[Frontiers in Bioscience 19, 1117-1150, June 1, 2014]

Emergence of life: from functional RNA selection to natural selection and beyond

Jeffrey Tze-Fei Wong

Division of Life Science and Applied Genomics Center, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, People's Republic of China

TABLE OF CONTENTS

1. ABSTRACT

This study tracks the rise, evolution and post-evolution of the genetic information system through emergence of life. The major stages traversed include prebiotic synthesis, functional RNA selection by metabolite, RNA World, peptidated RNA world, co-evolution of genetic code and amino acid biosynthesis, last universal common ancestor, Darwinian evolution and synthetic life.

2. INTRODUCTION

The origin of life has been a fascination as old as human history. During the past decades, advances in biochemistry and molecular biology have given a thorough account of much of cellular life based on the ground rules of the DNA-RNA-protein troika. However, as Monod (1) pointed out, the origin of life, the structure of the genetic code and the nature of neural memory represent three frontiers of biological science that are not resolvable by the workings of the troika. Enquiry into the emergence of life therefore has to be conducted on the foundation, but outside the bounds, of extant life.

Polanyi (2) proposed that the design of a machine controls its component physicochemical processes. It harnesses these processes to serve its mission, and is not deducible from these processes. Likewise, DNA sequence embodies the system design of an organism harnessing its molecular biological processes, and the design cannot be deduced from those processes. Furthermore, as pointed out by Pattee (3), the degree and type of order found in genetic information cannot be explained solely as the accumulation of evolutionary wisdom resulting from the action of natural selection on self-replicating systems, because the biological process of replication is itself dependent on the pre-existence of such order. It follows that replication and selection have to begin from well ordered sequences rather than random sequences. Consequently ordered macromolecular sequence information must be generated during the prebiotic phase.

Abel (4-6) examined the nature of *prescriptive information* (PI), which includes genetic information of modern organisms, that needs to arise prebiotically, and concluded that PI cannot be just intuitive or semantic

information, but has to be linear digital information in the form of 0's and 1's, or A, G, T and C useful in steering events and processes toward pragmatic benefit, selecting function over non-function. Without prebiotic PI, prebiotic development would founder, and genetic information would be devoid of an origin. How prebioitc PI might arise therefore represents a foremost challenge.

The aim of the present study is to track the rise, evolution and post-evolution of the prescriptive/genetic information encoding system through its prebiotic, biological and synthetic phases.

3. DEFINITION OF LIFE

To delineate the physicochemical events that shaped the emergence of life, the basic essence of living matter needs to be defined. The proposals for a definition of life that have been made over the years are striking in their diversity (7, 8):

FG Hopkins in 1913—A minimum requirement for life is a *"dynamic equilibrium in a polyphasic system."*

JBS Haldane in 1952—*a simple organism such as a bacterial virus contains about 100 bits of negative entropy or information and this is about the amount that would arise spontaneously in 10^9 years in the volume of the primitive ocean.*

NW Pirie in 1957—*"I argued twenty years ago that a rigid operational definition is not possible. This seems now generally to be accepted."*

David Abel—*"The more we can distil the essence of 'life', the greater the hope of elucidating the lost pathways of abiogenesis."*

Gustaf Arrhenius—The basic ingredients are self-organization, self-replication, evolution through mutation, metabolism and concentrative encapsulation.

Andre Brack—Self-reproduction, mutation and evolution.

David Brin—Energy flows downhill and order, information and manipulative ability rise steeply inside local bundles of space-time.

David Deamer—Semi-permeable boundary structure, a system of polymeric catalysts and a system of polymeric instructions.

Christian de Duve—*"Life is what is common to all living things."*

Klaus Dose—Membrane, metabolism, control of metabolism, replication and mutability making possible evolution.

Ricardo Guerrero and Lynn Margulis—Matter that makes choices, binds time and breaks gradients.

Romeu Cardoso Guimaraes—Metabolism, growth and reproduction with stability.

Robert Hazen—Metabolism and reproduction with variation, concept of a sequence of discrete emergent steps.

Gerald Joyce—Gives the NASA Working Definition of *"LIFE is a self-sustained chemical system capable of undergoing Darwinian evolution"* in which self-sustenance is supported by genetically instructed metabolism.

Vladimir Kompanichenko—Organized form of intensified resistance to self-propagating processes of destruction.

Hans Kuhn—*"Physical objects come into being that behave as if they have a purpose...Let us call physical objects with this fundamentally new property (not present in any ancestral form in a prebiotic universe) living."*

Stanley Miller—The origin of life is the origin of evolution. Darwinian evolution requires replication, mutation, selection.

Janet Siefert—*"One can begin to postulate on the defining moment or conditions for 'life' that arose here on Earth…I posit that life can be defined as the culmination and the eventual simultaneous occurrence"* of the four events of replication, translation, control and cell wall.

Eors Szathmary—Gives Tibor Ganti's definition based on metabolic network, template macromolecule and encapsulating membrane, all of which are autocatalytic.

Hubert Yockey—Having a genome and a genetic code.

Despite the diversity, there is also remarkable consensus in the recognition that that life has to be defined based on a combination of constituent attributes rather than any single attribute. The reason is straightforward: few of the single attributes are really unique to life. For instance, self-reproduction/replication is observed with salts crystallizing from supersaturated solutions seeded with a few crystals, there are many geochemical cycles on Earth that are even more ancient than metabolic cycles, all soap bubbles have membranes, and evolvability is displayed by ribozymes undergoing *in vitro* evolution. After all, if evolvability were an exclusive property of the living state, no prebiotic precell could have evolved to generate the living cell.

Among the attributes put forward as life-defining, those of metabolism, catalysis, membrane and genes/replication have attracted wide attention. Most of the biochemical reactions occurring inside the cell are catalyzed by enzymes, or at earlier times by ribozyme-like polymers. The generation time for a present day microbe is about 20 minutes, or 4×10^{-5} years. For RNA hydrolysis at least, ribozymes tend to be 10^3 fold slower than enzymes, but 10^{10} fold faster than uncatalyzed reaction. On this basis, the uncatalyzed replication of a primitive RNA genome is estimated to be of the order of 4×10^8 years. Thus not many more than ten uncatalyzed replications could be accommodated in the 4.6×10^9 years of Earth's history. Furthermore, because the multiplication of any RNA genome becomes possible only if its physical life time exceeds the time it takes to replicate, the *stability theorem* applies (8,9):

$$kLT < 1$$

where k is the rate constant for the hydrolysis of a phosphodiester bond in RNA, L the number of phosphodiester bonds in the genome and T the replication time. For a primitive genome with three 50-base RNA genes, L = 147 and $k = 1.5 \times 10^{-9}$ min^{-1} at pH 7. Therefore T must be less than 8.6 years, or else the RNA genome

would face extinction caused by gene degradation before it has undergone replication. Catalysis is essential to achieving such fast replication.

However, the nature of ribozymes demonstrates that replicators/genes can also fulfill the task of catalysts. Even in modern life, where ribozyme catalysis has mostly given way to enzyme catalysis, DNA/RNA as gene and mRNA still encode the biosynthesis of protein enzymes. Therefore in the RNA World as much as the Protein World, the biocatalysts need not be considered as independent players. Instead, they can be regarded as built-in appendages of catalysis-competent replicator/genes with the inherent capacity of bringing catalysis with them into the living system. Therefore, as suggested by Arrhenius, Dose and Ganti, membranes, replicators and metabolites are appropriate components for the living state, provided that the replicators are inherently catalysis-competent. The criterion for the attainment of life by an assembly of these components has to be performance-based with the goal line being the accomplishment of self-reproduction. Such a performance based definition of life can be stated as follows (8):

Life is the integration of membranes, metabolites and mutable, catalysis-competent replicators into a single entity. The defining moment of life is the moment of successful integration signaled by the self-reproduction of such an entity.

The birth of the living cell accordingly took place at its first successful self-reproduction. The inclusion of 'mutable' in the definition is based on the recognition that replicators are chemical in nature, and all chemical compounds are inevitably open to chemical change at above absolute zero temperature. On account of the replicators being mutable, any living lineage will unavoidably mutate to generate divergent lineages that compete against one another. Competition between different lineages in turn results in natural selection being the arbiter of population shifts and biological evolution. Thus biological evolution is an inherent property of life.

4. REPLICATOR-FIRST VS. METABOLITE-FIRST PATHWAYS

For the membranes, metabolites and mutable catalysis-competent replicators to integrate and produce primitive life, a pivotal question pertains to how a productive relationship between the metabolites and replicators might be brought about. There are two schools of thought on this question. The Replicator-First school is founded on the discovery that RNA molecules can serve as genes encoding replicable sequence, as well as catalysts as ribozymes. Accordingly, a proliferating population of ribozymes could mutate to generate novel ribozymes, and catalytically transform metabolites in the surroundings into novel metabolites. With hundreds of ribozymes producing hundreds of metabolites, the system would form, upon enclosure into lipid vesicles, precells on the way to living cells.

The crippling difficulty of this Replicator-First scenario, with ribozymes leading the development of replicator-metabolite assemblies, resides in the over-whelming problem posed by the massive presence of meaningless RNA sequences that would result from any prebiotic production of RNA or RNA-like oligomers unguided by prescriptive sequence information. Joyce and Orgel (10) estimated that a triple stem-loop RNA structure containing 40-60 nucelotides offers a reasonable hope of functioning as a replicase ribozyme, but an RNA library consisting of one copy each of all 10^{24} possible random 40-mers weighs about one kilogram. If two ribozymes are required, for example to bring about their cross-replication (11), 10^{48} RNAs with a mass comparable to that of the Earth would be needed. Moreover, while ribozymic RNA replicases might out-replicate random RNA sequences, this advantage cannot be extended to other ribozymes. Besides, primordial Earth was no PCR machine, and Wong (8) pointed out that once a single stranded RNA template has served as template strand in the formation of a product strand, further replication of the template-product duplex would be barred by lack of strand separation. As emphasized by Szostak (12), *"Thus, given an efficient and accurate copying chemistry, the conversion of a single-stranded RNA template into an RNA duplex would simply lead to a dead-end product."*

Given the Twin Pitfalls arising from the massive amounts of random RNA sequences needed to yield useful ribozymes, and their non-replicative dead-end duplexes, the Replicator-First Pathway to life is fundamentally unworkable. One's attention therefore turns to the Metabolite-First option (13, 14). In this regard, it has been suggested that cellular metabolism may be viewed as a shell structure, the core being the citric acid cycle and related reactions, 1st shell the syntheses of amino acids, 2nd shell sulfur incorporation into amino acids, and 3rd shell the synthesis of dinitrogen heterocycles. Through catalysis, a metabolite interconversion network such as the autocatalytic reductive TCA (tricarboxylic acid) cycle could grow in complexity, from low free energy to high free energy, and from autotrophy to heterotrophy, to generate a primitive metabolic system. Once these metabolites include the amino acids and nucleotides, primitive nucleic acids and proteins could develop and flourish. Autocatalytic chemical cycles such as peptide-cycles where small peptides catalyze their own synthesis from amino acids (15) might also collaborate with simple organic compound cycles to spur Metabolism-First development in the absence of informational replicators such as ribozymes.

However, important shortcomings of the Metabolism-First pathway to life have been identified (16, 17) based on the implausible assumptions made regarding the catalytic properties of minerals, and the ability of minerals to organize sequences of chemical reactions. Mineral-catalyzed chemical cycles are expected to lack the specificity of ribozymic or enzymic catalysis, and therefore result in problematic side reactions. For example, non-specific catalysts catalyzing the reductive carboxylation of succinic acid, a requisite of the reductive TCA cycle, would

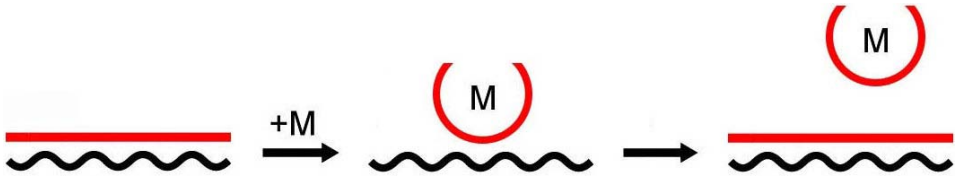

Figure 1. Replicator Induction by Metabolite (REIM). Dead-end duplex between aptamer strand (red line) and template strand (wavy line) reacts with metabolite M to form aptamer-M complex and free template strand, which thereupon causes formation of a new aptamer strand.

be prone to accept also malic acid as substrate, thereby disrupting the cycle. Therefore, the Metabolite-First Pathway to life is also an unlikelihood. Even if prebiotic metabolism should thrive and grow in complexity to bring in amino acid synthesis, nucleotide synthesis as well as polymerization of random RNA sequences, sooner or later the task of catalysis had to be transferred from mineral surfaces to primitive ribozymes. At that point, without any means to direct the accumulation of useful ribozymes vs. useless random RNA sequences, or any mechanism to replicate dead-end RNA duplexes, the Twin Pitfalls would inevitably block progress and abort the agenda of the Metabolite-First program. After all, what kind of life can the reductive TCA cycle together with its allied chemical cycles and peptide cycles possibly produce without the successful recruitment of functional replicators, the destined future genes?

Accordingly, Orgel (17) concluded, regarding both the Replicator-First and Metabolite-First approaches, *"solutions offered by supporters of geneticist or metabolist scenarios that are dependent on 'if pigs could fly' hypothetical chemistry are unlikely to help"*. In either case, the Twin Pitfalls must be overcome to bring into the system informational replicators that are competent of catalysis or capable of encoding other polymers armed with such competence, so that replicators, metabolites and membranes can be integrated into a single coordinated unit.

5. FUNCTIONAL RNA SELECTION BY REIM

If metabolism could not effectively lead replication and replication could not effectively lead metabolism, there could only be one possible solution to the dilemma: they must team up right from the start. For any partnership to succeed, each partner should have something to contribute to the other partner. The replicators could always contribute to metabolism by coming up with catalysts for novel metabolic reactions. On the other hand, how might the metabolites assist the replicators? For replicators, the central challenge must be to overcome the Twin Pitfalls, so that useful RNA sequences could be singled out from the sea of useless ones, and dead-end duplex RNA could undergo replication. Since metabolites are known to interact with their cognate aptamers and ribozymes, it is only logical to search for some mechanism whereby functional RNA-metabolite interactions might be recruited to build RNA-metabolite collaboration. On this basis, the REIM mechanism was proposed by Wong (8) as a prebiotic mechanism to eliminate the Twin Pitfalls.

Under prebiotic conditions, the untemplated polymerization of RNAs would lead to the production of a myriad of random single stranded RNA sequences, which in turn could serve as templates for a more speedy synthesis of RNA duplexes. Except for rare instances where diurnal temperature variation sufficed to bring about strand separation, upon templated polymerization the duplexes would be barred from further replication owing to lack of strand separation, and turned into dead-end structures. Where an RNA constituted a functional RNA in the form of an aptamer or ribozyme toward a cognate ligand, however, the presence of the cognate ligand would pull the functional RNA from its duplex with the template strand, thereby freeing the latter to induce another round of replication of the functional RNA (Figure 1). The nascent functional RNA upon dissociation from the cognate ligand could also initiate the templated synthesis of the template strand, and so on. As a result of such Replicator Induction by Metabolite (REIM), the functional RNA-containing duplexes would undergo autocatalytic replication while all the non-functional random RNA duplexes sat still as dead-end structures. With the turning over of the RNA pool, non-functional RNAs would diminish, and the once rare functional RNAs would become the predominant RNA species.

To assess the effectiveness of REIM, the chemical equilibria involved in template-aptamer-ligand interactions may be represented as follows:

$$D \text{ <=====> } S + A$$

$$K_D = S \times A/D \qquad \text{Eqn 1}$$

where S, A, and D represent single-stranded template, single-stranded aptamer, and the template-aptamer duplex respectively, S, A and D their concentrations, and K_D the duplex dissociation constant. Also,

$$C \text{ <===> } M + A$$

$$K_C = M \times A/C \qquad \text{Eqn 2}$$

where M and C represent metabolite and the complex between metabolite and A, M and C their concentrations, and K_C the complex dissociation constant. Combining Eqn 1 and Eqn 2 yields Eqn 3:

$$S/D = (K_D/ K_C) \times (M/C) \qquad \text{Eqn 3}$$

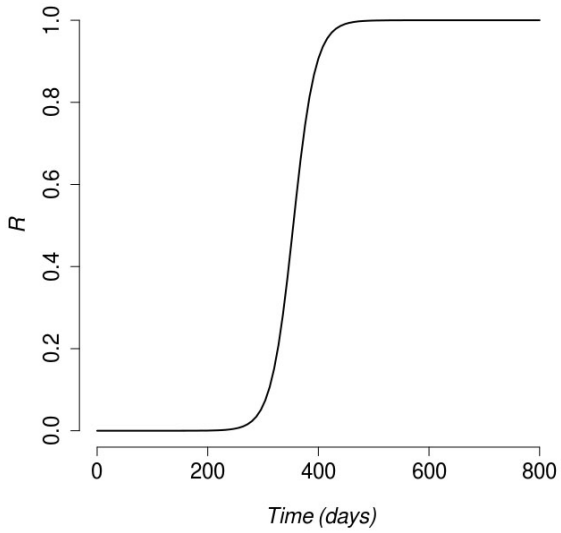

Figure 2. Accumulation of functional RNAs *R* in the presence of metabolite ligands according to Metabolic Expansion Equation. Parameters: $\alpha = 0.5$, $k = 0.1$ day^{-1}, $\sigma = 0$, $R_o = 2 \times 10^{-8}$. Results indicate a 5×10^7 fold rise of *R* from an initial level of 2×10^{-8} of the total RNA pool to a final level ~1.0, turning the entirety of the original random RNA pool into a functional RNA pool in the course of the graphed period, the timescale of which depends on the value of k.

Eqn 3 expresses *S/D*, the ratio between single-stranded template RNA and the template-aptamer duplex, i.e. between the active single-stranded form and the inactive duplex form of template RNA. K_C is typically <100 nM for the stronger aptamers (18). K_D, which is similar for RNA-RNA and RNA-DNA duplexes in the absence of formamide (19), is of the order of 10^{-6} M to 10^{-8} M for duplexes with Tm around 60°C (20, 21). Thus K_D/K_C range from 0.1 to 10, and it follows that *S/D* ranges from 0.1*M/C* and 10*M/C*.

Experimentally, targeted aptamers and ribozymes are isolated from populations of random RNA sequences at frequencies of 10^{-10} to 10^{-15} (22). Since a typical microbe is not much larger than 1 μ^3 and contains 5×10^5 RNA molecules (23), a volume of at least 2×10^4 μ^3 prebiotic aqueous medium containing random RNA at the same concentration would be required for the effective production of functional RNAs through REIM amplification, in which case total aptamer concentration $J = C + A$ would equal one molecule in 2×10^4 μ^3, or ~10^{-4} nM. Therefore, even when metabolite concentration *M* is as low as 1 μM, *M/C* is ~10^7, and it follows from Eqn 3 that most of the template strand exists in the active single stranded S form, ready to direct the templated synthesis of a nascent aptamer and signaling an efficient break-up of the aptamer-template duplex by the metabolite ligand. Since A also directs the template synthesis of S, REIM brings about the continuous template-directed production of both the aptamer and template strands, and such continuous production also holds for ribozymes, which binds its metabolite substrate in the course of catalysis.

Since S/D >> 1, total template strand $L \sim S$. With the *L/J* ratio represented by α, the rate constant of template-directed aptamer synthesis represented by k, and the concentration of ribonucleoside trisphosphate substrates available from a turnover of the random RNA pool and re-phosphorylation set at unity, the rate of single-stranded template dependent synthesis of new aptamers is given by the autocatalytic Eqn 4:

$$dJ/dt = kS = kL = k\alpha J \qquad \text{Eqn 4}$$

and summating over all functional RNAs, to be represented by R,

$$R = \int k\alpha R \, dt \qquad \text{Eqn 5}$$

During the early stages of functional RNA expansion, Eqn 5 shows the early expansion of R feeding on nucleotide building blocks from a turnover of the pool of random, mostly non-functional RNAs. However, when the accumulation of functional RNAs has transformed them from rarities in the pool into majority sequences, their further expansion has to depend increasingly on nucleotides derived from the degradation of the functional RNAs themselves. Therefore the generalized Eqn 6 has to include the nucleotide supply factor $(1 - R + \sigma)$, where $1-R$ represents nucleotide substrates from a turnover of residual non-functional RNA in the pool, and σ the influx of nucleotides from environment which may be much smaller than nucleotides from a turnover of the chemically labile RNA:

$$R = \int k\alpha R \, (1 - R + \sigma) \, dt \qquad \text{Eqn 6}$$

Because some of the unduplexed aptamer strands are complexed to their cognate ligands and therefore less available to encode the synthesis of new template strands, whereas the unduplexed template strands are fully available to encode the synthesis of new aptamer strands, aptamer strand synthesis may be expected to outpace template strand synthesis, hence $0 < \alpha < 1.0$. Eqn 6, referred to as the Metabolic Expansion Equation, describes the sigmoidal time course of functional RNA accumulation brought about by REIM: initially autocatalytic, but capped in the later stages by limited nucleotide supplies (Figure 2).

The presence of prebiotic metabolites, besides activating REIM, also made possible REAS, viz. Replicator Amplification by Stabilization, by binding to aptamers/ribozymes and thereby conferring increased resistance on these functional RNAs toward physical degradation relative to non-functional RNAs (8).

6. METABOLIC EXPANSION LAW

While random behavioral modes of single molecules are prone to give rise to unpredictable chance events, the behavior of a population of molecules often conforms to physicochemical laws. Such statistical, or population-valid, physicochemical laws are exemplified by the law of diffusion which describes the collective spatial distribution of a large population of solute molecules in

solution, even though the random movements of single solute molecules may be difficult to predict, and the Boltzmann distribution law that describes energy distribution over a population of molecules when the population is sufficiently large.

On account of the extremely wide range of ligands that aptamers can bind (24), REIM will amplify from the environmental random RNA pool aptamers cognate to most of the metabolites in the environment, as well as ribozymes active toward any of the metabolites. Furthermore, just as antibodies provide a versatile source of designer enzymes through structural modifications, aptamers can mutate to yield ribozymes. For example, a 38 nt malachite green-binding aptamer, isolated based on its binding capability, was found to display catalytic activity when presented with an appropriate substrate (25). Over generations of amplification, ribozymes also can mutate to yield variants with enhanced fitness (26) or altered specificities. Therefore the operation of REIM would lead to the appearance of novel ribozymes acting on the environmental metabolites to produce novel metabolites, which in turn will induce the amplification of their cognate aptamers and ribozymes from the random RNA pool to form yet another generation of metabolites.

Accordingly a **Metabolic Expansion Law** governing metabolic expansion as a predictable outcome of interactions between RNA-like replicators and metabolites can be formulated:

Under conditions of active synthesis of RNA-like replicators, accelerated template-directed synthesis of RNA-like replicators, and the presence of a huge population of random RNA-like duplexes in the environment, functional RNA-like aptamers/ribozymes will be selectively amplified by their cognate metabolites in the environment through the Replicator Induction by Metabolite (REIM) mechanism based on the Metabolic Expansion Equation, leading to the appearance of novel RNA-like ribozymes catalytically acting on the metabolites to form novel metabolites and thereby expand metabolism.

RNA-like replicators would include not only RNA, but also such polymers as threofuranosyl nucleic acid, pyranosyl-RNA etc. which are capable of serving as replicators-cum-catalysts and enter into dead-end duplexes with their complementary chains. Because this Law predicts that addition of a cognate metabolite ligand will selectively stimulate at non-melting temperatures the replication of duplexes containing functional RNAs capable of binding the ligand over duplexes containing non-functional RNAs, it is readily falsified if experiments should fail to demonstrate such predicted stimulus. As the only route for overcoming the Twin Pitfalls to secure the accumulation of functional RNAs out of a sea of random RNA sequences, this Law is essential to life's emergence.

Earlier, based on the essentiality of prescriptive information (PI) to prebiotic development, and the inability of known physicochemical laws to generate prebiotic PI, Abel posed the perceptive question of *"whether there might be some yet-to-be discovered new law of biology that will elucidate the derivation of prescriptive information and control"* (4). The Metabolic Expansion Law answers the question in the affirmative, and enables the emergence of prebiotic PI in the form of functional RNA aptamers and ribozymes. In modern organisms, the genetic information embodied in DNA genes originates from natural selection of the fittest ancestral genes, and information transfer from DNA to RNA to protein occurs through residue-by-residue sequence instruction. Prebiotically, the PI embodied in the functional RNAs originated from neither natural selection nor any form of residue-by-residue sequence instruction, but from physicochemical selection of functional RNA by cognate metabolite.

Later, when the precells gave rise to the first living cell, the functional RNAs became genes, and their PI became genetic information. Just as the replication and transcription of present day genes depend on the molecular logic of non-covalent hydrogen bondings between the complementary strands of a double helix, present day genetic information is traceable to the molecular logic of non-covalent physicochemical bondings between metabolites and functional RNAs three billion years ago.

7. CUMULONS AND PRECELLS

Under prebiotic type conditions, Miller (27) showed that electric-discharge through a primitive atmosphere could bring about the synthesis of amino acids and other organic compounds. In addition, other means including high energy radiation, heat, shock wave, mineral catalysis, hydrothermal vents, meteorites etc. are now known to produce metabolites including amino acids, nucleotides, sugars, carboxylic acids, hydroxy acids, acetyl-thioesters and lipids (28-32) that would induce the accumulation of their cognate functional RNAs through REIM. RNA polymerization also could occur in solution, on minerals such as Montmorillonites, in eutectic phase in water-ice, or on lipid bilayer lattices (12, 33-37).

The reaction mixture in which prebiotic accumulation of functional RNAs proceeded may be referred to as a *cumulon* mix. These cumulon mixes might be found in a 'warm little pond' as envisaged by Darwin (38), adhering to mineral surfaces (13), or within mineral micro-chambers (39). As shown by Deamer (29), the amphiphilic compounds available from the prebiotic environment could readily enclose the cumulon microdroplets to form *precells*, which embodied in vesicular form the pioneering concepts of Macallum and Oparin of a particulate forerunner of life (8). To ensure that the precell would contain a wide variety of functional RNAs for REIM selection to proceed, the enclosed cumulon must either include $\geq 10^{10}$ random RNA sequences in a volume some 20,000-fold the size of a large microbe if its RNA concentration was comparable to that of a microbe, or it had already proceeded to an advanced level of functional RNA accumulation prior to enclosure, in which case the precell could be much smaller, even microbe-size, in volume. The timing of vesicularization was also important with respect to the supply of high energy bonds. In the beginning, RNA synthesis and metabolism in the

cumulons likely had to rely on inorganic high-energy polyphosphates. However, as metabolism advanced, metabolic generation of high energy bonds appeared. As proposed by De Duve (40), amino acids and hydroxyacids in the prebiotic environment could convert to ketoacids, rendering thioesters an important primitive energy source as exemplified by present day pyruvate dehydrogenase, α-ketoglutarate dehydrogenase and thiolase reactions. Later on, ATP became the dominant high energy currency, completing a three-staged polyphosphate-thioester-ATP energy program to support prebiotic development (32). Shifting from membrane-impenetrant environmental polyphosphates to intravesicular metabolic production of thioesters and ATP could be a prerequisite to successful adaptation of the metabolite-replicator ensembles to enclosure within precells to launch membrane-metabolite-replicator collaboration. Since the precells were capable of growth and division, as demonstrated by the Luisi group (41), they were amenable to natural selection. Those precells that developed a wide and balanced array of functional RNAs enjoyed a competitive edge over less endowed ones, and began to manifest the characteristics of the self-sustaining autopoietic organization that is fundamental to life (42).

Inside the precells, each new ~70-mer ribozyme/aptamer added a PI content of some 140 bits (22), and expansion of information content was central in importance. Richly endowed geological sites as described by Mojsis *et al* (43), with access to metabolites derived from multiple sources e.g. lightning, clays, meteorites, hotsprings, volcanoes, and fumaroles, such as fire-and-ice-reactors (FAIR) in the vicinity of near shore hydrothermal vents, could furnish optimal settings for precell development (32). The simplest 'limping' life form to-day is estimated to contain at least 150 genes (23), close to the 182 genes found in *Carsonella ruddii* (44). Consequently, to acquire life-like complexity, the advanced precells needed to reach a level of complexity consisting of 150-180 functional RNAs (possibly minus the replicators relating to the DNA polymerase, ligase, DNA-dependent RNA polymerase etc. functions required to service the yet to arrive DNA, and the tRNAs and aminoacyl-tRNA synthetase functions required to service the yet to arrive Phase 2 amino acids). Without REIM, the acquisition of 150-180 functional RNAs would be an impossibility. Driven by REIM, reaching this target was a difficult but attainable goal.

Not every 150-180 replicator containing precell would comprise the same replicators. Precells that garnered a comprehensive replicator set with evolved regulatory controls would be favored by natural selection for faster growth and division. In time, the most evolved precells came to be poised for self-reproduction, and the chief remaining barrier for them to cross to become living cells would be the *cybernetic cut*, a theoretical chasm that separates the inanimate world, totally devoid of any utility orientation, from an organized utility-directed living state (6). The question is, how might utility directedness and its resultant purposeful behavior be acquired by the precells?

Consider microbe-1 which is equipped with two devices, a sugar sensor that senses sugar gradient and a flagellum for motility. Its regulatory control mechanisms couple the two devices to produce utility-directed swimming toward the presence of sugar. A microbe-2 is similarly equipped, but its control mechanisms bring about anti-utility swimming away from sugar. Microbe-1 soon divides to produce daughter-1 which retains the utility-directed swimming of microbe-1, and daughter-2 which on account of mutation loses all linkages between sensor and flagellum and swims around randomly in a non-utility manner. Eventually, the daughter-1 lineage thrives, but the lineages of microbe-2 and daughter-2 are eliminated by starvation. Since the utility-directed daughter-1, anti-utility microbe-2, and non-utility daughter-2 are all fully alive organisms, the utility-directedness of the sensor-flagellum coupling is clearly not essential to the state of aliveness. However, the survival of the daughter-1 lineage but not the lineages of daughter-2 and microbe-2 shows that, under natural selection, utility-directed phenotypes are strongly selected, which accounts for the universal presence of utility-directedness, ranging from enzyme induction by substrate and phototropism of plants to stinging defense of a hive and screening compounds for cancer drugs etc., among single cell organisms, multi-cell organisms as well as societies of organisms. The example illustrates how natural selection can introduce utility-directedness into the regulatory control mechanisms of precells and cells. Once the regulatory controls of a precell were amply ingrained with utility-directedness, it would be crossing the cybernetic gap heading for the living realm. This is confirmed by the success of whole genome transplantation in synthetic biology: when placed within competent membranes, an adequate assembly of test-tube produced genes copied from a template genome, thereby incorporating all the natural selection-honed utility-directed control mechanisms of the template genome, will automatically generate a viable organism unhindered by the cybernetic cut (45).

Accordingly, sooner or later, an advanced precell in the Peptidated RNA World if not RNA World, in possession of an adequate assembly of PI-laden replicators with natural selection-honed utility-directed control mechanisms, may be expected to sail right across the cybernetic cut, undergo self-reproduction and give birth to the living world.

8. PEPTIDATED RNA WORLD

With protein chemistry evidence pointing to the late arrival of DNA (46), RNA and RNA-like polymers were the likely bearer of prebiotic PI and early genetic information. These polymers were favored in this regard over other competing informational polymers by three important advantages:

Table 1. Covalent conjugates containing peptide or amino acid linked to nucleotides or nucleic acids

Conjugate	Ref.
Aminoacyl-tRNA in protein synthesis	
Peptidyl-tRNA in protein synthesis	
Glycyl-$tRNA_1$ in bacterial cell wall synthesis	182
UDP-N-acetylmuramyl-pentapeptide in bacterial cell well synthesis	183
Purine-6-carbamoyl-threonyl-amido group from tRNA post-transcriptional modification	184
Adenylated protein from post-translational modification	185
ADP-ribosylated protein from post-translational modification	186
Poliovirus RNA-protein	187
DNA nicking-closing enzyme-DNA	188
Bacteriophage phi X174 gene A protein-DNA	189
Tumor suppressor p53-RNA	190

(I) That ribozymes can perform as replicator-cum-catalysts resolves the question of whether replicators or biocatalysts came first;
(II) Ribosomal peptide synthesis is catalyzed by ribozyme;
(III) Because RNA replicates through complementary base-pairing forming dead-end duplexes, and aptamers/ribozymes can bind metabolite ligands, these RNA replicators are optimally suited to amplification by REIM.

Advantage I sidesteps a long standing debate, and advantage II led Orgel (33) to suggest, *"The demonstration that ribosomal peptide synthesis is a ribozyme-catyalyzed reaction makes it almost certain that there was once an RNA World"*. Advantages I and III apply to RNA as well as RNA-like polymers such as threofuranosyl nucleic acid (TNA) and pyranosyl-RNA (p-RNA) (47) as primordial PI-carriers. In contrast, clays and proteins would lose out as competitors owing to their inability to benefit from either advantage I or advantage III.

In time, however, precell/cellular development became constrained by the inherent weakness of ribozymes as catalysts. Nowadays, proteins can work with 20 canonical amino acids with a wide range of sidechains, but organisms still continue to introduce numerous extra sidechains through post-translational modifications to improve protein function. The selective pressure of this side-chain imperative is obviously too substantial to resist. For the RNA/RNA-like aptamers and ribozymes in the precells and early cells, their useful sidechains were limited to four kinds of nitrogenous bases, a sugar and phosphate. The impetus to incorporate variations into the nitrogenous bases or sugar moiety must be overpowering. Indeed, to-day more than 95 kinds of post-transcriptional modifications are found on RNAs. For the replicator-cum-ribozymes, the dilemma was, increasing the variety of nitrogenous bases on RNA would enhance catalytic function, but also increase error rate in base-pairing during replication. The choice for the RNAs was therefore uncompromising: either they retain their replicative role and give up their catalytic one, or vice versa.

Catalysis by either ribozyme or enzyme is mediated by the formation of a complex between catalyst E and substrate S, followed by transformation of substrate to reaction products with regeneration of free catalyst. The velocity (v) of the catalyzed reaction is described by the Michaelis-Mentin Equation:

$$v = k_{cat}[E_o][S]/([S] + K_m)$$

The catalytic rate constant k_{cat} determines the rate at which product is produced from the ES complex, and the Michaelis constant K_m equals the half-saturating substrate concentration. Both ribozymes and enzymes are outstanding in substrate binding, giving rise to low K_m values. However, ribozymes are prone to poorer performance compared to enzymes with respect to k_{cat}. For example, the enzymes RNases T1, T2 and A hydrolyze their substrates with k_{cat} of 5,000 min^{-1} to 180,000 min^{-1}, whereas ribozymes that cut RNA including both Group I ribozymes and hammerhead ribozymes typically display a k_{cat} of 100 min^{-1} or less (9, 48). Accordingly, owing to the rate limitations of ribozymes, the transition was eventually made whereby the ribozymes passed on most of their catalytic tasks to proteins. The question is, how was this transition brought about? As Cech remarked (49), "*The fun comes when we try to use our secure knowledge of the modern RNA world to infer what the primordial RNA world might have looked like*."

Amino acids were abundant in the prebiotic environment. Accordingly, when the primitive RNA or RNA-like polymers underwent post-replication modifications, some of the modifications would involve the covalent bonding of amino acids or peptides to the RNA. In view of the huge catalytic advantage that covalent post-replication modifications with amino acid and peptide moieties could confer on the ribozymes, Wong (50) proposed that, by forming peptidated (and aminoacylated) RNA, the resultant peptidyl prosthetic groups on the peptidated ribozymes would soon outperform the ribozymes themselves in catalysis to furnish the actual catalytic centers, as in the case of present day flavoproteins and hemoproteins where the prosthetic group constitutes the real active center. In time, the polypeptide prosthetic groups became detached from the ribozymes to act as stand alone proteins, and the role of RNA was largely confined to that of encoder of protein synthesis (Figure 3). This proposal is strongly supported by the pivotal discoveries of ribozymes that could form aminoacylated and peptidyl RNA (51-53), and by the rich variety of present day conjugates between amino acids/peptides and nucleotides/nucleic acids (Table 1). Notably, while 'peptidyl RNA' by usage usually implies o-peptide-RNA structures where the peptide is conjugated to the 3' terminus ribose of the RNA through an ester linkage, in 'peptidated RNA' the peptide can be conjugated to any position on the RNA.

Kurland (54) emphasized the inadequacy of the RNA World and the need for ribonucleoproteins in

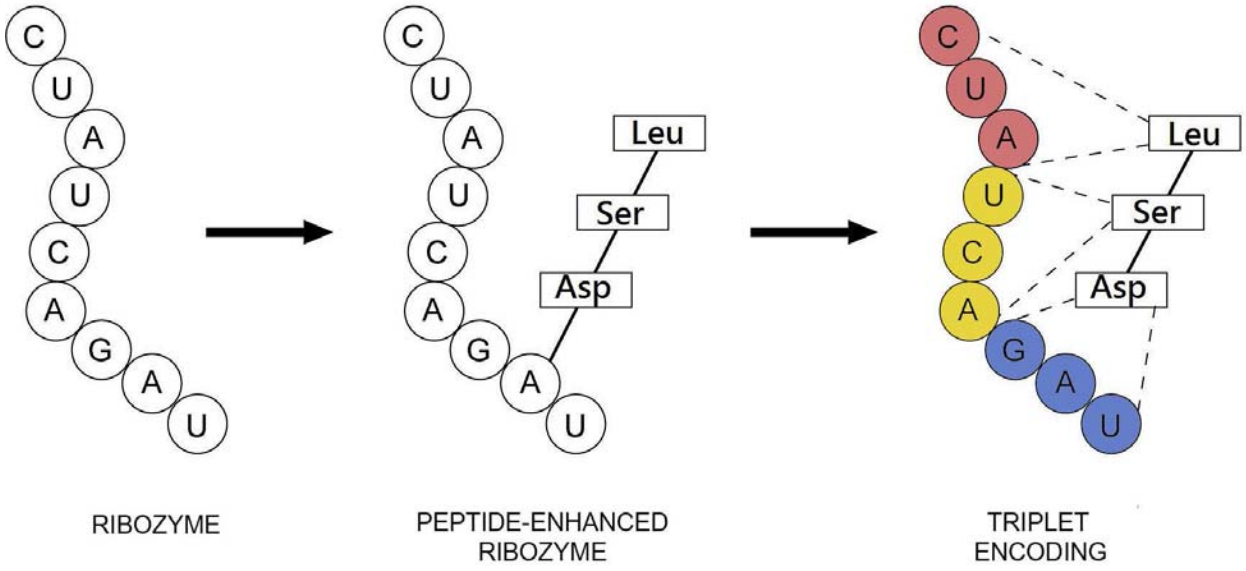

Figure 3. Progression from ribozyme segment to peptide-enhanced ribozyme segment and finally to triplet encoding of tripeptide.

primordial cells. Noller (55) pointed out that the structural versatility of RNA pales in comparison with that of proteins, and most modern functional RNAs including rRNA, ribonuclease-P RNA and spliceosomal RNA depend on proteins for their activity, suggesting that small peptides, which are known to induce large-scale structural changes in RNA, would be required to expand the structural repertoire of RNA. Di Giulio (56) postulated covalent linkages between proteins, which were preformed on coenzyme A-like molecules by non-ribosomal peptide synthetase type catalysts, and the 5' terminus of RNAs to produce RNA-coenzyme-polypeptides. These suggestions are in accord with the proposal of RNA peptidation with respect to the premise that no RNA World could accomplish much without peptides.

Since the catalytic peptidyl transferase center (PTC) on the ribosome is today a ribozyme, the possibility arises that an all-RNA ribosome might at first perform genetic encoding of protein synthesis by itself without calling on the assistance of polypeptides. However, this is ruled out by the findings of Caetano-Anolles and coworkers (57, 58) that rRNA and r-proteins interacted to start off ribosome evolution prior to appearance of PTC. Consequently S12 and S17, the oldest r-proteins that interacted with rRNA to start off ribosome evolution, as well as other protein domains that made a pre-PTC appearance including aaRS catalytic domains, had to be produced by non-PTC mechanisms. These findings are not consistent with the RNA World: *"In the "RNA world" view that prevails, nucleic acid structure precedes protein structure. This scenario is incompatible with phylogenomics, parsimony thinking, RNA and ribosomal biology, and data from molecular structure and function"*; instead, Kauffman-Dyson self-replicating peptides were considered as a possible source of the pre-PTC protein domains (58). However, examples of self-replicating peptides are few, and none with any catalytic activity besides that pertaining to their own replication. In contrast, the wide range of isolated ribozymes including ones that catalyze formation of peptidated RNA are all self-replicating, and the pre-PTC appearance of protein domains are totally in accord with the Peptidated RNA World. Moreover, even to-day, all organisms contain a far greater variety of peptidated RNAs than proteins: their ribosomes must produce 199 different 3'-terminus o-peptidated-tRNA intermediates of increasing lengths in order to make one 200-aa protein. Therefore, catalytically modern organisms are living in the Protein World employing enzymes that are proteins, but biosynthetically they are still living in the Peptidated RNA World manufacturing proteins through stepwise peptidation of tRNA. Some old habits never die.

The importance of RNA peptidation is founded on the rich catalytic potential of peptides: 10^6-fold fewer peptide sequences need to be searched to obtain an effective catalyst, suggesting that peptides are a million-fold fitter as catalysts than RNA (59). Peptides are equally valuable for membrane transport. Enwrapping replicators and metabolites in lipid vesicles was prerequisite to the development of precells, but brought with it the need for permeases to regulate the uptake of nutrients and ions. To-day, protein permeases perform their tasks as integral constituents of the cell membrane. In the RNA World, RNAs owing to their negative charges did not interact well with the vesicular membrane. However, covalent bonding of hydrophobic peptides to the RNA would readily anchor the RNA to membrane to implement permease function, just as myristoylation can redirect cytosol proteins to localize on cellular membranes (60). Not surprisingly, therefore, membrane transporters are among the earliest detected protein fold families (58).

The March of Progress in ribozyme catalysis described by Hughes and Ellington (61) comprised the

Table 2. Triple convergence.

	Gly	Ala	Ser	Asp	Glu	Val	Leu	Ile	Pro	Thr	Phe	Tyr	Arg	His	Trp	Asn	Gln	Lys	Cys	Met
Phase 1 or Phase 2	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
Irradiated synthesis	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	n	n
Meteoritic amino acid	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0

Evidence regarding prebiotic availability of different amino acids is provided by classification of Phase 1 vs Phase 2 amino acids based on coevolution of genetic code and amino acid biosynthesis (63, 65), synthesis using high energy particle irradiation (67, 68) or meteoritic composition (69). Production or presence is indicated by "+", and lack of production or absence by "0"; "n" indicates inapplicable on account of the absence of sulfur in the irradiated synthesis.

stages of (A) simple replicators, (B) trans-acting ribozyme ligases, (C) ribozyme assembly via tag sequences, (D) general polymerase, and (E) canonical RNA World. Transition to the Peptidated RNA World and the Protein World would then ensue wherein the ribozymes handed most of their catalytic roles to proteins, and their responsibilities were simplified to those of replicators and components of protein synthesis. Still later, when DNA was introduced, the RNAs gave up even their replicator role except in the case of RNA viruses, to focus above all on protein synthesis acting as mRNA, tRNA, and rRNA with PTC activity.

9. COEVOLUTION THEORY OF THE GENETIC CODE

RNA peptidation ushered in RNA-directed polypeptide synthesis, primitive tRNAs and aaRS, opening the door to the genetic code. Electric discharge through primitive atmospheres readily yielded some amino acids but not others, and the same was observed with other physical means including heat, shock wave and radiation. Notably, the amino acids produced are in all instances confined to those amino acids that serve as biosynthetic precursors to other amino acids in present day organisms (62). This together with the clustering in the genetic code of codons for biosynthetically related amino acids led to the Coevolution Theory as an explanation of the structure of the genetic code (63):

"The structure of the codon system is primarily an imprint of the prebiotic pathways of amino-acid formation, which remain recognizable in the enzymic pathways of amino-acid biosynthesis. Consequently the evolution of the genetic code can be elucidated on the basis of the precursor-product relationships between amino acids in their biosynthesis. The codon domains of most pairs of precursor-product amino acids should be contiguous, i.e., separated by only the minimum separation of a single base change."

There are four fundamental tenets to the Coevolution Theory. In the decades since its proposal, genomics science has advanced by leaps and bounds. The wealth of genomic information together with other advances has provided proofs for all four tenets (64).

Tenet 1: The prebiotic environment did not supply all twenty canonical amino acids at life's origin, but had to be complemented by sourcing through inventive biosynthesis.

The Coevolution Theory suggests that of the 20 canonical amino acids in the protein alphabet, Gly, Ala, Ser, Asp, Glu, Val, Leu, Ile, Pro and Thr represent Phase 1 amino acids that were early-arrivals supplied by the environment, and the remaining ten, Phe, Tyr, Arg, His, Trp, Asn Gln, Lys, Cys and Met, represent late-arrival, biosynthesis-derived Phase 2 amino acids. This 10 Phase 1:10 Phase 2 partition was proposed based on the results of prebiotic-type amino acid syntheses and the coevolution pathways of codon distribution (65, 66). Subsequent studies by Kobayashi *et al* (67, 68) using high energy particle irradiation of primitive atmosphere verified that exactly the ten Phase 1 amino acids and no Phase 2 amino acid were produced. Pizzarello (69) discovered exactly the same ten Phase 1 amino acids in carbonaceous chondrite meteorites including the recent detection of Ser and Thr on the GRA95229 Antarctica meteorite, and no Phase 2 amino acid. This Triple Convergence between Coevolution Theory, high energy particle irradiation and meteoritic composition (Table 2) provides exceptional evidence supporting the dual origins of prebiotic Phase 1 vs. biosynthetic Phase 2 amino acids proposed by the Coevolution Theory. The correlation between codon allocation and amino acid biosynthesis was shown to be unequivocal based on both the hypergeometric distribution (63) and Fisher's exact test (70). As well, the entry of Phase 1 amino acids into the code ahead of Phase 2 ones was confirmed by the temporal order of amino acid utilization in proteins, where the nine oldest amino acids identified, viz. Gly, Ala, Val, Asp, Pro, Ser, Glu, Leu and Thr are Phase 1 amino acids (71).

Besides the Triple Convergence and correlation between codon allocation and amino acid biosynthesis, Tenet 1 is proven by the thermal instability of at least two of the Phase 2 amino acids, viz. Gln and Asn. Glu and Asp are obtained in good yields under prebiotic-like conditions, and they can be amidated to form Gln and Asn. However, Gln and Asn are limited in thermal stability. In the prebiotic environment, the steady state concentration of Gln or Asn is given by the ratio v_f/k_d. The decomposition rate constant k_d is 1.7×10^{-2} yr^{-1} for Asn and 2.6×10 yr^{-1} for Gln at 20^oC (72). The maximum formation rate v_f of Asn cannot exceed that of Asp, and the maximum v_f of Gln cannot exceed that of Glu. Even under the highly optimistic assumption that all the UV radiation reaching planet Earth was absorbed by H_2S and available for prebiotic synthesis, giving rise to 20 M amino acids in the 10^{24} ml total hydrosphere on the planet,

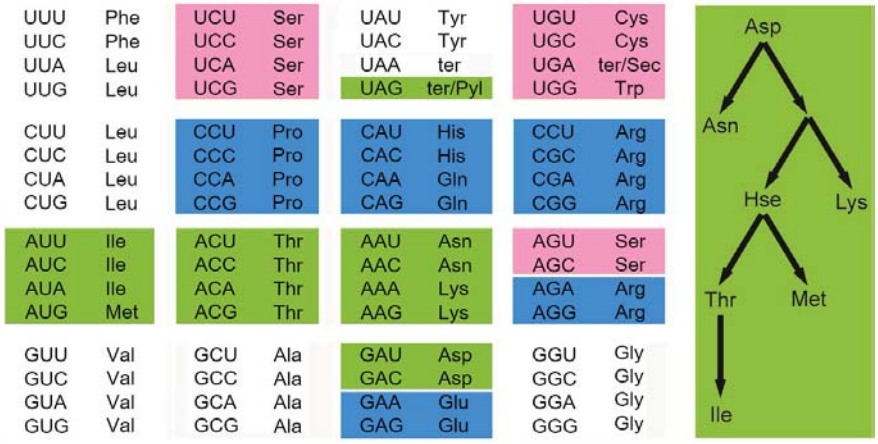

Figure 4. Clustering of amino acid biosynthetic family codons in the genetic code. Ser family, red; Glu family, blue; and Asp family, green. Panel on the right shows biosynthetic relationships in Asp family.

the steady-state concentration was only ≤24 nM for Asn, and ≤3.7 pM for Gln. It is therefore futile to look for availability of Gln and Asn from the prebiotic environment. In addition, the Phase 2 amino acids Cys, Met, Trp, His, Tyr and Phe are rapidly destroyed by UV (73), in accord with their exclusion from the Phase 1 amino acid roster.

Tenet 2. Pretran synthesis provided mechanisms for the initial encoding of some Phase 2 amino acids.

Since only Phase 1 amino acids were supplied by the environment, Phase 2 amino acids had to be synthesized from Phase 1 amino acids through inventive biosynthesis, and mechanisms must be provided to transfer codons from the Phase 1 to the Phase 2 amino acids. Such transfers caused the codons of biosynthetic product amino acids to be located overwhelmingly in contiguity with those of their respective precursors. Two kinds of mechanisms were suggested for these transfers (63). First, if a precursor-product amino acid pair are physically similar, the product could competitively attach to a tRNA that belonged to the precursor. Second, and more important because of its non-dependence on physicochemical similarity between product and precursor, a precursor-tRNA conjugate could undergo pretranslational modification to bring about the synthesis of a product-tRNA conjugate, in so doing handing to the product the codon on the tRNA that originally belonged to the precursor. At the time of the Coevolution Theory proposal, this *pretran synthesis* mechanism was known to convert Met-tRNA in bacteria to formyl-Met-tRNA, and Glu-tRNA to Gln-tRNA. Nowadays, the incorporations of Gln, Asn and Cys into proteins are known to proceed respectively via GlnRS, AsnRS and CysRS, or via pretran synthesis of Gln-tRNA from Glu-tRNA, Asn-tRNA from Asp-tRNA, or Cys-tRNA from phospho-Ser-tRNA. Sec-tRNA is only formed via pretran synthesis from phospho-Ser-tRNA.

Tenet 2 may be falsified if the use of aaRS is found to predate the use of pretran synthesis in all instances of aminoacyl-tRNA synthesis, demonstrating that the pretran pathway is merely a late invention unrelated to early genetic code development. Contrary to such an outcome, comparative phylogenetics indicate that use of pretran synthesis predates Gln-tRNA and Asn-tRNA (74). In the case of Cys, pretran synthesis of Cys-tRNA dates back to 3.5 billion years, whereas tRNA-independent Cys biosynthesis dates back only to 2.9 billion years. Therefore pretran synthesis of Cys-tRNA was older than both aaRS synthesis of Cys-tRNA and biosynthesis of Cys itself (75). Likewise, pretran synthesis of Asn-tRNA was older than biosynthesis of Asn (76, 77). *Methanopyrus,* the oldest known organism least evolved from the Last Universal Common Ancestor, is also devoid of GlnRS, AsnRS and CysRS (see below). That pretran synthesis rather than aaRS was utilized for the earliest production of Gln-tRNA, Asn-tRNA and Cys-tRNA for protein synthesis, and exclusively for the production of Sec-tRNA, proves Tenet 2.

Tenet 3. Biosynthetic relationships between amino acids were an important determinant of codon allocations.

A number of amino acids are derived from other amino acids via biosynthetic pathways that are mostly species-invariant. Such precursor-product amino acid pairs include Ser-Trp, Ser-Cys, Val-Leu, Thr-Ile, Gln-His, Phe-Tyr, Glu-Gln and Asp-Asn. The eight product amino acids named possess altogether 20 codons, and all 20 of them are in each instance contiguous with one of the codons belonging to the respective precursor. Such contiguity between the codons of precursor and codons of product yielded an aggregate random probability of less than 0.0002. Even if the Phe-Tyr and Val-Leu pairs are considered to be more biosynthetic siblings than precursor-product pairs and removed from consideration, the aggregate random probability is still only 0.0075 (63). Besides, the Glu-Pro, Glu-Arg, Asp-Lys pairs may also be regarded as precursor-product pairs, in which case the random probability will be further reduced.

This contiguity, or neighborliness between precursor and product codon locations (Figure 4) suggests that the earliest genetic code distributed triplet codons only to the Phase 1 biosynthetic precursors, and the Phase 2 products received their codons from a subdivision of the

erstwhile Phase 1 codon domains. This transfer of codons from precursor to product explains a range of special features of the genetic code:

(a) Since sibling product amino acids produced from the same precursor received their codons from a subdivision of the same original precursor codon domain, neighborliness between the codons of the siblings is expected as well. This explains why Ile and Met, both derived from homoserine, which is in turn derived from Asp, share the same AUN codon box; and why Cys and Trp, both derived from Ser, share the same UGN box.

(b) Sharing of the UGN box by Cys and Trp indicates that this box once belonged to Ser, which explains the otherwise strange separation of the non-contiguous UCN and AGY codon domains of Ser: UCN, UGN and AGY formed a contiguous Ser domain, but this was broken up when UGN was carved up and handed to Cys, Trp and ter, leaving behind the oddity of the disconnected UCN and AGY domains for Ser.

(c) The Coevolution Theory was proposed in 1975 to explain the structure of the 20 amino acid code. Years later, when Bock and coworkers made the astounding discovery of seleno-Cys (Sec) as the 21st encoded amino acid (78), it was surprising that Sec, biosynthesized from Ser, is encoded by the UGA codon, thus sharing the UGN box with the other Ser-derived siblings Cys and Trp and further confirming the erstwhile ownership of the UGN box by Ser.

(d) Still later, another astounding discovery followed suit, that of Krzycki and coworkers of pyrrolysine (Pyl) as the 22nd encoded amino acid (79). Pyl is a biosynthetic product of Lys, and its encoding by UAG, which is contiguous with the Lys codon AAG, adds yet another example of precursor-product codon contiguity to the genetic code.

(e) The Coevolution Theory pointed out that (63): *"Within this framework of coevolution of amino acids and their codons, it is expected that additional factors would help to determine the exact allocation of some of the codons.....codon contiguity between chemically similar amino acids could minimize the damage due to excessive mutations or coding errors."* As Figure 4 shows, the Ser-family codons are clustered in the UNN row in the code, the Glu-family codons in the CNN row, and the Asp-family codons in the ANN row. When these family domains were subdivided among different offspring amino acids, it would be advantageous to minimize errors by allocating codons from the family row to the offspring amino acids such that each offspring received codons from a column where physically similar amino acids were found in other rows. As a result, the clustering of physically similar amino acids within the same column is more pronounced than any physical clustering within the same row. Therefore the suggestion that the four columns, each enriched with physical similarities, rather than the four rows, each enriched with family kinships, furnished the primary divisions of codon domains (80) is groundless. Instead, the 10^5-fold greater importance of coevolution with amino acid biosynthesis than error minimization in universal code selection (see below in 'Selection of a Univeral Code') clearly indicates that the primary divisions of the code were based on row-centered family kinships, and the secondary subdivisions of the family codon domains were based on column-centered physical similarities established upon the entries of the Phase 2 amino acids.

(f) Asp and Glu are both biosynthetic precursors to a number of product amino acids, and ceded multiple codons to these products, keeping to themselves only the GAY and GAR codons respectively. Since the other GNN codons in the same row also encode Phase 1 amino acids, the suggestions have been made that the genetic code initially employed only the GNN codons (80, 81). Asp and Glu are also deaminated to form ketoacids that are metabolically linked through the TCA cycle. Thus it also has been suggested that the co-location of their codons in the GAN box, might be the result of their biosynthetic kinship (81). However, because both Asp and Glu are Phase 1 amino acids, they were both available from the prebiotic environment, and there was no evident need for Asp to derive its codons from Glu or vice versa. Such mutual biosynthetic non-dependence likewise holds for the Ala-Ser, Ser-Gly and Ala-Val pairs, all involving Phase 1 amino acids. The usage of only some of the codon rows but not others would generate numerous nonsense codons to perturb translation. Instead, the exclusive allocation of the GNN codons to Phase 1 amino acids, and the retention of the GAN codons by Asp and Glu suggest that translation of the GNN codons might be advantageously the most efficient and/or error-free in the primitive codes.

(g) The enrichments of Ser, Glu and Asp family codons in the UNN, CNN and ANN codon rows respectively raises the question of how these enrichments came about initially. In this regard, it is noteworthy that modern aaRS often recognize the anticodon bases on tRNAs as identity elements, e.g. the major identity elements on tRNA recognized by *Bacillus subtilis* TrpRS comprise the anticodon and the discriminator base-73 (82, 83). Accordingly, an effective mechanism for allocating the UNN, CNN and ANN codons to the Ser, Glu and Asp families respectively could be based on the preferences of primordial aaRS, viz. preference of a SerRS for A36, preference of a GluRS for G36, and preference of an AspRS for U36 at the 3rd anticodon position on the tRNA as an identity element. Remarkably, in accord with this plausible mechanism, modern CysRS and TrpRS of the Ser-family still employ A36, modern ProRS, GlnRS and ArgRS of the Glu-family still employ G36, and modern IleRS, MetRS, ThrRS, AsnRS and LysRS of the Asp-family all still employ U36, as one of the identity elements they recognize on their cognate tRNAs (84).

(h) The codon packages allocated by the code come in varying sizes: 6 codons to Leu and Ser, single codons to Trp and Met, and single shared (with ter signal) codons to Sec and Pyl. This is readily explained by biosynthesis as the foremost principle guiding codon allocations. For any amino acid to possess a large number of codons, it needs to be a Phase 1 amino acid, and not give away too many of its codons to offspring amino acids. Leu and Ser are both Phase 1 and possess 6 codons each; Asp and Glu are both Phase 1, but they have given away the majority of their codons. Trp, Met, Sec and Pyl are evidently late arrivals, and each of them manage to receive only one or half a codon. Just as countries of the world possess greatly varying sizes of territory and often odd looking boundaries because of events in human history, codon domains come in different sizes and are odd looking because of events in biosynthetic history.

(i) That many amino acids, e.g. hydroxyl-Pro, phospho-Tyr, caboxy-Glu etc. are given no codon and have to rely on post-translational modification for entry into proteins raises the question why Phase 2 code expansion stopped at 20+2. The analogy with human languages is evident (9). The human voice can make 40 different sounds, and correspondingly there are 40 letters in the International Phonetic Alphabet. However, there are 22 letters in the Hebrew alphabet, 23 in Latin, 26 in English, 33 in Cyrillic, and 39 in Hungarian runan (archaic). Thus the number of letters varies with the language group, but always has to ensure the effective representation of an adequate range of sounds. For the genetic code, too few encoded amino acids would provide insufficient sidechain versatility for protein function, whereas too many, e.g. to the point of requiring 4-codon boxes to encode four amino acids each, may increase translational errors. Notably, the adequacy of versatility would be tightly controlled by feedback from protein performance. When sidechain versatility was inadequate, the proteins in the translation machinery performed poorly causing high translational noise/errors, and the noise created by the introduction of a novel encoded amino acid into the code was well tolerated. However, once the encoded sidechain ensemble attained high versatility, translational noise was subdued, and the noise due to the introduction of yet another novel encoded amino acid would become too great a selective disadvantage, thereby freezing code expansion for the next 3 billion years. Code expansion to add Phase 2 amino acids was thus tantamount to a search for excellence in protein performance that was stoppable only when the highest performance standard has been achieved (85). The final standard is in fact so high that nowadays the kinetics of numerous enzymes are diffusion-controlled, catalyzing as fast as diffusion can bring enzyme and substrate together to form an enzyme-substrate complex to yield bimolecular rate constants of 10^6 - 10^9 $M^{-1}sec^{-1}$ (86), and humans employing the same antique but unantiquated ensemble of 20+2 amino acids are still constantly scaling unimaginable heights.

In pretran synthesis, a product amino acid acquires its allocated codons through a tRNA adapter that accepts its precursor amino acid. Thus Gln was allocated CAA and CAG because a Glu-tRNA carrying the anticodon to these codons came to be converted to Gln-tRNA by pretran synthesis, not because of error minimization or any stereochemical interaction between Gln and its codon/anticodon. Likewise, Asn received AAU and AAC from Asp, Cys received UGU and UGC from Ser, and Sec received part use of UGA from Ser, biosynthetically. Consequently biosynthesis was the sole determinant deciding the codon allocations to these amino acids, thereby proving Tenet 3.

Tenet 4. The amino acid ensemble encoded by the genetic code is mutable, allowing early code expansion to admit the Phase 2 amino acids.

Minor variations are known among organisms and organelles in the codons assigned to the 20 amino acids. What is universal, never mutated during the past 3 billion years, is the ensemble of 20 amino acids themselves, which raises the question of whether this amino acid alphabet encoded by the code is mutable at all. Since Coevolution Theory posits that the code was repeatedly reshaped by the entry of Phase 2 amino acids, it requires the encoded alphabet to be intrinsically mutable. Tenet 4 is thus falsified if prolonged experimental endeavour should fail to mutate the encoded amino acids. Accordingly experiments were carried out to mutate the genetic code of *Bacillus subtilis*, and the results described below under 'Synthetic Life' demonstrated unambiguously the intrinsic mutability of the code, proving Tenet 4. Such mutability allowed not only the entry of Phase 2 amino acids into the code, but also the possible participation of prebiotically available amino acids such as α-aminobutyric acid in the primitive code prior to their eventual elimination from the code.

10. SELECTION OF A UNIVERSAL CODE

There are an astronomical number of possible permutations of alternate codes, differing in the codon assignments for the amino acids. Although minor variations in codon assignments to amino acids are found in nuclear and organelle systems especially metazoan mitochondria, the usage of basically the same canonical code by all known organisms suggests that there are few if any substantially different alternate codes remaining in use in the living world. Permutations of codon packages of the canonical code yield $2x10^{19}$ possible alternative codes (48, 85). Since the universe is only 15 Gyr, or $4.7x10^{17}$ seconds old, forty kinds of life forms each bearing an alternate code would have to be competitively eliminated per second if competition between organisms with different codes represents the sole mechanism for removal of alternate codes. Therefore selection based solely on competition is impossible. Instead, code disallowance ruling out vast tracts of potential codes on account of inappropriate biosynthetic relations had to be invoked in the emergence of a unique code. Three important mechanisms participated in finalizing the code selection:

Error Minimization steers the genetic code toward error reduction through the selection of codes where physically similar amino acids are given contiguous codons. The canonical code was estimated to be 45.3% error minimized. This % can be increased to 49.4% just by switching the codons UAG and UGG between ter and Trp, and the code also can be rearranged by inspection to yield an alternate 72%-minimized code, indicating that there was only modest pressure toward error minimization in code evolution (87). Likewise, error minimization was found to contribute merely a factor of one-in-a-million, or 1x10-6 selection toward a unique code (88), although the methodology employed in this study was shown to be defective (89-91). Tellingly, when all 1280 neighboring codes of the canonical code were analyzed, 18% of them yielded more error minimization, pointing to local error minimization at the 82% level (91). These results thus establish the presence of finite but incomplete error minimization, the absence of strong evolution pressure to reduce errors, and that the canonical code is neither a local nor a global minimum on the error surface.

Stereochemical Interaction relies on amino acid-codon/anticodon interactions to guide codon assignments, and specific interactions of this nature have been experimentally demonstrated for a number of amino acids, which are estimated to contribute a factor of $4x10^{-4}$ toward the selection of a unique code (92).

Coevolution of the code with amino acid biosynthesis, by requiring the product amino acids to receive codons only from their respective precursors, disallows huge swaths of alternate codes. For example, when the UGN box belonging to Ser underwent subdivision, only offspring amino acids of Ser, viz. Cys, Trp and Sec qualified as recipients of codons from this box. All other Phase 2 amino acids were disallowed as recipients, e.g. Asn had to seek its codons from the erstwhile Asp codon domain instead. Such disallowance drastically reduces the number of allowable alternate codes. Random distribution of the 1-6 codon packages of the genetic code with or without disallowance by the coevolution process yields code counts of $N = 2.15 \times 10^8$ and $N = 2.29 \times 10^{19}$ respectively (85). Therefore coevolution contributes a selection factor of 10^{-11} toward a unique code.

Furthermore, when an equal-package model was analyzed consisting of the distribution of 60 codons equally to p initial biosynthetic family domains, followed by the subdivision of each domain into q equal portions to yield a total of 20 three-codon packages, the number of permutated altenate codes is given by:

$$\text{N, number of alternate codes} = p!(q!)^p$$

where $p.q = 20$. When $p = 20$ with all 20 amino acids receiving their three codons in parallel, or when $p = 1$ with all 20 amino acids each receiving a subdivided three-codon package from a single precursor amino acid, $N = 2.4 \times 10^{18}$. In contrast, when p is between 5 and 10, N reaches a value of only 5×10^8. These results therefore confirmed the high and low alternate code counts, and also showed that *least action* pathways of maximum efficiency were closely followed in the coevolution-based formation of the canonical code.

Accordingly, these three different mechanisms for alternate code reduction acting together are capable of a $4x10^{-21}$ selection, which is adequate for the selection of a unique code out of the $2x10^{19}$ possible alternate codes. This adequacy solves the mystery of how a universal code could have been selected out of $2x10^{19}$ possible codes. On this basis, the relative contributions made by these three mechanisms to the selection of a near universal code are:

Coevolution: Error Minimization: Stereochemical Interaction = 40,000,000: 400:1

Consequently, with coevolution of genetic code and amino acid biosynthesis being the predominant mechanism determining its evolved structure, the canonical genetic code stands as a lasting monument to the primordial history of amino acid biosynthesis.

11. HETEROTROPHY FIRST VS. AUTOTROPHY FIRST

There are two schools of thought on whether the first living cell was a heterotroph feeding on organic compounds available from environmental sources, or an autotroph using only CO_2 and one-carbon compounds from the environment to synthesize all other organic compounds. The experimental abiotic synthesis of a number of amino acids and nucleic acid constituents, along with the finding of meteorites as a rich source of organic compounds, have led to the suggestion of a heterotrophic origin (93). However, discoveries in recent years demonstrating the production of a range of organic compounds under hydrothermal vent conditions have favored the feasibility of an autotrophic beginning of life at these vents. This results in an ongoing Heterotrophy First vs. Autotrophy First debate (94-98).

For amino acids, the Triple Convergence of evidence from genetic code structure, prebiotic amino acid synthesis, and meteoritic composition provides strong confirmation that environmental availability was prerequisite to the admission of any amino acid into the Phase 1 genetic code. If the first cell was an autotroph, it would be expected to synthesize in-house all 20 canonical amino acids as in present day blue-green algae without restriction to only Phase 1 amino acids. On the other hand, if the first cell was a heterotroph, only Phase 1 but not Phase 2 amino acids were initially available from the environment for protein synthesis. Only later on, when the newly developed amino acid biosynthetic pathways produced the Phase 2 amino acids, would the latter be encoded to increase the chemical versatility of proteins. Consequently the Triple Convergence strongly favors the first living cell being a heterotroph (8).

12. LAST UNIVERSAL COMMON ANCESTOR

The wide range of properties that are common to different living organisms, including the universal genetic code with its 20 canonical amino acids, DNA with its T, C, A, G deoxyribonucleotides, RNA with its U, C, A, G ribonucleotides, ATP as major energy currency and coenzymes such as NAD and CoA point to the existence of a common ancestor. Methods to identify the nature of a Last Universal Common Ancestor, or LUCA, at the root of life depend on rooting phylogenetic trees by paralogs, or combined analysis of biopolymer sequences and structures. Since different rooting approaches may identify dissimilar basal nodes, LUCA needs to be defined by some specific criterion. In view of the fact that the DNA genome is a relatively late arrival yet shared by all living organisms, LUCA might be defined as the first organism to adopt a DNA genome.

Paralogous rooting of trees of life depends on sister sequences which stemmed from a duplication of the same gene but ended up serving different biochemical functions. Identical or nearly identical at the LUCA stage, these sister sequences diverged with time, and the extent of their divergence within a genome provides a measure of how far the genome has evolved from LUCA. The identification by Woese *et al* (99) of the three major biological domains of Archaea, Bacteria and Eukarya based on the SSU RNA phylogenetic tree is a landmark in biological science. However, SSU RNA has no paralogs in cells. Instead, early rootings of protein trees based of the EF-Tu/EF-G and ValRS/IleRS paralogous pairs containing only a few species located LUCA in the Bacteria domain (100-102), but it is now known that artifacts such as long branch attraction and horizontal gene transfers could easily invalidate such rootings based on just a few species. As a result, these early rootings were unreliable, and gave rise to pessimism questioning whether the root will ever be found (103,104). However, subsequent analysis of the ValRS/IleRS pair employing a wide range of species revealed major evolutionary disturbances in the IleRS tree, but rooted the ValRS tree in the Archaea based on both the maximum parsimony and neighbor joining methods (105).

Given the artifacts with protein paralogs, other biopolymers have to be looked to as well for sequence information. DNA and rRNAs lack paralogs. The tRNAs contain only about 75 bases each and some of the bases are semi-variant, but this limitation may be overcome by analyzing the entire tRNAome of a species, which comprises over two thousand bases.

Free living organisms contain 20 families of tRNAs cognate to the 20 different canonical amino acids. Analysis of the genetic distances between the tRNA sequences of each family from the alloacceptor sequences of the 19 other families within the same genome by Xue *et al* (106) located LUCA in the Archaea domain on the tRNA tree in the proximity of *Methanopyrus kandleri* (Mka) (Figure 5), on the basis that the most ancient tRNAomes display minimal average alloacceptor distances between the different tRNA families in conformance to the *cluster-dispersion model* of tRNA evolution, in which different tRNA families were originally clustered closely together in sequence space. As they evolved, their sequences were dispersed and became increasingly distant from one another, just as the stars in the sky travelled apart from one another following the Big Bang (Figure 6). On this basis, the greater similarity between e.g. the tRNAPhe and tRNATyr sequences in Archaea compared to Bacteria or Eukarya (Figure 7) points to the greater primitivity of Archaea than Bacteria and Eukarya. The location of LUCA closest to Mka was further confirmed by the findings that Mka also displayed among the 34 Bacteria, 18 Archaea and 8 Eukarya species analyzed the lowest genetic distance between the elongator and initiator tRNAs for Met, as well as the lowest genetic distance between paralogous aaRS pairs (107). These lines of evidence based on intragenomic tRNA and aaRS sequence comparisons are shown as Lines 1, 2 and 4 among the multiple lines supporting a location of LUCA either within the Archaea or close to *Methanopyrus* or both in Table 3. Based on intergenomic comparisons of information embedded in tRNA sequence and structure, Line 21 likewise indicates that Archaea were the most ancestral, followed by viruses, and then Eukarya and Bacteria (108).

The proximity of *Methanopyrus* to LUCA is not a total surprise. The transition from an oxygen poor to an oxygen rich atmosphere on Earth has forced extensive adaptations by most living lineages, erasing many genomic features inherited by them from the distant past. Only a small number of organisms including Mka escaped massive adaptations (109). The hydrothermal vents, home to Mka, represent one of the most continuously preserved ecological niches on Earth. Therefore Mka and other anaerobic vent inhabitants could escape the perturbation of an oxygen rich atmosphere, and survive with only minimal genomic departures from their earliest ancestors. It is this minimal departure that enables Mka to provide the best working likeness of LUCA. In vertebrate evolution, the hagfish lineage has stayed close to its earliest niches, and did not undergo a great deal of physiological changes; so a study of the hagfish tells us more about the vertebrate ancestor than a study of the mouse for example, even though both animal lineages have traversed the same time span of evolution since the days of the vertebrate ancestor. Likewise, although Mka is a modern microbe, its study can be more revealing about the nature of LUCA than the study of other known species. On this basis, LUCA is indicated by its phylogenetic proximity to Mka to be an archaeal-like anaerobic hyperthermophilic methanogen.

The utility of tRNA sequences in the search for LUCA stems from the slow and steady evolution of tRNA sequences compared to the faster changing protein sequences, so that the useful timescale based on tRNA dates further back in time. For example, *Buchnera aphidicola* (Bap) is believed to have descended from a free living Gram-negative bacterium. However, owing to its symbiosis with aphids beginning 160-280 Myr ago, its proteome has undergone extensive changes. In contrast, as shown in Figure 5, its tRNAome still retains close kinship with that of *E. coli* (Eco) on the tRNA tree.

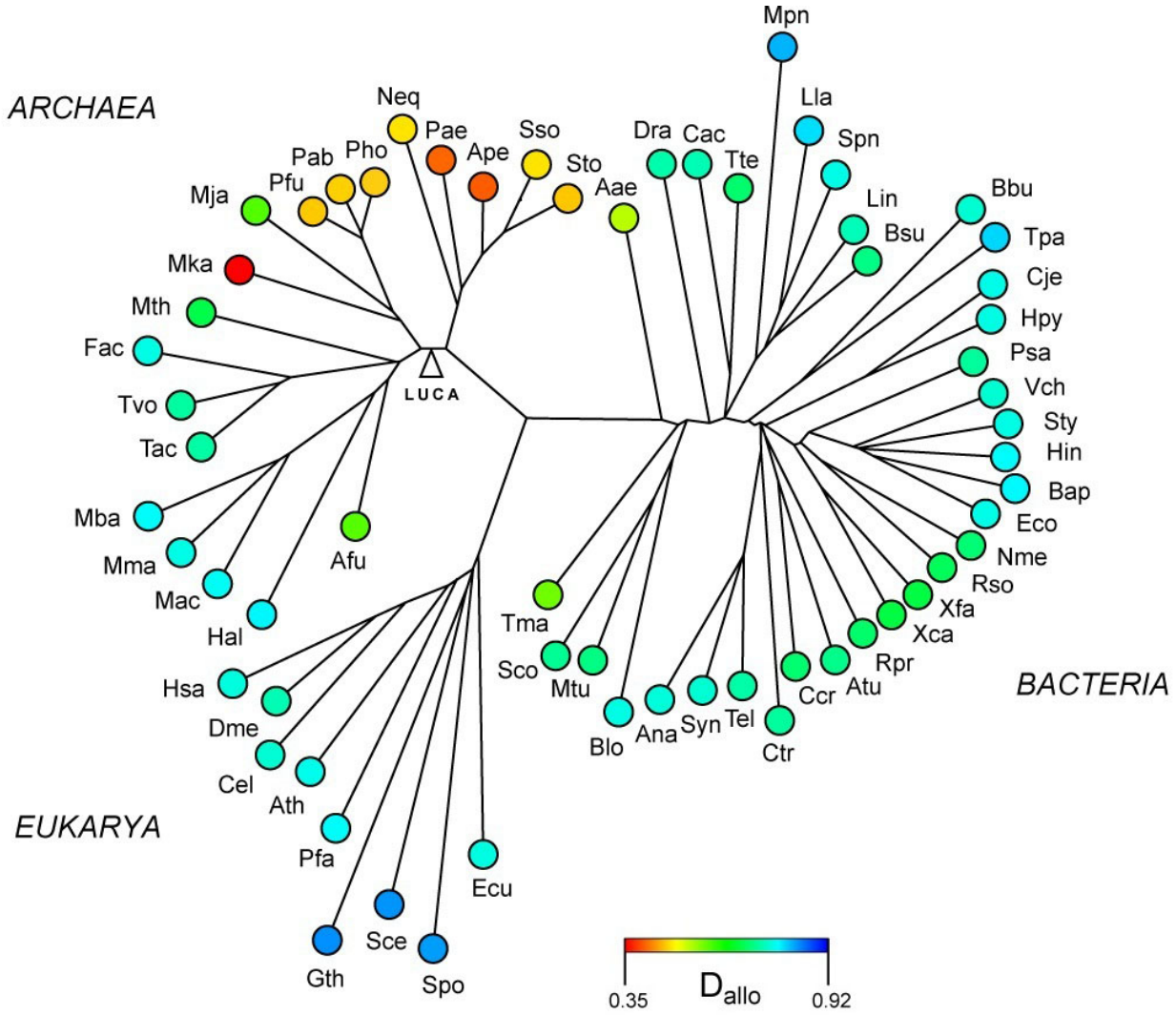

Figure 5. Transfer RNA phylogenetic tree. Average alloacceptor distance D_{allo} between tRNAs accepting different aamino acids for each species is indicated by thermal scale. Values ranged from a minimum 0.351 for Mka (*Methanopyrus kandleri*) to a maximum 0.839 for Sce (*Saccharomyces cerevisiae*). Estimated position of LUCA is indicated by triangle (106).

Line 3 in Table 3 is based on the striking simplicity of archaeal anticodon usages. For any species, the collection of tRNA genes determined from its complete genomic sequence reveal the nature of the anticodons it employs. In the genetic code there are thirteen standard codon boxes where the four codons in the box are allocated either all to the same amino acid, e.g. the family box of GUN for Val, or two each to two amino acids, e.g. AAU-AAC for Asn and AAA-AAG for Lys. In most bacterial and eukaryotic species these standard boxes are read by three or more different combinations of anticodons (Figure 8). In contrast, none of the archaeons analyzed uses more than two different combinations of anticodons for these boxes, and the majority of them use only a single combination. These results revealed the simplicity of archaeal anticodon usages, in stark contrast to the complex anticodon usages of Bacteria and Eukarya. Among the Archaea, extreme simplicity is displayed by *Methanopyrus*: its thirteen standard codon boxes are read uniformly by the same two anticodons GNN and UNN. The simplicity of archaeal anticodon usages favors the primitivity of Archaea, and the extra simple *Methanopyrus* usage affirms its extraordinary antiquity (110).

Line 5 shows the rooting of ValRS in the Archaea domain based on paralogous rooting by IleRS. Lines 10 and 11 are based on a composite protein tree pointing to the Euryarchaea-Crenarchaea separation as the most ancient biological event, and Mka as one of the deepest branching archaeons and species.

Lines 22-25 are based on the phylogenetic/ phylogenomic trees of 5S RNA, RNase P, protein folds and proteomes respectively, all pointing to the primitivity

Table 3. Lines of evidence of LUCA being archaeal (A) or Mka-proximal (M) or both

No.	Type of evidence	Evid-ence of	Ref.
1	Alloacceptor tRNA distances	A, M	106
2	Initiator-elongator $tRNA^{Met}$ distances	A, M	107
3	Anticodon usages	A, M	110
4	Aminoacyl-tRNA synthetase distances	A, M	107
5	Archaeal root of ValRS	A	105
6	Lack of GlnRS in Mka	M	105,191
7	Lack of AsnRS in Mka	M	105,191
8	Lack of CysRS in Mka	A, M	105,191
9	Lack of cytochromes in Mka	M	105,191
10	Early Euryarchaea-Crenarchaea separation	A, M	192
11	Mka as deep branching archaeon	M	192
12	Primitivity of methanogenesis	A, M	192-3
13	Primitivity of anaerobiosis	M	109,194
14	Primitivity of hyperthermophily	A, M	195-8
15	Primitivity of barophily	M	199
16	Primitivity of acidophily	M	200-1
17	Use of CO_2 as electron acceptor	A, M	202-3
18	Chemolithotrophy	M	105
19	Hydrothermal vents as appropriate home for LUCA	M	204-5
20	Minimalist regulations	M	191
21	tRNA evolution pattern	A	108
22	5S rRNA tree	A	206
23	Ribonuclease P tree	A	207
24	Protein fold tree	A	208
25	Proteome tree	A	208

of Archaea. Lines 12-16 indicate the primitivity of methanogenesis, anaerobiosis, hyperthermophily, barophily and acidophily, in accord with a *Methanopyrus*-proximal LUCA. Lines 17-18 emphasize the advantages of an Mka-like metabolism for LUCA, Line 19 supports the hydrothermal vents being the home of LUCA, and Line 20 is based on the simplicity of Mka regulatory mechanisms.

Missing genes provide powerful evidence for an archaeal-like *Methanopyrus*-proximal LUCA (Lines 6-9). In present day organisms, the Phase 2 amino acids Gln, Asn and Cys are incorporated into proteins either via GlnRS, AsnRS and CysRS, or via pretran synthesis from Glu-tRNA, Asp-tRNA and phospho-Ser-tRNA. In genetic code evolution, use of pretran synthesis instead of aaRS is a primitive trait. That the genes for GlnRS, AsnRS and CysRS are all missing from the Mka genome thus adds further evidence to the closeness of *Methanopyrus* to LUCA.

Cytochromes participate widely in electron transport in mitochondria, chloroplasts, sulfate-reducing organisms, and even the anaerobic methanogen *Methanosarcina*. It is therefore surprising that cytochrome genes are missing from the genomes of Mka, *Methanothermobacter thermautotrophicum* (Mth), *Methanococcus jannaschii* (Mja), *Pyrococcus furiosus* (Pfu), *Pyrococcus abyssi* (Pab) and *Pyrococcus horikoshii* (Pho), which are clustered together on the tRNA tree. Since this cytochrome-less group of genomes display some of the lowest alloacceptor tRNA distances, lowest tRNAMet-tRNAfMet distances, and lowest inter-aaRS paralog distances (106, 107), they constitute an Ancient Six group that are ultra-conservative in molecular evolution. Furthermore, since Mka, Mth and Mja consume H_2 and CO_2 metabolically whereas Pfu, Pab and Pho produce H_2 and CO_2, the cytochrome deficiency of the group does not stem from metabolic similarity, but from their phylogenetic proximity to a cytochrome-less LUCA (105).

The uncovering of cytochrome deficiency at the LUCA stage based on the Ancient Six suggests the use of these genomes to define the LUCA genome: genes that are present in all of these six genomes may be regarded at first approximation to be constituent genes of LUCA. However, because the Mka-Mth-Mja and Pab-Pfu-Pho groups have dissimilar modes of energy metabolism, the genes common to the two groups would lack the specific energy metabolism genes from either group. As LUCA could not survive without any energy metabolism genes, the LUCA gene set has to be supplemented with such genes. Since LUCA was closest to a hyperthermophilic methanogen, the methanogenesis genes common to Mka, Mth and Mja need to be added to the plausible LUCA genome based on the Ancient Six. Also, because these Ancient Six are all euryarchaeons, the genes common to them might include genes that were originally absent from LUCA but were added to the euryarchaeal branch. To counter this possibility, only the genes common to these Ancient Six as well as *Aeropyrum pernix* (Ape) and *Pyrobaculum aerophilum* (Pae), the two crenarchaeons with the lowest alloacceptor tRNA distances, are taken to represent plausible LUCA genes. On this basis, the minimum LUCA genome is estimated to contain altogether 463 genes including a proteome of 424 COGs and, based on the Mka genome, 39 structural RNA genes (111, 112).

Notably, the 424-COG minimum LUCA proteome includes 24 Group L COGs for DNA replication, recombination and repair, 26 Group K COGs for transcription, plus other DNA-related COGs in Groups B, D, F and R. Thus over 10% of the proteome is devoted to the DNA information machinery. This proteome accommodates a 20-amino acid genetic code, and provides for some post-transcriptional modifications. LUCA's 463 genes, not far from the gene numbers of 500-600 and 500-1,000 estimated from comparative genomics (113, 114), exceed the estimated gene number of 150-340 for minimal organisms (111, 112) and suggest that LUCA is not a minimal but a modestly complex organism.

13. RISE OF THE DNA GENOME

LUCA is an extinct species, and locating LUCA is tantamount to identifying its taxonomic affiliations. For this purpose, the polyphasic taxonomy approach depends on the integration of a maximum amount of information in order to achieve reliable identification. This approach is particularly important for the search of LUCA in view of the occurrence of horizontal gene transfers rendering evidence from any single genes especially protein genes open to ambiguity, including indel-based evidence that would be susceptible to misalignments as well (105, 115). To date diverse lines of evidence have been suggested favoring a bacterial, eukaryotic or archaeal LUCA, but so far the lines supporting a bacterial or eukaryotic LUCA are fewer than those summarized in Table 3 in support of an

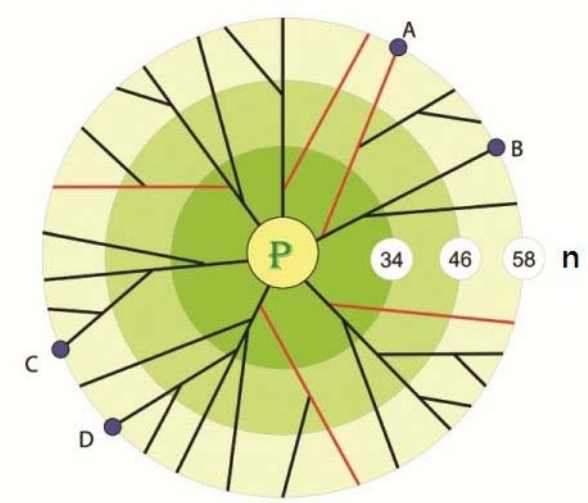

Figure 6. Cluster Dispersion Model of tRNA Evolution. In this model different acceptor tRNAs were originally closely clustered in sequence space. With the passage of time, these tRNAs (black lines) became dispersed and more distant from one another. During the dispersion, some of the tRNAs also became acceptors for novel amino acids (red lines). A-D denote present day sequences of four of the tRNAs. The number of tRNAs, n, in each species tends to increase with evolved distance. The ancestral tRNA sequence at P is approximated by the consensus sequence of the 34 Mka tRNAs as, with the *GNN* anticodon in italics: GCGGGCGCAGCUUAGCCUGGUCAGAGCGCGGGACU*GNN*GAUCCCGUGGCCCGGGUUCAAAUCCCGGCCGCCGCA

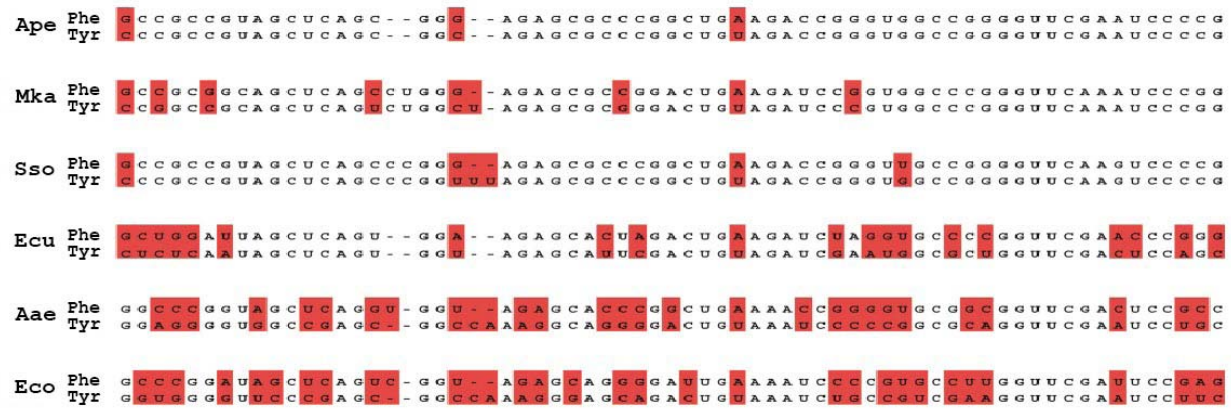

Figure 7. Paired tRNAPhe-tRNATyr sequences from the euryarchaeon Mka, the crenarchaeons Ape (*Aeropyrum pernix*) and SSO (*Sulfolobus solfataricus*), the eukaryote Ecu (*Encephalitozoon cuniculi*), and the bacteria Aae (*Aquifex aeolicus*) and Eco (*E. coli*). Base difference between the pair in each instance is marked in red.

archaeal LUCA close to Mka. Therefore, until the latter evidence is surpassed by the discovery of more lines supporting a bacterial, eukaryotic or another archaeal LUCA, the weight of evidence points to an Mka-proximal LUCA. In turn, this raises the question: given a heterotrophic first living cell, by what route could a hyperthertmophilic methanogen come to acquire its LUCAhood? There are at least two possible routes.

The first route is birthright. Ever since the discovery of living communities at the submarine hydrothermal vents, these vents have been considered as possible sites for the origin of life (116), where the geothermal energy released at the vents provided a hospitable environment for prebiotic evolution, supporting the autotrophic syntheses of a range of organic compounds. However, there are forbidding drawbacks to this scenario:

(a)Indication of a heterotrophic first cell by the Triple Convergence;

(b) Instability of RNA at elevated temperatures;

(c) Labile metabolites and intermediates that require metabolic channeling by evolved biocatalysts for protection at elevated temperatures; and

(d) With random RNAs existing as melted single strands at the elevated temperatures, the Twin Pitfalls cannot be eliminated by REIM.

		Mka	Tac*	Tvo	Fac	Lin	Eco	Sce			Mka	Tac*	Tvo	Fac	Lin	Eco	Sce			Mka	Tac*	Tvo	Fac	Lin	Eco	Sce			Mka	Tac*	Tvo	Fac	Lin	Eco	Sce
UUU	*F*								UCU	*S*							A																		
UUC	*F*	G	G	G	G	G	G	G	UCC	*S*	G	G	G	G	G	G																			
UUA	*L*	U	U	U	U	U	U	U	UCA	*S*	U	U	U	U	U	U	U																		
UUG	*L*		C	C	C	C	C	C	UCG	*S*		C	C	C	C	C	C																		
CUU	*L*				A				CCU	*P*							A	CAU	*H*								CGU	*R*					A	A	A
CUC	*L*	G	G	G		G	G	G	CCC	*P*	G	G	G	G		G		CAC	*H*	G	G	G	G	G	G	G	CGC	*R*	G	G	G	G			
CUA	*L*	U	U	U	U	U	U	U	CCA	*P*	U	U	U	U	U	U	U	CAA	*Q*	U	U	U	U	U	U	U	CGA	*R*	U	U	U	U			
CUG	*L*		C	C	C		C		CCG	*P*		C	C	C		C		CAG	*Q*		C	C	C		C	C	CGG	*R*		C	C	C	C	C	C
									ACU	*T*							A	AAU	*N*								AGU	*S*							
									ACC	*T*	G	G	G	G	G	G		AAC	*N*	G	G	G	G	G	G	G	AGC	*S*	G	G	G	G	G	G	G
									ACA	*T*	U	U	U	U	U	U	U	AAA	*K*	U	U	U	U	U	U	U	AGA	*R*	U	U	U	U	U	U	U
									ACG	*T*		C	C	C	C	C	C	AAG	*K*		C	C	C	C		C	AGG	*R*		C	C	C	C	C	C
GUU	*V*							A	GCU	*A*							A	GAU	*D*								GGU	*G*							
GUC	*V*	G	G	G	G	G	G		GCC	*A*	G	G	G	G		G		GAC	*D*	G	G	G	G	G	G	G	GGC	*G*	G	G	G	G	G	G	G
GUA	*V*	U	U	U	U	U	U	U	GCA	*A*	U	U	U	U	U	U	U	GAA	*E*	U	U	U	U	U	U	U	GGA	*G*	U	U	U	U	U	U	U
GUG	*V*		C	C	C			C	GCG	*A*		C	C	C				GAU	*E*		C	C	C			C	GGG	*G*		C		C		C	C

Figure 8. Anticodon usages in standard codon boxes. Usage of any codon in a box is indicated by 1st anticodon base shown aligned with its complementary codon. Different species are shown in top row: Archaea: Mka, Tac for *Thermoplasma acidophilum*, Tvo for *Thermoplasma volcanium*, Fac for *Ferroplasma acidarmanus*; Bacteria: Lin for *Listeria innocua*, Eco for *E. coli*; Eukarya: Sce for *S. cerevisiae*. The Mka anticodon usage is extreme in simplicity, employing the same GU two-anticodon combination (viz. GNN and UNN) to read all 13 standard codon boxes.*The Tac usage of a GUC three-anticodon combination in all 13 standard boxes is the predominant usage pattern for Archeae, shared by a wide range of crenarchaeons and euryarchaeons. In contrast, the bacterial and eukaryotic usages typically employ three or more different anticodon combinations in the 13 standard boxes.

Even a thermophilic origin at lower than vent temperatures cannot overcome all these drawbacks, e.g. the binding of 10-20mer RNA to template RNA and hence template-directed RNA polymerization would be difficult above 60°C without assistance by evolved biocatalysts. This leaves a mesophilic or psychrophilic first cell, in time giving rise to mainly mesophilic pre-Lucan life on account of slower cellular growth rates at low temperatures. Thus a hyperthermophilic LUCA descending directly from a hyperthermophilic first cell is highly unlikely.

The second route is conquest. The Latin languages were established over a large usage domain through military conquests by the Roman Empire in the short span of a few centuries. By the same token, a hyperthermophilic LUCA at the vents could yield descendants that spread to cooler zones, eliminating all indigenous competitor lineages in those zones, provided that LUCA had developed and bequeathed to its descendents some decisive biochemical weaponry, e.g. the DNA genome and/or 20-amino acid genetic code in LUCA's possession as revealed by its 463-gene genome. To enhance plausibility, the vents should play a role in facilitating the development of such weaponry.

In this regard, although the Triple Convergence points to the utilization of Phase 1 amino acids from prebiotic environment when genetic coding began, eventually the exponential multiplication of heterotophic mesophile pre-Lucans may be expected to outstrip the linear production of environmental organic compounds. In response, the pre-Lucans could either migrate or establish autotrophy. To migrate in the face of starvation is a most common practice for organisms with any measure of mobility.

Attractive migration destinations for the pre-Lucans included the hydrothermal vents with their abundant carbon dioxide and hydrogen supplies, and exergonic carbon fixation (117) supporting the prebiotic syntheses of organic compounds, especially when these volcanic formations could be far more numerous on early Earth than to-day. As competition among the pre-Lucans for the organic compounds in the vicinity of the vents turned fierce, they had to outdo one another by getting ever closer to the vents. In the process they needed to make their proteins more heat-resistant. In addition, the greater heat resistance of DNA relative to RNA became a key selective advantage, and the pre-Lucans that developed a DNA genome claimed the vents, perfected methanogenesis and gave rise to LUCA. At the vents LUCA and descendants so prospered that, as suggested by Kasting and Siefert (118, 119), their copious methane production generated approximately 10 degrees of greenhouse warming to postpone the first ice age even under a faint young sun, until rising oxygen levels emitted by photosynthesis reduced the photochemical lifetime of methane 1,000 fold from 10,000 years at low oxygen to 10 years to-day.

All cellular genomes are based on DNA even though the need of a free radical in the catalytic mechanism of ribonucleotide reductase suggests that DNA arrived late, subsequent to the arrival of proteins (46). There are distinct advantages of DNA over RNA that favor displacement of RNA by DNA as genetic material (120):

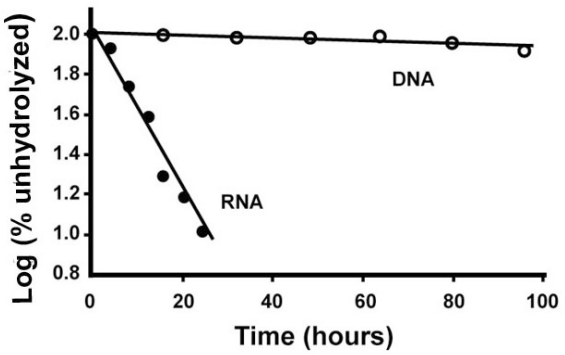

Figure 9. Thermal instability of RNA relative to DNA at 100°C, pH 7.0 (111).

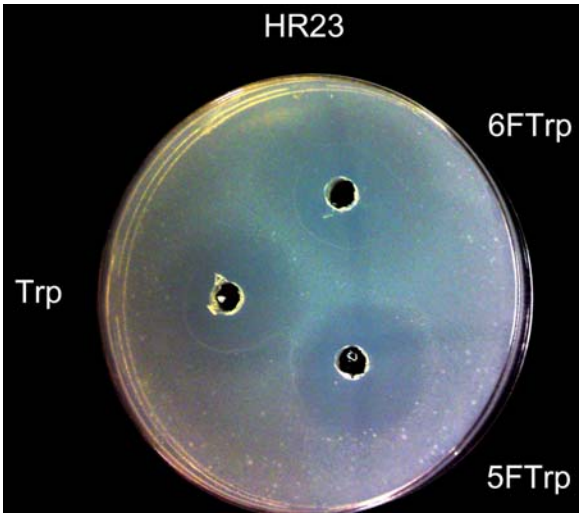

Figure 10. Inhibition of HR23 growth on 4FTrp by Trp, 5FTrp and 6FTrp, each of which produced an inhibition zone surrounding a well containing the inhibitor.

(a) Greater chemical stability of DNA;

(b) Absence of proofreading by RNA polymerase;

(c) Degradation of cytosine to uracil can be detected and repaired in DNA but not in RNA.

However, because more than 10% of the minimum LUCA proteome had to be devoted to the DNA information machinery, it would require a powerful evolutionary incentive sustained over thousands of generations to switch from RNA genes to DNA genes. It has been proposed that the switch might be brought about through cell-virus interactions (121, 122). Although cell-virus interactions are known to give rise to novel genetic alterations in both cells and viruses through such processes as restriction, lysogeny and transduction, these alterations are as a rule limited genic changes. Thus the capability of viruses to initiate large scale genomic developments in cells that require thousands of generations to accomplish is open to question.

In view of this, it is necessary to examine other possible evolutionary factors for the invention of DNA besides viruses. One such factor is adaptation to hyperthermophilic conditions. As shown in Figure 9, the half-life of RNA is so short that RNA genomes would quickly lose viability at the hydrothermal vents. For the mesophile pre-Lucans, migration to the vents for food provided a powerful enough incentive. Furthermore, because the migration was a gradual one as they moved slowly toward the vents, the transition from RNA genes to DNA genes could be accomplished over as many generations as necessary, allowing ample time to build the requisite multi-gene DNA information machine.

Accordingly, this *Hot Cross Scenario*, in which mesophilic heterotrophs crossed into hydrothermal vents to give rise to a hyperthermophilic methanogen LUCA endowed with a DNA genome and a 20 amino acid canonical genetic code, followed by the recrossing of its likewise endowed descendants back to the cooler zones to eliminate all less-endowed competitor lineages, can account for the rise of a hyperthermophilic methanogen LUCA, its acquisition of LUCAhood through the gift of superior inheritance to its descendants, as well as the development of a DNA information machine (111, 112).

Asteroid impacts were frequent on early Earth. It is estimated that the impact of a 440 km diameter projectile of the size of Vesta and Pallas would bring the oceans to a boil or near boil, and impact vaporization of the photic zone was probable as late as 3.8 Gyr. As well, smaller asteroid impacts might have briefly heated surface environments, leaving only thermophile/hyperthermophile survivors (123, 124). Such calamities could help a hyperthermophilic LUCA attain its LUCAhood by killing off the mesophiles and psychrophiles.

14. SYNTHETIC LIFE

The 20-amino acid alphabet encoded by the canonical genetic code has been in use for the past 3 billion years, raising the question of whether this alphabet is intrinsically mutable at all. To answer this question, experiments were conducted to mutate the genetic code of the Trp-auxotroph *B. subtilis* QB928 with the aim of conferring genetic coding on the fluorinated Trp analogue 4-fluoroTrp (4FTrp). In two mutant isolation steps, QB928 gave rise to strains LC8 and in turn LC33, both of which can propagate indefinitely on 4FTrp, with all the Trp residues in the *B. subtilis* proteome replaced by 4FTrp. Although two mutant isolation steps could be accompanied by more than two mutations, the number of mutations required to arrive at 4FTrp utilization for cell propagation was likely to be limited. Moreover, when LC33 was further mutated, it gave rise to HR15 in another two isolation steps, which propagates well on 4FTrp but not on Trp (125). In fact, for HR15 and its faster growing descendant strain HR23, Trp has become an inhibitory analogue (Figure 10). However, these cells can back mutate to enable Trp to regain its lost capacity to support cell propagation, as in the revertant strain TR7

Table 4. Synthetic life systems

Type	Altered Alphabet*	Altered Site	System
o-Synthetic	UNA	Proteome-wide	*B. subtilis* LC33, LC88 etc *E.coli* B7-3
m-Synthetic	UNA	Proteome-wide	*B. subtilis* HR15, HR23
o-Synthetic	UNA	Specific sites	*E. coli* with insertion of p-aminoPhe, p-azidoPhe etc
m-Synthetic	UNA	Specific sites	*E.coli thyA* R126L dependent on azaLeu insertion
o-Synthetic	UND	Genome-wide	*E.coli* CLU5

*UNA, unnatural amino acid; UND, unnatural deoxyribonucleotide

Table 5. Relative aminoacylation of different sources of tRNA by *E. coli* aaRS (135)

Source of tRNA	Phe	Leu	Asp	Lys	Arg	Tyr	Met	Val	Ser	Thr	His	Pro
Escherichia coli	100	100	100	100	100	100	100	100	100	100	100	100
Bacteria												
Bacillus subtilis	156	208	156	71	119	187	137	87	91	97	76	94
Micrococcus luteus	20	48	2	18	88	34	19	14	65	67	41	88
Rhodopseudomonas sphaeroides	54	100	12	75	58	4	61	44	63	108	6	86
Archaea												
Halobacterium cutirubrum	2	0.4	3	2	0.4	0.5	29	30	0	69	7	1
Eukarya												
Yeast	1	2	1	14	0.5	4	43	1	4	69	10	15
Wheat germ	8	2	2	5	62	3	25	2	1	56	24	5
Rat liver	2	0.4	0.5	1	16	0.3	6	3	0	2	7	2

(Figure 11). LC33, which grows on both Trp and 4FTrp, has also yielded through further mutations the LC88 strain, which grows on Trp, 4FTrp, 5-fluoroTrp (5FTrp) or 6-fluoroTrp (6FTrp), even though 4FTrp, 6FTrp and 5FTrp are potent inhibitors of bacterial growth (126). Since 4FTrp does not fluoresce like Trp, growth of 4FTrp utilizing mutants makes possible the preparation of proteins lacking in Trp fluorescence (127).

These mutations of the amino acid alphabet of the genetic code proved Tenet 4 of the Coevolution Theory that the genetic code is an intrinsically mutable code. That they were obtained through a small number of mutational steps suggested that the normal inability of 4FTrp to support QB928 propagation could be due to the incorporation of 4FTrp causing dysfunction of a small number of essential but 4FTrp-sensitive proteins. When appropriate mutations were introduced into LC8 and LC33, so the dysfunction was overcome, 4FTrp became supportive of cell propagation. Conversely, when some other essential proteins were mutated in HR15 and HR23, so that Trp incorporation resulted in protein dysfunction, Trp was rejected from the code as an incompetent building block. These results suggested that the remarkable stability of the 20 amino acid alphabet could be safeguarded by *oligogenic barriers* comprising analogue-sensitive proteins that turn dysfunctional when one of the canonical 20 is replaced by an analogue (73, 126). To test this possibility, the complete genomic sequences of QB928, LC8, LC33, HR23 and TR7 were determined, revealing only nine mutations between QB928 and LC8, and six mutations between LC8 and LC33 (which propagated 211-fold faster than LC8 on 4FTrp relative to Trp). While the conversion of LC33 to HR23 was accompanied by 40 mutations, the conversion of HR23 to TR7 restoring the capacity of Trp to support cell propagation was accompanied by only two mutations. These limited numbers of incurred mutations, including both secondary and incidental mutations, provided strong evidence for the important role of oligogenic barriers in safeguarding the extreme stability f the canonical amino acid alphabet encoded by the genetic code (128).

In synthetic biology, arrays of genes are assembled to produce designer genomes, but these genomes still utilize the canonical amino acid, ribonucleotide and deoxyribonucleotide alphabets. In the case of the HR15 strain, in contrast, the canonical amino acid alphabet itself has been altered. As Ellington observed in Code Breakers (129), *"HR15 is not just a picky eater, but an entirely new type of life."*

Accordingly, the term *synthetic life* has been proposed by Wong and Xue (130) to describe novel organisms such as HR15, where the canonical amino acid or nucleotide alphabet, or some other universal constituent, has been altered. As shown in Table 4, the resultant synthetic life forms can be either optional viz. *o-Synthetic*, where the organism can live on either the old or the new alphabet, or mandatory viz. *m-Synthetic*, where the new alphabet is essential for life. Since the first examples provided by *B. subtilis* LC8, LC33, HR15 and their derivatives, synthetic life forms have increased rapidly in number and become wide ranging in scope:

(a) *E. coli* has been mutated to yield an *unColi* B7-3 strain that can be propagated on 4FTrp without a detectible level of Trp incorporation into proteins (131, 132).

(b) The adaptation of the protein sequences of bacteriophage Q to growth on 6FTrp furnished evidence for genetic code divergence through ambiguous intermediate genomes that can simultaneously accommodate more than one amino acid at a given codon (133).

(c) Phenotypic suppression has been introduced as a method to induce propagation-dependence on an unnatural amino acid. *E. coli* was rendered thymidine auxotrophic by the Arg126Leu mutation of thymidylate synthase, and the resultant *E. coli* thyA R126L strain could be propagated without added thymidine only in the presence of azaLeu the incorporation of azaLeu at residue 126 restored an essential positive charge at that position for enzyme function (134).

:

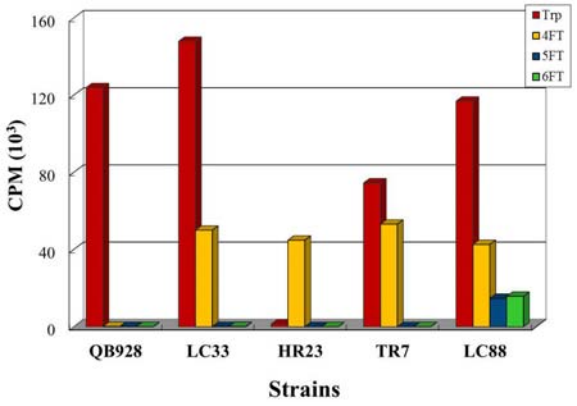

Figure 11. Propagation of *B. subtilis* QB928 and its genetic code mutants on Trp (red), 4FTrp (orange), 5FTrp (blue) and 6FTrp (green). Revertant TR7 propagates even better on Trp than on 4FTrp.

(d) As shown in Table 5, numerous aaRS are found to display strikingly low reactivities toward tRNA from other species, especially when the tRNA comes from the other side of a cross reactivity schism between the Archaea-Eukarya and Bacteria blocs (135). On the basis of such deficient cross-bloc reactivities, the strategy has been devised by Schultz and coworkers whereby an orthogonal tRNA-aaRS pair, e.g. archaeal TyrRS-tRNATyr from *Methanococcus jannaschii*, LeuRS-tRNALeu from *Methanobacterium thermoautotrophicum*, or GluRS from *Pyrococcus horikoshii* with a consensal archaeal tRNAGlu can be introduced into *E. coli* incurring minimal reaction with host aaRS and tRNAs (136). By fashioning the orthogonal pair to accept an unnatural amino acid and decode a special codon such as a nonsense codon, this *orthogonal aaRS-tRNA* approach has brought about the site-specific encoding of numerous unnatural amino acids, e.g. p-aminoPhe, p-azidoPhe etc (137, 138).

(e) In the case of orthogonally encoded p-aminoPhe, it has been possible to generate a biosynthetic pathway to render its incorporation into *E. coli* autonomous (139).

(f) Orthogonal suppressor tRNAs can be screened based on species-specific toxicity of suppressor tRNAs (140).

(g) Genetically encoded unnatural amino acids can be used as attenuator or activator of gene transcription (141).

(h) Genetically encoded photoreactive unnatural amino acids can be used to implement photoclick chemistry on proteins (142).

(i) Orthogonal ribosome has been evolved to efficiently direct the incorporation of unnatural amino acids in response to quadruplet codons (143).

(j) Genetic encoding of unnatural amino acids has been extended to multiple amino acids, e.g. homopropagylglycine, 4-azaTrp and (4s)-FPro, on the same target protein (144).

(k) Genetic encoding of unnatural amino acids has been extended to as many as 24 positions on the same target protein (145).

(l) Genetic encoding of unnatural amino acids has been extended to *S. cerevisiae* (146, 147), *Xenopus laevis* oocytes (148), mammalian cells (149, 150) and the multicellular animal *Caenorhabditis elegans* (151, 152).

Furthermore, synthetic life has been achieved not only with altered amino acid alphabets, but also with an altered DNA alphabet. Marliere *et al* (153) have evolved an *E. coli* thymidine auxotroph into a chemically modified organism *E. coli* CLU5 whose DNA genome is composed of the bases adenine, guanine, cytosine and the thymine analogue 5-chlorouracil. Thereby the DNA alphabet is transformed from the canonical A, G, C and T to A, G, C and 5-chloroU.

It has been pointed out that evolution has been restrained throughout by the use of the same old 20 canonical amino acids in proteins and 4 canonical bases in DNA from primordial times to the present. With synthetic life, however, a sequel has now been launched with altered protein alphabets (154), and recently an altered DNA alphabet as well. What other universal constituents of living organisms may also prove to be alterable will have to be determined experimentally.

15. DISCUSSION

The emergence of life covers a three billion year journey. It comprises eight milestone stages, each contributing a fundamental advance toward the development of the genetic information system.

15.1. Stage 1. Prebiotic synthesis

Prebiotic development was initiated by the prebiotic accumulation of a range of organic compounds through chemical synthesis, both on Earth and in space followed by delivery to Earth. A large number of biomolecules have been obtained from simulated prebiotic synthesis (32). Although there has been concern regarding the lack of simulated prebiotic synthesis of some of the canonical amino acids (155), the Triple Convergence indicates that only Phase 1 amino acids needed to be produced by prebiotic synthesis, and this need has been fulfilled experimentally with the production of all 10 Phase 1 amino acids (67, 68).

Some simulated prebiotic syntheses such as those of nucleotides and RNA polymers yet remain problematic, but important discoveries are continuing as in the case with shortage of prebiotic phosphate, once thought to be intractable but becoming solvable based on recent progress (156), deficiency of nucleotide synthesis which is being surmounted by new routes of synthesis (157), or difficulty of RNA polymerization where advances are forged through

innovative synthesis and exploration of precursor polymers of RNA (12, 33-37, 47).

15.2. Stage 2. Functional RNA selection by metabolite

Because prescriptive information (PI) prescribes utility (5), life's emergence would be aborted if there was no production of prebiotic PI. Since none of chaos theory, complexity theory, fractals, rugged fitness landscapes, Markov chains, hypercycles, dissipative structures, Shannon information theory, autopoiesis, evolutionary algorithms and directed evolution is found to generate any prebiotic or abiotic PI (4), the selection of PI-containing functional RNA by metabolites over random RNA through REIM represents so far a unique mechanism for prebiotic PI generation.

The preconditions for the amplification of functional RNA according to the Metabolic Expansion Law are stringent, requiring the simultaneous presence of a sizable range of metabolites and a huge number of random RNA duplexes. This very stringency lends substance to the long held view that life could be unique, or close to unique, in the universe. It is a singular achievement of research on prebiotic chemistry and functional RNA chemistry that these preconditions are now recognized to be within the realizable realm on primitive Earth. In fact, given the stringency of these preconditions, the room for any emergence of life on other planetary systems to depart from the Earth Model based on interactions between metabolites and REIM-responsive replicators, with eventual enclosure within membranes, could be limited. On the other hand, on whatever endless number of planets throughout the universe where its preconditions are fulfilled, the Metabolic Expansion Law will invariably propel prebiotic development toward the emergence of life.

15.3. Stage 3: RNA world

Ribozymes have resolved the vexatious chicken-or-egg dilemma of whether biocatalysts should precede replicators or vice versa. As well, random RNAs are found to be a reliable source for the experimental isolation of aptamers, and naturally occurring riboswitches regulate transcription or translation in response to a wide range of metabolites (158), both in accord with the prebiotic importance of not only ribozymes but also aptamers. Notably, functional RNAs could be selected out of a massive pool of random RNAs based on the Metabolic Expansion Law, and all functional RNAs are self-replicating through complementary base-pairing. In contrast, there is no general mechanism that could select functional polypeptides out of a massive pool of random polypeptides. Moreover, self-replicating proteins are rare, and ones that also catalyze metabolic reactions are either extremely rare or non-existent. Accordingly the RNA World has to be regarded as an indispensable stage in the emergence of life.

Aptamers and ribozymes were the bearers of prebiotc PI. Their versatility in binding ligands and metabolites, attested to by the experimental isolation of increasing numbers of these molecules, directly led to the accumulation of functional information, at first in the cumulons and later on in the precells and early cells. As functional information content increased, the variety of ribozyme-catalyzed reactions followed suit. These catalyzed reactions can be counted on to organize themselves, imbued with regulatory controls aligned with utility-directedness, under the joint direction of thermodynamics and natural selection to yield the pathways and cycles required by a living cell.

15.4. Stage 4: Peptidated RNA world

A direct transition of the RNA World, where RNAs served as biocatalysts, to the Protein World, where proteins serve as biocatalysts, was not a practicable possibility. The reason is, ribozymes might well develop a wide range of catalytic reactions, bring about a thriving metabolism, and proceed to construct a peptidyl transferase center (PTC) to make polypeptides, but the question is, what polypeptides? Since there is no REIM-type mechanism for functional polypeptide selection, the PTC would have to produce random protein sequences, so natural selection could weed out useless random sequences in order to select the functional ones. Besides the lack of evolutionary incentive for making random sequences, the problem of the random sequences itself would be overwhelming. Instead, it was necessary to have a Peptidated RNA World where RNA directed the development of functional protein folds and domains prior to ribosomal synthesis at PTC, so that the proteins produced by PTC were not random sequences, but conjoint assemblies of protein folds and domains whose functional value had already been proven by their actual collaboration with RNA.

The timelines of protein folds and ribosome development (57, 58) provide an important basis for delineation of the primordial world by establishing the timing of various events, including the appearance of protein domains prior to the evolution of ribosome, cooperation between r-RNA and r-proteins to start off ribosome evolution, and the delayed appearance of PTC long after the start of ribosome evolution. In this regard, the first appearance of PTC provides a useful time point dividing the timelines into pre-PTC and post-PTC periods, and the identified events in these periods are consistent with a multi-staged development of RNA peptidation:

(I) *Idiosyncratic peptidation.* Initially, individual ribozymes produced covalently attached aminoacyl and peptidyl prosthetic groups to assist function each in its own way, as has been demonstrated by the discoveries of different ribozymes that catalyze the formation of aminoacyl- and peptidyl-RNAs (51-53). That a tiny 5-nt-long ribozyme can catalyze the formation of multiple peptidyl-RNAs (159) indicates that RNA peptidation could be widespread among ribozymes. Furthermore, ribozymes can use aminoacyl-CoA thioesters to aminoacylate a non-terminus ribose 2'-OH (160), which vastly increases the number of aminoacylation and peptidation sites on the RNA, and potentially stabilizes the peptide attachment on account of the lack of a vicinal hydroxy group on the ribose.

(II) *Origin of mRNA.* The pre-PTC arrival of the catalytic domains of Class I and Class II aaRS enabled aaRS-

catalyzed formation of aminoacyl-CoA thioesters (161) for RNA peptidation and, together with the evidence of pre-PTC ribosome-tRNA interactions, pointed to the usage of aminoacyl-tRNAs to supply amino acids for RNA peptidation. Moreover, the development of mRNA decoding and helicase activities by the young pre-PTC ribosome was indicative of the occurrence of mRNA translation. The sequence of a ribozyme or aptamer, being constrained by the catalytic or metabolite binding task it must perform, could not double duty well as mRNA. Instead, a ribozyme or aptamer had to develop in time two different domains: a *functional domain* carried out catalysis or metabolite binding in collaboration with its polypeptide prosthetic group, and an *encoding domain* acted as mRNA to direct the amino acid sequence of the polypeptide (Figure 3). Evidence for an origin of mRNA stemming from primordial RNAs endowed with both a substrate/ligand binding domain and an encoding domain is provided by the finding that nearly all known riboswitch aptamers to-day are in fact located in non-coding regions of mRNAs (158). On this basis, each primordial ribozyme or aptamer would come to contain the specific mRNA sequence that encoded its own protein prosthetic group, which was of course devoted to the selfsame ribozymic or aptamer task as the functional domain on the RNA. Functional continuity and one-to-one correspondence between RNA and its protein product were thereby assured as the functional RNA eventually passed on its catalytic or metabolite binding responsibility to its protein product, and retreated to fulfill the role of a full time mRNA.

(III) *Protein folds and domains*. Protein studies have amply underlined the importance of protein folds and domains as key elements of protein structure and function. The Peptidated RNA World was the incubator, and the polypeptide prosthetic groups on the ribozymes and aptamers the precursors, of these elements. To-day, the non-ribosomal peptide synthetase (NRPS) A-domain catalyzes the formation of aminoacyl-AMP and its transfer to a thiol acceptor on the NRPS peptidyl carrier protein (PCP) (162). The pre-PTC appearance of this A-domain along with the aaRS domains expedited the formation of polypeptide prosthetic groups on RNAs. Wherever the polypeptide prosthetic groups were attached on the RNAs initially, it would be advantageous for them to be migrated in time to the 3' termini of the RNAs in labile ester bonds, for this would facilitate their subsequent transposition to the 3' termini of tRNAs at PTC, as well as their cleavage from the RNAs to form stand alone proteins.

(IV) *Centralized peptidation*. Although the young pre-PTC ribosome was equipped with mRNA decoding and helicase functions, it was devoid of PTC and therefore unable to conduct ribosomal protein synthesis. This suggests that the pre-PTC and post-PTC ribosomes could take on different responsibilities: the pre-PTC ribosome performed centralized (instead of idiosyncratic) peptidation of ribozymes and aptamers using the encoding domain of ribozyme or aptamer as mRNA, whereas the post-PTC ribosome performed ribosomal protein synthesis via peptidation of tRNA using dedicated single-purpose mRNA. Both of them likely employed tRNA-like adaptors and non-overlapping triplet codons in translation, thereby reducing any transitional discontinuity between them. It is noteworthy that the PTC has persisted as the ribosomal peptide-bond maker through the long Protein World ages unreplaced by any enzyme, even though PTC catalyzes peptide bond formation with only a modest rate constant of >300 sec^{-1}, much slower than many enzymes (163). This suggests that either there are important, not fully understood, advantages to the usage of a ribozymic peptidyl transferase over an enzymic one, or the transition from a ribozymic to an enzymic peptidyl transferase proved too complex for the cells to implement. Either factor would favor the usage of a ribozyme by the pre-PTC ribosome in peptide bond formation. The late post-PTC appearance of the peptide bond-making NRPS C(condensing) protein domain was also in keeping with this possibility.

Accordingly, prior to the advent of PTC, the Peptidated RNA World would already have developed for its own requirements aaRS, mRNA, tRNA, triplet codons, protein folds and domains, and a young mRNA-decoding ribosome. What transition to the Protein World still needed was confined mainly to developing PTC, and reorienting the translation apparatus from centralized RNA peptidation to protein synthesis. Without intermediation by Peptidated RNA World, such a seamless transition between RNA World and Protein World would be inconceivable.

On this basis, the three biocatalysis worlds, defined based on the nature of the dominant biocatalysts employed during the period, might be demarcated as follows:

(a) The RNA World began with REIM-selection of functional RNA and lasted until the first appearance of protein folds and domains; its historical role with respect to the development of translation was to expand functional information and initiate RNA peptidation.

(b) The Peptidated RNA World began with the first appearance of protein folds and domains and lasted until the appearance of PTC; its historical role was to produce aaRS, mRNA, tRNA, triplet codons and young pre-PTC ribosomes making polypeptide folds and domains on functional RNAs.

(c) The Protein World began at the appearance of PTC, and its historical role is to perfect PTC and ribosomal protein synthesis for hopefully never ending future eons.

Because the transition from precell to living cell had to be a highly demanding process, its occurrence might likely fall

within the Peptidated RNA World or Protein World more than the RNA World. In that case, historically there could be ribo-precells, peptidoribo-precells and peptidoribo-organisms, but not necessarily any ribo-organisms.

15.5. Stage 5. Coevolution of genetic code and amino acid biosynthesis

The Coevolution Theory predicted that while some of the 20 protein amino acids (Phase 1) were supplied by the prebiotic environment, others (Phase 2) were derived later from amino acid biosynthesis; pretran synthesis was a pivotal mechanism making possible the entry of Phase 2 amino acids into the genetic code; the entry of the Phase 2 amino acids, receiving codons from their biosynthetic precursors, was a key determinant of the structure of the genetic code; and the genetic code is a mutable code. It has taken three decades of advances in genomics analysis, prebiotic synthesis, meteoritic science and bacterial genetics to verify all of these predictions.

The coevolution case history on Earth also improves the probability of development of life on other planets. Based on the Earth Model, there is no need to begin prebiotic evolution with an optimal alphabet of building blocks. Instead, the process could start off with a suboptimal alphabet. Once started, the polymers consisting of these building blocks will expand catalysis and metabolism, leading to the appearance of additional biosynthetically derived building blocks. As the history on Earth illustrates, the search for novel building blocks will be unending until an optimal canonical alphabet has been achieved.

Pretran synthesis predates aaRS for the production of Asn-tRNA, Gln-RS and Cys-tRNA, and it is essential for the production of Sec-tRNA. In addition, it initiated the biosynthesis of Cys (75), and Asn (76, 77). That pretran synthesis introduced Cys and Asn into the living system makes biological sense, for these amino acids are far more important for protein structure and function than for general metabolism. As well, pretran synthesis and aminoacyl-tRNA participate in protein N-modifications, the ubiquitin pathway, 5-aminolevulinic acid and porphyrin biosynthesis, donation of Ser for antibiotic synthesis, cross-linking of peptidoglycan, and lysination of lipid bilayer (164-167). Clearly, pretran synthesis has given rise to biosynthetic routes important to ancient and present-day cells, and offers potential pathways for the production of novel amino acid-derived compounds in synthetic biology.

15.6. Stage 6. Last universal common ancestor

Since LUCA is strongly indicated to be an extinct species phylogentically close to *Methanopyrus*, how it could have acquired its LUCAhood has to be addressed. Although competition among organisms is always unrelenting, multiple species usually coexist in most habitats. That a single LUCA lineage could eliminate all competitor lineages on Earth to establish its LUCAhood is thus extraordinary, and suggests that LUCA might have developed some decisive biochemical innovation(s) that its contemporaries lacked. Because RNA and its constituents were developed in the RNA World, which would predate LUCA by eons, the innovation in question might not involve RNA. In contrast, both Gln-tRNA and Asn-tRNA in living organisms are produced by either pretran synthesis or aaRS; among 59 genomes analyzed, 34% use aaRS for Gln-tRNA, and 54% use aaRS for Asn-tRNA (66). Such mixed usages of two different pathways indicate that the genetic coding mechanisms were still under development at the time of LUCA. As well, some of the molecular mechanisms for DNA replication are non-homologous across the three biological domains with virus infections being a possible contributor to such non-homologies (122), likewise indicating that the DNA information machinery was still under development at that time. Accordingly, the 20 amino acid canonical code and/or the DNA genome were likely among the biochemical innovations that contributed to LUCA's rise.

The rise of a methanogenic LUCA is consistent with the early success of methanogens giving rise to a primitive Earth atmosphere enriched in biogenic methane (118, 119), which in turn suggests the potential significance of atmospheric methane detection in future astrobiological explorations of planets with habitable zones. The greater primitivity of Archaea compared to Bacteria and Eukarya is also in accord with the more extensive spread of Archaea into extremophilic environments. Compared to the Archaea, Bacteria and Eukarya would be later-arriving specialists that have excelled in narrower ranges of habitats.

When *Methanopyrus kandleri* from Guaymas Basin, Gulf of California was compared to the *Methanopyrus* isolates GC34 and GC37 from Pacific Ocean and KOL6, TAG1, TAG11 and SNP6 from Atlantic Ocean using minimal intragenomic genetic distance between the *ValRS* and *IleRS* genes as an indicator of primitivity, the Pacific lineages were found to be more primitive than the Atlantic ones. Both groups, however, were younger than environmental genomes from the Kairei Field of Central Indian Ridge. Therefore the most LUCA-proximal *Methanopyrus*, viz. the world's oldest organism, lives in the Indian Ocean (168).

15.7. Stage 7. Darwinian evolution

Stages 1 and 2 of life's emergence were physicochemical rather than natural selection-based. The governing physicochemical principles during these stages included those of equilibrium thermodynamics (14, 169), non-equilibrium thermodynamics (170) and Metabolic Expansion Law. Furthermore, advances in systems chemistry have provided important insights into the organization of prebiotic replicator and metabolic networks (171-178).

However, as soon as the replicators and metabolites in the cumulons were enclosed into lipoidal vesicles to form precells that underwent growth and imprecise division, natural selection began to act, to this day, as the foremost factor shaping the emergence of life. Accordingly, although Darwinian evolution was originally proposed to explain the speciation of organisms on the foundation of natural selection, the occurrences in Stages 3-6 were no less driven by natural selection.

For post-LUCA life, the usages of the DNA genome and 20 amino acid canonical code have remained stabilized, and Darwinian evolution has centered on DNA sequence variations. In the first century following the publication of *On the Origin of Species* in 1859, morphological changes in biological lineages as revealed by the fossil records of organisms large, small and microscopic have furnished the most incisive evolutionary records. In recent years, the genomic sciences are turning genomes into open books, and making possible analysis of biomolecular changes in the three domains of Archaea, Bacteria and Eukarya down to individual genera and species. While evolution has provided effective explanation of biological phenomena of wide ranging descriptions, important questions yet require further delineation, including the causes of macroevolutionary events and punctuated equilibria, the vastly different lifespans of different organisms, the >64 tRNA sequences encoded by a variety of eukaryotes (106), and the rarity if not absence of photosynthesis in Archaea (111) etc.

Regarding speciation events, a major question relates to the evolutionary causes underlying the deep-seated differences between Archaea, Bacteria and Eukarya. Adaptive advantages offered by the bacterial and eukaryotic cell plans over that of the archaeal cell plan likely contributed to the divergence of Bacteria and Eukarya from the ancestral Archaea. With the *Methanopyrus*-proximal hyperthermophilic methanogen LUCA suggested by present evidence, the upheaval accompanying the dispersion of its offsprings from the hydrothermal vents to diverse habitats in the cooler zones would be conducive to the appearance of many exploratory cell plans; two of these survived and flourished, one bacterial and one eukaryotic.

The bacterial and eukaryotic cell plans employ ester lipids instead of the archaeal ether lipids. They have also introduced a number of basic changes into the translation machinery:

(a) Both bacteria and eukaryotes have developed complex usages of different anticodon combinations to read different standard codon boxes within the same genome (110);

(b) Among 18 archaeons analyzed, only the CUN codon box of *Ferroplasma acidarmanus* employs a tRNA with an A at the first anticodon position; however, in this instance the AAG anticodon is employed along with UAG and CAG anticodons for the same box (110). Thus the Crick wobble rule utilizing an I anticodon base to read U, C and A codon bases is mainly a bacterial-eukaryotic convention rarely used if at all by archaeons.

(c) The elongation factor LepA (EF4), one of the most highly conserved proteins, is present in all bacteria, nearly all eukaryotes in the organelles, but not in Archaea. It enables back translocation during translation, and prevents ribosome stalling (179).

(d) Bacteria use formyl-Met-tRNA$_i$ for protein chain initiation.

(e) Eukaryotes perform translation with 40S and 60S rather than 30S and 50S ribosomal subunits.

Of these changes, the adaptive utility of LepA is evident. Complex anticodon usages and the Crick wobble rule pertaining to INN anticodons may also enhance the ability of the ribosome to fine-tune codon-anticodon base-pairings in different codon boxes. The potential advantages of formyl-Met-tRNA$_i$, 40S r-subunit or 60S r-subunit are unclear.

The temporal sequence of innovations in ribosomal translation in early Darwinian evolution is particularly evident in anticodon usages. Superwobbling where a single UNN anticodon reads all four codons in a codon box is practicable and might be important for the most primitive genetic codes, but reduces translational efficiency (180). Accordingly, LUCA-proximal *Methanopyrus kandleri* employs a GNN+UNN two-anticodon combination to read all its standard codon boxes. The majority of archaeons have advanced to the use of a GNN+UNN+CNN three-anticodon combination for all their standard codon boxes. Bacteria and Eukarya have gone one step further, and developed frequent use of three or more different anticodon combinations to read their standard codon boxes (110; and Figure 8).

Among the archaeons, *Methanosarcina acetivorans* has a particularly large 5.75 Mb genome, and it is known to enter into an optional multicellular state (181). This suggests that a genome size of >5.75 Mb might be important for multicellularity. Thus the development of a nuclear membrane by Eukarya facilitating the adoption of large genomes could be an underlying factor to the splendor of the flora and fauna of this domain.

15.8. Stage 8. Synthetic life

Both the genetically encoded 20-amino acid protein alphabet and the A, G, C, T DNA alphabet have been experimentally altered in proof of the intrinsic mutability of these alphabets, producing synthetic life of both the optional and mandatory varieties, and making possible the use of an altered alphabet on a proteome-wide, genome-wide or site-specific basis. Starting with the isolation of the first synthetic life forms of *Bacillus subtilis* LC8, LC33 and HR15 just 30 years ago, synthetic life is now undergoing rapid advances. Over the next 70 years, its accomplishments may be expected to be something of immense value. When developed under the strictest requisite safety guidelines, synthetic life forms will deepen understanding of life's emergence and give rise to medical and biotech innovations. The parameters of the living state that have enabled three billion years of uninterrupted expansion and diversification through functional RNA selection and natural selection will become fathomable. Fundamental but otherwise unapproachable questions such as the number and nature of amino acids or ribo- and deoxyribo-nucleotides that are compatible with life, and the replaceability of universally utilized lipids, carbohydrates and cofactors, will be brought into the realm of scientific enquiry.

16. ACKNOWLEDGMENTS

I wish to thank Prof. Hannah Hong Xue for valuable discussion, and Dr. Wai-Kin Mat and Xiaofan Ding for expert assistance.

17. REFERENCES

1. Jacques Monod. Chance and Necessity. Vintage Books, New York p. 138–148 (1972)

2. Michael Polanyi: Life's irreducible structure. *Science* 160, 1308-12 (1968)

3. Howard H. Pattee: On the origin of macromolecular sequences. *Biophys J* 1, 683–710 (1961)

4. David L. Abel: The capabilities of chaos and complexity. *Int J Mol Sci* 10, 247-291 (2009)

5. David L. Abel: The biosemiosis of prescriptive information. *Semiotica* 174, 1–19 (2009)

6. David L. Abel: The 'cybernetic cut': progressing from description to prescription in systems theory. *Open Cybernetics Systematics J* 2, 252-262 (2008).

7. Gyula Palyi, Claudia Zucchi, Luciano Caglioti: Short definitions of life. In: Fundamentals of Life. Eds: Palyi G, Zucchi C, Caglioti L. Elsevier, Paris p. 15-55 (2002)

8. J Tze-Fei Wong: Introduction. In: Prebiotic Evolution and Astrobiology. Eds: J Tze-Fei Wong, Antonio Lazcano. *Landes Bioscience, Austin* 1-9 (2009)

9. J Tze-Fei Wong, Hong Xue: Self-perfecting evolution of heteropolymer building blocks and sequences as the basis of life. In: Fundamentals of Life. Eds: Palyi G, Zucchi C, Caglioti L. Elsevier, Paris p. 473-494 (2002).

10. Gerald F. Joyce, Leslie E. Orgel: Prospects for understanding the origin of the RNA World. In: The RNA world 2nd ed. Eds: Gesteland RF, Cech TR, Atkins JF. Cold Spring Harbor Laboratory Press, p. 49-77 (1999)

11. Tracey A Lincoln, Gerald F Joyce: Self-sustained replication of an RNA enzyme. *Science* 323, 1229-32 (2009)

12. Jack W Szostak: The eightfold path to non-enzymatic RNA replication, *J Systems Chem* 3, 2 (2012)

13. Gunter Wachtershauser: Before enzymes and templates: theory of surface metabolism. *Microbiol Rev* 52, 452-484 (1988)

14. Harold J Morowitz, Jennifer D Kostelnik, Jeremy Yang, George D. Cody: The origin of intermediary metabolism. *Proc Natl Acad Sci USA* 97, 7704-7708 (2000)

15. Stuart A. Kauffman: Autocatalytic sets of proteins. *J Theor Biol* 119, 1-24 (1986)

16. Frank AL Anet: The place of metabolism in the origin of life. *Curr Opinion Chem Biol* 8, 654-659 (2004)

17. Leslie E. Orgel: The implausibility of metabolic cycles on the prebiotic Earth. *PLoS Biol* 6, e18 (2008)

18. James M. Carothers, Stephanie C. Oestreich, Jonathan H. Davis, Jack W. Szostak: Informational complexity and functional activity of RNA structures. *J Am Chem Soc* 126, 5130–5137 (2004)

19. Malcolm M. Anderson: Hybridization strategy in gene probes 2, Eds: B. D. Hames and S. J. Higgins, Oxford Univ. Press, p.8 (1995)

20. Veena Vijayanathan, Thresia Thomas, Leonard H Sigal, T. J. Thomas: Direct measurement of the association constant of HER2/neu antisense oligonucleotide to its target RNA sequence using a molecular beacon. *Antisense Nucl Scid Drug Develop* 12, 225-233 (2002)

21. H Thomou, NK Katsanos: The theory of DNA-RNA hybridization reaction *Biochem J* 153, 241-247 (1976)

22. Jack W Szostak: Molecular messages. *Nature* 423, 689 (2003)

23. Markus Pech, Knud H. Nierhaus. The minimal cell. In: Prebiotic Evolution and Astrobiology. Eds J. Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin p10-17 (2009)

24. Eun Jeong Cho, Andrew D. Ellington: Optimization of the biological component of a bioelectrochemical cell. *Bioelectrochemistry* 70, 165-72 (2007)

25. David M. Brackett□, Thorsten Dieckmann: Aptamer to ribozyme: the intrinsic catalytic potential of a small RNA. *Chembiochem* 7, 839-43 (2006)

26. Gerald F Joyce: Evolution in an RNA World. *Cold Spring Harbor Symp Quant Biol* 74, 17-23 (2009)

27. Stanley L. Miller: The formation of organic compounds on the primitive Earth. *Annals NY Acad Sci* 69, 260-275 (1957)

28. S Yuasa, D Flory, B Basile, J. Oro: Abiotic synthesis of purines and other heterocyclic compounds by the action of electric discharges. *J Mol Evol* 21, 76-80 (1984)

29. David W Deamer. Prebiotic amphiphilic compounds. In: Origins Genesis, Evolution and Diversity of Life. Ed: Joseph Seckbach. Kluwer p.77-89 (2004)

30. Ahmed I Rushdi, Bernd R T Simoneit: Lipid formation by aqueous Fischer-Tropsch-Type synthesis over a temperature range of 100 to 400 °C. *Orig Life Evol Biosph* 31, 103-118 (2001)

31. C Cheng, C Fan, R Wan, C Tong, Z Miao, J Chen, Y Zhao: Phosphorylation of adenosine with trimetaphosphate

under simulated prebiotic conditions. *Orig Life Evol Biosph* 32, 219-224 (2002)

32. J Tze-Fei Wong: Biomolecules. In: Prebiotic Evolution and Astrobiology. Eds J Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin p.65-75 (2009)

33. Leslie E Orgel: Prebiotic chemistry and the origin of the RNA World. *Critical Review Biochem Mol Biol* 39, 99-123 (2006)

34. James P Ferris: Montmorillonite-catalyzed formation of RNA oligomers: the possible role of catalysis in the origin of life. *Phil Tran R Soc B* 361, 1777-1786 (2006)

35. Pierre-Alain Monnard, Anastassia Kanavarioti, David W. Deamer: Eutectic phase polymerization of activated ribonucleotide mixtures yields quasi-equimolar incorporation of purine and pyrimidine nucelobases. *J Am Chem Soc* 125, 13734-13740 (2003)

36. S Rajamani, A Vlassov, S Benner, A Coombs, F Olasagasti, D Deamer: Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Orig. Life Evol Biosph* 38, 57-74 (2007)

37. Pierre-Alain Monnard: The dawn of the RNA world. In: Prebiotic Evolution and Astrobiology. Eds J. Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin p.76-86 (2009)

38. Andre Brack: The chemistry of life's origins. In: Origins, Genesis, Evolution and Diversity of Life. Ed: Seckbach J. Kluwer. Academic Publishers, p.61-73 (2004)

39. Eugene V Koonin, William Martin: On the origin of genomes and cells within inorganic compartments. *Trends* Genet 21, 647-54 (2005)

40. Christian De Duve: Blueprint for a Cell. Neil Patterson Publishers. p. 113-116 (1991)

41. Pasquale Stano, Pier Luigi Luisi: Precellular evolution. In: Prebiotic Evolution and Astrobiology. Eds J Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin p.94-105 (2009)

42. Humberto R Maturana, Francisco J Varela: Autopoiesis and cognition: the realization of the living. Springer. (1980)

43. Stephen J Mojzsis, Ramanarayanan Krishnamurthy, Gustaf
Arrhenius. Before RNA and after: geophysical and geochemical constraints on molecular evolution. In: The RNA World, 2nd ed, Eds: R. F. Gesteland, T. R. Cech, J. F. Atkins. Cold Spring Harbor Laboratory Press 1-47 (1999)

44. Atsushi Nakabachi, Atsushi Yamashita, Hidehiro Toh, Hajime Ishikawa, Helen E Dunbar, Nancy A Moran, Masahira Hattori: The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* 314, 267 (2006)

45. Carole Lartigue, John I Glass, Nina Alperovich, Rembert Pieper, Prashanth P Parmer, Clyde A Hutchison III, Hamilton O Smith, J Craig Venter: Genome transplantation in bacteria: changing one species to another. *Science* 317, 632-638 (2007)

46. Stephen J Freeland, Robin D Knight, Laura F Landweber: Do proteins predate DNA? *Science* 286, 690-692. (1999)

47. Albert Eschenmoser: Etiology of potentially primordial biomolecular structures: from vitamin B12 to the nucleic acids and an inquiry into the chemistry of life's origin: a retrospective. *Angew Chem Int Ed Engl* 50, 12412-72 (2011)

48. J Tze-Fei Wong. Genetic code. In: Prebiotic Evolution and Astrobiology. Eds J Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin, p. 110-119 (2009)

49. Thomas R Cech: The RNA Worlds in context. *Cold Spring Harb Perspective Biol* 4(7):a006742 (2012)

50. J Tze-Fei Wong: Origin of genetically encoded protein synthesis: a model based on selection for RNA peptidation. *Orig Life Evol Biosph* 21, 165-76 (1991)

51. Mali Illangasekare, Michael Yarus: Small molecule-substrate interactions with a self-aminoacylating ribozyme. *J Mol Biol* 268, 631-639 (1997)

52. B Zhang, TR Cech: Peptidyl-transferase ribozymes: trans reactions, structural characterization and ribosomal RNA-like features. *Chem Biol* 5, 539-553 (1998)

53. N Lee, Y Bessho, K Wei, JW Szostak, H Suga: Ribozyme-catalyzed tRNA aminoacylation. *Nat Struct Biol* 7, 28-33 (2000)

54. Charles G Kurland: The RNA dreamtime: modern cells feature proteins that might have supported a prebiotic polypeptide world but nothing indicates that RNA world ever was. *Bioessays* 32, 866-871 (2010)

55. Harry F Noller: The driving force for molecular evolution of translation. *RNA* 10, 1833-1837 (2004)

56. Massimo Di Giulio: On the RNA world: evidence in favor of an early ribonucleopeptide world. *J Mol Evol* 45, 571-578 (1997)

57. Ajith Harish, Gustavo Caetano-Anolles: Ribosomal history reveals origins of modern protein synthesis. *PLoS ONE* 7, e32776 (2012)

58. Gustavo Caetano-Anolles, Kyung Mo Kim, Derek Caetano-Anolles: The phylogenomic roots of modern biochemistry: origins of proteins, cofactors and protein synthesis. *J Mol Evol* 74, 1-34 (2012)

59. Steven A. Benner, Petra Burgstaller, Thomas R. Battersby, Simona Jurczyk. Did the RNA World exploit

an expanded genetic alphabet? In: The RNA World 2nd ed. Eds: R.F. Gesteland, T.R. Cech, J.F. Atkins. Cold Spring Harbor Laboratory Press; 163-181 (1999)

60. Gonzalo L Vilas, Maria M. Corvi, Greg J Plummer, Andrea M. Seime, Gareth R. Lambkin, Luc G. Berthiaume: Posttranslational myristoylation of caspase-activated p21-activated protein kinase 2 (PAK2) potentiates late apoptotic events. *Proc Natl Acad Sci USA* 103, 6542-6547 (2006)

61. Randall A. Hughes and Andrew D. Ellington. Ribozymes and the evolution of metabolism. In: Prebiotic Evolution and Astrobiology. Eds: J. Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin (2009)

62. J Tze-Fei Wong: Evolution of the genetic code. *Microbiol Sci* 5, 174–181 (1988)

63. J Tze-Fei Wong: A coevolution theory of the genetic code. *Proc Natl Acad Sci USA* 72, 1909-1912 (1975)

64. J Tze-Fei Wong: Question 6: Coevolution theory of the genetic code: a proven theory. *Orig Life Evol Biosph* 37, 403-408 (2007)

65. J Tze-Fei Wong: Coevolution of the genetic code and amino acid biosynthesis. *Trends in Biochem Sc*i 6, 33-35 (1981)

66. J Tze-Fei Wong: Coevolution theory of the genetic code at age thirty. *BioEssays* 27, 416-425 (2005)

67. Kensei Kobayashi, Masahiko Tsuchiya, Tairo Oshima, Hiroshi Yanagawa: Abiotic synthesis of amino acids and imidazole by proton irradiation of simulated primitive earth atmosphere. *Orig Life Evol Biosph* 20, 99-109 (1990)

68. Kensei Kobayashi, Takeo Kaneko, Takeshi Saito, Tairo Oshima: Amino acid formation in gas mixtures by high energy particle irradiation. *Orig Life Evol Biosph* 28, 155-165 (1998)

69. Sandra Pizzarello: Meteorites and the chemistry that preceded life's origin. In: Prebiotic Evolution and Astrobiology. Eds: J. Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin, p. 46-51. (2009)

70. Massimo Di Giulio, Umberto Amato: The close relationship between the biosynthetic families of amino acids and the organization of the genetic code. *Gene* 435, 9-12 (2009).

71. Edward N Trifonov, Idan Gabdank, Danny Barash, Yehoshua Sobolevsky: Primordia Vita. Deconvolution from modern sequences. Orig Life Evol Biosph 36, 559-565 (2006)

72. J Tze-Fei Wong, Patricia M. Bronskill: Inadequacy of prebiotic synthesis as origin of proteinous amino acids. *J Mol Evol* 13, 115-125 (1979)

73. J Tze-Fei Wong. Evolution and mutation of the amino acid code. In: Dynamics of Biochemical Systems. Eds: Ricard J, Cornish-Bowden A. Plenum New York, 247-257 (1984)

74. Patrick O'Donoghue, Anurag Sethi, Carl R. Woese, Zaida A. Luthey-Schulten: The evolutionary history of Cys-tRNACys formation. *Proc Natl Acad Sci USA* 102, 19003-19008 (2005)

75. Hong-Yu Zhang, Tao Qin, Ying-Ying Jiang, Gustavo Caetano-Anolles: Structural phylogenomics uncovers the early and concurrent origins of cysteine biosynthesis ansd iron-sulfur proteins. *J Biol Struct Dynamics* 30, 542-545 (2012)

76. Herve Roy, Hubert Dominique Becker, Joseph Reinbolt, Daniel Kern: When contemporary aminoacyl-tRNA synthetases invent their cognate amino acid metabolism. *Proc Nalt Acad Sci USA* 100, 9837-42 (2003)

77. Christopher Francklyn: tRNA synthetase paralogs: Evolutionary links in the transition from tRNA-dependent amino acid biosynthesis to de novo biosynthesis. *Proc Nalt Acad Sci USA* 100, 9650-52 (2003)

78. Stephane Commans, August Böck: Selenocysteine inserting tRNAs: an overview. *FEMS Microb Rev* 23, 335-351 (1999)

79. Gayathri Srinivasan, Carey M. James, Joseph A. Krzycki: Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science* 296, 1459-1462 (2002)

80. Paul G Higgs: A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biology Direct* 4, doi:10.1186/1745-6150-4-16 (2009)

81. Massimo Di Giulio: An extension of the coevolution theory of the genetic code. *Biology Direct* 3:37, doi:10.1186/1745-6150-3-37 (2008)

82. Hong Xue, Wenyan Shen, Giege R. J Tze-Fei Wong: Identity elements of tRNA(Trp). Identification and evolutionary conservation. *J Biol Chem* 268, 9316-22 (1993)

83. Qing Guo, Qingguo Gong, Ka-Lok Tong, Bente Vestergaard, Annie Costa, Jean Desgres, Mansim Wong, Henri Grosjean, Guang Zhu, J Tze-Fei Wong, Hong Xue: Recognition by tryptophanyl-tRNA synthetases of discriminator base on tRNATrp from three biological domains. *J Biol Chem* 277, 14343-9 (2002)

84. Richard Giege, Marie Sissler, Catherine Florentz: Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acid Res* 26, 5017-5035 (1998)

85. J Tze-Fei Wong: The evolution of a universal genetic code. *Proc Natl Acad Sci USA* 73, 2336-2340 (1976)

86. J Tze-Fei Wong. Kinetics of Enzyme Mechanisms. Academic Press p. 200-201 (1975)

87. J Tze-Fei Wong: Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc Natl Acad Sci USA* 77, 1083-1086 (1980)

88. Stephen J Freeland, Laurence D Hurst: The genetic code is one in a million. *J Mol Evol* 47, 238-248 (1998)

89. Massimo Di Giulio, Mario Medugno: The level and landscape of optimization in the origin of the genetic code. *J Mol Evol* 52, 372-382 (2001)

90. Massimo Di Giulio: The origin of the genetic code cannot be studied using measurements based on the PAM matrix because this matrix reflects the code itself, making any such analysis tautologous. *J Theo Biol* 208, 141-144 (2001)

91. David Morgens, Andre Cavalcanti: An alternative look at code evolution: using non-canonical codes to evaluate adaptive and historical models for the origin of the genetic code. *J Mol Evol*, 76:71-80 (2013).

92. Rob Knight, Laura Landweber, and Michael Yarus: Tests of a stereochemical genetic code. In: Translation Mechanisms. Eds. J. Lapointe, L. Brakier-Gingras. Landes Bioscence 115-12 (2003)

93. Antonio Lazcano, Stanley L. Miller: On the origin of metabolic pathways. *J Mol Evol* 49, 424-431 (1999)

94. Carl Zimmer. How and where did life on Earth arise? *Science* 309, 89 (2005)

95. Jeffrey L Bada, Bruce Fegley Jr, Stanley L Miller, Antonio Lazcano, H. James Cleaves, Robert M. Hazen, John Chalmers. Debating evidence for the origin of life on Earth. *Science* 315, 937-938 (2007)

96. Günter Wächtershäuser, Claudia Huber: Response to "Debating evidence for the origin of life on Earth". *Science* 315, 938-939 (2007)

97. David S Ross: A quantitative evaluation of the iron-sulfur world and its relevance to life's origins. *Astrobiology* 8, 267-272 (2008)

98. Frederick A Kundell: A suggested pioneer organism for the Wachtershauser origin of life hypothesis. *Orig Life Evol Biosph* 41, 175-198 (2011)

99. Carl R Woese, Otto Kandler, Mark L Wheelis: Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87, 4576-79 (1990)

100. N Iwabe, K Kuma, M Hasegawa, S Osawa, T Miyata: Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA* 86, 9355-59 (1989)

101. James R Brown, W Ford Doolittle: Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci USA* 92, 2441-45 (1995)

102. Carl R Woese: Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci USA* 97, 8392-96 (2000)

103. E Pennisi: Is it time to uproot the tree of life? *Science* 284, 1305-07 (1999)

104. Yuri I Wolf, Igor B Rogozin, Nick V Grishin, Eugene V Koonin: Genome trees and the tree of life. *Trends Genetics* 18, 472-479 (2002)

105. J Tze-Fei Wong, Jianhuan Chen, Wai-Kin Mat, Siu-Kin Ng, Hong Xue: Polyphasic evidence delineating the root of life and roots of biological domains. *Gene* 403, 39-52. (2007)

106. Hong Xue, Ka-Lok Tong, Christian Marck, Henri Grosjean, J Tze-Fei Wong: Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. *Gene* 310, 59-66 (2003)

107. Hong Xue, Siu-Kin Ng, Ka-Lok Tong, J Tze-Fei Wong: Congruence of evidence for a Methanopyrus-proximal root of life based on transfer RNA and aminoacyl-tRNA synthetase genes. *Gene* 360, 120-130 (2005)

108. Feng-Jie Sun, Gustavo Caetano-Anollés: Evolutionary patterns in the sequence and structure of transfer RNA: early origins of archaea and viruses. PLoS Computational Biol 4:e1000018.(2008)

109. Jason Raymond, Daniel Segrè: The effect of oxygen on biochemical networks and the evolution of complex life. Science 311, 1764-67 (2006)

110. Ka-Lok Tong, J Tze-Fei Wong: Anticodon and wobble evolution. Gene 333, 169-177 (2004)

111. Wai-Kin Mat, Hong Xue, J Tze-Fei Wong: The genomics of LUCA. Frontiers Biosc 13, 5605-5613 (2008)

112. J Tze-Fei Wong. Root of life. In: Prebiotic Evolution and Astrobiology. Eds J. Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin, p. 120-144 (2009)

113. Eugene V Koonin: Comparative genomics, minimum gene sets and the last universal common ancestor. *Nat Rev Microbiol* 1, 127-136 (2003)

114. A Mushegian: Gene content of LUCA, the last universal common ancestor. *Frontiers Biosc* 13, 4657-66 (2008)

115. Massimo Di Giulio: The evidence that the tree of life is not rooted within the archaea is unreliable: a reply to Skophammer *et al. Gene* 194, 105-106 (2007).

116. John B. Corliss, John A. Baross, Sarah E. Hoffman: A hypothesis concerning the relationship between submarine hot springs and the origin of life on Earth. *Oceanol Acta* 4, 59-69 (1981)

117. Everett L. Shock: Chemical environment in submarine hydrothermal systems. *Orig Life Evol Biosphere (suppl.)* 22, 67-107 (1992)

118. James F. Kasting, Janet L. Siefert: Life and the evolution of Earth's atmosphere. *Science* 296, 1066-1068 (2002)

119. James Kasting: The primitive earth. In: Prebiotic Evolution and Astrobiology. Eds J. Tze-Fei Wong, Antonio Lazcano. Landes Bioscience, Austin, p. 57-64 (2009)

120. A Lazcano, R Guerrero, L Margulis, J. Oro: The evolutionary transition from RNA to DNA in early cells. *J Mol Evol* 27, 283-290. (1988)

121. Patrick Forterre: The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie* 87, 793-803(2005)

122. Patrick Forterre: Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domains. *Proc Natl Acad Sci USA* 103, 3669-3674 (2006)

123. Norman H. Sleep, Kevin J Zahnle, James F Kasting, Harold J Morowitz: Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* 342, 139-142. (1989)

124. Norman H Sleep: The Hadean-Archaean environment. *Cold Spring Harb Perspective Biol* 2, a002527 (2010)

125. J Tze-Fei Wong: Membership mutation of the genetic code: loss of fitness by tryptophan. *Proc Natl Acad Sci USA* 80, 6303–6306 (1983)

126. Wai-Kin Mat, Hong Xue, J Tze-Fei Wong: Genetic code mutations: the breaking of a three billion year invariance. *PLoS ONE* 5(8):e12206 (2010)

127. Patricia M. Bronskill, J Tze-Fei Wong: Suppression of fluorescence of tryptophan residues in proteins by replacement with 4-fluorotryptophan. *Biochem J* 249, 305-308. (1988)

128. Chi-Shing Yu, Aldrin Kay-Yuen Yim, Wai-Kin Mat, Amy Hin-Yan Tong, Si Lok, Hong Xue, Stephen Kwok-Wing Tsui, J Tze-Fei Wong, Ting-Fung Chan: Mutations enabling displacement of tryptophan by 4-fluorotryptophan as a canonical amino acid of the genetic code. *Genome Biol Evol* doi:10.1093/gbe/evu044 (2014)

129. Tina Hesman: Code breakers. Scientists are altering bacteria in a most fundamental way. *Science News* 157, 360–362 (2000)

130. J Tze-Fei Wong, Hong Xue: Synthetic genetic codes as the basis of synthetic life. In: Chemical Synthetic Biology. Eds: Pier Luigi Luisi, Cristano Chiarabelli. Wiley, p. 178-199 (2010)

131. Jamie M Bacher, Andrew D. Ellington: Selection and characterization of Escherichia coli variants capable of growth on otherwise toxic tryptophan analogues. *J Bacteriol* 183: 5414-5425 (2001)

132. Jamie M Bacher, Randall A Hughes, J Tze-Fei Wong, Andrew D Ellington: Evolving new genetic codes. *Trends Ecol Evol* 19, 69-75 (2004)

133. JM Bacher, JJ Bull, AD Ellington: Evolution of phage with chemically ambiguous proteomes. *BMC Evol Biol* 3, 24 (2003)

134. Beatrice Lemeignan, Pierre Sonigo, Philippe Marliere: Phenotypic suppression by incorporation of an alien amino acid. *J Mol Biol* 231, 161-166 (1993)

135. Yau Kwok, J Tze-Fei Wong: Evolutionary relationships between Halobacterium cutirubrum and eukaryotes determined by use of aminoacyl-tRNA synthetases as phylogenetic probes. *Can J Biochem* 58, 213-218 (1980)

136. Stephen W Santoro, J Christopher Anderson, Vishva Lakshman, Peter G Schultz: An archaealbacteria-derived glutamyl-tRNA synthetase and RNA pair for unnatural amino acid mutagenesis of proteins in Escherichia coli. *Nucleic Acid Res* 31, 6700-6709 (2003)

137. Chang C Liu, Peter G Schultz: Adding new chemistries to the genetic code. *Annu Rev Biochem* 79, 413-44 (2010)

138. Michael Georg Hoesl, Nediljko Budisa: Recent advances in genetic code engineering in Escherichia coli. *Curr Opinion Biotech* 23, 751-757 (2012)

139. RA Mehl, JC Anderson, SW Santoro, L Wang, AB Martin, DS King, DM Horn, PG Schultz: Generation of a bacterium with a 21 amino acid genetic code. *J Am Chem Soc* 125, 935-939 (2003)

140. Hong Tian, Danni Deng, Jie Huang, Dongning Yao, Xiaowei Xu, Xiangdong Gao: Screening system for orthogonal suppressor tRNAs based on the species-specific toxicity of suppressor tRNAs. *Biochimie* doi: 10.1016/j. biochip. 2012.12.010 (2012)

141. Chang C Liu, Lei Qi, Charles Yanofsky, Adam P Arkin: Regulation of transcription by unnatural amino acids. *Nat Biotechnol* 29, 164-168 (2011)

142. Jiangyun Wang, Wei Zhang, Wenjiao Song, Yizhong Wang, Zhipeng Yu, Jiasong Li, Minhao Wu, Lin Wang, Jianye Zang, Qing Lin: A biosynthetic route to photoclick chemistry on proteins. *J Am Chem Soc* 132, 14812-18 (2010)

143. Kaihang Wang, Wolfgang H. Schmied, Jason W Chin: Reprogramming the genetic code: from triplet to quadruplet codes. *Angew Chem Int Ed Engl* 51, 2288-2297 (2012)

144. S Lepthien, L Merkel, N Budisa: *In vivo* double and triple labeling of proteins using synthetic amino acids. *Angew Chem Int Ed Engl* 49, 5446-50 (2010)

145. Lars Merkel, Melina Schauer, Garabed Antranikian, Nediljko Budisa: Parallel incorporation of different fluorinated amino acids: on the way to "Teflon" proteins. Chembiochem 11, 1505-07 (2010)

146. Susan M Hancock, Rajendra Uprety, Alexander Deiters, Jason W Chin: Expanding the genetic code of yeast for incorporartion of diverse unnatural amino acids via a pyrrolysyl-tRNA synthetase/tRNA pair. J Am Chem Soc 132, 14819-24 (2010)

147. Sebastian Nehring, Nediljko Budisa, Birgit Wiltschi: Performance analysis of orthogonal pairs designed for an expanded eukaryotic genetic code. PLoS One 7(4): e31992 (2012)

148. Shixin Ye, Morgane Riou, Stephanie Carvalho, Pierre Paoletti: Expanding the genetic code in Xenopus laevis oocytes. *Chembiochem* 14, 230-235 (2013)

149. Bin Shen, Zheng Xiang, Barbara Miller, Gordon Louie, Wenyuan Wang, Joseph P Noel, Fred H Gage, Lei Wang: Genetically encoding unnatural amino acids in neural stem cells and optically reporting voltage-sensitive domain changes in differentiated neurons. *Stem Cells* 29, 1231-40 (2011)

150. Gabrielle Nina Thibodeaux, Xiang Liang, Kathryn Moncivais, Aiko Umeda, Oded Singer, Lital Alfonta, Zhiwen Jonathan Zhang: Transforming a pair of orthogonal tRNA-aminoacyl-tRNA synthetase from Archaea to function in mammalian cells. *PLoS One* 5(6):e11263 (2010)

151. Sebastian Greiss, Jason W Chin: Expanding the genetic code of an animal. *J Am Chem Soc* 133, 14196–14199 (2011)

152. Angela R Parish, Xingyu She, Zheng Xiang, Irene Coin, Zhouxin Shen, Stephen P Briggs, Andrew Dillin, Lei Wang: Expanding the genetic code of Caenorhabditis elegans using bacterial aminoacyl-tRNA synthetase/tRNA pairs. *ACS Chem Biol* 7, 1292-1302 (2012)

153. Philippe Marliere, Julien Patrouix, Volker Doring, Piet Herdewijn, Sabine Tricot, Stephane Cruveiller, Madeline Bouzon, Rupert Mutzel: Chemical evolution of a bacterium's genome. *Angew Chem International* 50, 7109-7114 (2011)

154. Philip Cohen: Life the Sequel. *New Scientist* 30 September, 167, 32-36 (2000)

155. Gene D McDonald, Michael C Storrie-Lombardi: Biochemical constraints in a protobiotic earth devoid of basic amino acids: the "BAA(-) world". *Astrobiology* 10, 989-1000 (2010)

156. Alan W Schwartz: Phosphorus in prebiotic chemistry. *Phil Trans R Soc B* 361, 1743-1749 (2006)

157. MW Powner, B Gerland, JD Sutherland: Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* 459, 239-242 (2009)

158. Ronald R Breaker: Riboswitches and the RNA world. Cold Spring Harb Perspect Biol 4, doi: 10.1101 /cshperspect.a003566

159. Rebecca M Turk, Nataliya V Chumachenko, Michael Yarus: Multiple translational products from a five-nucleotide ribozyme. *Proc Nat Acad Sci USA* 107, 4585-89 (2010)

160. Na Li, Faqing Huang: Ribozyme-catalyzed aminoacylation from CoA thioesters. *Biochemistry* 44, 4582-90 (2005)\

161. H Jakubowski: Amino acid selectivity in the aminoacylation of coenzyme A and RNA minihelices by aminoacyl-tRNA synthetases. *J Biol Chem* 275, 34845-48 (2000)

162. Gene H Hur, Christopher R Vickery, Michael D Burkart: Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat Prod Rep* 29, 1074-98 (2012)

163. P Bieling, M Beringer, S Adio, MV Rodnina: Peptide bond formation does not involve acid-base catalysis by ribosomal residues. *Nat Struct Mol Biol* 13, 423-428 (2006)

164. Antoine Danchin: Homeotropic transformation and the origin of translation. *Prog Biophys Mol Biol* 54, 81-86 (1989)

165. Michael Ibba, Alan W Curnow, Dieter Soll: Aminoacyl-tRNA synthesis: divergent routes to a common goal. *Trends Biochem Sci* 22, 39-42 (1997)

166. L Randau, S Schauer, S Ambrogelly, JC Salazar, J Moser, S Sekine, S Yokoyama, D Soll, D Jahn: tRNA recognition by glutamyl-tRNA reductase. *J Biol Chem* 279, 34931-7 (2004)

167. CS Franklyn, A Minajigi: tRNA as an active chemical scaffold for diverse chemical transformations. *FEBS Lett* 584, 366-375 (2010)

168. Zhiliang Yu, Ken Takai, Alexei Slesarev, Hong Xue, J Tze-Fei Wong: Search for primitive Methanopyrus based on genetic distance between Val- and Ile-tRNA synthetases. *J Mol Evol* 69, 386-394 (2009)

169. Robert Pascal, Laurent Boiteau: Enerygy flows, metabolism and translation. *Philos Trans R Soc Lond B Biol Sci* 366, 2949-58 (2011)

170. I Prigogine, G Nicolis: Biological order, structure and instabilities. *Q Rev Biophys* 4, 107-48 (1971)

171. Manfred Eigen, Peter Schuster: A principle of natural self-organization. *Naturwissenschaften* 64, 541-565 (1977)

172. Stuart A Kauffman: The origins of order: Self-organization and selection in evolution. Oxford University Press, Oxford (1993)

173. Helmut H Zepik, Eveline Blochliger, Pier Luigi Luisi: A chemical model of homeostasis. *Angew Chem Int Ed* 40, 199-202 (2001)

174. Zehavit Dadon, Nathaniel Wagner, Gonon Ashkenasy: The road to non-enzymatic molecular networks. *Angew Chem Int Ed* 47, 6128-36 (2008)

175. Jacques Ricard: Are biochemical networks possible ancestors of living systems? *C R Biol* 333, 769-778 (2010)

176. Nilesh Vaidya, Michael L Manapat, Irene A Chen, Ramon Xulvi-Brunet, Eric J Hayden, Niles Lehman: Spontaneous network formation among cooperative RNA replicators. *Nature* 491, 72-77 (2012)

177. Omer Markovitch, Doron Lancet: Excess mutual catalysis is required for effective evolvability. *Artif Life* 18, 243-266 (2012)

178. Addy Pross: How does biology emerge from chemistry? *Orig Life Evol Biosph* 42, 433-435 (2012)

179. Yan Qin, Norbert Polacek, Oliver Vesper, Eike Staub, Edda Einfeldt, Daniel N. Wilson, Knud H. Nierhaus: The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome. *Cell* 127, 721-733 (2006)

180. Marcelo Rogalski, Daniel Karcher, Ralph Bock: Superwobbling facilitates translation with reduced tRNA sets. *Nature Struct Mol Biol* 15, 192-198 (2008)

181. AJ Macario, E Conway de Macario: The molecular chaperon system and other anti-stress mechanisms in archaea. *Frontiers Biosc* 6, d263-283 (2001)

182. Stamatina Giannouli, Athanasios Kyritsis, Nikolaos Malissovas, Hubert Dominique Becker, Constantinos Stathopoulos: On the role of an unusual tRNAGly isoacceptor in Staphylococcus aureus. Biochime 91, 344-351 (2008)

183. Utz Kohlrausch, Joachim-Volker Höltje: One-step purification procedure for UDP-N-acetylmuramyl-peptide murein precursor from Bacillus cereus. *FEMS Microbiol Let*t 62, 253-257 (1991)

184. DW Smith, AL McNamara, M Rice, DL Hatfield: The effects of a post-transcriptional modification on the function of tRNALys isoaccepting species in translation. *J Biol Chem* 256, 10033-36 (1981)

185. Mark Odell, Verl Sriskanda, Stewart Shuman, Dimitar B Nikolov: Crystal structure of eukaryotic DNA ligase-adenylate illuminates the mechanism of nick sensing and strand joining. *Mol Cell* 6, 1183-93 (2000)

186. SM Hengel, DR Goodlett: A review of tandem mass spectrometry characterization of adenosine diphosphate-ribosylated proteins. *Int J Mass Spectrom* 312, 114-121 (2012)

187. James B Flanegan, Ralf F Pettersson, Victor Ambros, Martinez J. Hewlett, David Baltimore: Covalent linkage of a protein to a defined nucleotide sequence at the 5'-terminus of virion and replicative intermediate RNAs of poliovirus. *Proc Natl Acad Sci USA* 74, 961-965 (1977)

188. James J Champoux: DNA is linked to the rat liver DNA nicking-closing enzyme by a phosphodiester bond to tyrosine. *J Biol Chem* 256, 4805-9 (1981)

189. Monica J Roth, David R Brown, Jerard Hurwitz: Analysis of bacteriophage phi X174 gene A protein-mediated termination and reinitiation of phi X DNA synthesis. II. Structural characterization of the covalent phi X A protein-DNA complex. *J Biol Chem* 259, 10556-68 (1984)

190. Akhtar Samad, Robert B Carroll: The tumor suppressor p53 is bound to RNA by a stable covalent linkage. *Mol Cell Biol* 11, 1598–1606 (1991)

191. Alexei I Slesarev *et al*: The complete genome of hyperthermophilic Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proc Natl Acad Sci USA* 99, 4644-49 (2002)

192. Fabia U Battistuzzi, Andreia Feijao, S Blair Hedges: A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* 4, 44-57 (2004)

193. Massimo Di Giulio: A methanogen hosted the origin of the genetic code. *J Theoret Biol* 260, 77-82 (2009)

194. Marco Archetti, Massimo Di Giulio: The evolution of the genetic code took place in an anaerobic environment. *J Theoret Biol* 245, 169-174 (2007)

195. Karl O Stetter: Hyperthermophilic prokaryotes. *FEMS Microbiol Rev* 18, 149-158 (1996)

196. Massimo Di Giulio: The universal ancestor and the ancestor of bacteria were hyperthermophiles. *J Mol Evol* 57, 721-730 (2003)

197. Eric A Gaucher, Sridhar Govindarajan and Omjoy K. Ganesh: Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451, 704-708 (2008)

198. Mathieu Groussin, Manolo Gouy: Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea. Mol Biol Evol 28, 2661-74 (2011)

199. Massimo Di Giulio: A comparison of proteins of Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. *Gene* 346, 1-6 (2005)

200. Massimo Di Giulio: Structuring of the genetic code took place at acidic pH. *J Theoret Biol* 237, 219-226 (2005)

201. Harold S Bernhardt, Warren P Tate: Primordial soup or vinaigrette: did the RNA world evolve at acidic pH? *Biol Direct* 10.1186/1745-6150-7-4 (2012)

202. John A Leigh: Evolution of energy metabolism. In: Biodiversity of Microbial Life, eds. Staley JT & Reysenbach A-L. Wiley-Liss p. 103-120 (2001)

203. Paul G Falkowski: Tracing oxygen's imprint on Earth's metabolic evolution. *Science* 2006; 311, 1724-1725 (2006)

204. E.L. Shock: High-temperature life without photosynthesis as a model for Mars. *J Geophys Res* 102, 23687-94 (1997)

205. Ken Takai, Kentaro Nakamura: Archaeal diversity and community development in deep-sea hydrothermal vents. *Curr Opin Microbiol* 14, 282-291 (2011)

206. Feng-Jie Sun, Gustavo Caetano-Anollés: The evolutionary history of the structure of 5S ribosomal RNA. *J Mol Evol* 69, 430-443 (2009)

207. Feng-Jie Sun, Gustavo Caetano-Anollés: The ancient history of the structure of ribonuclease P and the early origins of Archaea. *BMC Bioinformatics* 11, 153 (2010)

208. Kyung Mo Kim, Gustavo Caetano-Anollés: The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other kingdoms. *BMC Bioinformatics* 12, 13 (2012)

Key Words: Metabolic Expansion Law, RNA World, Peptidated RNA World, Coevolution Theory, Last Universal Common Ancestor, synthetic life

Send correspondence to: Jeffrey Tze-Fei Wong, Division of Life Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, People's Republic of China, Tel: +852-23587288, Fax: +852-23581552, E-mail: bcjtw@ust.hk