Title

Understanding the Early Major Transitions in Evolutionary History Part 1: Stages in the Emergence of Complex Life

Authors and affiliations

Primary Author:

Aaron D. Goldman, e:agoldman@oberlin.edu, t:440-775-8749, Oberlin College, Department of Biology, Blue Marble Space Institute of Science, Prebiotic Chemistry and Early Earth Environments Consortium

Co-authors:

Greg Fournier, e:g4nier@mit.edu, t:617-324-6164, Massachusetts Institute of Technology, Department of Earth, Atmospheric, and Planetary Science

Johann Peter Gogarten, e:gogarten@uconn.edu, t:860-465-6267, University of Connecticut, Department of Molecular and Cell Biology

Anton S. Petrov, e:anton.petrov@biology.gatech.edu, t:G404-894-8338, Georgia Institute of Technology, School of Chemistry and Biochemistry, Center for the Origins of Life

Lynn Rothschild, e: lynn.j.rothschild@nasa.gov, t:650-604-6525, NASA Ames Research Center, Space Science and Astrobiology Division

Daniel Segrè, e:dsegre@bu.edu, T:617-358-2301, Boston University, Department of Biology, Bioinformatics Program, Biological Design Center, Microbiome Initiative

Eric Smith, e:dsmith470@gatech.edu, t:720-364-6394, Georgia Institute of Technology, School of Biological Sciences, Center for the Origins of Life

Loren Williams, e:loren.williams@chemistry.gatech.edu, T:(404) 385-6258, Georgia Institute of Technology, School of Chemistry and Biochemistry, Center for the Origins of Life, Prebiotic Chemistry and Early Earth Environments Consortium

Recommendation We recommend that early evolution research, the excavation of the historical record of biology, continue to be a high priority of the Astrobiology Program because of its role in understanding the origin and diversity of life on Earth, its importance in guiding prebiotic chemistry research, and contribution toward understanding the ability of life to adapt to diverse environments. Early evolution research is key to, and must be integrated into the design of, life detection strategies. Finally, there remain keys to alternate trajectories that life could have taken in early evolved clades, providing insights into the possibility of other life forms elsewhere.

Motivation All extant life on Earth shares a common biochemistry based on a relatively small set of organic molecules (Kluyver and Donker, 1926), the same mechanisms of information storage and inheritance (Woese, 1965), RNA, DNA and protein, enzymatic cofactors and ATP energy currency, a dependence on water, a related cellular organization, and a handful of core metabolic pathways with reactions performed by proteins with a shared ancestry. These central features of life as we know it all evolved in the period between the origin of life and the last universal common ancestor (LUCA) of all extant life. Understanding their evolutionary history has an important role in astrobiology research, and has increasing potential thanks to improved research strategies and cross-disciplinary collaborations, as well as advances in molecular and cellular biology, increasing power of computational resources, and breadth of bioinformatics databases.

Extant biology contains detailed and interpretable records of pre-biological processes and molecules and of early biology. The early history of life on Earth provided by these "top-down" approaches provides unique information and important perspectives for NASA's astrobiology and exobiology programs. These approaches have revealed extensive information on the geochemistry of Earth, chemical evolution, pre-LUCA biological phenomena and early evolution of Bacteria, Archaea and Eukarya. Top-down approaches guide bottom-up approaches, and directly inform astrobiological questions such as planetary conditions for origins and habitability.

Understanding how and under what circumstances life increased in complexity, became capable of populating new environments, and altering the planet and its atmosphere, can inform our expectations of when a planet or moon may be inhabited. Furthermore, in the search for life elsewhere, we only have one biosphere to inform our search for extraterrestrial life. That biosphere was very different in early evolutionary history. Thus, understanding the early biosphere gives us an extended set of features which may be informative of biological processes across a broader time window; currently, we cannot distinguish universal features of life in the universe from those that are quirks of evolutionary history as it occurred on Earth. But this distinction is essential for designing sufficiently broad life-detection methods. Below are several examples of current research areas and future directions within the field of early evolution that center on the identification of ancient proteins and proteomes, the functions they performed in ancient life, and their implications for ancient organismal ancestors.

Early Evolution of Translation *Can we use phylogenetic, structural and biophysical information to resurrect the decoding-competent LUCA ribosome and more primitive pre-coding ribosomes?*

The translation system is an ancient molecular fossil that provides a telescope to the origins of life. Made from RNA, protein and inorganic cations, the ribosome translates mRNA to coded protein in all living systems. Universality, economy, centrality and antiquity are ingrained in translation, which dominates the set of genes that are universally shared as orthologues across the tree of life. In fact, the lineage of the translation system defines the universal tree of life.

The ribosome as a nursery of protein evolution. Can we infer the role of analogy and homology in the fold repertoire of the translation system at and before LUCA; can we infer the basic principle of the sequence regularities, required to form the simplest folds; can we estimate the amount of information required to code for them; can we propose a model for a simple translation system, which is capable of this task? Our understanding of the evolution of the ribosome, and inspection of universal ribosomal proteins (rProteins) reveals a reaction coordinate for the evolution of protein folding that appears imprinted within the ribosome with ribosomal rRNA acting as a molecular cast. This reaction coordinate within the ribosome suggests that the maturation of condensation chemistry initially resulted in the production of randomly coiled oligopeptides, followed later by a selection for oligomers capable of forming of β-hairpins. In the next step polymers were selected based on formation of globular β-domains. Globular proteins required at least primitive coding, participation of tRNA and the SSU (Kovacs, 2017). Folding progressed to thermodynamically stable β-structures, and then to kinetically trapped α-structures. The latter were enabled by a long, mature exit tunnel that partially offset the general thermodynamic tendency of all polypeptides to form β-sheets (Bowman JC, 2020).

Extant proteins utilize a small fraction of the combinatorial sequence space available for 20 proteinogenic amino acids. The emergence of the most ancient domains must have involved an exploration of sequence space that allowed the evolution of proteins with incremental folding ability. The ribosome (within its own proteins) contains a structural record of the earliest history of protein folding. At least four major biological systems must have taken shape roughly contemporaneously within this pre-LUCA era: the ribosome, the ancestral aminoacyl-tRNA synthetases, the genetic code, and metabolism coupled to bioenergetics. Looking forward, we ask whether a causal interdependency among these four origins of order can be found.

Co-evolution of rProteins and tRNA synthetase: Can we identify correlations between evolution of the translational machinery and tRNA synthetases? Can we develop a synchronized evolutionary model for both coding and decoding and map it onto a single timeline? A small number of protein families have a detectable homology tracing back before LUCA, through the presence of related protein families (paralogs) that evolved by gene duplication in the pre-LUCA lineage. These pre-LUCA paralogs are especially valuable for Astrobiology and Origins research, as they permit comparative evolutionary biology and phylogenetics techniques to push beyond the “event horizon” of species-tree coalescence at LUCA. Some of the most valuable of these protein families are aminoacyl-tRNA synthetases (aaRS). These proteins recognize cognate tRNA and aminoacylate them with the correct amino acid, to be used in translation and polypeptide synthesis in the ribosome. aaRS come in two major, unrelated classes (Class I and

Class II), each roughly having 10 constituent protein families, each for a specific amino acid-tRNA pairing. While the sequence and structure of protein families within each class varies considerably, elements of sequence and structure homology are detectably conserved, and can provide the basis for comparative evolutionary studies, and reconstructing the pre-LUCA history of aminoacylation [O'Donoghue & Luthey-Schulten, 2003]. Current and past work on reconstructing pre-LUCA aaRS history has provided novel insights into this very early period in life's evolution. For example, the observation that aaRS that are cognate for similar amino acids tend to be more closely related suggests that the early stages of aminoacylation evolution may be tied to aaRS diversification and sub-functionalization. Reconstructing the likely sequence composition of pre-LUCA ancestors also informs genetic code evolution. Previous work suggests that, by the time the earliest Class I aaRS proteins were diverging, their sequences were already making use of all 20 amino acids, providing strong evidence of an origin of the genetic code before synthetases had their modern specificities (Founier et al., 2011). In the next decade, with further improvements in ancestral sequence reconstruction algorithms and experimental methods, pre-LUCA aaRS ancestors may yield even more secrets. The earliest tRNA-aaRS interaction dynamics may be elucidated, which, subsequently, could inform the origin of tRNA modifications, code specificity, and codon partitioning. Complementary studies on evolving ribozymes capable of tRNA aminoacylation could provide insight into how and why proteins took over this critical role in the cell.

Origin of the genetic code: Can we develop a comprehensive theory of the genetic code that accounts for all evolutionary aspects of the translation system, charging system, and protein folding described above? Numerous speculations on the origins and evolution of the genetic code have been advanced. There is evidence that the code, along with selection of biopolymer building blocks and linkage chemistries, are products of evolution. Amino acids appear to be assigned to codons so as to minimize phenotypic effects of mutation and mistranslation. Random sets of amino acids that cover chemical space better than the proteinaceous amino acids are rare and energetically costly. The causes behind the origins and evolution of the genetic code will become increasingly clear as a molecular reconstruction of more ancient stages of the translation system is accomplished. From the reconstructed history of the molecular components (including tRNA synthetases), factors contributing to the genetic code and details of the codon assignments can be deciphered.

The observation that the genetic code was essentially complete by the time of LUCA, including the "modern" machinery for its syntax and implementation, is in itself a starting point for important theoretical exploration of origins and habitability. Regardless of the later evolutionary elaborations across the tree of life, the universal genetic code was entirely sufficient to provide the "firmware" for all subsequent protein evolution, from photosystems to feathers. Why should this be the case? Potential answers all suggest profound astrobiological implications, for both the likelihood of protein-based life emerging, and its subsequent survival and diversification into potentially detectable biospheres. Is this evidence of an extensive period

of pre-LUCA evolution, where the code and amino acid alphabet evolved through both positive and purifying selection to be a generalized, universal "solution machine" for evolution? Or are many such amino acid alphabets that life may discover "sufficient" to provide generalized evolutionary solutions? In the next decade, further theoretical and experimental investigations into alternative amino acid libraries and codes will help to resolve these questions.

Phylogenetic analysis of gene families prior to the LUCA The study of molecular evolution is not limited to tracing back organismal lineages. Gene duplications, symbiosis, and gene transfer between divergent organisms traditionally were seen as problems in applying the molecular evolutionary history to reconstructing organismal evolution. While this is true to some extent, the study of ancient gene duplications and gene transfer events allows us to reconstruct the evolutionary history that preceded LUCA. The study of pre-LUCA aaRS ancestors above represents a prime example of the promise of this avenue of research. In the next decade, recently developed techniques for characterizing these pre-LUCA ancestral proteins (e.g. Founier and Alm 2015) will lead to a better understanding of life at this very early stage. The potential for greater taxonomic representation will make it possible to discover new protein families that diverged prior to the LUCA (e.g. Harris and Goldman, 2018). And, most compellingly, better methods for accurate phylogenetic reconstruction of very ancient protein families may allow the study of pre-LUCA protein families to be extended to protein domains, which would open up an entirely new window into the evolutionary history of life before the LUCA.

Characteristics of the LUCA The root of the tree of life identifies the ancestral, organismal point of coalescence of all known life on Earth. However, little is known about this last universal common ancestor (LUCA) or its relation to the Origin of Life (OoL). Comparative genomics and cell biology suggest that the organism(s) represented by LUCA were likely cellular and contained many genes, proteins, and biological functions present within modern lineages (e.g. Becerra et al., 2007). However, this minimalist reconstruction lacks the power to discriminate between competing hypotheses about prebiotic chemistry and the OoL, the relative timing of LUCA, the earliest metabolisms, the origin of cellularity and Darwinian selection, and other issues critical for placing the emergence of life on Earth within an astrobiological context.

Further biological investigation of LUCA will require advances in paleogenomics and molecular evolutionary biology as a complement to ongoing theoretical and experimental research in geochemistry, organic chemistry, and planetary science. Recent increases in the growth of bioinformatics databases, the sophistication of evolutionary and ecological models, the availability of high performance computing resources, and the emerging capabilities of synthetic biology, permit new investigations of life at this stage of evolution at greater detail and with greater confidence than has been previously attempted or even considered possible.

Every model of life's origins proposed so far involves a prebiotic chemical system that achieves a level of complexity permitting self-replication. The transition to organismal individuality and vertical inheritance (Woese, 2002), that is, the "Darwinian threshold", is not

well understood, but may have been a pre-requisite for the subsequent evolution of ecological complexity and the metabolic diversity from which Earth's biosphere is constructed. This transition, the timing and nature of LUCA, and the emergence of the first branches on the tree of life can, therefore, be understood in mutual context. In this way, a greater understanding of life at the root of the evolutionary tree will inform our expectations for life throughout the universe. Despite serious progress over the last two decades in understanding the genome, metabolism, and molecular physiology of the LUCA, many important questions remain.

Does the root of the tree of life represent a single individual, a species, or a population of species? It is not yet known whether the root of the tree represents a single organism or many, and if the latter, whether or not these organisms lived in a single environment at the same time. This issue may be informed by ecological modeling and resolved through greater scrutiny of the phylogenies of individual gene families present in LUCA. A popular evolutionary model for the earliest life is that it was dominated by horizontal gene transfer so that individual lineages cannot be distinguished from one another. A transition from this regime to one of vertical inheritance and the establishment of distinct lineages would, in retrospect, result in the appearance of a singular common ancestor (LUCA). Alternatively, this transition could have occurred earlier, with vertical inheritance, speciation, and distinct microbial lineages existing before LUCA. In this schema, pre-LUCA lineages could still have acquired many genes via horizontal gene transfer, as organismal lineages still do today. Evolutionary modeling may help explain why the transition toward organismal individuality and vertical inheritance became predominant, and whether we should expect a similar transition for other forms of life elsewhere in the universe. Further, the reliance on specific geochemistries and other environmental factors (e.g., solar spectrum) must be explored if these studies are to be applied to the search for life elsewhere.

Which gene families had evolved by the time of the root of the evolutionary tree? In what order did these earliest functions emerge? A minimal LUCA genome can and has been created by comparing the gene complements of extant species across the evolutionary tree (Goldman *et al.,* 2013*)*. In some cases, the emergence of certain gene families prior to LUCA can be ordered. Further refinement of these techniques is required to achieve higher confidence predictions of LUCA's genome and pre-LUCA gene evolution, and to possibly infer other biological properties that may not have endured to be present in extant lineages. The growth of metabolic databases and metabolic pathway prediction based on organismal gene complements permits the prediction of metabolic pathways in extant organisms. This same approach is now being applied to LUCA gene complements, which should allow researchers in the near future to constrain possible biological environment(s) for life at the root of the evolutionary tree.

What can the LUCA tell us about the emergence of core biological systems? Evidence for a DNA genome in LUCA is present, but sparse. The evolutionary history of DNA-related proteins can be understood in tandem with better estimates of LUCA's gene complement. Evidence of the relative ages of RNA and DNA-associated proteins may directly bear on the hypothesis of prebiotic evolution in an "RNA world", and whether or not a transition to DNA is

predicted to occur before or after the emergence of the first living individual cells. Similar to the DNA genome, the nature of cellular membranes at the root of the tree of life remains unclear. Important membrane-related protein families were present in LUCA, including those responsible for inserting other proteins into the membrane. The ability of LUCA to use the membrane as a controlled boundary may reveal characteristics pertaining to organismal individuality and its relationship to the environment and other organisms within it. Chemical differences between the membranes within the major Domains of life pose an additional challenge in reconstructing the nature of the earliest cellular membranes, and may provide clues to the mechanisms and causes of early physiological diversification.

What was the geological and evolutionary context of the LUCA's emergence? It is unknown if LUCA existed during the Hadean, or later within the Archaean eon. The answer to this question is important for determining the kind of environmental selection that may have shaped early life, before, during, and after its initial diversification, and thus is relevant to extraterrestrial life detection. With respect to evolutionary history, the interval of time and evolutionary change between the first living cells and LUCA remains largely mysterious (Cantine & Fournier, 2018). The answer to this question is of profound astrobiological importance, as it directly impacts our estimate for the probability of life arising on planets with suitable initial conditions. Information about protein families inferred to be present at the time of LUCA may yield clues to geological context, for example the nature of the early metalloproteome, or the use of cofactors that mimic geochemical catalysts.

Is there any genomic evidence for interplanetary panspermia (e.g., Mars to Earth) at this early time? Transfer of cellular material between the planets of the inner solar system via impact ejecta is a statistical near-certainty over long periods of geological time, and would be even more likely early in Earth's history when impact rates were far higher. Therefore, the possibility that extant life on Earth originated on Mars cannot be reasonably excluded. It is possible that adaptations and selections consistent with an exoplanetary origin may be preserved and detectable within reconstructions of the LUCA or pre-LUCA genome.

From FECA to LECA The subsequent evolution of complex life requires the emergence of higher orders of cellular organization that expand life's ability to form cooperative structures, from colonies, to bodies with integrated tissues. The evolution of the eukaryotic cell makes all of this possible. The Last Eukaryotic Common Ancestor (LECA) is connected to its prokaryotic ancestors by a long branch in most molecular phylogenies, along which tremendous evolutionary changes took place (Poole & Penny, 2007). The last common ancestor of this lineage and its closest prokaroytic relative has become known as the First Eukaryotic Common ancestor (FECA). The recent discovery of an archaeal group from sequencing of environmental metagenomes, provides key information about what types of organisms may have been closely related to LECA (Spang et al., 2018). However, the evolutionary history of the eukaryal stem lineage is long, and remains almost entirely cryptic. The wealth of molecular data provided

through metagenomics and single cell genome sequencing (through in vitro amplification without the need to cultivate the cells to be sequenced), may permit the reconstruction of this transition in unprecedented detail. In particular, the roles of environment, symbiosis (the mitochondrial endosymbiont), cytological complexification, gene duplications and gene transfers will continue to illuminate this pivotal transition. These studies will also assess the roles of the ecological setting, molecular parasites, and constructive neutral evolution in this transition, and thereby provide an understanding of evolution that integrates selection acting at different levels of evolution. From an astrobiological perspective, the origin of eukaryotes may-or may not- be a more path-dependent, historical, and unlikely event than the origin of life itself; understanding the circumstances that led to these evolutionary events therefore may be one of the most important questions in exploring how rare or common complex life may be in the universe.

Conclusions Understanding early evolutionary history and the major evolutionary transitions that shaped our biosphere has been an important research goal of NASA's Astrobiology Program and the Exobiology Program that preceded it. It must continue to play an indispensable role in origin of life research and a more prominent role in the design of life detection strategies, which together constitute the majority of astrobiology research overall. The next decade will likely see major advances due to a number of recent innovations. These include, but are not limited to, a better understanding of best practices in determining whether a protein family is ancient, better phylogenetic methods that permit highly accurate trees of very old protein families as well as accurate reconstructions of ancestral protein sequences along those trees, and innovative approaches developed in the previous decade that will lead to a deeper understanding of those ancestral proteins and the molecular functions they performed. Early evolution research in the United States has primarily been funded through NASA's Astrobiology Program and it must continue to be, not only for its own sake, but for the benefit of all areas of astrobiology research.

References

Becerra A, et al. (2007) Annu Rev Ecol Evol. Syst. 38:361-379
Bowman JC, et al. (2020) Chem Rev 120: 4848–78.
Cantine MD & Fournier GP. (2018) Orig Life Evol Biosph. 48:35-54.
Fournier GP & Alm EJ (2015) J Mol Evol, 80:171- 185.
Goldman AD, et al. (2013) Nucleic Acids Res. 41:D1079-82.
Harris AJ & Goldman AD (2018) J Mol Evol. 86:277-282.
Kluyver AJ & Donker HJL (1926) Chem Zelle Gewebe, 13:134–190
Kovacs NA, et al. (2017) Mol. Biol. Evol. 34:1252-1260.
O'Donoghue P & Luthey-Schulten Z. (2003) Microbial Mol Biol Rev. 67(4):550-73.
Poole AM & Penny D. (2007) Bioessays. 29(1):74-84.
Spang Aet al. (2018) PLoS Genet. 14(3):e1007080.
Woese CR (1965) Proc Natl Acad Sci USA, 54:1546–1552
Woese CR (2002) Proc Natl Acad Sci USA. 99:8742-7.