schwartz and Henderson-Sellers 1983, quoting Spitzbergen as a modern counterpart) could be preconcentrated by eutectic freezout in the ice (Sanchez et al. 1966). Oligophosphates condensed in molten volcanic rock have been observed by Yamagata et al. (1991) to be carried into the hydrosphere by supercritical steam and to be effectively concentrated by bilateral surface-active minerals such as double-layer metal hydroxides and clays (Arrhenius et al. 1997). Condensed phosphates may also be produced at contact heating by lava flows over sedimentary deposits containing the protonated Mg-Ca phosphate that is expected as the abiotic product of phosphate precipitation (Sect. 6.1). Surface-active minerals

---

**Figure 3** Physiographic diagram of a “Darwin pond” environment, visualizing proposals made to solve problems in biopoesis by authors referred to in the text. Condensed phosphates are emerging in hydrothermal steam and solution and also produced by lava flows in contact with sediment containing the condensed phosphate mineral, whitlockite, perhaps dominant in an abiotic environment.
forming from aqueous solution are capable of phosphorylation, selective concentration, and oligomerization of aldehydes and carboxylic acid anions (Pitsch et al. 1995; Kolb et al. 1997; De Graaf et al. 1998). Aside from the accumulation and reaction of prebiotically important organic molecules in lakes, porous groundwater aquifers with slowly diffusing and percolating solutions would offer voluminous and chemically diverse environments for selective concentration and reaction. Such a reacting groundwater environment would be important particularly for relevant chemistry taking place on bilateral surface-active minerals such as double-layer metal hydroxides and clays. In the “genetic takeover” concept of Cairns-Smith (1975, 1982), this would be the primary medium for emergence of the earliest life forms, with reactive minerals providing an integrated scaffolding, and a part of life itself.

The conceptual notions presented in Figure 3 should not be taken too literally from a geographic point of view, but they illustrate principles proposed by various authors in order to overcome some of the important geochemical obstacles perceived in molecular evolution in an Archean environment.

2.3 Synthesis of Source Molecules in the Interstellar Medium and the Atmospheric Electric Circuit

The nature of the space medium has increasingly been seen as controlling the production of compounds of prebiotic interest considered difficult to synthesize in an “infertile” terrestrial atmosphere. The favorable conditions for building large organic molecules in the space medium are due to the extremely low kinetic temperatures (10–50 K) and densities (~<10^6 cm^-3) with resulting long collision lifetimes (hours–days), coupled with high excitation, ion, and electron temperatures (10^3–10^6 K). The molecular products of such plasma reactions are gathered in solid condensates and brought to the planets by infalling solids. Such compounds, observed in meteorites, comets, and the interstellar medium, include (in meteorites) ethylene glycol, dihydroxyacetone, glycerol, and glycric acid (Cooper 1996), hydroxyacids (Peltzer et al. 1984), phosphonates and amino acids (Cronin and Pizzarello 1997), in one case with chiral excess. These all are found in extremely low concentration (ppm range or even less), which emphasizes the problem of bringing these materials to Earth at sufficiently high concentrations for biopoesis.

Extended cold plasma regions in the atmosphere include the auroral electrojet (Fig. 4), the subauroral electron deposition region in the upper atmosphere (Fig. 5) and the transient sprites (Fig. 6) and jets in the tropo-
Figure 4  Electric current structure in the aurora. (Courtesy of C.G. Fälthammar.)

Figure 5  Boreal auroral oval seen from space. (NASA.)
sphere-stratosphere. Sprites and jets are luminous transient plasma discharges that, because of the lack of visibility from the ground, have only recently been discovered and explored (Sentman and Wescott 1995; Sentman et al. 1995; Dowden et al. 1996; Hampton et al. 1996). Sprites extend to altitudes of 90–100 km above thunderclouds, forming a link in the atmospheric electric circuit; organic synthesis and transport of products in sprites may have played a role in the upper atmosphere of the Archean. The yield from organic reactions may have been limited by the short pulse duration, which may, however, contribute to efficient quenching of product molecules.

In the auroral oval, the currents feeding into and leaving the circuit are powered by the solar wind-magnetosphere dynamo; electrostatic double layers accelerate electrons into the upper stratosphere (<100 km altitude) with energies often exceeding 10 keV. In the high-energy range of the trajectories, the electron excitation of atmospheric molecules results in disso-
cation and positive ion formation; at thermalization, the electrons attach
themselves to molecules, forming hydrated negative cluster ions. These,
together with solid condensates, dominate the chemical reactions in the
polar stratosphere, today including the formation and destruction of ozone.
In an anoxic atmosphere, they presumably controlled the formation of
cyanide, formaldehyde, glycolaldehyde, and other organic compounds.

In contrast to these cold plasma phenomena, lightning at the high gas
density of the troposphere is associated with high kinetic temperatures and
high energy density. Lightning is relatively effective in activating exo-

ermic reactions such as the oxidation of nitrogen, but ineffective in
synthesizing metastable polyatomic molecules. The high-voltage, low-
current spark technique employed by Miller (1953) and numerous later
workers as a simulation of lightning in fact bears little similarity to this
natural process, and owes its relative efficiency to the narrowness of the
discharge, thus enabling quenching of the numerous polyatomic species
produced in a range of reducing gas mixtures (Schlesinger and Miller
1983). It is notable that the proportions of amino acids found in mete-
orites, with nitrile precursors most likely produced by ion–molecule reac-
tions in the interstellar medium, are similar to the relative amounts found
by Miller (1953, 1987) in the different spark discharge regime. Under
present-day conditions, organic compounds deposited in the ionosphere
and upper atmosphere by interstellar dust and vaporizing meteorites are
destroyed by oxidation such that only the ashes of the meteors remain in
this layer in the form of magnesium, calcium, iron, and other silicate
mineral-derived ions and their condensates. However, in an Archean dys-
oxic atmosphere, meteoritic and cometary ablation products and atmos-
pheric gases in the ionosphere and in the upper atmosphere may have
provided significant sources for organic molecule formation.

The low density and kinetic temperature of the auroral plasma (<350
K), and the darkness during the polar night, would tend to protect reaction
products against thermal and UV decomposition, which overcomes some
problems faced by other production schemes. The polar night cone would
then provide an efficient regime for molecular synthesis in an atmosphere
dominated by N\textsubscript{2} and CO\textsubscript{2} with small mixing fractions of atmospheric re-
ductants such as CO and CH\textsubscript{4}, and with added organic pyrolysis products
from vaporization of extraterrestrial material. Laboratory model experi-
ments (Melton and Ropp 1958) indicate that excited atomic nitrogen
under these conditions reacts with species such as HCO\textsubscript{2} to form cyanide,
an ion of crucial importance in biopoesis. The classic cold plasma model
experiments by Löb (1906) cited above demonstrated the formation of
formaldehyde and formic acid as main products, with glycolaldehyde as
an additional component under similar conditions, and the amino acid glycine in the presence of ammonia. An important difference between the present Earth and the early Archean, is that the auroral current may have been increased by a factor of two if the inferred primordial 12-hour spin rate correspondingly increased the Earth’s magnetic dipole field and beyond this if the magnetospheric current density was enhanced by coupling to an intensified solar wind.

2.4 The Carbon Dioxide Cycle and Greenhouse Effects in the Early Atmosphere

A question of importance for the radiation balance and thermal state of early Earth is how large a fraction of the total carbon reservoir of the planet would at any given time have been present as carbon dioxide in the atmosphere. As realized already in the last century (Högbohm 1894; Arrhenius 1896; J.J. Ebelmen in Berner and Maasch 1996), carbon dioxide plays a primary role in the weathering of rocks. Carbonic acid in rain continually attacks exposed surfaces, which in the case of igneous rocks, consist of silicates with cations mainly from the alkali-alkaline earth and transition element groups. Weathering releases calcium, magnesium, divalent iron, and other cations into solution; in this process the silicates are converted to the corresponding carbonates together with free silica and low-energy hydrated silicates in the hydrosphere. As a result, the store of carbon dioxide in the atmosphere gradually becomes bound as sedimentary deposits of calcium, magnesium, and iron carbonates, which now represent about 80% of the total crustal reservoir of carbon. In the return branch of the geochemical cycle, carbonate rocks, brought to high temperature in the interior of the Earth, dissociate into carbon dioxide, metal oxides, and silicate. The cycle becomes complete when carbon dioxide is returned to the atmosphere by volcanic exhalations.

During the major part of the first 500 Ma, the time of emergence of life on Earth, pH in the hydrosphere has been estimated at 5–7 (Fig. 7), based on the speciation of minerals preserved in the oldest sedimentary precipitates such as banded-iron formations. A number of factors other than the partial pressure of CO₂ come into play in this process, in particular the total area of exposed continental rocks susceptible to weathering, and the increased rate of turnover of oceanic crust and its sediment load by enhanced sea floor spreading, releasing heat and volatiles and subduction of one crustal plate beneath another. Competition to CO₂ as a weathering agent is also offered by other stronger acids present in the primordial atmosphere related to increased volcanism, particularly by gaseous...
HCl and HF released by volcanism. These have been estimated by H.D. Holland (pers. comm.) to have kept pH of the initial hydrosphere as low as perhaps 2 or 3, with a titration time toward neutrality of perhaps a few million years.

Early greenhouse effects have to be considered in the context of the early history of the sun (Owen and Cess 1979). Studies of stellar evolution have made it possible to tie the formation of planets to a specific stage of evolution of young stars, represented by the T-Tauri objects, and to generalize stellar evolution in terms of luminosity. It is on this basis likely that the sun, during the early stages of the history of Earth before 4 billion years ago, would have been about 30% fainter than it is today (Newman and Rood 1977; Gough 1981). As an explanation of the postulated
warmth of the early Earth regardless of the low incident solar radiation, Kasting and Ackerman (1986) have evaluated the effect of an enhanced CO$_2$ concentration. Without the compensating thermal blanketing by photoactive gases, the low solar power could have led to a completely frozen Earth with such a high albedo that the resulting cryosphere might have been prevented from ever returning to liquid state. This frozen world would have persisted to the present unless some later thermal event could have delivered sufficient energy to thaw out (Bada et al. 1994). This cycle is suggested to have been active in assisting in the emergence of life by permitting the production, accumulation, and reaction of hydrothermally produced organic molecules below the ice cover as proposed by Bada et al. (1994). Since the geological record appears to exclude sustained catastrophic impacts later than about 3900 Ma, such a freezing and thawing, which per se does not conflict with any geophysical principles, would need to be relegated to the earliest, unrecorded Archean (Hadean).

Kasting, as illustrated in Figure 8, estimates the initial partial pressure of CO$_2$ to have been somewhere between 0.1 and 10 bar. The lower value for early Archean pCO$_2$ corresponds to the estimated minimum requirement for keeping the Earth from irreversibly freezing over, and thereby increasing the albedo of the planet and locking it into a permanent “ice house.” With those less catastrophic accretion models that lead to moderate average temperatures of the Earth’s surface, the reaction of the crust with rainout of dissolved carbon dioxide could have proceeded in pace with the late phases of accretion and resulted in atmospheric pressure of carbon dioxide in the lower range of Figure 8.

Oxygen isotope measurements (Knauth and Lowe 1978) and petrological observations (Costa et al. 1980) have been taken as indications of high water temperatures associated with the Archean banded-iron formations. The nature of the oldest known sedimentary rocks contributes testimony to a liquid hydrosphere persisting from the earliest recorded stages of the history of the Earth 3850 million years ago, or only 600 million years after the formation of the planet (Dymek and Klein 1988; Mojzsis 1997; Nutman et al. 1997), and giving rise to the conditions necessary for an RNA World that probably developed much earlier.

3. GEOCHEMISTRY OF MOLECULAR EVOLUTION

3.1 The Case for RNA

Since RNA can perform both catalytic and information-transfer functions, and is utilized in the most primitive organisms, it suggests molecules of similar construction as simpler predecessors. This observation provides
the chemical paradigm for research into an RNA World and potential processes for the production and interaction of its components. The current structural design of RNA entails a backbone of pentose (furanosyl ribose) sugars linked by phosphodiester bonds, with nucleobases serving as letters in the genetic alphabet. A structure of this complexity is considered by many unlikely to appear by random molecular evolution. Simple (unidentified) nitriles may possibly have served as precursors of nucleobases in early forms of self-recognizing molecules. Organic phosphates

![Figure 8](image)

**Figure 8** Range of model atmospheric partial pressures of CO₂ (*left vertical scale*) and concentration (*right-hand scale*; present atmospheric concentration as unit) as function of change in solar luminosity with time (units of 10⁹ years [1000 Ma]; horizontal axis). With a 30% reduction in solar luminosity in the early Archean, CO₂ partial pressures of 10⁻¹ bar 4500 million years ago would represent the minimum CO₂ concentration necessary to warm Earth to 0°C at that time. This ignores any thermal increment due to enhanced heat flow from Earth’s interior. The initial CO₂ pressure of 10 bars refers to the model by Dymek and Klein (1988), assuming an ocean covered Earth, and corresponding to a calculated temperature of 85°C at 30% reduced solar luminosity. The figures 5°C–20°C are estimates by Kasting for maximum and minimum temperatures during the Precambrian ice ages, 5°C being a probable limit below which the entire hydrosphere might freeze due to feedback effects; 20°C is the temperature estimate, derived from oxygen isotope measurements, for the time, about 30 million years ago, when the Cenozoic ice caps started to expand. (Courtesy of J.F. Kasting, Pennsylvania State University.)
are pervasive in a wide variety of present-day biochemical processes, implying, because of their unique chemical capabilities, similar functions already in ancient biochemistry (Westheimer 1987) and also in prebiotic chemistry (Arrhenius et al. 1997; Sect. 6.1). Processes for the formation of simple aldehydes as well as for their phosphorylation, concentration, and surface-induced oligomerization to form sugar phosphates have been accomplished experimentally. However, the plausible formation of nucleobases and their attachment to the sugar phosphate backbone under abiotic conditions remains complicated, considering the interference in cyanide chemistry by formaldehyde. In contrast, the next step, the oligomerization (and potentially, replication) of nucleic acids, has been successfully approached (Orgel 1986; Ferris 1987, 1993; Bolli et al. 1997). Insight into the geochemical aspects of these precursor processes is being sought by laboratory model experiments and space observations, and may possibly be obtained from the oldest sedimentary systems on Mars and Earth. Mars is the only planet in our solar system that may have offered conditions for molecular evolution comparable to those on our planet, and it could contain a record of that critical time.

3.2 Search for Source Compounds

Simple source molecules, which could have served as building blocks for RNA or its possible precursor forms, include formaldehyde and glycolaldehyde, active phosphate species capable of phosphorylation, as well as reduced nitrogen species such as nitriles which serve as preferred components of hydrogen-bonding recognition molecules. For some compounds, like formaldehyde and phosphates, acquisition in concentrations that permit their use in natural synthesis reactions appears compatible with geochemical constraints. However, in other cases, like cyanide and ammonia, the accumulation to reactive concentration presents an enigma as far as terrestrial processes are concerned. Similar conflicts appear in the subsequent stages of molecular evolution between, on the one hand, the biopotic wish for unlimited supplies of highly reduced organic raw materials, and on the other, the austere limits imposed by geochemical and geophysical preferences outlined here.

3.2.1 Oxiranecarbonitrile

Any compound expected to be reactive in a natural environment needs to overcome the threat of dilution in the hydrosphere. One of the most efficient ways to selectively achieve the necessary concentration is by excess
Figure 9  (a) Gas chromatogram of the (derivatized) large number of sugars and related compounds induced by a small amount of glycolaldehyde in the indiscriminate formose reaction, which would also require geochemically implausible reaction conditions. The inefficiency of this avenue to formation of specific sugars such as ribose, as well as the relative instability of unprotected sugars, has often been quoted in the literature as a major obstacle to the rise of an RNA World. (Reprinted, with permission, from Decker et al. 1982 [copyright Elsevier Science].)
charge, and this is most effectively accomplished by phosphorylation. Thus, the smallest and simplest (two-carbon unit) aldehyde phosphate (glycolaldehyde phosphate, GAP) has been investigated as a starting material in RNA-related biopoiesis experiments. Studies have also been made of the analogous phosphonate acetaldehyde ester (De Graaf et al. 1997, 1998).

The epoxide oxiranecarbonitrile ($\text{ocn}$) has been considered as a source for GAP in nature (Pitsch et al. 1994). This compound has been shown experimentally to yield GAP by efficient reaction with orthophosphate in aqueous solution and could possibly react analogously on surface-active phosphate minerals or with the more soluble condensed phosphates. If effectively produced in the interstellar cloud medium, $\text{ocn}$ could theoretically become incorporated in cometary condensates and subsequently be transported to Earth. It could also possibly have formed in the upper atmosphere of the Archean Earth, particularly in the electron deposition region below the auroral ring current (Fig. 4) in the Earth’s nightcone, where constructive negative ion reactions are induced and the products are photoprotected. Atmospheric sprites and jets are other cold plasma regions where similarly favorable reaction conditions may have existed (Sect. 2.3).

In view of the chemical attractiveness of the $\text{ocn}$ scheme, a radio telescope search of the interstellar medium for its microwave signature has been initiated; the early results of this investigation show that the abundance is below the detection limit of the technique so far used (Dickens et al. 1996); condensation of low vapor pressure products like $\text{ocn}$ may contribute to their paucity in the gas-plasma phase of the space medium, and their transfer to the interstellar dust component.

### 3.2.2 Cyanhydrin and Hexacyanoferroate

Because of the presumed prevalence of carbon dioxide and water vapor, formaldehyde is likely, next to carbon monoxide, to have been one of the

---

(b) In contrast to the formose reaction, the mineral (in this case hydrotalcite) catalyzed oligomerization of glycolaldehyde phosphate is highly selective, generating mainly erythrose-2,4-diphosphate as a precursor of altrose-2,4,6-triphosphate, the most abundant hexose phosphate product. This reaction also approaches more closely natural conditions; the mineral concentrates the reactant from ~10 µM solution and operates at near neutral pH (~7.5); the products, sorbed in the mineral, are stable under the reaction conditions. Together with glyceraldehyde phosphate, the mineral-induced reaction produces pentose phosphates in high yield, primarily ribose-2,4-diphosphate and arabinose diphosphate. The availability of this reaction path removes a frequently voiced reservation against the concept of an RNA World.
most abundant reduced carbon molecules in the Archean atmosphere–hydrosphere system. Activation mechanisms facilitating formaldehyde formation that have been evaluated experimentally include cold plasmas (Löb 1906), electric sparks (Miller 1953), UV irradiation (Hubbard et al. 1971; Pinto et al. 1980), and thermal shock (Chyba and Sagan 1992). A high abundance relative to hydrogen cyanide places formaldehyde potentially in control of Archean nitrile chemistry and thus of pathways to nucleobases (Arrhenius et al. 1994). Formaldehyde reacts spontaneously with hydrogen cyanide to form cyanohydrin, a well-known reaction that has vexed workers in the field of prebiotic chemistry relying on an unencumbered availability of HCN in high concentration to form a plethora of evolved molecules. Specifically, hydrogen cyanide in hypothetical isolation from formaldehyde has been implicated as a precursor of purines and pyrimidines, a scheme that appears realistic only if conditions can be found that provide protection against formaldehyde.

The initially forming monomeric cyanohydrin (glycolonitrile), when present in high concentration such as in eutectic freezeout and under basic conditions, undergoes rapid oligomerization to a dimer and a trimer (Arrhenius et al. 1994, 1997). Standard mechanistic analysis of the trimerization of cyanohydrin predicts three 6-membered heterocycles; an s-triazine, an oxazoline, and an amino-pyrimidine. The kinetically formed products comprise a dimer hydrate and a trimer based on the 5-membered oxazoline skeleton; the main hydrolysis product is glycolamide. Computational investigations predict that the pyrimidine isomer is thermodynamically favored and that the oxazoline actually found is a kinetic side product. These cyanohydric oligomerization products are of prebiotic interest because they could represent a geochemically plausible pathway to nitrogen bases circumventing the trap that formaldehyde sets for reactions of HCN with itself.

Under certain conditions, depending on pH and concentration, HCN would be expected to react with Fe**, present in comparatively high concentration in the Archean hydrosphere due to the past low partial pressure of free oxygen (Keefe and Miller 1996). The high negative charge of the hexacyanoferrate ion [Fe(CN)$_6$]$^{3-}$ thus formed favors effective sorption in the interlayer of catalytically active double-layer metal hydroxide minerals (Arrhenius et al. 1997 and references therein). These minerals have been found to effectively concentrate charged molecules, such as phosphate, carbonate, and cyanide ions, believed to be significant in nucleoside and nucleotide synthesis, and also sulfur species and negatively charged amino acid monomers and oligomers such as polyaspartate, polyglutamate, and serine phosphate. Ferrocyanide (hexacyanoferrate)
ion, when sorbed into the interlayer of the double-layer mineral green rust ([Fe²⁺Fe³⁺(OH)₆][Aⁿ⁻~4H₂O]), undergoes further reaction with Fe (III) to form ferriferrocyanide (Prussian blue) (Arrhenius et al. 1993), or with Fe (II) in the green rust structure to form ferroferrocyanide (P.S. Braterman, pers. comm.). Such a concentrated mineral reservoir of cyanide could possibly have provided an escape from the dilemma of cyanide scavenging by formaldehyde.

3.2.3 Aldehydes and Sugars

As discussed above, formaldehyde is a natural precursor for selective synthesis of aldehydes because it was presumably, together with formic acid, the most abundant reduced carbon compound that could have formed from the Archean atmosphere. Based on known production mechanisms, formaldehyde was probably generally in excess over hydrogen cyanide, so HCN would have consumed only a minor fraction of the formaldehyde as cyanohydrin. Formaldehyde is a major monomeric component in the coma and tail of comets presumably originating from polyoxymethylene polymers in the cometary nucleus. H-C-O compounds, including formaldehyde, are also produced by photoreduction of carbonate ion with Fe²⁺ in solution (Åkermark et al. 1980).

The classic formose reaction is often invoked in metaphoric schemes as a pathway to sugar formation from formaldehyde. However, it requires relatively high concentration, strongly alkaline conditions, or a suitable catalyst, and produces an indiscriminate mixture of sugars and sugar derivatives of which ribose is a minuscule fraction (Fig. 9a). In contrast to the formose reaction, the bilateral mineral-catalyzed aldolization reaction (Fig. 9b; Sec. 3.2.5) is highly selective and operates from dilute, neutral solution. The dimerization of formaldehyde to glycolaldehyde is an important step (Fig. 10), since the latter permits phosphorylation and glycolaldehyde phosphate has been shown to give rise to RNA-related sugar phosphates (Sect. 3.2.5). The demonstration of glycolaldehyde as a product of cold plasma reactions in addition to formaldehyde and formic acid was one of the early results of the electric discharge studies by Löb (1906). Photochemical pathways have also been shown for the formation of glycolaldehyde (Schwartz 1993; Schwartz and De Graaf 1993).

A remaining impediment to the further reactions of these simple aldehydes to form the sugar component of RNA is that the uncharged state of these molecules renders them unprotected against dilution in the hydrosphere. However, recent experiments have demonstrated efficient phosphorylation of glycolaldehyde in dilute aqueous solution (Krishnamurthy
The attachment of negative charge has proven to be an effective process for selective concentration of aldehydes by sorption on cationic minerals. As pointed out in a classic paper by Westheimer (1987), phosphorylation is nature’s favored solution to the need for charge.

### 3.2.4 Phosphate Condensation, Activation, and Esterification

With the assumption of a fundamental role of soluble, condensed, naturally produced phosphates through early stages of molecular evolution, it becomes necessary to identify mechanisms by which phosphate ion can be coupled to the evolving molecular systems. Extensive studies pointing at such processes have been carried out, particularly by Orgel and collaborators (see, e.g., Österberg and Orgel 1972); an overview of these results is given by Kolb et al. (1997). A recent result is the achievement of efficient phosphate ester formation at near-neutral pH and low concentration of the reactants (Krishnamurthy et al. 1998).

Living organisms voraciously consume available phosphate and interfere with the inorganic phosphate reaction equilibria that must have controlled abiotic phosphate geochemistry. As outlined in Sect. 6.1, the thermodynamically stable phosphate mineral that forms from sterile solutions with the high ratio of magnesium to calcium characteristic of natural waters is the phosphate mineral, whitlockite (HMgCa$_9$ (PO$_4$)$_7$). As a consequence of the protonation, whitlockite readily condenses upon heating and dehydration, forming a series of linear oligophosphates with enhanced solubility compared to the orthophosphate source. Because of the increasing charge with oligomer size, the condensed phosphate species are preferentially concentrated from dilute aqueous solution by the posi-

---

**Figure 10** Reaction between glycolaldehyde phosphate (GAP) and glyceraldehyde-2-phosphate (G2P) in varying proportions in the presence of mixed valence double layer metal hydroxide minerals. Substitution of the metal cation in the main hydroxide sheets of the mineral can be made by divalent metal species (Fe$^{++}$, Mg$^{++}$, Mn$^{++}$...). The interlayer water content is generally in the range of 4–6 molecules per anion, giving an effective anion concentration of the order of 10 M in the reactive interlayer solution.
tively charged double-layer hydroxide minerals, found to be catalytically active in aldol phosphate condensation.

### 3.2.5 Sugar Phosphate Formation by Abiotic Processes

Considering the unique biochemical advantages of phosphate esters in synthesis, conservation, and function of information-carrying and energy-transducing biomolecules, Westheimer (1987) has pointed out the lack of any ionic group that could compete with phosphate species in fulfilling these functions. From the point of view of RNA synthesis, the initial phosphorylation of simple aldehydes would appear to be of central importance, and several arguments point to an early selection of aldose diphosphates in molecular evolution toward RNA-like molecules. One reason is evolutionary: Ribose diphosphate is now the backbone element in RNA, extending phylogenetically back in time to the most primitive organisms known. It must be said that although labeled “primitive,” these organisms at the base of the tree are highly evolved biochemically and must have a long evolutionary pedigree, now seemingly lost from the geologic record on Earth.

Illustrating the efficiency of selective sugar phosphate formation, Müller et al. (1990) demonstrated the formation of rac.-pentose-2,4-diphosphates and hexose-2,4,6-tri phosphates in strongly basic aqueous solution starting from glycolaldehyde phosphate and formaldehyde, resulting in rac.-ribose-2,4-diphosphate as the major product. It was shown that glycolaldehyde phosphate (GAP) is rapidly and efficiently sorbed in the interlayer of bilaterally active double-layer hydroxide minerals from dilute (≥ 10 µM) solution to form a reactive, concentrated (~10 M), quasi-two-dimensional solution constituting an isolated, diffusive, and exchangeable Stern layer in these minerals. Some of these minerals are widespread in nature and would have been particularly abundant in an anoxic Archean hydrosphere. GAP in this catalytic reaction space is rapidly (2 days, RT, pH 7.5) and in high yield (~60%) selectively aldolized to form tetrose-2,4-diphosphates. By further condensation (Fig. 9b), hexose- (predominantly altrose-) 2,4,6-tri phosphates, in the external solution are produced (Pitsch et al. 1995). The sugar phosphates, intercalated in the mineral interlayer, are stable against hydrolysis (~5% in 10 months) and thus form a robust system, in contrast to unphosphorylated sugar in free solution.

The formation of pentose-2,4-diphosphates (up to 25% yield) was observed when a dilute aqueous solution of glycolaldehyde phosphate and glyceraldehyde-2-phosphate was exposed to the double-layer hydroxide
minerals, again at near neutral pH of the external solution (Fig. 10). In this study it was also observed that different divalent metal ions in the double-layer hydroxide mineral had a different effect specifically on the yield of ribose-2,4-diphosphate. For example, with Mn\(^{++}\) as the divalent metal cation, the yield of ribose-2,4-diphosphate was as high as 48% of the pentose phosphates. The stability of the pentose phosphates parallels that of the hexose phosphates within the mineral interlayer.

3.2.6 The State of Metaphors for Natural Synthesis of RNA

As described above, there has been significant progress in the synthesis of the backbone units of RNA under conditions that bear some semblance to those expected in an abiotic world. Still missing are some critically important links in the chain leading to RNA. The natural synthesis of nitrogen bases or some other, undiscovered recognition molecules, is an open question; their attachment to the sugar phosphate backbone molecules is another unanswered problem. Further down the line another stretch of successful synthesis achievements follows: the oligomerization of (activated) nucleotides by J.P. Ferris and his group, the template-directed formation of complementary oligonucleotide strands by L.E. Orgel and collaborators, the demonstration by T. Cech and S. Altman of catalytic activity of RNA bound in ribozymes, and their evolutionary qualities shown by G.F. Joyce and J.W. Szostak. For most of the missing pieces there are obvious experimental search strategies, and the challenge of a natural synthesis of RNA now appears less daunting than other crucial problems in the emergence of life. One of these has been the widely accepted assumption that the terrestrial planets were heavily bombarded, sterilizing Earth as late as 3700 to 3800 Ma, a concept discussed below.

4. GIANT IMPACTS: A THREAT TO EARLY LIFE

4.1 Emergence of Life and the Late Heavy Bombardment

The extent and nature of impacts by asteroids, meteorites, and comets on Earth that occurred between the time of accretion and the first hints of the geological record in the early Archean has until recently been a subject relegated to speculation. This is based on events that happened late in solar system history (after 3900 Ma) recorded on the surface of the moon, which in contrast to Earth, has not been substantially altered and preserves some record of these earliest times. The overlap in time of proposed impacts on the early Earth, with the emergence of life, creates a problem for the sustainability of early ecosystems.
Following the planet-forming accretion period beginning at about 4600 Ma, the deposition of extraterrestrial matter has continued on all surfaces of the planets and their satellites, although today at an extremely attenuated rate. If uniformly extrapolated back in time on the basis of present rates, the amount of material of extraterrestrial origin that would have accumulated following the end of formation of the planets would be represented by only an approximately 10-meter-thick layer covering the Earth. Impacts by asteroid-size bodies have occasionally occurred with detrimental global effects on the biosphere, the latest notable event of this kind being the giant impact at the Cretaceous-Tertiary (K-T) boundary 65 million years ago. Relatively minor increases above the rate of small meteorite influx that we experience as normal are recorded at several occasions in the geological history. One such instance of enhanced concentration of cosmic dust and preserved meteorites in marine sediments is observed at 500 Ma during the Ordovician period and is correlated with a radiochemically dated breakup event in the asteroid belt that must have scattered debris into Earth-crossing orbits (Schmitz et al. 1997).

The chronological record on the moon, established from samples returned by the Apollo and Luna missions, shows that large impacts with high mass and at many times higher rates of impact than today occurred as late as 3800 Ma (Pappanastassiou and Wasserburg 1969; Tera et al. 1974); this phenomenon is often quoted as the “late heavy bombardment” (see, e.g., Chyba 1993). A currently prevalent interpretation of the lunar impact record assumes that the solar system during this time period was invaded by a swarm of marauding objects, bombarding all the planets and their satellites, including Earth and Mars, but with a retrievable record so far available only on the moon. A possible mechanism for this kind of assault has been presented by Kaula and Bigeleisen (1975) and Fernandez and Ip (1983), who suggested interactions with the outer planets that could have caused perturbation of asteroids into highly eccentric and ultimately planet-crossing orbits. Such a swarm of asteroids would, in a relatively short time period, hit all planets and satellites in the inner solar system. Analyses of meteorites for melting, shock, and exposure events provide evidence of collisions in the asteroid belt that extend from the time of formation of the moon at 4450 Ma up to as late as, in a few cases, 4000 Ma and throughout the history of the solar system.

On Earth, the presumed effects of energetic cometary and asteroidal bombardments up to and past 3800 Ma have been estimated to range from evaporation of the entire hydrosphere and melting the surface of the planet, to partial evaporation of the ocean (Sleep et al. 1989; Chyba et al. 1990). In the most catastrophic scenarios, the emergence of life would be
prohibited and any existing life would be extinguished. In milder variants of this impact frustration theory, those life forms could have been spared that resided in the deepest reaches of the ocean and crust. The hyperthermophiles surviving thermal pulses from impacts by hiding in deep-sea vents could then, after the several extinction events implied, emerge as sole survivors in the last preserved ecosystem.

The extrapolation of the lunar impact record to the Earth with consequences for life stands in contrast to new observations on Earth and Mars. The early Archean banded iron formations consist of laminated sedimentary sequences testifying to undisturbed deposition (Sect. 5). Those studied in detail so far bear no evidence of interruptions by catastrophic events such as local impacts and their associated crushed-up rock formations (breccias), tsunamis from remote impacts, or complete or partial evaporation of the hydrosphere. Rather than being sterile, the sediments, extending in time beyond 3850 Ma, contain ubiquitous “chemofossils” (Mojzsis et al. 1996; Mojzsis and Arrhenius 1998) with isotopic composition suggestive of highly evolved enzyme systems—life seems to have been developing for an unknown, but probably considerable length of time before 3850 Ma (Fig. 1; Sect. 5).

On Mars, the record now available is limited to a few meteorites and crater size and density distributions correlated with the dated lunar record (Neukum and Hiller 1981). It is interesting to note that the igneous rock that forms the major part of the martian meteorite ALH84001 of potential biological renown (McKay et al. 1996), has a crystallization age of about 4500 Ma (Jagoutz et al. 1994). This suggests that the corresponding impact region of the martian crust from which this meteorite was derived has not been remelted in the same cataclysmic bombardment that affected the moon late in time. The late heavy bombardment ought to have been even more intense on Mars because of its larger gravitation and proximity to the asteroid belt. The same meteorite bears evidence of shock at 4000 Ma, somewhat earlier than the major phase of lunar bombardment peaking at 3850 Ma. This shock age of ALH84001 has been used to argue the case for impacting of Mars coincident with the terminal bombardment of the moon and thus the rest of the planets of the inner solar system (Ash et al. 1996).

Two interpretations of the late bombardment are possible. (1) The late lunar impactors came from external sources and affected the entire inner solar system, including Earth and Mars, but the impact events were sufficiently episodic so that the observed long undisturbed sedimentary sequences harboring life could be accumulated between catastrophic events and their aftermaths that have so far escaped attention. According to this
In a different interpretation, and on the basis of presently available evidence from the oldest known terrestrial record, the late bombardment of the moon could reflect the process of its own accretion or of its collision with other objects in orbit around the Earth. Such possible events, having dynamically limited effects on Earth, are discussed below.

4.2 Uncertainties about Lunar Origin, Impact Environments, and Hazards to Emerging Life

A currently popular scenario for the formation of the moon assumes that Earth in a relatively advanced state of accretion, including or followed by core formation, would have collided with a hypothetical planet in an Earth-crossing orbit and with a mass substantially larger than that of Mars (Wänke et al. 1984; Cameron and Benz 1991). Ejecta from the impact would have been placed in prograde orbit around Earth and eventually coalesced to form the moon.

Assuming this scenario to be within the realm of possibility, there are several reasons to believe that the last stages of accretion of the moon from its source ring of material could have extended the time of formation of the major impact features on the face of our satellite past 3800 Ma and perhaps thus explain the paradox of heavy bombardment and the presence of early life forms occurring simultaneously. One such reason for retardation of the terminal accretion would be the metastable retention of accretional material by 1:1 resonance in the Lagrangian points 60° before and after the moon in its orbit. Such resonance locking is a feature seen, for example, in Jupiter’s orbit, where two clusters of asteroids, the Trojans, are bound in its two Lagrangian potential wells. In the case of the moon, these wells, because of its lower mass, would be much more shallow, and the rocks contained in them could be scattered out of resonance at an early time such as during a period ending past 3850 Ma. Their ghosts in the form of remaining matter, the Kordylevsky clouds, are claimed to have been confirmed by space observations (Roach 1975).

This explanatory legend is tied to the specific theory for the impact origin and evolution of the moon. In theories for the origin of the moon by capture (Gerstenkorn 1955, 1969; Singer 1972), alternative explanations are offered for the late lunar bombardment. One suggestion, based on the persistent generation of regular systems of proportionately small
Figure 11  Three schematic views of the late lunar bombardment, and its relationship in time to the undisturbed 3800 Ma Isua sediments, southern West Greenland. (Diagram courtesy of K. Zahnle, NASA Ames Research Center, and S.J. Mojzsis, UCLA.)
satellites in the equatorial plane around all magnetized planets, assumes
that Earth would also originally have been endowed with a satellite sys-
tem of perhaps half a dozen moonlets swept up by the receding, abnor-
mally large, and differently formed moon in its early recession from the
Earth (Alfvén and Arrhenius 1969, 1972). Collision velocities of one or
two km s⁻¹ would have substantial cosmetic effects on the surface of our
moon but would create ejecta generally slower than its escape velocity of
2.4 km s⁻¹. Such slow ejecta could have been retained in near-lunar orbit
and recaptured by the moon with only a minor fraction overcoming the
substantial energy barrier for diversion to Earth-impacting orbits.

All of these theories of lunar origin and collision history remain spec-
culative. However, there are avenues toward verification of related cir-
cumstances. One consists in detailed examination of as long sequences as
possible of the oldest laminated sediments such as those described from
the Isua district in Greenland (Sect. 5), searching for and identifying de-
positional irregularities and embedded meteorites and interplanetary dust
in them. Such quantitative elemental and mineralogical measurements are
likely to give information about the existence, rate, and frequency of
bursts of extraterrestrial accretion, and about their source.

The second approach toward verification involves analysis of lunar
samples, chronologically identified with, and serving as the basis for, con-
clusions about the late bombardment on the moon. If the projectiles were
interplanetary marauders, they would be expected to consist of undiffer-
entiated cometary or meteorite type material, conferring their characteris-
tic excess of platinum group elements (PGE; Ir, Os, Pd, Pt, Ru, Rh) to the
ejecta, with allowance for dilution with lunar target material. If they were
to be identified with partly differentiated asteroids, some impact material
would be depleted in PGEs, but others, enriched in metal, would have a
correspondingly enhanced PGE content. If, however, such a signature is
found to be consistently absent in the samples representing these events,
the most likely interpretation would be that the impacts were caused by
metal-depleted low-density material of the kind that formed the moon and
possibly original Earth satellites.

In conclusion, the necessarily speculative views of the hardships or
even survival of life during the first quarter of the existence of Earth are
now being replaced by observations in the oldest preserved records on our
planet by the study of probably even older samples from ancient terranes
on Mars, and by extended analysis of the lunar samples from the Apollo
missions. Life on Earth instead of (as is often quoted) originating 3.5 to
3.8 billion years ago, is found to have developed to a high degree of au-
totrophic sophistication already before 3850 Ma. Carbon isotope fraction-
ation mechanisms indicated by the geological record of even the oldest rocks are matched today only by the action of enzymes such as the ribose- and ribulose-phosphate carboxylases that mediate the first step in carbon fixation (Mojzsis et al. 1996).

5. THE RECORD OF LIFE’S APPEARANCE ON EARTH

5.1 Liquid Water and Life

Once the need for mutually supportive criteria for early Earth environments and organic reactions has been established, the emergence of life on Earth then becomes a geological issue as well as a biochemical one, and can be considered among the most important developments in the geochemical history of the surface of the planet after its formation. The initiation of biological processes could have occurred in the surficial layer of the planet comprising the hydrosphere, atmosphere, and lithosphere, so long as conditions persisted that satisfied the stability of liquid water on million- to billion-year time scales. Bioessential element cycling (H, C, N, O, P, S) is accomplished in the hydrosphere, by life. Therefore, the earliest evidence for liquid water at the surface of the Earth may be taken as the principal constraint on the timing of the emergence of life, and on the environmental conditions necessary for its sustainability, propagation, and evolution (Chang 1988).

After formation, the preservation of water oceans over geologic time scales is the main characteristic that separates the geophysical and geochemical evolution of Earth from the other terrestrial planets. The presence on a planet of a stable hydrosphere defines habitability. Liquid water on a planetary body is maintained under conditions such as atmospheric density, proximity to the central star, or the effects of another energy source such as tidal heating, and is used to define the habitable zone of the solar system within which a planet or satellite could harbor life (Kasting 1993; Sect. 2). A liquid water ocean hydrothermally alters oceanic crust during and following mid-oceanic ridge volcanism. This crust is then eventually drawn into subduction zones to form hydrated silica-saturated melts, the building material of granites and, ultimately, of continents. The effect of water, and the resulting generation of continental crust, on the bulk chemistry of the hydrosphere-atmosphere-lithosphere and biosphere system cannot be overstated. That a liquid water ocean has remained stable on Earth for the length of the known geological record implies that surface temperatures at the earliest recorded times were within the presently observed range. As has been emphasized
previously, water has played a profound role in mediating tectonic styles on Earth, which makes possible the continued element cycling needed to maintain chemical equilibrium of the hydrosphere and atmosphere over billions of years. In one theory (Woese 1979; Woese and Wächtershäuser 1990), liquid water in the form of droplets is postulated as the locus for the origin of life in the cloud cover over a hot, uninhabitable crust, long before the rainout of a homogeneous hydrosphere. Liquid water, rather than steam, ice, or supercritical fluid, is essential for the origin and development of life. Early Archean sedimentary precipitates that must have formed in water and are older than 3800 Ma, such as cherts and banded-iron formations, signify that the terrestrial processes of accumulation and saturation of dissolved species like silica, carbonate, ferrous iron, and phosphate were operative at and before that time (Mojzsis et al. 1996; Mojzsis 1997; Nutman et al. 1997). The accumulation and concentration of such dissolved species, as reflected in the rocks, reveals the bioavailability of carbon and phosphorus in the early hydrosphere and has opened avenues to the search for evidence of life in the oldest known rocks.

The Earth’s sedimentary rocks contain a relatively detailed record both of the planetary surface environment and of biological evolution only during the last 600 million years or so. The Earth lacks a record of its own accretion and of its earliest subsequent history, because of the actions of plate tectonics, weathering, and denudation. These processes, activated by water, recycle and modify crustal materials, and (until recently) the oldest known rocks of a clearly sedimentary origin extended back only into the range 3600–3700 Ma. Newly discovered deposits, as discussed below, extend into an age older than 3850 Ma. This is also the time at which the Earth was previously speculated to be uninhabitable due to heavy bombardment by objects raining in from space. New perspectives on this early period in the history of the Earth have been opened up by the finding instead of undisturbed sediment sequences containing evidence of biochemically advanced life forms in the oldest rocks. These recent observations have consequences for constraining environments for life also for the other planets and satellites of the outer planets in our solar system, and for the rapidity of life arising under the right conditions elsewhere. Due to the Earth’s continuous self-digestion of crust, it is unlikely that a geologic record penetrating much deeper into the infancy of life can be found. Much hope is therefore attached to the exploration of the ancient surface of Mars, where remote sensing indicates minimal crustal recycling and the presence of an early, and in some regions, vigorous, hydrological cycle of weathering and deposition in liquid water. Such regimes on ancient Mars gave rise to extensive
sedimentary deposits that are the principal targets for sample return missions, and are expected to provide crucial clues to the nature of the earliest planetary surfaces and prospects for past life, or perhaps even events leading to life on Mars.

5.2 The Incomplete Geologic Record

Age calculations based on radiogenic isotope distributions in primitive meteorites, the least evolved objects in the solar system, indicate that the events leading to the condensation, aggregation, and differentiation of the source material for the planets took place in an interval of approximately several tens of millions of years between 4600 and 4500 Ma. The impacting of the early Earth by meteorites and comets and the recycling mechanisms active from the outset add to the difficulty of arriving at an agreeable age of the first terrestrial crust. In related scenarios, assuming an early accretion of an iron-rich core (heterogeneous accretion) followed by growth of a mainly silicate mantle at low average surface temperature, the planet could have become habitable during or soon after planetary formation up to four and a half billion years ago, permitting several hundred million years for the emergence of life. Heterogeneous accretion theories for the formation of the Earth thus connect with biopoiesis scenarios, which for intuitive reasons need the longest possible time available for the emergence of life on Earth. It is worth emphasizing that there is no factual evidence for the length of time actually required for biopoiesis (Orgel 1998). Heterogeneous accretion of the planets is also favorable for models seeking to maximize the rate of generation of methane and hydrogen from the reducing power of available metallic iron while minimizing (by other means) their destruction and escape rate (Miller et al. 1998).

In contrast, homogeneous accretion theories assume accretion of mixed silicates and iron metal and subsequent core formation by gravitational settling of the metal to the center of the planet. The liberation of gravitational energy would be of such magnitude that the entire planet would melt, and all volatiles would form a dense atmosphere. A dense atmosphere would delay substantially, perhaps forever, the time when the surface would become inhabitable.

All known remnants of the oldest sedimentary record are locked in ancient (3600–4000 Ma) granitic gneiss complexes. In these terranes, the oldest sediments appear as strongly deformed rafts and enclaves between the much more voluminous igneous rocks, which have been severely altered by chemical changes induced by heat and pressure in the process of metamorphism. Sediments and volcanic rocks that are recognizable from
the earliest Archean of Greenland (which are as old as 3900 Ma) include quartzites, originally precipitated as cryptocrystalline silica ("chert") directly from solution onto the early Archean seafloor; banded iron-formation (BIF) as alternating bands of silica (as quartz) and magnetite (Fe$^{++}$Fe$_3^{+++}$O$_4$), conglomerates probably formed in debris flows from river channels or submarine landslides; quartzitic sandstones that may have been the remains of stream and river run-off channel deposits; and pillow basalts and other volcanic rocks that are formed in a marine system of active volcanism, erosion, and sedimentation (McGregor 1973; Nutman et al. 1984).

Definite sediments older than 3600 Ma have not yet been discovered outside of southern West Greenland. The ages of these oldest Greenlandic sedimentary rocks range from about 3900 Ma (Nutman et al. 1997) to about 3600 Ma (McGregor 1973; Moorbath et al. 1973). A large section of the sedimentary and volcanic rocks of the Isua district of southern West Greenland consists of units that may have been formed in, or adjacent to, volcanic arcs (Dymek and Klein 1988), as exemplified in the present-day Western Pacific. Such an island arc sedimentary environment is a typical case preserved in the rocks of the extensive Archean Greenstone Belts (Eriksson et al. 1997) found throughout the world as massive accumulations and intercalations of sediments, volcanic rocks, and their metamorphosed remains. These preserved rocks probably represent a common style of marine sedimentation in the Archean and are the oldest preserved remnants of ecosystems yet recognized. Some types of sediments found in the oldest rocks of West Greenland may be the weathering products of emergent land masses such as volcanic islands and/or perhaps microcontinents, similar in some ways to present-day New Zealand. Such land masses, in an early CO$_2$-rich atmosphere, with water abundant and with no vegetative cover, would have experienced massive physical and chemical weathering contributing to the inventory of sediments (Sect. 2.4), including the carbonaceous remains of organisms accumulating on the ancient ocean floor. Many areas of the world, however, contain igneous rocks and their metamorphic equivalents with ages in excess of 3600 Ma. Such rocks are found, for example, in the northeastern Sino-Korean Craton (Song et al. 1996), the Napier Complex in Antarctica (Black et al. 1986), the Acasta Lake area, Northwest Territories, Canada (Bowring et al. 1989), and southern Africa (see, e.g., Compston and Kröner 1988). Some of these regions, like Antarctica, are little explored and are likely to contain exposures of as-yet undiscovered very old rocks, possibly including those hosting sediments and chemical traces of early life.
The oldest relatively unaltered sediments are found in the approximately 3500 Ma Warrawoona Group of the Pilbara Craton, Western Australia (Pidgeon 1978; Barley et al. 1979; DiMarco and Lowe 1989) and the Onverwacht Group of Swaziland, South Africa (see, e.g., Lowe and Bryerly 1986). These sediments were deposited in a shallow marine setting with abundant and well-preserved structures of diverse type, including those representative of shallow water deposition at continental and/or island margins. These rocks suggest ice-free oceanic conditions in the early Archean before 3500 Ma, at least in their local environment of deposition, even given that the sun was less luminous at that time. Well-preserved stromatolitic structures and yet rarer microfossils, recognizable by their shapes (morphofossils) in early Archean cherts of the Apex Basalt, Warrawoona Group (Schopf and Packer 1987) are strong evidence for advanced microbial (possibly even cyanobacterial) life then (Awramik et al. 1988; Schopf 1993).

Alternative views to the interpretation of the age relationships of these rocks and the genetically differentiated nature of fossilized organisms are given by Buick (1990) and Doolittle et al. (1996), respectively. In addition to the morphological evidence for life put forth from the interpretation of microfossils by micropaleontologists, stable carbon isotopic analyses reported from these rocks have, as in the older Greenland rocks, provided strong evidence for life in the early Archean that is independent of the presence of microstructures (for review, see Hayes et al. 1983; Strauss and Moore 1992; Sect. 6.2). It has long been recognized that the early Archean morphofossils represent life forms that must have required a considerable period of time to have evolved from inchoate life to the metabolically sophisticated and structurally complex organisms inferred by Awramik and coworkers (1988) and Schopf (1993) solely from their shapes. These observations must therefore relegate the emergence and first appearance of life on Earth, and certainly the hydrosphere, to the time period before 3800 Ma (Schidlowski 1988; Mojzsis et al. 1996; Nutman et al. 1997).

6. THE PHOSPHATE-GRAPHITE ASSEMBLAGE AS A BIOMARKER

6.1 The Importance of Being Phosphate

Phosphate is an essential constituent of all life, and its controlling importance in biochemical functions is likely to extend back to the origin of life itself. Aside from its role in the structure of RNA and DNA, the unique chemistry of phosphate is utilized in a vast variety of biochemical functions (Westheimer 1987). As a consequence, phosphate is a prominent
component of bioorganic matter in marine sediments, and phosphatic biominerals appear in intimate association with isotopically light carbonaceous matter, consistent with its biological origin. This characteristic association of the remains of organisms with phosphate, primarily in the form of basic calcium phosphate (the mineral apatite), is found in presently forming sediments and throughout the geologic record to the oldest preserved sedimentary rock so far identified.

To understand the interaction between abiotic phosphate mineral formation and phosphatic biomineralization, knowledge of the solubilities and thermodynamic stabilities is crucial. The notable disequilibrium effects exerted by organisms on phosphate mineralization may be utilized for indicating the presence or absence of cellular life: Controlled synthesis experiments suggest that the deviation from equilibrium expressed by the formation of apatite \((\text{Ca}_5(\text{PO}_4)_3 \text{(OH, F, Cl)})\) instead of the equilibrium phase whitlockite \((\text{HMgCa}_9(\text{PO}_4)_7)\) in natural waters is due to the ability of cellular organisms to dynamically segregate \(\text{Mg}^{++}\) into solution while selectively retaining calcium, and in this environment to nucleate apatite by interaction with specific cellular polypeptides. In controlled sterile experiments, whitlockite is formed at \(\text{Mg/Ca} > 0.04\) (LeRoux et al. 1963), at \(\text{pH} < 8.5\), and at 0–100˚C. Whitlockite, rather than apatite, is the thermodynamically stable phase of the calcium phosphates under these conditions (Verbeeck et al. 1986; Driessens and Verbeeck 1990; Elliot 1994; Gedulin et al. 1993). Absence of life would therefore be expected to be reflected in the appearance of whitlockite (at \(\text{pH} > 8.5\) together with apatite) as the sedimentary phosphate mineral, and without association with bioorganic carbonaceous remains in an ancient sedimentary record (Arrhenius et al. 1993). Additional evidence is provided by the fact that biogenic apatite, in modern as well as ancient sediments, is found to be intergrown with residual organic matter from the organisms mediating the deposition of the mineral. In sedimentary rocks from the Archean, including the oldest known metasediments on Earth, the organic matter component of these microaggregates has been completely altered to crystalline graphite, enveloping or enclosed in the recrystallized apatite component (Fig. 12). Such characteristic apatite–graphite aggregates are found to be ubiquitous in sediments of increasing age such as the Gunflint chert (2100 Ma), Hamersley BIF (2500 Ma), Isua Supracrustal Belt (3600–3810 Ma), and what is the oldest currently known metasediment, the BIF of Akilia island in Greenland (older than 3850 Ma; Fig. 1). It would thus appear that evolved cellular life has existed on Earth well beyond 3850 Ma, leaving an interval of no more than 600 Ma for the emergence and early evolution of life on Earth following
planetary formation and the time of the first isotopic evidence for biological activity.

6.2 Carbon Isotopic Record of Life

The most extensive alteration by life arises from the release of free oxygen by the splitting of water during oxygenic photosynthesis, which is catalyzed by phosphate-based enzymatic mechanisms. This creates the reducing power needed for the synthesis of biogenic carbonaceous matter, which becomes enriched in $^{12}$C to a unique extent in this process. A record

*Figure 12*  
At elevated temperature and high pressure during metamorphism of rocks, the individual clusters of apatite microcrystals merge into a single crystal of the mineral, and the carbonaceous material remains of organisms convert to polycrystalline graphite (elemental carbon). Due to its surface tension on the phosphate, the originally disseminated carbonaceous matter coalesces into rounded inclusions in, and surface coatings on, the host apatite crystals. This scanning electron micrograph of a polished section from the >3850-million-year-old Akilia Island sedimentary rock shows a void, left after dissolving an apatite crystal by 30-min etching in 2% HNO$_3$, revealing an acid-resistant carbonaceous inclusion. This carbon is a chemofossil typical of those analyzed in Fig. 13 as “oldest terrestrial sediments,” and with a carbon isotope composition indicating a bioorganic origin. The cross-section of the inclusion is $\sim$10 µm: (Photograph by S.J. Mojzsis, UCLA.)
of this alteration and partial takeover of the global carbon cycle by life becomes captured in the carbonaceous remains of organisms in sedimentary rocks where they can be identified from their stable carbon isotopic signature, frequently in close association with biominerals. Earlier analyses (Fig. 13) (Schidlowski 1988) provided suggestive evidence on this basis for a biogenic origin of carbonaceous remains in some sediments of the Isua district in Greenland (3810 Ma), but were regarded as inconclusive because of the low resolution of the technique used and uncertainties about the sources of carbon described. More precise data have since been obtained by analyzing microscopic carbonaceous inclusions in apatite crystals in the sedimentary Akilia island rocks with an age >3850 Ma, objects believed to be of cellular origin (Fig. 12). The minute dimensions (5–15 µm) of the carbonaceous inclusions require that the mass spectrometric measurements to determine their isotopic compositions be carried out by high-resolution ion microprobe analysis.

The stable isotopes of carbon ($^{12}C$ and $^{13}C$) are partitioned as a result of both equilibrium exchange reactions and kinetic effects due to the metabolic mechanisms of organisms as well as inorganic processes. Inorganic processes such as evaporation, diffusion, and condensation are of far less magnitude in fractionating the isotopes than biogenically induced fractionations. In the standard delta notation, $\delta^{13}C$ is defined as the ratio of the $^{13}C/^{12}C$ of the sample relative to the ratio of the conventional PeeDee Belemnite (PDB) standard expressed in parts per thousand (per mil): $\delta^{13}C = \left[\left(\frac{^{13}C}{^{12}C}\right)_{\text{sample}} / \left(\frac{^{13}C}{^{12}C}\right)_{\text{PDB}}\right] - 1 \times 1000$. High-resolution mass spectrometric measurements indicate a strong fractionation of $^{13}C/^{12}C$ ($\delta^{13}C = -23$ to $-50\%$), a range uniquely limited to bioorganic carbon and well separated from the range (around 0–10%), of inorganic sources on Earth (Fig. 13) and in consonance with the bioorganic origin suggested by the phosphate mineral phase association and crystal chemistry (Mojzsis et al. 1996).

7. CONCLUSION: THE WEB OF STRATEGIES

Problems related to biopoiesis, the creation of life, are approached in the literature in three stages with growing constraints. At the most fundamental level, an inquiry is made about probabilistic aspects and the generation and transfer of information. Although inspiration at this stage is derived from the chemical structure of the modern replication system, no detailed inferences are necessarily made about structures or reactions participatory in the origin of life; in the words of one of the leaders in the field: “the chemists will take care of that.”
Figure 13  The distribution of carbon isotopic compositions in nature, showing the enrichment of the light isotope $^{12}\text{C}$ by autotrophic organisms (expressed as negative values of $^{13}\text{C}$). This marked fractionation takes place at the enzymatically mediated transfer from the pool of inorganic carbon (mainly CO$_2$) and identifies bioorganic carbon in fossilized carbonaceous remains of organisms, such as in Fig. 12. Reduced carbon in sedimentary and metamorphic rocks is largely of such organic origin.
The organic chemists thus enter at the next stage in the interpretation of what it takes to make life by attempting to identify any stepwise organic chemical process that, coupled with various energy sources, could possibly have led from simple source molecules to complex structures. Such complex structures must have become functionally similar to those active in present-day biosystems. In order not to miss any opportunities at this level, most actors in the origins-of-life field purposely avoid side glances at complicating environmental constraints. Thus, if an appealing but clearly unnatural process is found, adaptation to natural conditions is left behind as a secondary problem to be dealt with downstream. Experimental proof is generally demanded for proposals at this level to be taken seriously.

At the third level, the question is raised whether chemically attractive reactions are compatible with environmental constraints imposed by geophysical considerations. Since natural processes are frequently highly complex, and since there is no cohesive geological record from the first 500 million years on Earth, a rather wide latitude for chemical speculation remains yet possible. However, with the combined criteria of the laws of physics and relative plausibilities of geochemical models, in addition to the important and increasing access to planetary materials and the oldest rocks, a more limited set of likelihoods is produced for the conditions present on Earth at the time of the origin of life. A typical example is presented by past chemical theories requiring a high atmospheric partial pressure of molecular hydrogen for desired reactions to proceed. This condition is fulfilled on massive gas-giant planets of the outer solar system like Jupiter. However, the light gases hydrogen and helium rapidly escape from the Earth’s weaker gravitational field, and hydrogen generation sources on a planetary scale, capable of counterbalancing the escape in a sustained fashion, are unknown.

We have followed the record of the Archean Earth and prebiotic chemical schemes within the context of early life and the postulated RNA World. We see that the accretion history and early thermal state of the Earth remain subjects of widely varying hypotheses and conjectures, based only on fragmentary observational evidence that provides few concrete constraints. The body of such evidence is, however, increasing, deriving from measurement of magnetization and chemical separation in the interstellar medium, the observation of planet formation around protostars, radiochemical timing of the formation and development of the internal zones of Earth and Mars, and uncovering of increasingly older sedimentary records on Earth. All of these fields are in rapid development based on deliberate research efforts aimed at understanding the early
Earth and the origin of life. In the present situation it is, however, possible to claim initial thermal states of the Earth ranging from a low average temperature established already during the later stages of accretion, to a planetary ocean of molten magma blanketed by a hot and dense atmosphere, and requiring some kind of corrective acts to convert to a habitable crust and hydrosphere. These events have to take place in time to permit the evolution of biochemically advanced life to appear less than 650 million years after the accretion of the Earth; whether the earliest hints of life so far recognized were part of an RNA World remains unknown.

In this situation it is tempting for organic chemists, pursuing all chemically possible but otherwise unrestricted alternatives, to claim that as long as geologists, planetary physicists, and astrochemists cannot make up their minds, and continue to vacillate between such extremes as discussed above, there should be complete liberty not only for exploring all conceivable chemical pathways, but also to claim, without reservation, their validity on “the primitive Earth.” However, as discussed above, additional and much more specific restrictions arise from consideration of the early state of the atmosphere and hydrosphere based on geophysical constraints. This, together with the question of life struggling against all odds in a fragile peace with impacts or other environmental catastrophes in the earliest times brings out the unsolved problem of how much time is needed for life to follow the “cosmic imperative” of de Duve (1995); to emerge from an inorganic state that, according to Descartes “does not act” but merely exists.

ACKNOWLEDGMENTS

The authors acknowledge support from NASA grants NAG5-4563 and NAGW-1031 (G.A.), and National Science Foundation grant EAR-9704651 (S.J.M.). Comments on the manuscript by C.E. Salomon are gratefully appreciated. We have benefited from discussions with P.S. Brateman, C.F. Chyba, M. Eigen, C. Engrand, A. Eschenmoser, M. Fayek, H.D. Holland, J.F. Kasting, G. Marklund, L.E. Orgel, and K. Zahnle. We also thank T. Cech for constructive editorial comments, and the invitation to participate in this volume.

REFERENCES


Arrhenius S. 1896. On the influence of carbonic acid in the air upon the temperature of the ground. *Phil. Mag.* **41**: 237–276.


