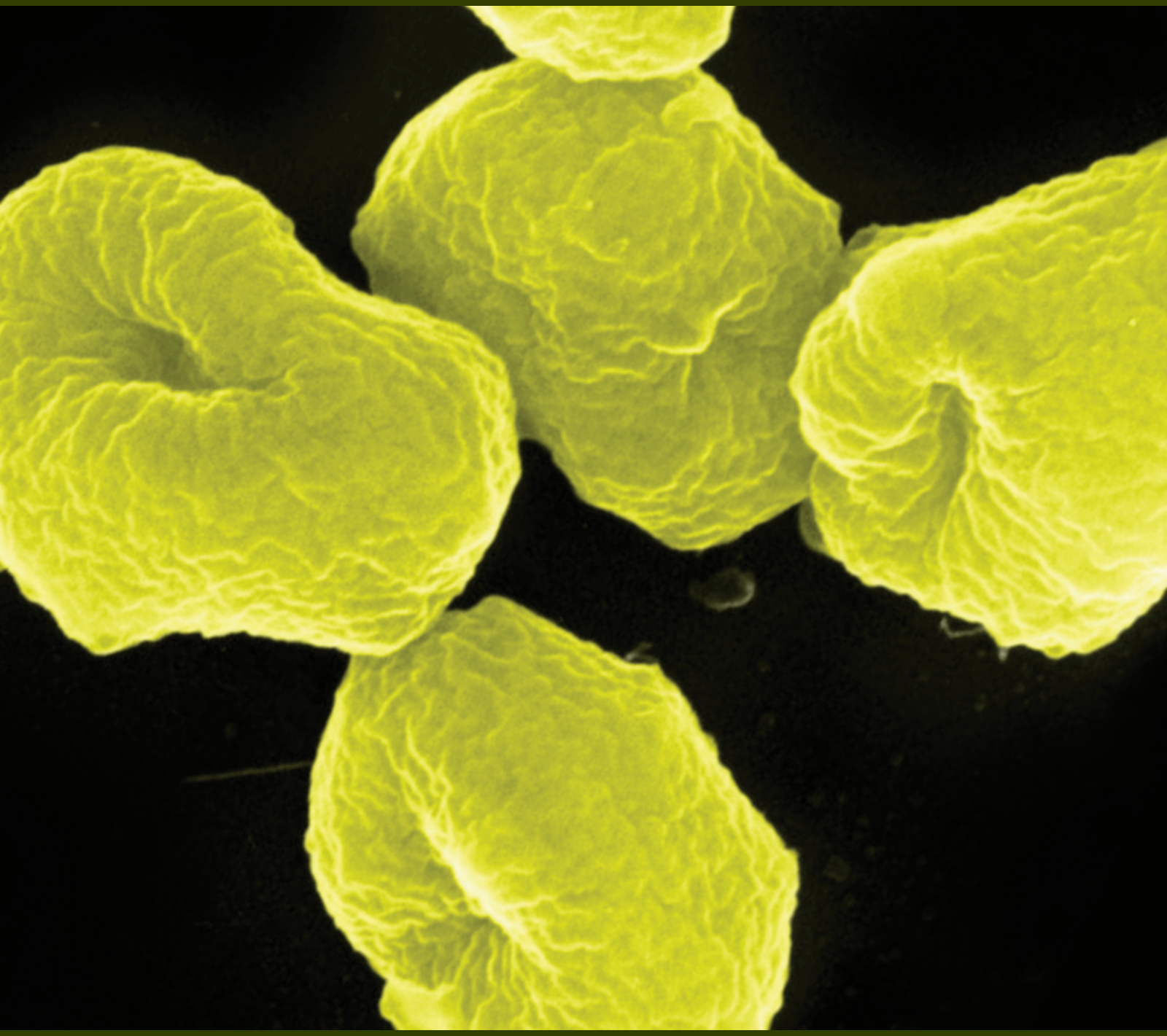
Archaea

The Origin and Evolution of the Archaeal Domain

Guest Editors: Gustavo Caetano-Anollés and Kyung Mo Kim



The Origin and Evolution of the Archaeal Domain

Archaea

The Origin and Evolution of the Archaeal Domain

Guest Editors: Gustavo Caetano-Anollés and Kyung Mo Kim

Copyright © 2014 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "Archaea." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editorial Board

Maqsudul Alam, USA
Sonja-Verena Albers, Germany
Ricardo Amils, Spain
Haruyuki Atomi, Japan
Nils K. Birkeland, Norway
Paul H. Blum, USA
Elizaveta A. B.-O., Russia
Mara J. Bonete, Spain
Giovanna Cacciapuoti, Italy
Isaac K. O. Cann, USA
Danil Charlier, Belgium
Uwe Deppenmeier, Germany
Nejat Dzgnes, USA
Jerry Eichler, Israel
Harald Engelhardt, Germany
Michael W. Friedrich, Germany
Toshiaki Fukui, Japan
Roger Garrett, Denmark
Dennis W. Grogan, USA
Robert P. Gunsalus, USA
Reinhard Hensel, Germany
Li Huang, China
Michael Ibba, USA
Yoshizumi Ishino, Japan
Toshio Iwasaki, Japan
Zvi Kelman, USA
Hans-Peter Klenk, Germany
Paola Londei, Italy
Peter A. Lund, UK
Giuseppe Manco, Italy
William W. Metcalf, USA
Marco Moracci, Italy
Masaaki Morikawa, Japan
Volker Mueller, Germany
Biswarup Mukhopadhyay, USA
Katsuhiko Murakami, USA
Alla Nozhevnikova, Russia
Francesca M. Pisani, Italy
Marina Porcelli, Italy
David Prangishvili, France
Reinhard Rachel, Germany
Anna-Louise Reysenbach, USA
Frank T. Robb, USA
Francisco Rodriguez-Valera, Spain
Roberto Scandurra, Italy
Kevin R. Sowers, USA
Stefan Spring, Germany
Michael Thomm, Germany
Herman van Tilbeurgh, France
Antonio Ventosa, Spain
William B. Whitman, USA
Masafumi Yohda, Japan
Chuanlun Zhang, USA
Christian Zwieb, USA
Servé W. M. Kengen, The Netherlands

Contents

Hindawi Publishing Corporation
Archaea
Volume 2014, Article ID 915828, 2 pages
http://dx.doi.org/10.1155/2014/915828

Editorial

The Origin and Evolution of the Archaeal Domain

Gustavo Caetano-Anollés[1] and Kyung Mo Kim[2]

[1] *Evolutionary Bioinformatics Laboratory, University of Illinois, Urbana, IL 61801, USA*
[2] *Microbial Resource Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Republic of Korea*

Correspondence should be addressed to Gustavo Caetano-Anollés; gca@illinois.edu and Kyung Mo Kim; ksnuo2@gmail.com

Received 7 May 2014; Accepted 7 May 2014; Published 4 June 2014

Copyright © 2014 G. Caetano-Anollés and K. M. Kim. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With this special issue on the origin and evolution of Archaea we honor and celebrate the life and impactful contributions of Carl Woese (July 15, 1928-December 30, 2012). Carl was born and raised in Syracuse, New York. His undergraduate studies were in Amherst College and his graduate training in Yale. Sol Spiegelman brought him to the University of Illinois at Urbana-Champaign, where he unfolded a brilliant career. Carl was inspired by the originality of his mentor, Ernest C. Pollard, the tradition of biological form of D'arcy Thompson, the charisma of Francis Crick, the evolutionary tempo-mode perspective of G. G. Simpson, and the foresight of Darwin and Wallace. He understood the centrality of evolution in our understanding of biology and championed this perspective as he explored the molecular makeup of the translation machinery. His insightful mind is responsible for the discovery of the archaeal domain and for transforming comparative views of microbial diversity into an overarching evolutionary framework. Archaea constitutes the third domain of life, a remarkable group of akaryotic microbes with unique biochemical and genomic features, some of which resemble those of eukaryotes. Their habitats and lifestyles are very diverse, from extremophiles living in harsh environments to soil and marine mesophiles, from free-living microbes to gut-inhabiting methanogens and symbionts.

Carl's work did not only result in the definition of a new "urkingdom," originally named by him as "archaebacteria," but his insights prompted an appreciation (and respect) for the incredible microbial diversity of the biosphere. He battled the establishment to make way to a redefined microbiological science that treasured evolutionary thinking and acknowledged the centrality of microbes in the global ecosystems of our planet. He was also a harsh critic of the field of biology in general. He felt our biological views are still governed by reductionistic biases inherited from the genetic and genomic revolutions of last century, which could not identify any important questions left to answer. Furthermore, he strongly felt the biological sciences were devoted and defined by the application side, that is, by focusing on providing "service to society" through bioengineering instead of acting as "society's teacher" of man's place in the universe. A number of unsolved problems that are central to understanding life remain to be answered, and Carl posed some of the basic questions from the very beginning. What were life's origins? How did molecular and organismal complexity unfold? What are the ultimate governing principles of life? He recognized the limitations of the primacy of a genetic, molecular biology and mechanistic outlook that was gene-centered and prompted an exploration of biological complexity and emergence of biological organization within an evolutionary and physics framework. He recognized the importance of the proteinaceous backbone of life and how its design and function is delimited by the genetic code, translation, and its complex regulatory control.

In this special issue we bring back some of Carl's basic unanswered questions. While it is becoming clear that the archaeal domain may have an independent evolutionary history, its origin and links to the other two domains of cellular complexity remain contentious, as well as its placement in the tree of life. The question demands urgent attention. Several contributions of this special issue tackle important aspects of the origin, diversity, and evolution of the archaeal domain.

A review article by A. Spang et al. comprehensively describes current hypotheses on the relationships of the three domains and evaluates archaeal diversity and evolution

using recent genomic data (e.g., metagenomes and single-cell genomes). P. Forterre also evaluates the contemporary scenarios for the origins of the three domains. Archaeal ancestor scenarios and the fusion hypothesis are criticized. Interestingly, he brings the evolutionary role of the virosphere to explain the diversification of the three domains from the last universal common ancestor of life. A. Nasir and G. Caetano-Anollés explore a novel comparative genomic framework that makes the vertical horizontal evolutionary contributions explicit, and G. Caetano-Anollés et al. advance structural phylogenomic analyses of protein and nucleic acid structures and their associated functions. These approaches reveal that Archaea is the most ancient domain, which prompts a careful reevaluation of current phylogenetic methodologies and our understanding of the rooting of the tree of life.

D. S. Shin et al. review the robustness of archaeal proteins against extremophilic environments at the protein 3-dimensional structural level. C. J. Reed et al. describe how archaeal species can be adapted into thermophilic, psychrophilic, piezophilic, and halophilic environments by characterizing the biophysical property of archaeal proteins. Both studies emphasize the importance of archaeal structural biology for understanding human biology with medical and industrial impacts.

G. Borrel et al. present a bioinformatics analysis of three genomes from a newly identified order of methanogens and find the pyrrolysine (22nd amino acid) coding system. The phylogenetic analysis indicates that this genomic feature is conserved in both archaeal methanogens and bacteria, which can be an example of continuing evolution of the genetic code directed by metabolic requirements. On another front, L. S. Yafremava et al. study amino acid substitution patterns in the protein domains of nonbarophilic and barophilic *Pyrococcus* species and reveal that barophily is a very ancient trait that unfolded with the early evolution of the genetic code during early adaptation to deep ocean environments.

J. R. Peterson et al. use many different state-of-the-art approaches (e.g., SiMPull and RNA-Seq) to quantitatively characterize the methanogenesis pathways and translational machinery of the methanogen *Methanosarcina acetivorans*. This bioinformatics modeling can be a first step to establish new archaeal model systems, very much as *E. coli* is used for bacteria.

Taken together, articles highlight patterns and processes responsible for archaeal diversity at genetic, genomic, biochemical, physiological, and ecological levels. It is our intention that the work presented here will stimulate further evolutionary thinking, following Carl's pioneering and unorthodox spirit.

Gustavo Caetano-Anollés
Kyung Mo Kim

Hindawi Publishing Corporation
Archaea
Volume 2014, Article ID 590214, 26 pages
http://dx.doi.org/10.1155/2014/590214

Review Article

Archaea: The First Domain of Diversified Life

Gustavo Caetano-Anollés,[1] Arshan Nasir,[1] Kaiyue Zhou,[1] Derek Caetano-Anollés,[1] Jay E. Mittenthal,[1] Feng-Jie Sun,[2] and Kyung Mo Kim[3]

[1] *Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, Institute for Genomic Biology and Illinois Informatics Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[2] *School of Science and Technology, Georgia Gwinnett College, Lawrenceville, GA 30043, USA*
[3] *Microbial Resource Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Republic of Korea*

Correspondence should be addressed to Gustavo Caetano-Anollés; gca@illinois.edu

Received 30 September 2013; Revised 15 February 2014; Accepted 25 March 2014; Published 2 June 2014

Academic Editor: Celine Brochier-Armanet

Copyright © 2014 Gustavo Caetano-Anollés et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The study of the origin of diversified life has been plagued by technical and conceptual difficulties, controversy, and apriorism. It is now popularly accepted that the universal tree of life is rooted in the akaryotes and that Archaea and Eukarya are sister groups to each other. However, evolutionary studies have overwhelmingly focused on nucleic acid and protein sequences, which partially fulfill only two of the three main steps of phylogenetic analysis, formulation of realistic evolutionary models, and optimization of tree reconstruction. In the absence of character polarization, that is, the ability to identify ancestral and derived character states, any statement about the rooting of the tree of life should be considered suspect. Here we show that macromolecular structure and a new phylogenetic framework of analysis that focuses on the parts of biological systems instead of the whole provide both deep and reliable phylogenetic signal and enable us to put forth hypotheses of origin. We review over a decade of phylogenomic studies, which mine information in a genomic census of millions of encoded proteins and RNAs. We show how the use of process models of molecular accumulation that comply with Weston's generality criterion supports a consistent phylogenomic scenario in which the origin of diversified life can be traced back to the early history of Archaea.

1. Introduction

"*Imagine a child playing in a woodland stream, poking a stick into an eddy in the flowing current, thereby disrupting it. But the eddy quickly reforms. The child disperses it again. Again it reforms, and the fascinating game goes on. There you have it! Organisms are resilient patterns in a turbulent flow—patterns in an energy flow*"— Carl Woese [1].

Understanding the origin of diversified life is a challenging proposition. It involves the use of ideographic thinking that is historical and retrodictive, as opposed to nomothetic explorations that are universal and predictive [2]. Experimental science for the most part is nomothetic; the search for truth comes from universal statements that can be conceptualized as being of general predictive utility. Nomothetic explorations are in general both philosophically and operationally less complex to pursue than any ideographic exploration. In contrast, retrodictions speak about singular or plural events in history that must be formalized by "transformations" that comply with a number of evolutionary axioms [3] and interface with a framework of maximization of explanatory power [4]. The fundamental statement that organismal diversity is the product of evolution is supported by an ensemble of three nested primary axioms of the highest level of universality [3]: (i) evolution occurs, including its principle that history of change entails spatiotemporal continuity (*sensu* Leibnitz), (ii) only one historical account of all living or extinct entities of life and their component parts exists as a consequence of descent with modification, and (iii) features of those entities (characters) are preserved through generations via genealogical descent. History must comply with the "principle of continuity," which crucially supports evolutionary thinking. The axiomatic rationale of "*natura non facit saltum*" highlighted by Leibnitz, Linnaeus, and Newton must be considered a generality of how natural things change and a "fruitful principle of discovery." We note

that this axiomatic generality, which we have discussed in the context of origin of life research [5], encompasses rare punctuations (e.g., quantum leap changes such as genome duplications and rearrangements or the rare evolutionary appearance of new fold structures) embedded in a fabric of gradual change (e.g., changes induced by point mutations). Both gradual changes and punctuations are interlinked and are always expressed within spatiotemporal continuity (e.g., structural punctuations in the mappings of sequences into structures of RNA [6]). This interpretative framework can explain novelty and complexity with principles of scientific inquiry that maximize the explanatory power of assertions about retrodictions.

Phylogenetic theories are embodied in evolutionary "trees" and "models." Trees (phylogenies) are multidimensional statements of relationship of the entities that are studied (phylogenetic taxa). Models are evolutionary transformations of the biological attributes examined in data (phylogenetic characters), which define the relationships of taxa in trees. The tripartite interaction between characters, models, and trees must occur in ways that enhance retrodictive power through test and corroboration [4]. In other words, it must follow the Popperian pillars of scientific inquiry or suitable philosophical analogs. We note that retrodictive statements allow drawing inferences about the past by using information that is extant (i.e., that we can access today) and is necessarily modern. The challenge of travelling back in time rests on not only making inferences about archaic biology with information drawn from modern biological systems but also interpreting the retrodictive statements without conceptual restrictions imposed by modernity. This has been an important obstacle to historical understanding, starting with grading hypotheses inspired by Aristotle's great chain of being, the *scala naturae*.

It was Willi Hennig in the fifties who first formalized retrodiction in quantitative terms [7]. Since then, his "phylogenetic systematics" has benefitted from numerous conceptual and bioinformatics developments, which are now responsible for modern phylogenetic analysis of systems of any kind: from molecules and organisms to language and culture, from engineering applications to astrophysics. Astrocladistics, for example, focuses on the evolution and diversification of galaxies caused by transforming events such as accretion, interaction, and mergers (e.g., [8]). While major views have emerged in the "discovery operations" (*sensu* [2]) of the phylogenetic systematics paradigm, including maximum parsimony and the frequentist and uncertainty views of maximum likelihood and Bayesian thinking, the major technical and philosophical challenges persist [9]. More importantly, as we will explain below, technical and philosophical aspects of the ideographic framework in some cases have been turned into landscapes of authoritarianism and apriorism [3]. This insidious trend is pervasive in the "rooting of the tree of life" field of inquiry [10] that underlies the origins of biochemistry and biodiversity we here discuss.

In this opinion paper we address the challenges of finding an origin to biodiversity and propose a new framework for deep phylogenetic analysis that focuses on the parts of biological systems instead of the whole. We review the application of this framework to data drawn from structural and functional genomics and argue that the origin of cellular life involved gradual accretion of molecular interactions and the rise of hierarchical and modular structure. We discuss our findings, which provide strong support to the very early rise of primordial archaeal lineages and the emergence of Archaea as the first domain of diversified cellular life (superkingdom). The term "domain" of life stresses the cohesiveness of the organism supergroup, very much like domains in proteins and nucleic acids stress the molecular cohesiveness of their atomic makeup. Instead, the term "superkingdom" (superregnum) makes explicit the fact that there is a nested hierarchy of groups of organisms, many of which share common ancestors (i.e., they are monophyletic). We propose that the rise of emerging lineages was embedded in a primordial "evolutionary grade" (*sensu* Huxley [11]), a group of diversifying organisms (primordial archaeons) in active transition that were initially unified by the same and archaic level of physiological complexity. Our discussion will attempt to reconcile some divergent views of the origin of diversified life and will provide a generic scenario for "turning points" of origin that may be recurrent in biology.

2. A Tripartite World of Organismal Diversity

Carl Woese and his colleagues of the Urbana School were responsible for the groundbreaking discovery that the world of organisms was tripartite; that is, it encompassed not two but three major "domains" of cellular life (Archaea, Bacteria, and Eukarya). Two of the three "aboriginal" lines of decent were initially conceptualized as "urkingdoms" of deep origin that were microbial and qualitatively different from eukaryotic organisms [12]. They corresponded to Archaea and Bacteria. The discovery of Archaea challenged the established akaryote/eukaryote divide (we use the term "akaryote" to describe a cell without a bona fide nucleus. This term complements the word "eukaryote" ("eu," good, and "karyon," kernel), which is ahistorical. The new term takes away the time component of the widely used "prokaryote" ("pro" before) definition, which may be incorrect for many organisms of the microbial domains) that supported "ladder" scenarios of gradual evolution from simplistic microbes to "higher" organisms, which were tenaciously defended by molecular biologists and microbiologists of the time. Woese and Fox [12] made it clear: *"Evolution seems to progress in a "quantized" fashion. One level or domain of organization gives rise ultimately to a higher (more complex) one. What "prokaryote" and "eukaryote" actually represent are two such domains. Thus, although it is useful to define phylogenetic patterns within each domain, it is not meaningful to construct phylogenetic classifications between domains: Prokaryotic kingdoms are not comparable to eukaryotic ones."* The discovery was revolutionary, especially because *scala naturae* deeply seated the roots of the akaryote/eukaryote divide and microbes were considered primitive forms that did not warrant equal standing when compared to the complex organization of Eukarya (see [13] for a historical account). The significance of the tripartite world was quickly realized and vividly resisted

by the establishment. Its resistance is still embodied today in new proposals of origins, such as the archaeon-bacterium fusion hypothesis used to explain the rise of Eukarya (see below). It is noteworthy that the root of the universal tree of cellular organisms, the "tree of life" (ToL), was initially not the driving issue. This changed when the sequences of proteins that had diverged by gene duplication prior to a putative universal common ancestor were analyzed with phylogenetic methods and the comparisons used to root the ToL [14, 15]. Paralogous gene couples included elongation factors (e.g., EF-Tu and EFG), ATPases (α and β subunits), signal recognition particle proteins, and carbamoyl phosphate synthetases, all believed to be very ancient (reviewed in [16]). In many cases, bacterial sequences were the first to branch (appeared at the base) in the reconstructed trees, forcing archaeal and eukaryal sequences to be sister groups to each other. This "canonical" rooting scheme of the ToL (Figure 1(a)) was accepted as fact and was quickly endorsed by the supporters of the Urbana School [17]. In fact, the acceptance of the "canonical" rooting in Bacteria became so deep that it has now prompted the search for the origins of Eukarya in the molecular and physiological constitutions of the putative archaeal sister group [18]. For example, Embley and coworkers generated sequence-based phylogenies using conserved proteins and advanced algorithms to show that Eukarya emerged from within Archaea [19–21] (refer to [22] for critical analysis). Importantly, these analyses suffer from technical and logical problems that are inherent in sequence-based tree reconstructions. For example, proteins such as elongation factors, tRNA synthetases, and other universal proteins used in their analyses are prone to high substitution rates [23]. Mathematically, it leads to loss of information regarding the root of the ToL as shown by Sober and Steel [24] (refer to [25] for more discussion). On the other hand, paralogous rootings sometimes contradicted each other and were soon and rightly considered weak and unreliable [23, 26, 27]. The validity of paralogy-based rooting methodology has proven to be severely compromised by a number of problems and artifacts of sequence analysis (e.g., long branch attraction, mutational saturation, taxon sampling, horizontal gene transfer (HGT), hidden paralogy, and historical segmental gene heterogeneity). Consequently, there is no proper outgroup that can be used to root a ToL that is built from molecular sequences, and currently, there are no proper models of sequence evolution that can provide a reliable "evolutionary arrow." Because of this fact, archaeal and eukaryal rootings should be considered equally probable to the canonical bacterial rooting (Figure 1(c)). This is an important realization that needs to be explicitly highlighted, especially because it affects evolutionary interpretations and the likelihood of scenarios of origins of diversified life.

3. Mining Ancient Phylogenetic Signal in Universal Molecules

Woese's crucial insight was the explicit selection of the ribosome for evolutionary studies. The universality of the ribosomal ensemble and its central role in protein synthesis ensured it carried an ancient and overriding memory of the cellular systems that were studied. This was made evident in the first ToL reconstructions. In contrast, many of the proteins encoded by paralogous gene couples (e.g., translation factors) likely carried convoluted histories of protein domain cooption or important phylogenetic biases induced, for example, by mutational saturation in their protein sequences. The ribosome is indeed the central feature of cellular life: the signature of "ribocells." However, its embedded phylogenetic signatures are also convoluted. The constitution of the ribosome is heterogeneous. The ribosome represents an ensemble of 3-4 ribosomal RNA (rRNA) and ~70 protein (r-protein) molecules, depending on the species considered, and embodies multiple interactions with the cellular milieu needed for function (e.g., assembly and disassembly; interactions with the membrane of the endoplasmic reticulum). A group of 34 r-proteins is present across cellular life [28]. Ribosomal history has been shown to involve complex patterns of protein-RNA coevolution within the evolutionarily conserved core [29]. These patterns are expressed distinctly in its major constituents. While both of its major subunits evolved in parallel, a primordial core that embodied both processive and catalytic functions was established quite early in evolution. This primordial core was later accessorized with structural elements (e.g., accretion of numerous rRNA helical segments and stabilizing A-minor interactions) and r-proteins (e.g., the L7/L12 protein complex that stimulates the GTPase activity of EFG) that enhanced its functional properties. This included expansion elements in the structure of rRNA that were specific not only to subunits but also to individual domains of life. Figure 2, for example, shows a phylogenomic model of ribosomal molecular accretion derived from the survey of protein domain structures in genomes and substructures in rRNA molecules. The accretion process of component parts of the universal core appears to have been a painstakingly slow process that unfolded during a period of ~2 billion years and overlapped with the first episodes of organismal diversification [30]. Despite of this complexity, the focus of biodiversity studies was for decades the small subunit rRNA molecule [31]. This focus has not changed much in recent years. Consequently, the history of organisms and populations is currently recounted by the information seated in the small subunit rRNA molecule. In other words, the historical narrative generally comes from only ~1% of ribosomal molecular constitution. This important and unacknowledged bias was already made explicit in early phylogenetic studies. For example, de Rijk et al. [32] demonstrated that phylogenies reconstructed from the small and large subunit rRNA molecules were different and that the reconstructions from the large subunit were more robust and better suited to establish wide-range relationships. The structure of the small and large subunits was also shown to carry distinct phylogenetic signatures [33]. However, only few evolutionary studies have combined small and large subunit rRNA for history reconstruction. Remarkably, in all of these cases phylogenetic signal was significantly improved (e.g., [34]).

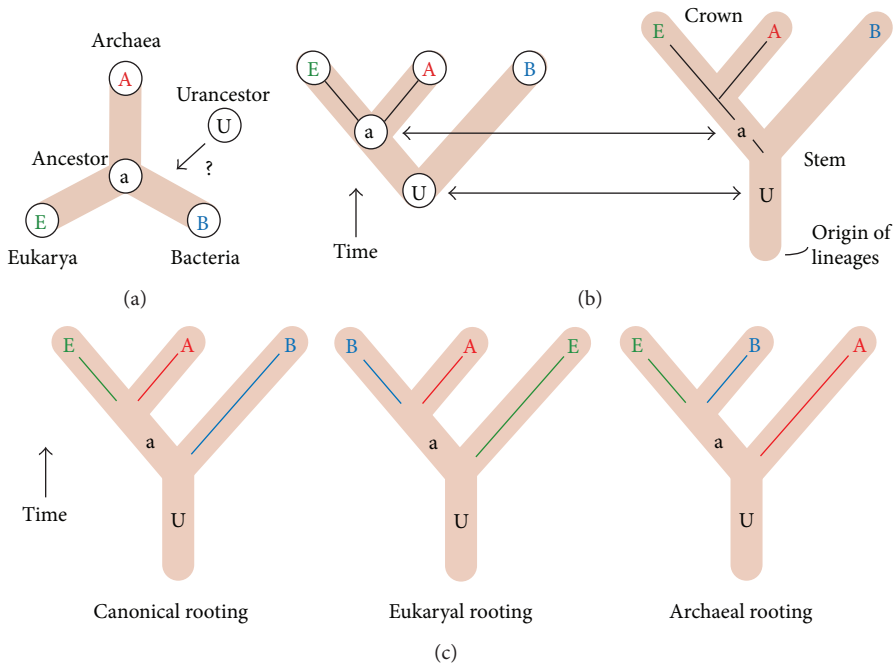


Figure 1: Rooting the tree of life (ToL): an exercise for "tree-thinkers." (a) A node-based tree representation of the ToL focuses on taxa (sampled or inferred vertices illustrated with open circles) and models edges as ancestry relationships. The unrooted tree describes taxa as conglomerates of extant species for each domain of life, Archaea (A), Bacteria (B), and Eukarya (E) (abbreviations are used throughout the paper). Adding a universal ancestor vertex (rooting the ToL) implies adding a reconstructed entity (urancestor) that roots the tree but does not model ancestry relationships. The vertex is either a "most recent universal ancestor" if one defines it from an ingroup perspective or a "last universal common ancestor" (LUCA) if the definition relates to an outgroup perspective. Rooting the tree enables defining "total lineages," which are lists of ancestors spanning from the ancestral taxon to the domain taxa. (b) A stem-based view focuses on edges (branches), which are sampled, and inferred ancestral taxa are viewed as lineages under the paradigm of descent with modification. Vertices correspond to speciation events. Terminal edges represent conglomerates of lineages leading to domains of life (the terminal nodes of node-based trees) and the ancestral stem represents the lineage of the urancestor (U) (double arrowhead line). Total lineages are simply a chain of edges that goes back in time and ends in the ancestral stem. Both node-based and stem-based tree representations are mathematically isomorphic but they are not equal [154]. They change the concept of monophyletic and paraphyletic relationships. A node-based clade starts with a lineage at the instant of the splitting event, incorporating the ancestor into the makeup of the clade. In contrast, a stem-based clade originates with a planted branch on the tree, where the branch represents a lineage between two lineage splitting events. Planting an ancestral stem defines an origin of lineages and the first speciation event in the record of life. This delimits a crown clade of two domain lineages in the stem-based tree (labeled with black lines) that includes the ancestor of the sister groups, a, and a stem domain group at its base. (c) The three possible rootings of the ToL depicted with stem-based tree representations. Terminal edges are labeled with thin lines and these conglomerate lineages can include stem and crown groups.

4. Building a Tower of Babel from a Comparative Genomic Patchwork of Sequence Homologies

Molecular evolutionists were cognizant of the limitation of looking at the history of only few component parts, which by definition could be divergent. When genomic sequences became widely available, pioneers jumped onto the bandwagon of evolutionary genomics and the possibility of gaining systemic knowledge from entire repertoires of genes and molecules (e.g., [35–39]). The genomic revolution, for example, quickly materialized in gene content trees that reconstructed the evolution of genomes directly from their evolutionary units, the genes (e.g., [37, 39, 40]), or the domain constituents of the translated proteins [35, 41]. The sequences of multiple genes were also combined or concatenated in attempts to extract deep phylogenetic signal [42–44]. The results that were obtained consistently supported the tripartite world, backing-up the claims of the Urbana School. However, clues about ancestors and lineages leading to extant taxa were still missing.

With few exceptions that focused on the structure of RNA and protein molecules (see below), analyses based on genomic sequences, gene content, gene order, and other genomic characteristics were unable to produce rooted trees without the help of outgroups; additional *ad hoc* hypotheses that are external to the group of organisms being studied and generally carry strong assumptions. An "arrow of time" was not included in the models of genomic evolution that were used. As with sequence, ToLs that were generated were unrooted (Figure 1(a)) and generally rooted *a posteriori* either claiming that the canonical root was correct or making assumptions about character change that may be strictly incorrect. In a recent example, distance-based approaches

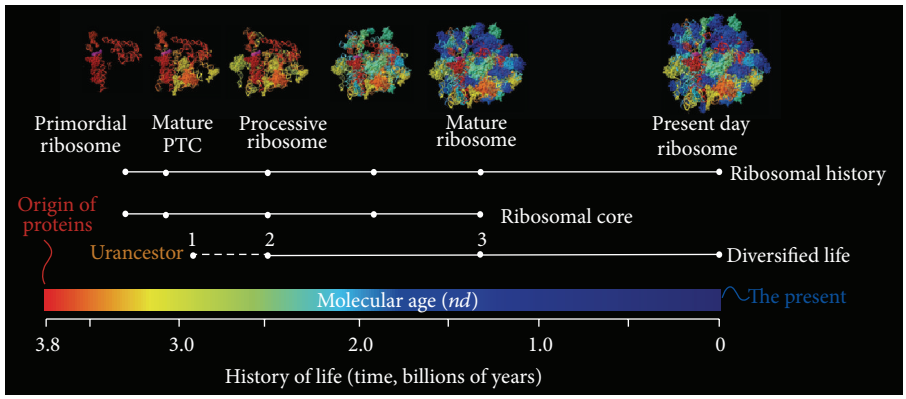


Figure 2: The evolutionary history of ribosome was traced onto the three-dimensional structure of its rRNA and r-protein components. Ages of components were colored with hues from red (ancient) to blue (recent). Phylogenomic history shows that primordial metabolic enzymes preceded RNA-protein interactions and the ribosome. The timeline of life derived from a universal phylogenetic tree of protein domain structure at fold superfamily level of complexity is shown with time flowing from left to right and expressed in billions of years according to a molecular clock of fold structures. The ribosomal timeline highlights major historical events of the molecular ensemble. History of the conserved ribosomal core [29] overlaps with that of diversified life [48], suggesting that episodes of cooption and lateral transfer have pervaded early ribosomal evolution.

were used to build universal network trees from gene families defined by reciprocal best BLAST hits [45]. These networks showed a midpoint rooting of the ToL between Bacteria and Archaea. However, this rooting involves a complex optimization of path lengths in the split networks and critically assumes that lineages evolved at roughly similar rates. This diminishes the confidence of the midpoint rooting, especially when considering the uncertainties of distances inferred from BLAST analyses and the fact that domains in genes hold different histories and rates of change. Another promising but ill-conceptualized case is the rooting of distance-based trees inferred by studying the frequency of l-mer sets of amino acids in proteins at the proteome level [46]. The compositional data generated a ToL that was rooted in Eukarya when randomized proteome sequences were used as outgroups. The assumption of randomness associated with the root of a ToL is however unsupported or probably wrong, especially because protein sequence space and its mappings to structure are far from random [47]. More importantly, a large fraction of modern proteins had already evolved protein fold structures when the first diversified lineages arose from the universal common ancestor of cellular life [48]. These evolved repertoires cannot be claimed to be random.

The inability of genomics to provide clear answers to rooting questions promoted (to some extent) parsimony thinking and the exploration of evolutionary differences of significance that could act as "anchors" and could impart "polarity" to tree statements [49]. As we will describe below, we applied parsimony thinking to the evolution of protein and RNA structures [41, 50], and for over a decade we have been generating rooted ToLs that portray the evolution of proteomes with increasing explanatory power. This work however has been for the most part unacknowledged. Instead, molecular evolutionists focused almost exclusively on molecular sequences in their search for solutions to the ToL problem. For example, analysis of genomic insertions and deletions (indels) that are rare in paralogous gene sets rooted the ToL between the group of Actinobacteria and Gram-negative bacteria and the group of Firmicutes and Archaea [51]. Unfortunately, little is known about the dynamics of indel generation, its biological role in gene duplication, and its influence on the structure of paralogous protein pairs (e.g., [52]). If the dynamics are close to that of sequences, indels should be considered subject to all the same limitations of sequence analyses, including long branch attraction artifacts, character independence, and inapplicable characters. Given a tree topology, parsimony was also used to optimize the evolutionary transformation of genomic features. For example, when considering the occurrence of protein domains that are abundant, variants of domain structures accumulate gradually in genomes; operationally the domain structure per se cannot be lost once it had been gained. This enabled the use of a variant of the unrooted Dollo algorithmic method (e.g., [53]); the Dollo parsimony model [54] is based on the assumption that when a biological feature that is very complex is lost in evolution it cannot be regained through vertical descent. The method however could not be applied to akaryotic genomes, as these are subject to extensive lateral gene transfer, and until recently [55] was not extended to ToL reconstructions.

Parsimony thinking was also invoked in "transition analysis," a method that attempts to establish polarity of character change by, for example, examining homologies of proteins in proteolytic complexes such as the proteasome, membrane and cell envelope biochemical makeup, and body structures of flagella, sometimes aided by BLAST queries [56, 57]. The elaboration however is restricted to the few molecular and cellular structures that are analyzed, out of thousands that populate the akaryotic and eukaryotic cells. The approach is local, has not yet made use of an objective

analytic phylogenetic framework, and does not weigh gains, losses, and transfers with algorithmic implementations. Thus, many statements can fall pray of incorrect optimizations or processes of convergence of structure and function, including cooptions that are common in metabolic enzymes [58]. Transitions also fail to consider the molecular makeup of the complexes examined (e.g., evolution of the photosynthetic reaction centers), which often hold domains with heterogeneous histories in the different organismal groups. For example, the F_1/F_o ATP synthetase complex that powers cellular processes has a long history of accretion of domains that span almost 3.8 billion years of history, which is even older than that of the ribosome [59]. Finally, establishing the validity of evolutionary transitions in polarization schemes can be highly problematic; each transition that is studied requires well-grounded assumptions [60], some of which have not been yet properly elaborated.

The inability to solve the rooting problem and the insistence on extracting deep phylogenetic signal from molecular sequences that are prone to mutational saturation raised skepticism about the possibility of ever finding the root of the ToL. Bapteste and Brochier [60] made explicit the conceptual difficulties claiming scientists in the field had adopted Agrippa's logic of doubt. Misunderstandings on how to conduct ideographic inquiry in evolutionary genomics had effectively blocked the reerection of a Tower of Babel that would explain the diversification of genomes (Figure 3). Instead, there was "Confusion of Doubts." Unfortunately, the impasse was aggravated by aprioristic tendencies inherited from systematic biology [3, 10], disagreements about the evolutionary role of vertical and lateral inheritance and the problem of homology [61], and currently disagreements about the actual role of Darwinian evolution in speciation and the ToL problem [13, 62].

Figure 3: The "Confusion of Tongues," an engraving by Gustave Doré (1832–1883), was modified by D. Caetano-Anollés to portray the event of diversification that halted the construction of the Tower of Babel. In linguistics, the biblical story inspired tree thinking. We take the metaphor as the fall of the urancestor of cellular life and the replacement of "*scala naturae*" by branching processes of complexity.

During the decade of evolutionary genomic discovery, the effects of HGT on phylogeny [63] took front cover. HGT was invoked as an overriding process. However, little attention was paid to alternative explanations such as differential loss of gene variants, ancient or derived, and there was little concern for other sources of reticulation. HGT is certainly an important evolutionary process that complicates the "tree" concept of phylogenetic analysis and must be carefully studied (e.g., [64–66]). In some bacterial taxonomic groups, such as the proteobacteria, HGT was found to be pervasive and challenged the definition of species [67]. Cases like these prompted the radical suggestion that the ToL should be abandoned and that a "web of life" should be used to describe the diversification of microbes and multicellular organisms [68]. However, the problem has two aspects that must be separately considered.

(i) Mechanics. The widespread and impactful nature of HGT must be addressed. Does its existence truly compromise the validity of phylogenetic tree statements? While HGT seems important for some bacterial lineages [69], its evolutionary impacts in Archaea and Eukarya are not as extensive (e.g., [61, 70]) and its global role can be contested [71]. In turn, the role of viruses and RNA agents in genetic exchange continues to be understudied (e.g., [72]) and could be a crucial source of reticulation. Furthermore, the relationship between differential gene loss and HGT has not been adequately formalized, especially for genes that are very ancient. Establishing patterns of loss requires establishing polarity of change and rooted trees, which as we have discussed earlier remains unattained for ToLs reconstructed from molecular sequences.

(ii) Interpretation. HGT generally materializes as a mismatch between histories of genes in genomes and histories of organisms. Since genomes are by functional definition collections of genes, reticulations affect history of the genes and not of the organisms, which given evolutionary assumptions result from hierarchical relationships (e.g., [73]). Thus and at first glance, reticulation processes (HGT, gene recombination, gene duplications, and gene loss) do not obliterate vertical phylogenetic signal. The problem however remains complex. A ToL can be considered an ensemble of lineages nested within each other [74]. Protein domain lineages will be nested in gene lineages; gene lineages will be nested in lineages of gene families; gene families will be nested in organismal lineages; organisms will be nested in lineages of higher organismal groupings (populations, communities, ecosystems, and biomes), and so on. All lineage levels are defined by biological complexity and the hierarchical organization of life and follow a fractal pattern distribution, each level contributing vertical and lateral phylogenetic signal

to the whole. However, sublineages in one hierarchical level may hold different histories compared to the histories of higher and lower levels. Historical mismatches introduce, for example, lineage sorting and reticulation problems that represent a complication to phylogenetic analysis. For example, mismatches between gene and organismal phylogenies exist in the presence of differential sorting of genomic components due to, for example, species extinctions or genomic rearrangements. These mismatches violate the fundamental cladistic assumption that history follows a branching process, but can be explained by *homoplasy* (convergent evolution), the horizontal "trace" of the homology-homoplasy yin-yang of phylogenetic signal. The problem of homoplasy cannot be solved by conceptually preferring one particular level of the hierarchy [74]. A focus on the organismal level, for example, will not solve the problems of hybridization that are brought by strict or relaxed sexual reproduction or viral-mediated genetic exchanges, or the problems of ancient events of fusions or endosymbiosis. However, homoplasy and the optimization of characters and trees of cladistic analysis provide a rigorous framework to discover the magnitude and source of nonvertical processes in evolution. Thus, cladistics or the ToL are not invalidated by network-like signals. In fact, recent discrete mathematical formulations that test the fundamental axioms of tree construction have also proven that the bifurcating history of trees is preserved despite evolutionary reticulations [75]. While there is no "confusion of doubts" at this level, the babel of confusion has not stopped.

Genomics revealed evolutionary patchiness, which incited hypotheses of chimeras and fusions. Phylogenies of genes were found highly discordant. Organisms that were being sequenced shared very few gene sequences, and more troublingly, gene trees that were generated possessed different topologies, especially within akaryotic organisms. With time, the number of nearly universal genes decreased and the patchiness and discordancy increased. The comparative genomic patchwork of sequence homologies showed that there were groups of genes that were only shared by certain domains of life. Thus, the makeup of genomes appeared chimeric. A number of hypotheses of chimeric origins of eukaryotes were proposed based on the fact that eukaryotes shared genes expressing sister group phylogenetic relationships with both Bacteria and Archaea [64, 76]. One that is notable is the rise of eukaryotes from a "ring of life" fusion of an archaeon and a bacterium (e.g., [77]), which in a single blow defeated both the tree-like and the tripartite nature of life. Under this school, homology searches with BLAST against a database of ~3.8 million akaryotic sequences allowed to assign archaeal, bacterial, or ambiguous ancestries to genes in the human genome and explain homology patterns as relics of the akaryotic ancestors of humans [78]. Remarkably, archaeal genes tended to be involved in informational processes, encoded shorter and more central proteins, and were less likely to be involved in heritable human diseases. The chimeric origin of eukaryotes by fusions has been rightfully contested; there is no proper evidence supporting its existence [79]. Unfortunately, chimerism opened a flood of speculations about reticulation. The history of life was seen by many through the lens of a "forest" of gene histories [80]. Reticulations and ultimately HGT were forced to explain chimeric patterns. Fusion hypotheses diverted the central issue of the rooting of the ToL and prompted the separate analysis of eukaryotic origins and akaryotic evolution.

Dagan and Martin [64] denounced that the ToL was a "tree of one percent" because only a small fraction of sequences could be considered universal and could be mined for deep phylogenetic signal. The rest would account for lateral processes that confounded vertical descent. The claim came fundamentally from networks constructed using BLAST heuristic searches for short and strong matches in genomic sequences. Phylogenetic "forests" of akaryotic genes later on boosted the claim [81]. These networks were built from 6,901 trees of genes using maximum likelihood methods. Only 102 of these trees were derived from nearly universal genes and contributed very little vertical phylogenetic signal. The initial conclusion was striking: "the original tree of life concept is obsolete; it would not even be a tree of one percent" [81]. However, the fact that vertical signal was present in the forests merited reevaluation: "replacement of the ToL with a network graph would change our entire perception of the process of evolution and invalidate all evolutionary reconstruction" [82]. We note that in these studies an ultrametric tree of akaryotes was recovered from vertical signal in a supertree of nearly universal genes. The rooted tree, which was used to simulate clock-like behavior, revealed the early divergence of *Nanoarchaeum equitans* and then Archaea [81]. While inconsistency of the forest supertree increased at high phylogenetic depths, its associated supernetwork showed there were no reticulations between Archaea and Bacteria. Thus, the deep rooting signal of archaeal diversification may be bona fide and worthy of further study.

While those that value tree thinking have contested on many grounds the idea that the ToL is "obsolete" (e.g., [13]), our take in the debate is simple. Homologies between gene sequences established through the "emperor's BLAST" (*sensu* [71]) are poor substitutes to phylogenetic tree reconstructions from gene sequences. By the same token, phylogenies reconstructed from sequences are poor substitutes to phylogenies that consider the molecular structure and function of the encoded products. Gene sequences are not only prone to mutational saturation but they generally come in pieces. These pieces represent evolutionary and structural modules that host the functions of the encoded molecules. Protein domains, for example, are three-dimensional (3D) arrangements of elements of secondary structure that fold autonomously [83], are compact [84], and are evolutionarily conserved [85, 86]. The landscape of evolutionary exploration changes when the role of protein domains in function and evolution is considered [87]. Changes at sequence level, including substitutions, insertions, and deletions, can have little impact on structure, and vice versa; sequence changes in crucial sites can have devastating consequences on function and fitness. However, there is no detailed analysis of historical mismatches between gene sequences and domain structures at the ToL level. Remarkably, while HGT seems rampant at sequence level, its impact at the domain structural level is limited [88]. This makes trees derived from domains

effectively "trees of 99 percent" and their use very powerful. The fact that domains diversify mostly by vertical descent (e.g., [89–91]) suggests that gene reticulations simply reflect the pervasive and impactful combinatorial effect of domain rearrangements in proteins [92] and perhaps little else. This important claim must be carefully evaluated. It offers the possibility of dissecting how levels of organization impact processes of inheritance in biology.

5. Out of the Impasse: Parts and Wholes in the Evolving Structure of Systems

The rooting of the ToL is clearly muddled by the high dynamic nature of change in protein and nucleic acid sequences and by the patchiness and reticulation complexities that exist at gene level. However, it is also possible that the problems of the ToL are ultimately technical. We have made the case that the use of molecular sequences is problematic on many grounds, including mutational saturation, definition of homology of sites in sequence alignments, inapplicable characters, taxon sampling and tree imbalance, and different historical signatures in domains of multidomain proteins [93]. In particular, violation of character independence by the mere existence of atomic structure represents a very serious problem that plagues phylogenetic analysis of sequences. We here present a solution to the impasse. We show that the ToL can be rooted with different approaches that focus on structure and function and that its root is congruently placed in Archaea.

Epistemologically, phylogenetic characters must comply with symmetry breaking and the irreversibility of time [94]. In other words, characters must establish transformational homology relationships and serve as independent evidential statements.

(i) Characters Must Be Homologous and Heritable across Tree Terminal Units (Taxa). Character homology is a central and controversial concept that embodies the existence of historical continuity of information [74]. Characters are "basic" evidential statements that make up columns in data matrices used for tree reconstruction. They are conjectures of perceived similarities that are accepted as fact for the duration of the study, are strengthened by Hennigian reciprocal illumination, and can be put to the test through congruence with other characters, as these are fit to the trees. To be useful, characters must be heritable and informative across the taxa rows of the data matrix. This can be evaluated, for example, with the cladistics information content (CIC) measure [95]. Finding informative characters can be particularly challenging when the features that are studied change at fast pace and when taxa sample a wide and consequently deep phylogenetic spectrum. When building a ToL, the highly dynamic nature of change in the sequence makeup of protein or nucleic acid molecules challenges the ability to retrieve reliable phylogenetic signatures across taxa, even if molecules are universally distributed and harbor evolutionarily conserved regions with deep phylogenetic signal (e.g., rRNA). The reason is that given enough time, functionally or structurally constrained regions of the sequence will be fixed (will be structurally canalized [6]) and will offer little if any phylogenetically meaningful signal to uncover, for example, the universal rRNA core. In turn, mutational saturation of unconstrained regions will quickly erase history.

The mutational saturation problem was made mathematically explicit by Sober and Steel [24] using "mutual information" and the concept of time as an information destroying process. Mutual information $I(X, Y)$ between two random variables X and Y is defined by

$$I(X,Y) = \sum_{x,y} P(X = x, Y = y) \log\left(\frac{P(X = x, Y = y)}{P(X = x)\,P(Y = y)}\right). \quad (1)$$

When $I(X, Y)$ approaches 0, X and Y become independent and no method can predict X from knowledge of Y. Importantly, mutual information approaches 0 as the time between X and Y increases in a Markov chain. Regardless of the use of a parsimony, maximum likelihood, or Bayesian-based framework of analysis, $I(X, Y)$ will be particularly small when sequence sites are saturated by too many substitutions due to high substitution rates or large time scales. Ancestral states at interior nodes of the trees cannot be established with confidence from extant information even in the most optimal situation of knowing the underlying phylogeny and the model of character evolution. Under simple models, the problem is not mitigated by the fact that the number of terminal leaves of trees and the sources of initial phylogenetic information increases with time. Since a phase transition occurs when substitution probabilities exceed a critical value [96], one way out of the impasse is to find features in sufficient number that change at much slower pace than sequence sites and test if mutual information is significant and overcomes Fano's inequality. These features exist in molecular biology and have been used for phylogenetic reconstruction. They are, for example, the 3D fold structures of protein molecules [87] or the stem modules of RNA structures [97], features that change at very slow rate when compared to associated sequences. For example, protein 3D structural cores evolve linearly with amino acid substitutions per site and change at 3–10 times slower rates than sequences [98]. This high conservation highlights the evolutionary dynamics of molecular structure. Remarkably, rates of change of proteins performing a same function are maintained by functional constraints but accelerate when proteins perform different functions or contain indels. In turn, fold structural diversity explodes into modular structures at low sequence identities probably triggered by functional diversification. Within the context of structural conservation, the fact that fold structures are structural and evolutionary modules that accumulate in proteomes by gene duplication and rearrangements and spread in biological networks by recruitment (e.g., [99]) also provides a solution to the problems of vanishing phylogenetic signal. Since fold accumulation increases with time in the Markov chain, mutual information must increase, reversing

the "data processing inequality" that destroys information and enabling deep evolutionary information.

(ii) Characters Must Show at Least Two Distinct Character States. One of these two states (transformational homologs) must be ancestral (the "plesiomorphic" state) and the other must be derived (the "apomorphic" state) [74]. Only shared and derived features (synapomorphies) provide vertical phylogenetic evidence. Consequently, determining the relative ancestry of alternative character states defines the polarity of character transformations and roots the underlying tree. This is a fundamental property of phylogenetic inference. Polarization in tree reconstruction enables the "arrow of time" (*sensu* Eddington's entropy-induced asymmetry), solves the rooting problem, and fulfills other epistemological requirements.

Cladistically speaking, character polarity refers to the distinction between the ancestral and derived states and the identification of synapomorphies. However, an evolutionary view of polarity also refers to the direction of character state transformations in the phylogenetic model. Historically, three accepted alternatives have been available for rooting trees [100–102], the *outgroup comparison,* the *ontogenetic method,* and the *paleontological or stratigraphic method*. While the three methods do not include assumptions of evolutionary process, they have been the subject of much discussion and their interpretation of much controversy. The first two are however justified by the assumption that diversity results in a nested taxonomic hierarchy, which may or may not be induced by evolution. We will not discuss the stratigraphic method as it relies on auxiliary assumptions regarding the completeness of the fossil record. The midpoint rooting criterion mentioned earlier will not be discussed either. The procedure is contested by the existence of heterogeneities in rates of change across the trees and problems with the accurate characterization of phylogenetic distances.

In outgroup comparison, polarity is inferred by the distribution of character states in the ingroup (group of taxa of interest) and the sister group (taxa outside the group of interest). In a simple case, if the character state is only found in the ingroup, it must be considered derived. Outgroup comparison is by far the method of choice because phylogeneticists tend to have confidence in the supporting assumptions: higher-level relationships are outside the ingroup, equivalent ontogenetic stages are compared, and character state distributions are appropriately surveyed. Unfortunately, the method is "indirect" in that it depends on the assumption of the existence of a higher-level relationship between the outgroup and the ingroup. Consequently, the method cannot root the ToL because at that level there is no higher-level relationship that is presently available. Moreover, the method in itself does not polarize characters. It simply connects the ingroup to the rest of the ToL [100].

The ontogenetic criterion confers polarity through the distribution of the states of homologous characters in ontogenies of the ingroup, generally by focusing on the generality of character states, with more widely distributed states being considered ancestral. Nelson's rule states *"given an ontogenetic character transformation from a character observed to be more general to a character observed to be less general, the more general character is primitive and the less general advanced"* [103]. This "biogenetic law" appears powerful in that it depends only on the assumption that ontogenies of ingroup taxa are properly surveyed. It is also a "direct" method that relies exclusively on the ingroup. Consequently, it has the potential to root the ToL. Unfortunately, Nelson's "generality" has been interpreted in numerous ways, especially as it relates to the ontogenetic sequence, leading to much confusion [102]. It also involves comparison of developmentally nested and distinct life history stages, making it difficult to extend the method (originally conceptualized for vertebrate phylogeny) to the microbial world. However, Weston [100, 104] made it clear that the ontogenetic criterion embodies a wider "generality criterion" in which the taxic distribution of a character state is a subset of the distribution of another. In other words, character states that characterize an entire group must be considered ancestral relative to an alternative state that characterizes a subset of the group. Besides the centrality of nested patterns, the generality criterion embeds the core assumption that every homology is a synapomorphy in nature's nested taxonomic hierarchy and that homologies in the hierarchy result from additive phylogenetic change [100]. Weston's more general rule therefore states that *"given a distribution of two homologous characters in which one, x, is possessed by all of the species that possess its homolog, character y, and by at least one other species that does not, then y may be postulated to be apomorphous relative to x"* [104]. The only assumption of the method is that relevant character states in the ingroup are properly surveyed. This new rule crucially substitutes the concept of ontogenetic transformation by the more general concept of homology and additive phylogenetic change, which can be applied to cases in which homologous entities accumulate "iteratively" in evolution (e.g., generation of paralogous genes by duplication). Since horizontally acquired characters (xenologs) are not considered synapomorphies, they contribute towards phylogenetic noise and are excluded after calculation of homoplasy and retention indices (i.e., measures of goodness of fit of characters to the phylogeny).

We have applied the "generality criterion" to the rooting of the ToL through polarization strategies that embody axioms of evolutionary process. Figure 4 shows three examples. A rooted phylogeny describing the evolution of 5S rRNA molecules sampled from a wide range of organisms was reconstructed from molecular sequence and structure [105]. The ToL that was recovered was rooted paraphyletically in Archaea (Figure 4(a)). The model of character state transformation was based on the axiom that evolved RNA molecules are optimized to increase molecular persistence and produce highly stable folded conformations. Molecular persistence materializes in RNA structure in vitro, with base pairs associating and disassociating at rates as high as $0.5\,s^{-1}$ [106]. The frustrated kinetics and energetics of this folding process enable some structural conformations to quickly reach stable states. This process is evolutionarily optimized through structural canalization [6], in which evolution attains molecular functions by both increasing the average life and stability of

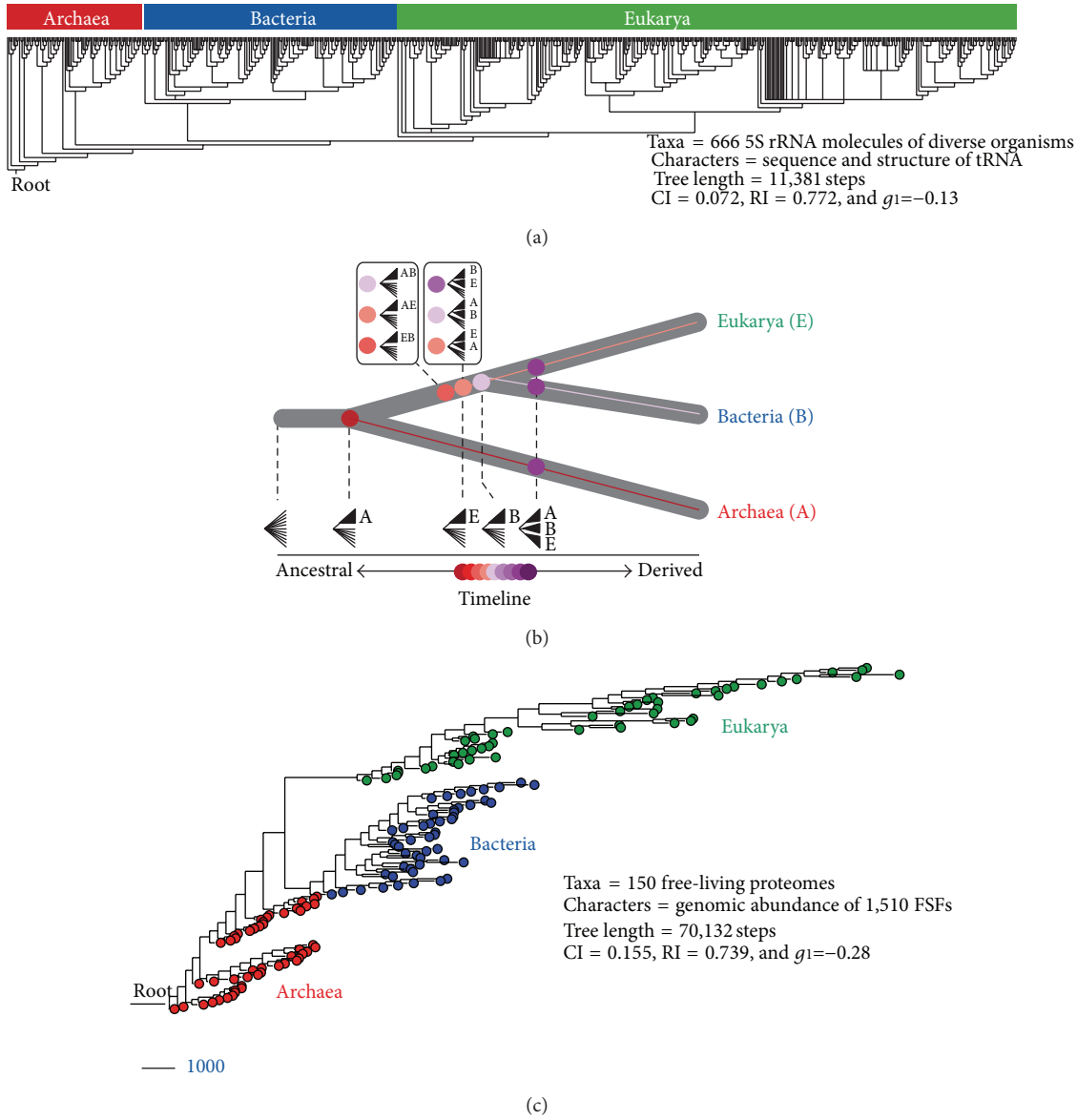


Figure 4: Trees of life generated from the structure of RNA and protein molecules congruently show a rooting in Archaea. (a) A rooted phylogenetic tree of 5S rRNA reconstructed from both the sequence and the structure of the molecules (from [105]). (b) Global most-parsimonious scenario of organismal diversification based on tRNA (from [107]). A total of 571 tRNA molecules with sequence, base modification, and structural information were used to build a ToL, which failed to show monophyletic groupings. Ancestries of lineages were then inferred by constraining sets of tRNAs into monophyletic groups representing competing (shown in boxes) or noncompeting phylogenetic hypotheses and measuring tree suboptimality and lineage coalescence (illustrated with color hues in circles). (c) A ToL reconstructed from the genomic abundance counts of 1,510 FSFs as phylogenetic characters in the proteomes of 150 free-living organisms sampled equally and randomly from the three domains of life (data taken from [122, 123]). Taxa were labeled with circles colored according to superkingdom. CI = consistency index, RI = retention index, and g_1 = gamma distribution parameter.

selected conformations and decreasing their relative number. Thus, conformational diversity measured, for example, by the Shannon entropy of the base-pairing probability matrix or features of thermodynamic stability act as "evo-devo" proxy of a generality criterion for RNA molecules, in which the criterion of similarity (e.g., ontogenetic transformation) is of positional and compositional correspondence. Using a different approach, we recently broadened the use of phylogenetic constraint analysis [107, 108], borrowing from a formal cybernetic method that decomposes a reconstructable system into its components [109], and used it to root a ToL derived from tRNA sequence and structure (Figure 4(b)). The ToL was again rooted in Archaea. The number of additional steps required to force (constrain) particular taxa into a monophyletic group was used to define a lineage coalescence distance (S) with which to test alternative hypotheses of monophyly. These hypotheses were then ordered according to S value in an evolutionary timeline. Since S records

the relative distribution of character states in taxic sets, it also embodies the generality criterion of rooting. Finally, we generated rooted ToLs that describe the evolution of proteomes directly from a census of protein fold structures in proteomes (Figure 4(c)). This ToL shows a paraphyletic rooting in Archaea. The method extracts phylogenetic signal from proteomic abundance of protein fold structures and considers that the most abundant and widely distributed folds are of ancient origin when defining transformation series [41]. This polarization scheme, which results in the gradual growth of the proteome repertoire, again represents an embodiment of the generality criterion in which statements of homology (fold structures) result from additive phylogenetic change (increases in abundance). It is noteworthy that these ToL reconstructions take into account the genomic abundance of each and every fold structure in each proteome and across the entire matrix, thereby generating a frustrated system. In general, very high abundance of only few folds will not attract taxa (species) to the derived branches in the ToL. The ancestry of taxa is determined by both the abundance and interplay among fold structure characters. For example, metabolic folds such as those involved in ATP hydrolysis are widespread in cells and considered to be very ancient. These are also the most ancient folds in our phylogenies. In comparison, some popular eukaryote-specific folds (e.g., immunoglobulin superfamilies) are highly abundant but appear in a derived manner in our phylogenies. Thus we reason that there is no circularity involved in the character polarization scheme.

The compositional schemes extend the concept of rooting with paralogous sequences to the entire proteome complements, from gene family level [100] to structural hierarchies. The three examples make use of different rooting rationales but provide a congruent scenario of origins of diversified life. Technically, roots are inferred by the Lundberg method [110] that does not require any outgroup taxa specification. This method roots the trees *a posteriori* by attaching the hypothetical ancestor to that branch of the unrooted network that would yield minimum increase in the tree length (thus preserving the principle of parsimony).

(iii) Characters Must Serve as Independent Evolutionary Hypotheses. Valid phylogenetic optimization requires that characters be independent pieces of evidence. Characters should not depend on other characters. When the assumption of independence is violated, characters are overweighed in the analysis and the resulting phylogeny fails to represent true history [111]. Possible dependencies could be of many kinds, from structural to functional, from developmental to ecological. These dependencies distort and obscure phylogenetic signal and must be either avoided or coded into the phylogenetic model through parameters or weight corrections.

As we will now elaborate, the problem of character independence is about *parts* and *wholes* in the hierarchical fabric of life and in the nested hierarchies of the ToL. Biological systems are by definition made of parts regardless of the way parts are defined. In evolution, diversification and integration of parts unify parts into cohesive entities, modules, which then diversify [112]. This process and the rise of modules may explain evolutionary waves of complexity and organization and the emergence of structure (defined broadly) in biology that is hierarchical. The hierarchical makeup is made evident in the structure of protein molecules, where lower level parts of the polymer (the amino acid residues) interact with each other and establish cohesive higher-level modular parts, which also establish interaction networks and are crucial for molecular function and for interaction of proteins with the cellular environment. Consequently, the structure of proteins can be described at increasing hierarchical levels of structural abstraction: sequences, motifs, loops, domains, families, superfamilies, topologies, folds, architectures, and classes. Two accepted gold standards of protein classification, structural classification of proteins (SCOP) [113] and class architecture topology homology (CATH) [114], use parts of this incomplete scheme to describe the atomic complexity of the molecules. We note that these classifications do not consider unrealized structural states, such as protein folds that are possible but that have never been identified in the natural world of protein structures. We also note that modules sometimes engage in combinatorial games. For example, protein domains are rearranged in evolution by fusions and fissions producing the enormous diversity of alternative domain rearrangements that exist in multidomain proteins [92].

Since biological systems evolve and carry common ancestry, parts of these systems by definition evolve and by themselves carry common ancestry. In other words, the histories of parts are embodied in the ensemble of lineages of the ToL. The focus of phylogenetic analysis however has been overwhelmingly the organism as the biological system, as testified by the effort devoted to taxonomical classification, systematic biology, and building of the ToL. The genomic revolution has provided a wealth of parts in the amino acid sequence sites of proteins and nucleic acid molecules, which have been used as phylogenetic characters for analysis of molecules representing organisms. This has maintained the focus of reconstructing trees of systems (the wholes). For example, amino acid sites of a protein are generally fit into a data matrix (alignment), which is then used to build trees of genes and organisms (Figure 5(a)) using modern algorithmic implementations [115]. However and as we have mentioned, proteins have complicated structures that result from interactions between amino acids at the 3D atomic level. These intramolecular interactions, or at least some of them [116], induce protein folding and delimit molecular function and protein stability. They are responsible for protein secondary, supersecondary, domain, and tertiary structure, and, by definition, their mere existence induces violation of character independence. Penny and Collins [117] proposed the simple thought experiment in which the bioinformatician exchanges rows of sequence sites in the alignment matrix and asks what was lost in the process. Randomization of characters (columns) in the data matrix does not change the phylogenetic tree. However, randomization destroys the structure of the molecule and very likely its function. This confirms that reconstruction of trees from information in sequence sites violates character independence, and in the process ignores structure and biology. The effects of violation

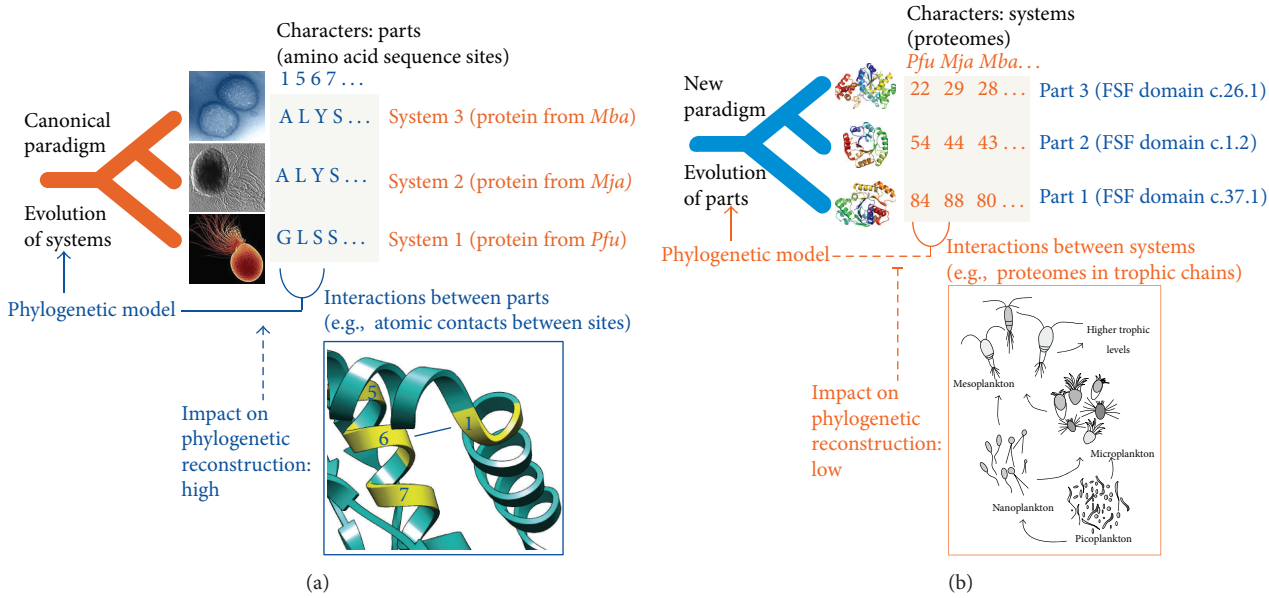


FIGURE 5: A new phylogenetic strategy simplifies the problems of character independence. (a) The canonical paradigm explores the evolution of systems, such as the evolution of organisms in the reconstruction of ToLs. For example, the terminals of phylogenetic trees can be genes sampled from different organisms (e.g., *Pyrococcus furiosus* (*Pfu*), *Methanococcus jannaschii* (*Mja*), and *Methanosarcina barkeri* (*Mba*)), and phylogenetic characters can be amino acid sequence sites of the corresponding gene products. Character states can describe the identity of the amino acid at each site. Since characters are molecular parts that interact with other parts when molecules fold into compact 3D structures, their interaction violates the principle of character independence. Consequently, the effects of covariation must be considered in the phylogenetic model used to build the trees. (b) The new paradigm explores the evolution of parts, such as the evolution of protein domains in proteomes. For example, the terminals of phylogenetic trees can be domains defined at fold superfamily level of structural complexity and the characters used to build the trees can be proteomes. Character states can be the number of domains holding the FSF structure. Since proteomes interact with other proteomes when organisms establish close interactions, interactions that could affect the abundance of domains in proteomes should be considered negligible (unless there is an obligate parasitic lifestyle involved) and there is no need to budget trophic interactions in the phylogenetic model.

of character independence may be minimal for trees of sequences that are closely related. However, trees describing deep historical relationships require that sequences be divergent and this maximizes the chances of even wider divergences in molecular structure that are not being accounted for by the models of sequence evolution.

The genomic revolution has also provided a wealth of models of 3D atomic structure. These structures are used as gold standards to assign with high confidence structural modules to sequences. As mentioned earlier, proteomes embody collections of protein domains with well-defined structures and functions. Protein domain counts in proteomes have been used to generate trees of protein domains (Figure 5(b)) using standard cladistic approaches and well-established methods (reviewed in [87]). These trees describe the evolution of protein structure at global level. They are effectively trees of parts. While domains interact with each other in multidomain proteins or establish protein-protein interactions with the domains of other proteins, these interactions of parts do not violate character independence. This is because phylogenetic characters are actually proteomes, systems defined by structural states that exist at much higher levels than the protein domain, not far away from the organism level. Remarkably, no information is lost when character columns in the matrix are randomized in the thought experiment. The order of proteome characters in the matrix does not follow any rationale. Characters are not ordered by lifestyles or trophic levels of the organisms. Interactions between free-living organisms will seldom bias their domain makeup, and if so, those characters can be excluded from analysis. Even the establishment of symbiotic or obligate parasitic interactions, such as the nodule-forming symbioses between rhizobia and legumes, may have little impact on character independence, as long as the joint inclusion of the host and the symbiont is avoided.

6. Evidence Supporting the Archaeal Rooting of the Universal Tree

Figure 4 shows examples of rooted ToLs generated from the sequence and structure of RNA and protein molecules. The different phylogenomic approaches arrive at a common rooted topology that places the stem group of Archaea at the base of the ToL (the archaeal rooting of Figure 1(c)). However, the evolutionary interrelationship of parts and wholes prompt the use of trees of parts, a focus on higher-level structure, and a decrease in confidence in the power of trees of systems. These concepts have been applied to the study of the history of nucleic acid and protein structures for over a decade and

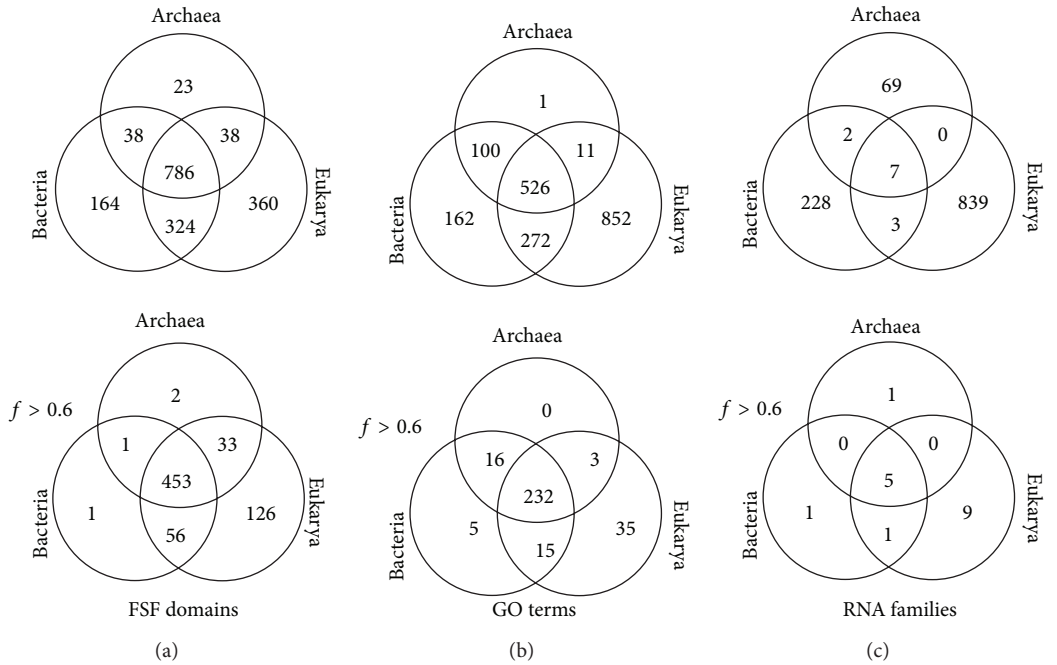


Figure 6: Venn diagrams displaying the distributions of 1,733 FSF domains (a), 1,924 terminal GO terms (b), and 1,148 RNA families (c) in the genomes of the three domains of life. FSF domain data was taken from Nasir et al. [122, 123] and included 981 completely sequenced proteomes from 70 Archaea, 652 Bacteria, and 259 Eukarya. Terminal GO terms corresponding to the "molecular function" hierarchy defined by the GO database [118] were identified in 249 free-living organisms, including 45 Archaea, 183 Bacteria, and 21 Eukarya (data taken from [123]). The Venn diagram of RNA families and the distribution of Rfam clans and families in organisms were taken from Hoeppner et al. [129] and their Dataset S1. Shown below are distribution patterns for FSFs, GO terms, and RNA families that are present in more than 60% of the organisms examined ($f > 0.6$). All distributions highlight maximum sharing in the ancient ABE and BE taxonomic groups and minimal sharing in archaeal taxonomic groups.

provide additional evidence in support of the archaeal rooting scenario.

6.1. Comparative Genomic Argumentation. The distribution of gene-encoded products in the genomes of sequenced organisms and parsimony thinking can reveal global evolutionary patterns without formal phylogenetic reconstruction. We will show how simple numerical analyses of protein domains, molecular activities defined by the gene ontology (GO) consortium [118], and RNA families that are domain-specific or are shared between domains of life can uncover the tripartite division of cellular life, exclude chimeric scenarios of origin, and provide initial insight on the rooting of the ToL (Figure 6).

The number of unique protein folds observed in nature is very small. SCOP ver. 1.75 defines only ~2,000 fold superfamilies (FSFs), groups of homologous domains unified on the base of common structural and evolutionary relationships, for a total of 110,800 known domains in proteins [113, 119]. FSFs represent highly conserved evolutionary units that are suitable for studying organismal diversification [87]. Yafremava et al. [120] plotted the total number of distinct FSFs (FSF diversity) versus the average reuse of FSFs in the proteome of an organism (FSF abundance). This exercise uncovered a scaling behavior typical of a Benford distribution with a linear regime of proteomic growth for microbial organisms and a superlinear regime for eukaryotic organisms (Figure 8 in [120]). These same scaling patterns are observed when studying the relationship between open reading frames and genome size [121]. Remarkably, archaeal and eukaryal proteomes exhibited both minimum and maximum levels of FSF abundance and diversity, respectively. Bacterial proteomes however showed intermediate levels. We note that the general scaling behavior is consistent with a scenario in which evolutionary diversification proceeds from simpler proteomes to the more complex ones in gradual manner, supporting the principle of spatiotemporal continuity and revealing the nested phylogenetic hierarchies of organisms. Under this scenario (and results of [120]), the streamlined archaeal proteomes represent the earliest form of cellular life. Remarkably, archaeal species harboring thermophilic and hyperthermophilic lifestyles encoded the most streamlined FSF repertoires (Figure 7). Clearly, modern thermophilic archaeons are most closely related to the ancient cells that inhabited planet Earth billions of years ago (also read below).

FSF domain distributions in the genomes of the three domains of life provide further insights into their evolution. Nasir et al. [122, 123] generated Venn diagrams to illustrate FSF sharing patterns in the genomes of Archaea, Bacteria, and Eukarya (Figure 6(a)). These diagrams display the total number of FSFs that are unique to a domain of life (taxonomic groups A, B, and E), shared by only two (AB, BE, and AE), and

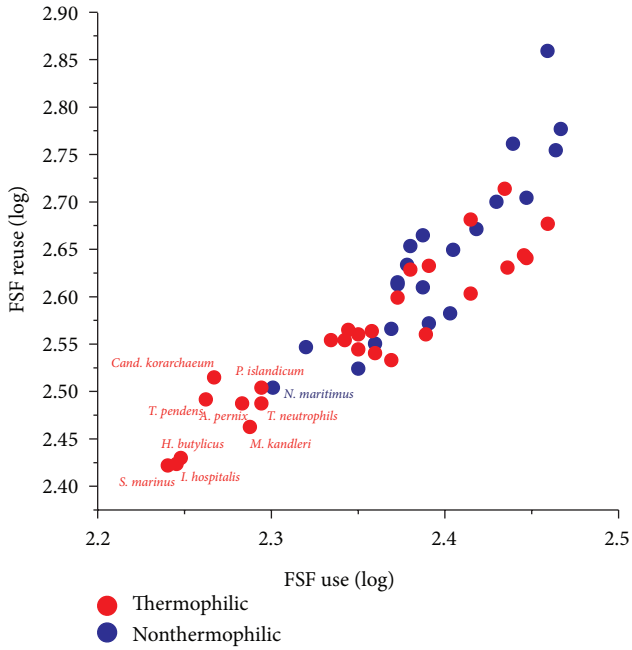


Figure 7: A plot of FSF use (diversity) against FSF reuse (abundance) reveals a linear pattern of proteomic growth in 48 archaeal proteomes. Thermophilic archaeal species occupy positions that are close to the origin of the plot. They also populate the most basal branch positions in ToLs (see discussion in the main text). Both axes are in logarithmic scale.

those that are universal (ABE). About half of the FSFs (786 out of 1,733) were present in all three domains of life, 453 of which were present in at least 60% of the organisms that were examined ($f > 0.6$, where f is the distribution index ranging from 0 to 1). The fact that about 70% of widely shared FSFs (672) belong to the ABE taxonomic group strongly supports a common evolutionary origin for cells. Evolutionary timelines have confirmed that a large number of these universal FSFs were present in the ancestor of the three domains of life, the urancestor (Figure 1), and were for the most part retained in extant proteomes [48].

Interestingly, the number of BE FSFs was ~10-fold greater than AE and AB FSFs (324 versus 38 and 38) (Figure 6(a)). The finding that Bacteria and Eukarya encode a significantly large number of shared FSF domains is remarkable and hints towards an unprecedented strong evolutionary association between these two domains of life. Moreover, a significant number of BE FSFs is widespread in the bacterial and eukaryotic proteomes; 56 of the 324 FSFs are shared by more than 60% of the bacterial and eukaryal organisms that were analyzed (Figure 6(a)). This is strong evidence of an ancient vertical evolutionary trace from their mutual ancestor (anticipated in [123]). This trace uniquely supports the archaeal rooting of the ToL. In turn, the canonical and eukaryotic rooting alternatives are highly unlikely as none alone can explain the remarkable diversity of the BE taxonomic group (as well as its ancient origin; [123]). The remarkable vertical trace of the ABE taxonomic group and the negligible vertical trace of the AB group (only one FSF shared by more than 60% of organisms) also falsify fusion hypotheses responsible for a putative chimeric makeup of Eukarya. In light of these findings, the most parsimonious explanation of comparative genomic data is that Archaea is the first domain of diversified cellular life.

The patterns of FSF sharing are further strengthened by the genomic distributions of terminal-level molecular function GO terms (Figure 6(b)). The three domains of life shared about a quarter (526) of the 1,924 GO terms that were surveyed. A total of 232 of these ABE terms were shared by 60% of organisms analyzed. Again, the fact that about 76% of widely shared GO terms (306) belong to the ABE taxonomic group strongly supports a common vertical trace, while the finding that Bacteria and Eukarya share a significantly large number of GO terms (272) supports again the archaeal rooting of the ToL. An alternative explanation for the very large size of BE could be large-scale metabolism-related gene transfer from the ancestors of mitochondria and plastids to the ancestors of modern eukaryotes [124]. However, we note that the BE group is not restricted to only metabolic functions. It also includes FSFs and GOs involved in intracellular and extracellular processes and regulation and informational functions [125]. Thus the very large size of BE is a significant outcome most likely shaped by vertical evolutionary scenarios and cannot solely be explained by parasitic/symbiotic relationships that exist between Bacteria and Eukarya [123]. Moreover, the simplicity of archaeal FSF repertoires is not due to the paucity of available archaeal genomic data. We confirmed that the mean FSF coverage (i.e., number of proteins/annotated to FSFs/GOs out of total) for Archaea, Bacteria, and Eukarya was largely comparable (e.g., Table S1 in [122]). Finally, as we will describe below, these comparative patterns and tentative conclusions are confirmed by phylogenomic analysis [91, 122, 126]. Congruent phylogenies were obtained with (Figure 4(c)) and without equal and random sampling of taxa. Thus, the relatively low number of archaeal genomes is also not expected to compromise our inferences.

We note that Eukarya shares many informational genes with Archaea and many operational genes with Bacteria. While informational genes have been thought more refractory to HGT than noninformational genes, we have confirmed at GO level that this is not the case. A ToL built from GO terms showed that in fact noninformational terms were less homoplasious, while statistical enrichment analyses revealed that HGT had little if any functional preference for GO terms across GO hierarchical levels. We recently compared ToLs reconstructed from non-HGT GO terms and ToLs reconstructed from informational GO terms that were extracted from non-HGT GO terms (Kim et al. ms. in review; also read below). In both cases, Archaea appeared as a basal paraphyletic group of the ToLs and the common origin of Bacteria and Eukarya was maintained. Thus, the 272 GO terms shared by Bacteria and Eukarya harbor a strong vertical trace.

Another observation often used as support for Archaea-Eukarya kinship is the discovery of few eukaryote-specific proteins (e.g., actin, tubulin, H3, H4, ESCRT, ribosomal proteins, and others) in some archaeal species [127] and

their complete absence from Bacteria (except for documented tubulin HGT from eukaryotes to bacterial genus *Prosthecobacter*; [128]). This suggests either that eukaryotes arose from an archaeal lineage [127] or that the ancestor of Eukarya and Archaea was complex and modern Archaea are highly reduced [25]. Indeed, few eukaryote-specific proteins have now been found in some archaeal species (in some cases just one species!). We argue that this poor spread cannot be taken as evidence for the Archaea-Eukarya sister relationship. This needs to be confirmed by robust phylogenetic analysis, which is unfortunately not possible when using protein sequences. Recently, we described a new strategy for inferring vertical and horizontal traces [123]. This method calculates the spread of FSFs and GOs that are shared between two-superkingdom groups (i.e., AB, AE, and BE). Balanced distributions often indicate vertical inheritance while biased distributions suggest horizontal flux. For example, penicillin binding molecular activity (GO: 0008658) was present in 100% of the sampled bacterial proteomes but was only present in 11% of the archaeal species [123]. Thus, presence of GO: 0008658 in Archaea was attributed to HGT gain from Bacteria. Using this simple method we established that both Bacteria and Eukarya were united by much stronger vertical trace than either was to Archaea. In fact, strong reductive tendencies in the archaeal genomes were recorded [123]. Thus, in our opinion, presence of eukaryote-specific proteins in only very few archaeal species could in fact be an HGT event that is not detectable by sequence phylogenies.

In turn, many arguments favor now the Bacteria-Eukarya sisterhood in addition to the structure-based phylogenies and balanced distribution of molecular features. For example, Bacteria and Eukarya have similar lipid membranes that can be used as argument for their evolutionary kinship. Moreover, Archaea fundamentally differ from Eukarya in terms of their virosphere. Viruses infecting Archaea and Eukarya are drastically different, as recently discussed by Forterre [25].

The genomic distribution of RNA families was taken from Hoeppner et al. [129] and also shows a vertical evolutionary trace in five crucial Rfam clans that are universal, including tRNA, 5S rRNA, subunit rRNA, and RNase P RNA. These universal RNA groups are likely minimally affected by HGT. However, 99% of Rfam clans and families are specific to domains of life and only 11 of the 1,148 groups were shared at $f > 0.6$ levels. This clearly shows that the functional complexity of RNA materialized very late during organismal diversification and that it is not a good genomic feature for exploring the rooting of the ToL. Only five RNA families are shared between two domains of life, and only one of these does so at $f > 0.6$, the G12 pseudoknot of the 23S rRNA, which is present in bacterial and eukaryotic organellar rRNA. While the large subunit rRNA scaffold supports the G12 pseudoknot with its vertical trace, all five interdomain RNA families can be explained most parsimoniously by HGT. Thus, only a handful of ancient and universal RNA species can be used to root the ToL.

We end by noting that the Venn diagrams consistently show that Archaea harbors the least number of unique (A) and shared (AB and AE) FSFs, GO terms or RNA families. This trend supports an early divergence of this domain of life from the urancestor and the possibility that such divergence be shaped by evolutionary reductive events. We reason that such losses would be more parsimonious early on in evolution than in the later periods. This is because genes often increase their abundance in evolutionary time (by gene duplications, HGT, and other processes). Thus it is reasonable to think that loss of an ancient gene would be more feasible very early in evolution relative to losing it very late.

6.2. Phylogenomic Evidence from the Sequence and Structure of RNA. Phylogenetic analyses of the few RNA families that are universal and display an important vertical evolutionary trace (Figure 6(c)) provide compelling evidence in favor of an early evolutionary appearance of Archaea [105, 107, 108, 130–135]. Here we briefly summarize evidence from tRNA, 5S rRNA, and RNase P RNA. Unpublished analyses of rRNA sequence and structure using advanced phylogenetic methods also show that Archaea was the first domain of life.

tRNA molecules are generally short (~73–95 nucleotides in length) and highly conserved. Consequently, their sequence generally contains limited amount of phylogenetic information. These limitations have been overcome by analyzing entire tRNomes [136], which extend the length of an organismal set of tRNAs to over 2,000 bases. Xue et al. [130, 131] analyzed the genetic distances between tRNA sequences as averages between alloacceptor tRNAs from diverse groups of tRNomes using multiple molecular base substitution models. The distances were mapped onto an unrooted phylogeny of tRNA molecules (Figure 8(a)). The "arrow of time" assumption in these studies is that ancient tRNA paralogs closely resemble each other when lineages originate close to the time of the gene duplication. Remarkably, the results revealed a paraphyletic rooting of the ToL in Archaea. The root was specifically located close to the hyperthermophilic methanogen *Methanopyrus kandleri* (Figure 8(a)). The hypothesis of this specific rooting scenario has been supported by several other studies [134, 137], including a study of genetic distances between paralogous pairs of aminoacyl-tRNA synthetase (aaRS) proteins [131]. A remarkable match between distance scores of tRNA and pairs of aaRS paralogs (Figure 8(b)) not only confirms the early appearance of Archaea but also suggests a coevolutionary trace associated with molecular interactions that are responsible for the genetic code. In fact, a recent exhaustive phylogenomic analysis of tRNA and aaRS coevolution explicitly reveals the origins and evolution of the genetic code and the underlying molecular basis of genetics [59]. Di Giulio [132, 133] has also proposed an archaeal rooting of the tree of life, specifically in the lineage leading to the phylum of Nanoarchaeota. This rooting is based on unique and ancestral genomic traits of *Nanoarchaeum equitans*, split genes separately codifying for the 5′ and 3′ halves of tRNA and the absence of operons, which are considered molecular fossils [132]. However, this claim needs additional support as contrasting evidence now recognizes *N. equitans* as a highly derived archaeal species [138].

In addition to sequences, structural features of tRNA molecules also support the archaeal rooting of the ToL.

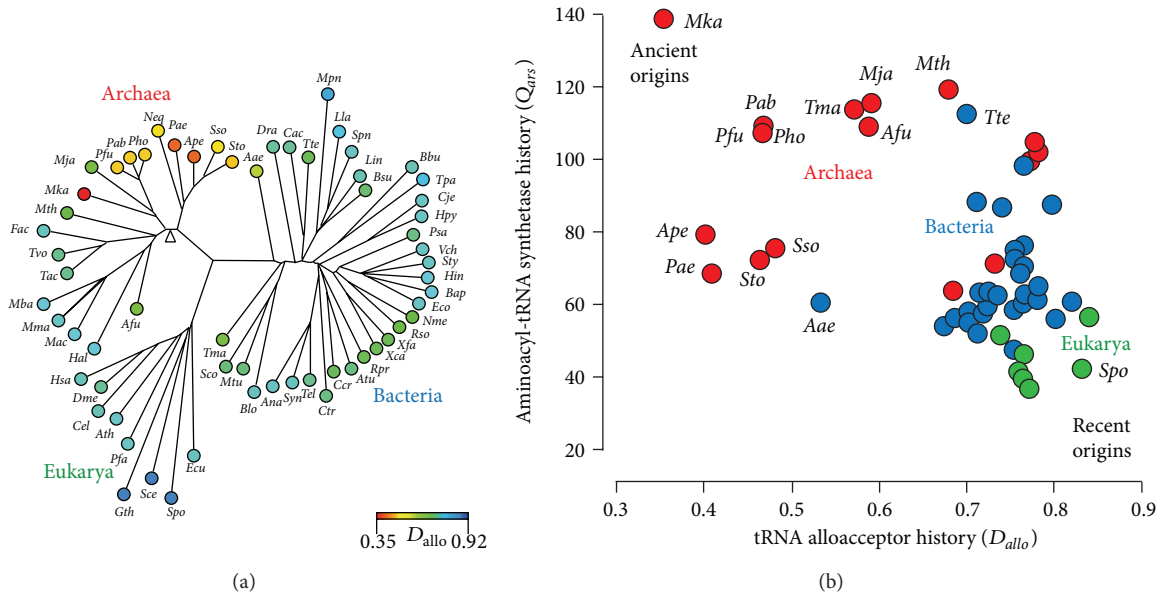
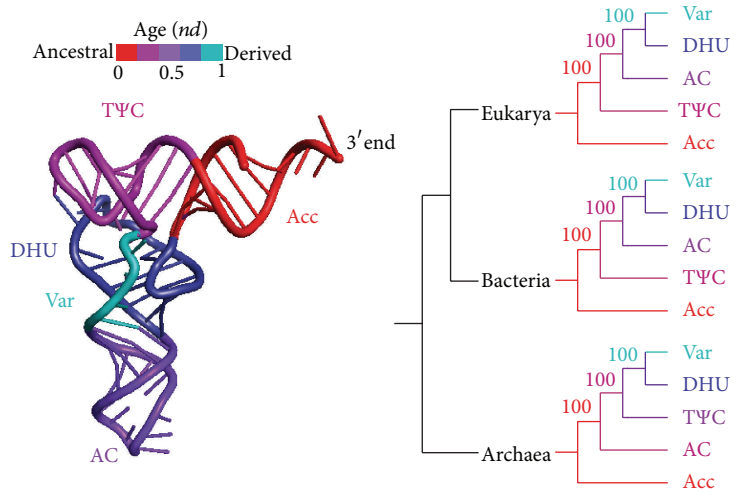


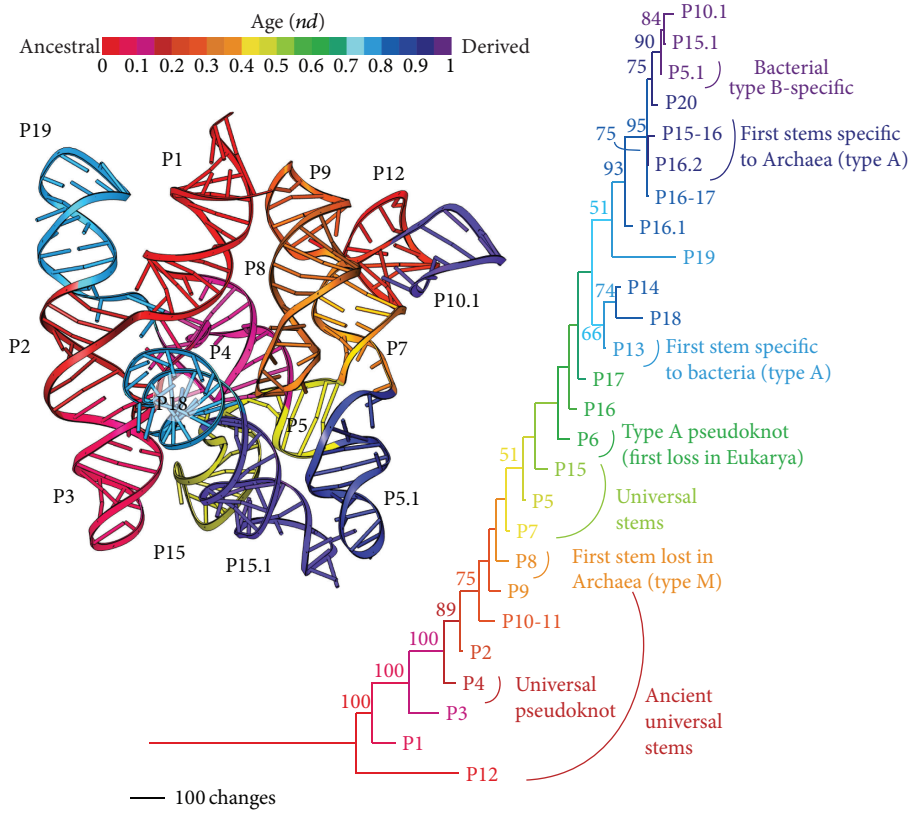
FIGURE 8: Ancient phylogenetic signal in the sequence of tRNA and their associated aminoacyl-tRNA synthetase (aaRS) enzymes. (a) Unrooted ToL derived from tRNA sequences with their alloacceptor D_{allo} distances traced in thermal scale (from [130]). D_{allo} is average of pairwise distances for 190 pairs of tRNA isoacceptors. These distances measure pairwise sequence mismatches of tRNAs for every genome and their values increase for faster evolving sequences of species of more recent origin. (b) Coevolution of aaRSs and their corresponding tRNA. Genetic distances between the top 10 potentially paralogous aaRS pairs estimated using BLASTP define a measure (Q_{ars}) of how closely the proteins resemble each other in genomes (from [131]). Larger Q_{ars} scores imply more ancestral and slowly evolving protein pairs. The plot of Q_{ars} scores against D_{allo} distances reveals a hidden correlation between the evolution of tRNA and aaRSs and the early origin of Archaea (larger Q_{ars} and lower D_{allo} distances).

The application of RNA structural evidence in phylogenetic studies [50, 139–141] has multiple advantages over sequence data when studying ancient events, especially because RNA structures are far more conserved than sequences. This has been demonstrated in a phylogenetic approach that uses RNA structural information to reconstruct evolutionary history of macromolecules such as rRNA [29, 50], tRNA [107, 108], 5S rRNA [105], RNase P RNA [135], and SINE RNA [142]. Geometrical and statistical properties of structure (e.g., stems or loops commonly found in the secondary structures of RNA molecules) are treated as linearly ordered multistate phylogenetic characters. In order to build rooted trees, an evolutionary tendency toward conformational order is used to polarize change in character state transformations. This defines a hypothetical ancestor with which to root the ingroup using the Lundberg method. Reconstructed phylogenies produce trees of molecules and ToLs (e.g., Figure 4(a)) or trees of substructures that describe the gradual evolutionary accretion of structural components into molecules (Figure 9). For example, phylogenetic trees of tRNA substructures define explicit models of molecular history and show that tRNA originated in the acceptor stem of the molecule [108]. Remarkably, trees reconstructed from tRNA drawn from individual domains of life demonstrate that the sequence of accretion events occurred differently in Archaea than in Bacteria and Eukarya, suggesting a sister group evolutionary relationship between the bacterial and eukaryotic domains (Figure 9(a)). A similar result obtained from trees of 5S rRNA substructures revealed different molecular accretion sequences of archaeal molecules when these were compared to bacterial and eukaryal counterparts [105], confirming again in a completely different molecular system the history of the domains of life.

An analysis of the structure of RNase P RNA also provides similar conclusions [135]. While a ToL reconstructed from molecular structure placed type A archaeal molecules at its base (a topology that resembles the ToL of 5S rRNA), a tree of RNase P RNA substructures uncovered the history of molecular accretion of the RNA component of the ancient endonuclease and revealed a remarkable reductive evolutionary trend (Figure 9(b)). Molecules originated in stem P12 and were immediately accessorized with the catalytic P1–P4 catalytic pseudoknotted core structure that interacts with RNase P proteins of the endonuclease complex and ancient segments of tRNA. Soon after this important accretion stage, the evolving molecule loses its first stem in Archaea (stem P8), several accretion steps earlier than the first loss of a stem in Eukarya or the first appearance of a Bacteria-specific stem. These phylogenetic statements provide additional strong support to the early origin of the archaeal superkingdom prior to the divergence of the shared common ancestor of Bacteria and Eukarya. As we will discuss below, the early loss of a structure in the molecular accretion process of a central and ancient RNA family is significant. It suggests



(a)



(b)

Figure 9: The history of accretion of tRNA and RNase P RNA substructures reveals the early evolutionary appearance of Archaea. (a) Rooted trees of tRNA arm substructures reveal the early appearance of the acceptor arm (Acc) followed by the anticodon arm (AC) in Archaea or the pseudouridine (TΨC) arm in both Bacteria and Eukarya (from [108]). The result confirms the sister group relationship of Bacteria and Eukarya. (b) Trees of molecular substructures of RNase P RNAs were reconstructed from characters describing the geometry of their structures (from [135]). Branches and corresponding substructures in a 3D atomic model are colored according to the age of each substructure (*nd*, node distance). Note the early loss of stem P8 in Archaea immediately after the evolutionary assembly of the universal functional core of the molecule.

that the emerging archaeal lineages were subjected to strong reductive evolutionary pressures during the early evolution of a very ancient RNA molecule.

6.3. Phylogenomic Evidence from Protein Domain Structure. G. Caetano-Anollés and D.Caetano-Anollés [41] were the first to utilize protein domain structures as taxa and to reconstruct trees of domains (ToDs) describing their evolution. Figure 10 shows a rooted ToD built from a census of FSF structures in 981 genomes (data taken from [122, 123]). These trees are unique in that their terminal leaves represent a finite set of component parts [87]. These parts describe at global level the structural diversity of the protein world. When building trees of FSFs, the age of each FSF domain structure can be calculated from the ToDs by simply calculating a distance in nodes between the root of the tree and its corresponding terminal leaf. This is possible because ToDs exhibit highly unbalanced (pectinate) topologies that are the result of semipunctuated processes of domain appearance and accumulation. This node distance (*nd*), rescaled from 0 to 1, provides a relative timescale to study the order of FSF appearance in evolutionary history [126, 143]. Wang et al. [144] showed that *nd* correlates linearly with geological time and defines a global molecular clock of protein folds. Thus, *nd* can be used as a reliable proxy for time. Plotting the age of FSFs in each of the seven taxonomic groups confirmed evolutionary statements we had previously deduced from the Venn diagrams of Figure 6. The ABE taxonomic group included the majority of ancient and widely distributed FSFs. This is expected. In the presence of a strong vertical trace, molecular diversity must delimit a nested taxonomic hierarchy. The ABE group was followed by the evolutionary appearance of the BE group, which preceded the first domain-specific structures, which were Bacteria-specific (B). Remarkably, Archaea-specific (A) and Eukarya-specific (E) structures appeared concurrently and relatively late. These general trends, captured in the box plots of Figure 10, have been recovered repeatedly when studying domain structures at various levels of structural complexity, from folds to fold families [91, 126], when using CATH or SCOP structural definitions [122, 145] or when exploring the evolution of terminal GO terms.

While the early rise of BE FSFs supports the early divergence of Archaea from the urancestor, the very significant trend of gradual loss of structures occurring in the lineages of the archaeal domain and the very late appearance of Archaea-specific structures (e.g., [126]) demand explanation. Since ancient BE FSFs are widely distributed in proteomes (Figure 10), they cannot arise from separate gains of FSFs in Bacteria and Eukarya or by processes of horizontal spread of structures. This was already evident from the Venn diagrams of Figure 6. Moreover, the BE sisterhood to the exclusion of Archaea was further supported by the inspection of FSFs involved in lipid synthesis and transport (Table 1). Membrane lipids are very relevant to the origins of diversified life ([146] and references therein). Bacteria and Eukarya encode similar lipid membranes while archaeal membranes have different lipid composition (isoprenoid ethers). To check if lipid synthesis was another BE synapomorphy, we identified 17 FSFs that were involved in lipid metabolism and transport

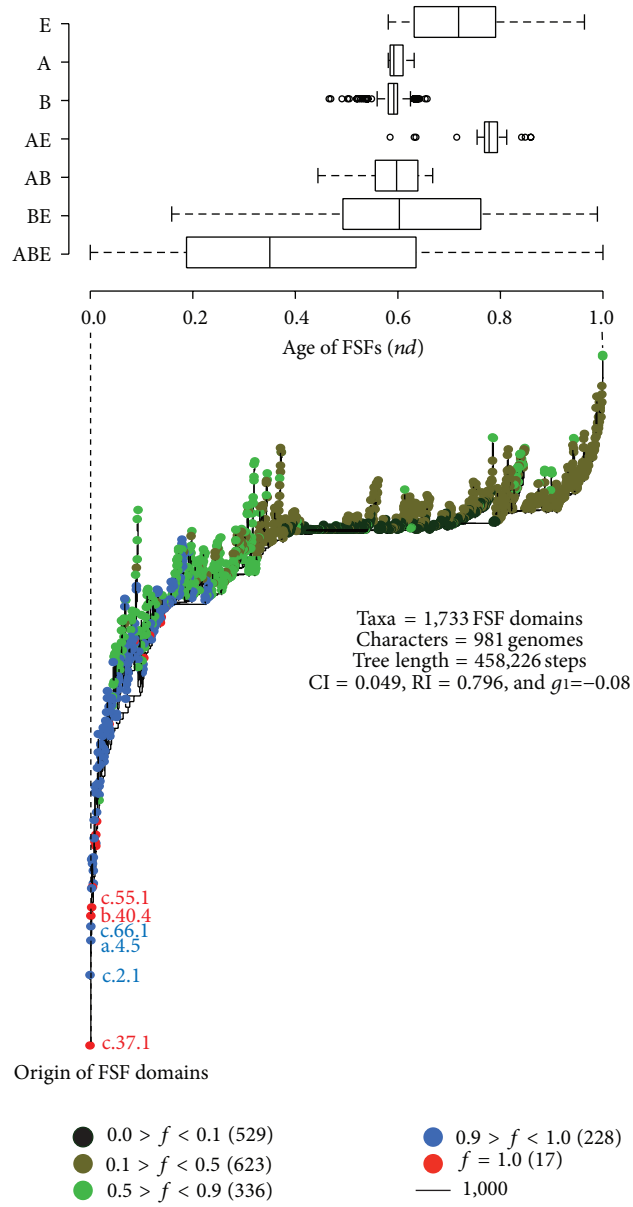


Figure 10: Phylogenomic tree of domains (ToD) describing the evolution of 1,733 FSF domain structures. Taxa (FSFs) were colored according to their distribution (f) in the 981 genomes that were surveyed and used as characters to reconstruct the phylogenomic tree (data taken from [122, 123]). The most basal FSFs are labeled with SCOP alphanumeric identifiers (e.g., c.37.1 is the P-loop containing nucleoside triphosphate hydrolase FSF). Boxplots display the age (*nd* value) distribution of FSFs for the seven possible taxonomic groups. *nd* values were calculated directly from the tree [122] and define a timeline of FSF innovation, from the origin of proteins (*nd* = 0) to the present (*nd* = 1). The group of FSFs that are shared by the three domains of life (ABE) is the most ancient taxonomic group, which spans the entire time axis and their FSFs are widely distributed in genomes. The appearance of the BE group coincides with the first reductive loss of an FSF in Archaea. FSF structures specific to domains of life appear much later in evolution.

Table 1: List of FSFs involved in lipid metabolism and transport along with taxonomic distribution (data taken from [122, 123]).

Group	SCOP Id	FSF Id	FSF description
ABE	89392	b.125.1	Prokaryotic lipoproteins and lipoprotein localization factors
ABE	53092	c.55.2	Creatinase/prolidase N-terminal domain
ABE	49723	b.12.1	Lipase/lipooxygenase domain (PLAT/LH2 domain)
ABE	54637	d.38.1	Thioesterase/thiol ester dehydrase-isomerase
ABE	63825	b.68.5	YWTD domain
BE	47027	a.11.1	Acyl-CoA binding protein
BE	48431	a.118.4	Lipovitellin-phosvitin complex, superhelical domain
BE	55048	d.58.23	Probable ACP-binding domain of malonyl-CoA ACP transacylase
BE	56968	f.7.1	Lipovitellin-phosvitin complex; beta-sheet shell regions
BE	58113	h.5.1	Apolipoprotein A-I
BE	47162	a.24.1	Apolipoprotein
BE	56931	f.4.2	Outer membrane phospholipase A (OMPLA)
B	82220	b.120.1	Tp47 lipoprotein, N-terminal domain
E	47699	a.52.1	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin
E	82936	h.6.1	Apolipoprotein A-II
E	57190	g.3.10	Colipase-like
E	49594	b.7.4	Rab geranylgeranyltransferase alpha-subunit, insert domain

(Table 1). Remarkably, the majority of these FSFs (7 out of 17) were unique to the BE group. In comparison, none were present in either AE or AB groups. The ABE group included five universal FSFs, while one was unique to Bacteria and four were eukarya-specific. In turn, no FSF was unique to Archaea. The BE FSFs cannot be explained by modern effects impinging on variations in proteomic accumulation in FSFs or by processes of domain rearrangement, since these appear in the protein world quite late in evolution [92]. The only and most-parsimonious explanation of the patterns of FSF distribution that unfold in the ToD is the very early (and protracted) rise of the archaeal domain by processes of reductive evolution, possibly triggered by the adaptation of urancestral lineages to harsh environments and survival modes. Under extremophilic conditions typical of hyperthermophilic environments, considerable investments of matter-energy and information must be made for protein persistence [120]. This puts limits on viable protein structures [147]. Extremophilic environments will thus poise the maintenance of a limited set of FSFs for persistence of emergent diversified lineages. This would induce a primordial episode of reductive evolution in the growing FSF repertoire, explaining why hyperthermophilic and thermophilic archaeal species hold the most reduced proteomes (Figure 7). It would also explain the biases that exist in FSFs, GO terms and RNA families (Figure 6), and the placement of hyperthermophilic and thermophilic archaeal species at the base of ToLs. Since Archaea populate the oceans and sometimes rival in number Bacteria in those environments, we further interpret the late appearance of Archaea-specific FSFs as the result of late colonization of these mild environments by both ancient archaeons and emerging eukaryotes. This relaxes primordial extremophilic pressures on protein structures and enables the late archaeal exploration of structural flexibility and functional novelty.

6.4. Full Circle: Evidence from Trees of Proteomes and Functionomes and a Tree Derived from the Distribution of Viral Replicons in Superkingdoms. While we distrust trees of systems, especially ToLs built from sequences, the use of molecular structure at high levels of structural abstraction has the potential to mitigate some limitations of sequence analysis [93]. For example, rooted ToLs built from abundance counts of domain structures and terminal GO terms in the genomes of free-living organisms describe the evolution of proteomes (e.g., [91]). All ToL reconstructions of these kinds approximate the physiology of living organisms, dissect the three primary domains of life, and reveal the early paraphyletic origin of extremophilic archaeal lineages, followed by the late appearances of monophyletic Bacteria and Eukarya. These patterns have been reliably recovered with datasets of varying sizes irrespective of the structural classification scheme [91, 122, 126, 145]. Even a tree reconstructed from the distribution of 2,662 viral replicons in superkingdoms from an exhaustive comparative genomic analysis of viral genomes showed the basal placement of Archaea and the sister taxa relationship between Bacteria and Eukarya (Figure 11).

6.5. Additional Evidence from Comparative Genomics. The uneven distribution of protein domain structures in the world

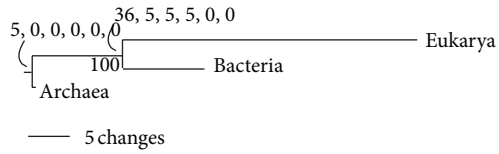


Figure 11: One of three optimal phylogenetic tree reconstructions with identical topologies recovered from an exhaustive maximum parsimony search (58 steps; CI = 1; RI = 1; HI = 0; RC = 1; $g_1 = -0.707$) of the abundance of viral replicon types of dsDNA, ssDNA, dsRNA, ssRNA(+), ssRNA(−), and retrotranscribing viruses. Abundance was scored on a 0–20 scale and ranged from 0 to 759 viral replicons. Vectors of abundance reconstructions in internal nodes are given as percentage of total abundance of replicons in superkingdoms. Boostrap support values are shown below nodes.

of proteomes (Figure 6) is preserved as we climb in the structural hierarchy. This was recently made evident when studying the evolution of CATH domains [145]. The Venn diagrams of Figure 12 show how domain structures in all taxonomic groups decrease in numbers with increases in evolutionary conservation. At the highest CATH architectural level, there were 32 universal architectures but no domain-specific architectures. The four architectures shared by Bacteria and Eukarya were present in at least 60% of the proteomes that were surveyed. The other two interdomain architectures were topological designs of considerable complexity that were poorly shared between proteomes and were evolutionarily derived. They were the *clam* architectures lost in Eukarya and the *box* architectures of nucleotide excision repair shared by Archaea and Eukarya. The most parsimonious corollary of these distribution patterns is that the BE taxonomical group must arise by loss of structures in Archaea. Indeed, ToDs describing the evolution of CATH domain structures again confirm the early appearance of BE structures and consequently their loss in Archaea.

7. Paraphyletic Origins: Grades and Clades in Archaeal History

Saying that "a ToL is rooted in a domain of life" is an incorrect statement that comes from phylogenetic methodology (the use of outgroups) and the tendency to look at the past with modern eyes. Clades in ToLs have been rooted relative to each other by generating unrooted trees and by defining extant organisms as outgroup taxa. The ToL however must be considered rooted in the urancestor of cellular life (Figure 1). This planted edge that connects to the ingroup of extant organisms represents a cellular state in which productive diversification (in the sense of successful lineages) was absent. The primordial urancestral edge leads to a "phase transition," the last universal cellular ancestor (LUCA), of which little is known. The physiology of the urancestor cannot be considered linked to that of any extant organism, even if it shared a common molecular core with all of them. The urancestor was not an archeon, a bacterium, or a eukaryote [148]. It was not necessarily thermophilic. Perhaps it was a communal entity or a megametaorganism in the sense of a modern syncytium (the result of multiple cell fusions) and a modern coenocyte (the result of multiple cell divisions). The organismal boundaries were likely present, judging by the number of widely distributed protein domains that associate with membranes and appear at the base of our ToDs and by the universal existence of acidocalcisome organelles [149]. However, the molecular makeup of the urancestral cells was most likely fluid and quasistatistical; the repertoire distributed unequally in the urancestral populations of communal parts, of course, within confines delimited by persistence. This urancestral population is therefore consistent with the idea of a primordial stem line proposed by Kandler and Woese [150, 151]. However, it was relatively richer in molecular structures and functions as opposed to the simple cellular systems hypothesized by Woese. This richness is confirmed by modern analyses of proteomes and functionomes that reveal vast number of universally shared protein domains and GOs among three superkingdoms. While each syncytial/coenocyte element of the megaorganism exchanged component parts in search of cellular stability and persistence, the process could not be equated with modern HGT. The exchanging community of primordial cells was not cohesive enough to make the horizontal exchange meaningful. Macromolecules most likely established loose and diverse associations with each other and with smaller molecules, limited by the short average life of their unevolved structural conformations. With time, molecules with better-optimized properties engaged in more durable interactions, stabilizing the emergent cells and providing increased cellular cohesiveness. This poised the urancestral community towards a phase transition (a crystallization; [148]), a point in which cellular groups had distinct properties and could be individuated. We believe this was the time of the origin of the archaeal lineage 2.9 billion years ago [48].

At the base of the ToLs that were reconstructed from genomic data, basal archaeal taxa arise as paraphyletic lineages (Figures 4(a) and 4(c)). These lineages likely arose from subgroups of the urancestral population that pervasively lost crucial domain structures and molecular functions. This represents an evolutionary grade under the scenario described above. The emerging lineages shared with the urancestral community a unifying condition that was related to archaic biochemistry. In other words, the urancestral and emerging archaeal lineages expressed fundamental structural and functional equivalences in terms of their repertoires, but revealed in each emerging and durable paraphyletic lineage a handful of distinct newly developed traits. These traits could be global, such as increased thermostability of some crucial members of the protein repertoire or change in the membrane makeup, or local, such as the selective loss of crucial structures and functions. Figure 13 uses the tree paradigm to portray the structural and functional equivalences of the urancestral and emerging archaeal lineages and the slow progression from grades to clades.

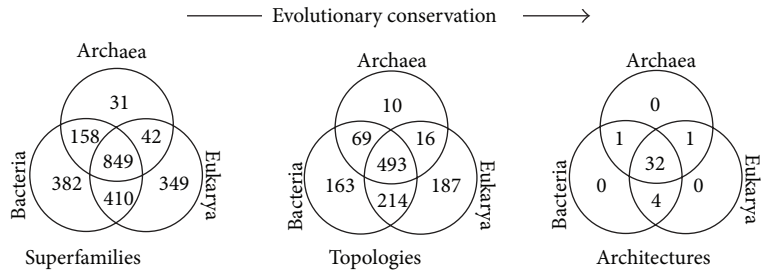


Figure 12: Venn diagrams displaying the distributions of 2,221 homologous superfamilies, 1,152 topologies, and 38 architectures of CATH domains in the proteomes of 492 fully sequenced genomes (from [145]). All distributions highlight maximum sharing in the ancient ABE and BE taxonomic groups and minimal sharing in archaeal taxonomic groups.

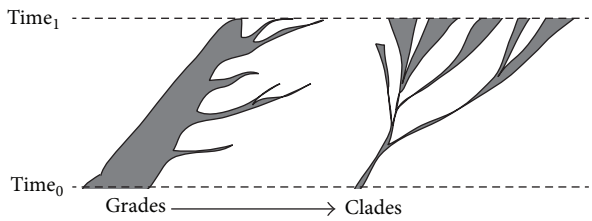


Figure 13: From grades to clades. The cartoon describes a possible progression of modes of organismal diversification during the rise of primordial archaeal lineages. The width of emerging lineages is proportional to uniquely identifying features of physiological and molecular complexity.

8. Through the Wormhole: The Makeup of the "Megaorganism" and the Emerging Archaeal Lineages

Character state reconstructions of proteome repertoires derived from ToLs coupled to the timelines of ToDs provide an effective way to define the ancestral protein domain complement of the urancestor [48] and, consequently, the likely makeup of the emerging archaeal lineages. The urancestral proteome possessed a lower bound of ~70 FSF domain structures, 75% of which were composed of α/β and $\alpha + \beta$ proteins. About 50% of FSFs were part of metabolic enzymes, including a rich toolkit of transferases and enzymes of nucleotide metabolism. The rest of domains were involved in functions related to information (translation, replication, and repair), intracellular processes (transport, protein modification, and proteolytic activities), regulation (kinases/phosphatases and DNA binding functions), and small molecule binding. The urancestor had a limited repertoire of aaRSs and translation factors. It contained a primordial ribosome with a limited core of universal ribosomal proteins. It had numerous membrane proteins necessary for transport, including a relatively advanced ATP synthetase complex, and structures necessary for cellular organization (filaments and primordial cytoskeletal structures). The cells lacked enzymes for deoxyribonucleotide production, so it is likely that the cellular urancestor itself did not harbor a DNA genome. The cells lacked functions related to extracellular processes (cell adhesion, immune response, and toxins/defense) and cellular motility, suggesting an ancient living world without competitive strategies of survival.

9. Conclusions

The rooting of the ToL has been always controversial in evolutionary biology [26, 152, 153]. While it is popularly accepted that the ToL based on sequence phylogenies is rooted in the akaryotes and that Archaea and Eukarya are sister groups to each other, only two of the three main steps of phylogenetic analysis [104] have been partially fulfilled with sequences. This includes selecting an appropriate statistical or nonstatistical evolutionary model of character change and an optimization method for phylogenetic tree reconstruction. However, no adequate method exists for character polarization that identifies ancestral and derived character states in sequences. In the absence of robust polarization methodology, any statement about the rooting of the ToL should be considered suspect or subject of apriorism. Here we show that information derived from a genomic structural and functional census of millions of encoded proteins and RNAs coupled with process models that comply with Weston's generality criterion provide the means to dissect the origins of diversified life. The generality criterion is fulfilled in these studies by focusing on the accumulation of modules such as protein domain structures, elements of RNA substructures, or ontogenetic definitions of molecular function. In general, these features are the subject of accretion processes that comply with additive phylogenetic change within the nested taxonomic hierarchy and result in changes of abundance. These processes include those responsible for the growth of molecules (e.g., multidomain proteins), molecular ensembles (e.g., the ribosome), and molecular repertoires (e.g., proteomes). The new methods unfold a consistent evolutionary scenario in which the origin of diversified life traces back to the early history of Archaea. Remarkably, the archaic origin of this microbial urkingdom now does justice to its name.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank members and friends of the GCA laboratory for fruitful discussions. This research has been supported by grants from the National Science Foundation (MCB-0749836 and OISE-1132791) and the United States Department of Agriculture (ILLU-802-909 and ILLU-483-625) to Gustavo Caetano-Anollés and from KRIBB Research Initiative Program and the Next-Generation BioGreen 21 Program, Rural Development Administration (PJ0090192013), to Kyung Mo Kim.

References

[1] C. R. Woese, "A new biology for a new century," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 2, pp. 173–186, 2004.

[2] T. Grant, "Testing methods: the evaluation of discovery operations in evolutionary biology," *Cladistics*, vol. 18, no. 1, pp. 94–111, 2002.

[3] E. O. Wiley, "Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists," *Systematic Zoology*, vol. 24, no. 2, pp. 233–243, 1975.

[4] J. S. Farris, "Parsimony and explanatory power," *Cladistics*, vol. 24, no. 5, pp. 825–847, 2008.

[5] G. Caetano-Anollés, K. M. Kim, and D. Caetano-Anollés, "The phylogenomic roots of modern biochemistry: origins of proteins, cofactors and protein biosynthesis," *Journal of Molecular Evolution*, vol. 74, no. 1-2, pp. 1–34, 2012.

[6] W. Fontana, "Modeling "evo-devo" with RNA," *BioEssays*, vol. 24, no. 12, pp. 1164–1177, 2002.

[7] W. Hennig, *Phylogenetic Systematics*, University of Illinois Press, Urbana, Ill, USA, 1999.

[8] D. Fraix-Burnet, T. Chattopadhyay, A. K. Chattopadhyay, E. Davoust, and M. Thuillard, "A six-parameter space to describe galaxy diversification," *Astronomy and Astrophysics*, vol. 545, article A80, 24 pages, 2012.

[9] E. Sober, "The contest between parsimony and likelihood," *Systematic Biology*, vol. 53, no. 4, pp. 644–653, 2004.

[10] E. K. Lienau and R. DeSalle, "Is the microbial tree of life verificationist?" *Cladistics*, vol. 26, no. 2, pp. 195–201, 2010.

[11] J. S. Huxley, "Evolutionary processes and taxonomy with special reference to grades," *Uppsala Universitets Årsskrift*, vol. 6, pp. 21–39, 1958.

[12] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.

[13] P. Forterre, "The universal tree of life and the last universal cellular ancestor: revolution and counterrevolutions," in *Evolutionary Genomics and Systems Biology*, G. Caetano-Anollés, Ed., pp. 43–62, Wiley-Blackwell, Hoboken, NJ, USA, 2010.

[14] J. P. Gogarten, H. Kibak, P. Dittrich et al., "Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 17, pp. 6661–6665, 1989.

[15] N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata, "Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 23, pp. 9355–9359, 1989.

[16] J. P. Gogarten and L. Olendzenski, "Orthologs, paralogs and genome comparisons," *Current Opinion in Genetics and Development*, vol. 9, no. 6, pp. 630–636, 1999.

[17] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[18] J. Martijn and T. J. Ettema, "From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell," *Biochemical Society Transactions*, vol. 41, no. 1, pp. 451–457, 2013.

[19] C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, and T. M. Embley, "The archaebacterial origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 51, pp. 20356–20361, 2008.

[20] P. G. Foster, C. J. Cox, and T. M. Embley, "The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1527, pp. 2197–2207, 2009.

[21] T. A. Williams, P. G. Foster, T. M. Nye, C. J. Cox, and T. M. Embley, "A congruent phylogenomic signal places eukaryotes within the Archaea," *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1749, pp. 4870–4879, 2012.

[22] S. Gribaldo, A. M. Poole, V. Daubin, P. Forterre, and C. Brochier-Armanet, "The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse?" *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 743–752, 2010.

[23] H. Philippe and P. Forterre, "The rooting of the universal tree of life is not reliable," *Journal of Molecular Evolution*, vol. 49, no. 4, pp. 509–523, 1999.

[24] E. Sober and M. Steel, "Testing the hypothesis of common ancestry," *Journal of Theoretical Biology*, vol. 218, no. 4, pp. 395–408, 2002.

[25] P. Forterre, "The common ancestor of Archaea and Eukarya was not an archaeon," *Archaea*, vol. 2013, Article ID 372396, 18 pages, 2013.

[26] P. Forterre and H. Philippe, "Where is the root of the universal tree of life?" *BioEssays*, vol. 21, no. 10, pp. 871–879, 1999.

[27] S. Gribaldo and H. Philippe, "Ancient phylogenetic relationships," *Theoretical Population Biology*, vol. 61, no. 4, pp. 391–408, 2002.

[28] J. K. Harris, S. T. Kelley, G. B. Spiegelman, and N. R. Pace, "The genetic core of the universal ancestor," *Genome Research*, vol. 13, no. 3, pp. 407–412, 2003.

[29] A. Harish and G. Caetano-Anollés, "Ribosomal history reveals origins of modern protein synthesis," *PLoS ONE*, vol. 7, no. 3, Article ID e32776, 2012.

[30] G. Caetano-Anollés and M. J. Seufferheld, "The coevolutionary roots of biochemistry and cellular organization challenge the RNA world paradigm," *Journal of Molecular Microbiology and Biotechnology*, vol. 23, no. 1-2, pp. 152–177, 2013.

[31] N. R. Pace, "Mapping the tree of life: progress and prospects," *Microbiology and Molecular Biology Reviews*, vol. 73, no. 4, pp. 565–576, 2009.

[32] P. de Rijk, Y. van de Peer, I. van den Broeck, and R. de Wachter, "Evolution according to large ribosomal subunit RNA," *Journal of Molecular Evolution*, vol. 41, no. 3, pp. 366–375, 1995.

[33] G. Caetano-Anollés, "Tracing the evolution of RNA structure in ribosomes," *Nucleic Acids Research*, vol. 30, no. 11, pp. 2575–2587, 2002.

[34] J. Mallatt and C. J. Winchell, "Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes," *Molecular Biology and Evolution*, vol. 19, no. 3, pp. 289–301, 2002.

[35] M. Gerstein, "Patterns of protein–fold usage in eight microbial genomes: a comprehensive structural census," *Proteins: Structure, Function, and Bioinformatics*, vol. 33, no. 4, pp. 518–534, 1998.

[36] M. Gerstein and H. Hegyi, "Comparing genomes in terms of protein structure: surveys of a finite parts list," *FEMS Microbiology Reviews*, vol. 22, no. 4, pp. 277–304, 1998.

[37] B. Snel, P. Bork, and M. A. Huynen, "Genome phylogeny based on gene content," *Nature Genetics*, vol. 21, no. 1, pp. 108–110, 1999.

[38] F. Tekaia, A. Lazcano, and B. Dujon, "The genomic tree as revealed from whole proteome comparisons," *Genome Research*, vol. 9, no. 6, pp. 550–557, 1999.

[39] Y. I. Wolf, S. E. Brenner, P. A. Bash, and E. V. Koonin, "Distribution of protein folds in the three superkingdoms of life," *Genome Research*, vol. 9, no. 1, pp. 17–26, 1999.

[40] J. O. Korbel, B. Snel, M. A. Huynen, and P. Bork, "SHOT: a web server for the construction of genome phylogenies," *Trends in Genetics*, vol. 18, no. 3, pp. 158–162, 2002.

[41] G. Caetano-Anollés and D. Caetano-Anollés, "An evolutionarily structural universe of protein architecture," *Genome Research*, vol. 13, no. 7, pp. 1563–1571, 2003.

[42] S. L. Baldauf, A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle, "A kingdom-level phylogeny of eukaryotes based on combined protein data," *Science*, vol. 290, no. 5493, pp. 972–977, 2000.

[43] J. R. Brown, C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope, "Universal trees based on large combined protein sequence data sets," *Nature Genetics*, vol. 28, no. 3, pp. 281–285, 2001.

[44] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life," *Science*, vol. 311, no. 5765, pp. 1283–1287, 2006.

[45] T. Dagan, M. Roettger, D. Bryant, and W. Martin, "Genome networks root the tree of life between prokaryotic domains," *Genome Biology and Evolution*, vol. 2, no. 1, pp. 379–392, 2010.

[46] S. Jun, G. E. Sims, G. A. Wu, and S. Kim, "Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 1, pp. 133–138, 2010.

[47] E. A. Schultes, P. T. Hraber, and T. H. LaBean, "No molecule is an island: molecular evolution and the study of sequence space," in *Algorithmic Bioprocesses*, A. Condon, D. Harel, J. N. Kok, A. Salomaa, and E. Winfree, Eds., pp. 675–704, Springer, Berlin, Germany, 2009.

[48] K. M. Kim and G. Caetano-Anollés, "The proteomic complexity and rise of the primordial ancestor of diversified life," *BMC Evolutionary Biology*, vol. 11, no. 1, article 140, 2011.

[49] V. A. Albert, *Parsimony, Phylogeny, and Genomics*, Oxford University Press, Oxford, UK, 2005.

[50] G. Caetano-Anollés, "Evolved RNA secondary structure and the rooting of the universal tree of life," *Journal of Molecular Evolution*, vol. 54, no. 3, pp. 333–345, 2002.

[51] J. A. Lake, R. G. Skophammer, C. W. Herbold, and J. A. Servin, "Genome beginnings: rooting the tree of life," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1527, pp. 2177–2185, 2009.

[52] B. Guo, M. Zou, and A. Wagner, "Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication," *Molecular Biology and Evolution*, vol. 29, no. 10, pp. 3005–3022, 2012.

[53] M. K. Basu, I. B. Rogozin, O. Deusch, T. Dagan, W. Martin, and E. V. Koonin, "Evolutionary dynamics of introns in plastid-derived genes in plants: saturation nearly reached but slow intron gain continues," *Molecular Biology and Evolution*, vol. 25, no. 1, pp. 111–119, 2008.

[54] J. S. Farris, "Phylogenetic analysis under Dollo's Law," *Systematic Biology*, vol. 26, no. 1, pp. 77–88, 1977.

[55] H. Fang, M. E. Oates, R. B. Pethica et al., "A daily-updated tree of (sequenced) life as a reference for genome research," *Scientific Reports*, vol. 3, article 2015, 2013.

[56] T. Cavalier-Smith, "The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification," *International Journal of Systematic and Evolutionary Microbiology*, vol. 52, no. 1, pp. 7–76, 2002.

[57] T. Cavalier-Smith, "Rooting the tree of life by transition analyses," *Biology Direct*, vol. 1, article 19, 2006.

[58] H. S. Kim, J. E. Mittenthal, and G. Caetano-Anolles, "Widespread recruitment of ancient domain structures in modern enzymes during metabolic evolution," *Journal of Integrative Bioinformatics*, vol. 10, no. 1, article 214, 2013.

[59] G. Caetano-Anollés, M. Wang, and D. Caetano-Anollés, "Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility," *PLoS ONE*, vol. 8, no. 8, Article ID e72225, 2013.

[60] E. Bapteste and C. Brochier, "On the conceptual difficulties in rooting the tree of life," *Trends in Microbiology*, vol. 12, no. 1, pp. 9–13, 2004.

[61] A. M. Poole, "Horizontal gene transfer and the earliest stages of the evolution of life," *Research in Microbiology*, vol. 160, no. 7, pp. 473–480, 2009.

[62] D. Raoult, "The post-Darwinist rhizome of life," *The Lancet*, vol. 375, no. 9709, pp. 104–105, 2010.

[63] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science*, vol. 284, no. 5423, pp. 2124–2128, 1999.

[64] T. Dagan and W. Martin, "The tree of one percent," *Genome Biology*, vol. 7, no. 10, article 118, 2006.

[65] W. F. Doolittle and E. Bapteste, "Pattern pluralism and the tree of life hypothesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 7, pp. 2043–2049, 2007.

[66] E. V. Koonin, "Towards a postmodern synthesis of evolutionary biology," *Cell Cycle*, vol. 8, no. 6, pp. 799–800, 2009.

[67] T. Kloesges, O. Popa, W. Martin, and T. Dagan, "Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths," *Molecular Biology and Evolution*, vol. 28, no. 2, pp. 1057–1074, 2011.

[68] S. Halary, J. W. Leigh, B. Cheaib, P. Lopez, and E. Bapteste, "Network analyses structure genetic diversity in independent genetic worlds," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 1, pp. 127–132, 2010.

[69] S. S. Abby, E. Tannier, M. Gouy, and V. Daubin, "Lateral gene transfer as a support for the tree of life," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 13, pp. 4962–4967, 2012.

[70] N. Glansdorff, Y. Xu, and B. Labedan, "The conflict between horizontal gene transfer and the safeguard of identity: origin of

meiotic sexuality," *Journal of Molecular Evolution*, vol. 69, no. 5, pp. 470–480, 2009.

[71] C. G. Kurland and O. G. Berg, "A hitchhiker's guide to evolving networks," in *Evolutionary Genomics and Systems Biology*, G. Caetano-Anollés, Ed., pp. 361–396, Wiley-Blackwell, Hoboken, NJ, USA, 2010.

[72] L. P. Villarreal and G. Witzany, "The DNA habitat and its RNA inhabitants: at the dawn of RNA sociology," *Genomics Insights*, vol. 6, pp. 1–12, 2013.

[73] S. Gribaldo and C. Brochier, "Phylogeny of prokaryotes: does it exist and why should we care?" *Research in Microbiology*, vol. 160, no. 7, pp. 513–521, 2009.

[74] B. D. Mishler, "The logic of the data matrix in phylogenetic analysis," in *Parsimony, Phylogeny and Genomics*, V. A. Albert, Ed., pp. 57–69, Oxford University Press, New York, NY, USA, 2005.

[75] A. Dress, V. Moulton, M. Steel, and T. Wu, "Species, clusters and the "tree of life": a graph-theoretic perspective," *Journal of Theoretical Biology*, vol. 265, no. 4, pp. 535–542, 2010.

[76] N. Lane and W. Martin, "The energetics of genome complexity," *Nature*, vol. 467, no. 7318, pp. 928–934, 2010.

[77] M. C. Rivera and J. A. Lake, "The ring of life provides evidence for a genome fusion origin of eukaryotes," *Nature*, vol. 431, no. 7005, pp. 152–155, 2004.

[78] D. Alvarez-Ponce and J. O. McInerney, "The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences," *Genome Biology and Evolution*, vol. 3, no. 1, pp. 782–790, 2011.

[79] A. Poole and D. Penny, "Eukaryote evolution: engulfed by speculation," *Nature*, vol. 447, no. 7147, p. 913, 2007.

[80] E. V. Koonin and Y. I. Wolf, "The fundamental units, processes and patterns of evolution, and the tree of life conundrum," *Biology Direct*, vol. 4, no. 1, article 33, 2009.

[81] P. Puigbò, Y. I. Wolf, and E. V. Koonin, "Search for a "tree of life" in the thicket of the phylogenetic forest," *Journal of Biology*, vol. 8, no. 6, article 59, 2009.

[82] P. Puigbò, Y. I. Wolf, and E. V. Koonin, "Seeing the tree of life behind the phylogenetic forest," *BMC Biology*, vol. 11, article 46, 2013.

[83] D. B. Wetlaufer, "Nucleation, rapid folding, and globular intrachain regions in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 70, no. 3, pp. 697–701, 1973.

[84] J. S. Richardson, "The anatomy and taxonomy of protein structure," *Advances in Protein Chemistry*, vol. 34, pp. 167–339, 1981.

[85] M. Riley and B. Labedan, "Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module," *Journal of Molecular Biology*, vol. 268, no. 5, pp. 857–868, 1997.

[86] J. Janin and S. J. Wodak, "Structural domains in proteins and their role in the dynamics of protein function," *Progress in Biophysics and Molecular Biology*, vol. 42, pp. 21–78, 1983.

[87] G. Caetano-Anollés, M. Wang, D. Caetano-Anollés, and J. E. Mittenthal, "The origin, evolution and structure of the protein world," *Biochemical Journal*, vol. 417, no. 3, pp. 621–637, 2009.

[88] J. Gough, "Convergent evolution of domain architectures (is rare)," *Bioinformatics*, vol. 21, no. 8, pp. 1464–1471, 2005.

[89] K. Forslund, A. Henricson, V. Hollich, and E. L. L. Sonnhammer, "Domain tree-based analysis of protein architecture evolution," *Molecular Biology and Evolution*, vol. 25, no. 2, pp. 254–264, 2008.

[90] S. Yang and P. E. Bourne, "The evolutionary history of protein domains viewed by species phylogeny," *PLoS ONE*, vol. 4, no. 12, Article ID e8378, 2009.

[91] K. M. Kim and G. Caetano-Anollés, "The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms," *BMC Evolutionary Biology*, vol. 12, article 13, 2012.

[92] M. Wang and G. Caetano-Anollés, "The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world," *Structure*, vol. 17, no. 1, pp. 66–78, 2009.

[93] G. Caetano-Anollés and A. Nasir, "Benefits of using molecular structure and abundance in phylogenomic analysis," *Frontiers in Genetics*, vol. 3, article 172, 2012.

[94] K. Brading and E. Castellani, *Symmetries in Physics: Philosophical Reflections*, Cambridge University Press, Cambridge, UK, 2003.

[95] J. A. Cotton and M. Wilkinson, "Quantifying the potential utility of phylogenetic characters," *Taxon*, vol. 57, no. 1, pp. 131–136, 2008.

[96] E. Mossel and M. Steel, "A phase transition for a random cluster model on phylogenetic trees," *Mathematical Biosciences*, vol. 187, no. 2, pp. 189–203, 2004.

[97] M. H. Bailor, X. Sun, and H. M. Al-Hashimi, "Topology links RNA secondary structure with global conformation, dynamics, and adaptation," *Science*, vol. 327, no. 5962, pp. 202–206, 2010.

[98] K. Illergård, D. H. Ardell, and A. Elofsson, "Structure is three to ten times more conserved than sequence—a study of structural response in protein cores," *Proteins: Structure, Function and Bioinformatics*, vol. 77, no. 3, pp. 499–508, 2009.

[99] G. Caetano-Anollés, S. K. Hee, and J. E. Mittenthal, "The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 22, pp. 9358–9363, 2007.

[100] P. H. Weston, "Methods for rooting cladistic trees," in *Models in Phylogeny Reconstruction*, D. J. Siebert, R. W. Scotland, and D. M. Williams, Eds., pp. 125–155, Oxford University Press, Oxford, UK, 1994.

[101] H. N. Bryant, "The polarization of character transformations in phylogenetic systematics: role of axiomatic and auxiliary assumptions," *Systematic Biology*, vol. 40, no. 4, pp. 433–445, 1991.

[102] H. N. Bryant and G. Wagner, "Character polarity and the rooting of cladograms," in *The Character Concept in Evolutionary Biology*, G. P. Wagner, Ed., pp. 319–338, Academic Press, New York, NY, USA, 2001.

[103] G. Nelson, "Ontogeny, phylogeny, paleontology, and the biogenetic law," *Systematic Biology*, vol. 27, no. 3, pp. 324–345, 1978.

[104] P. H. Weston, "Indirect and direct methods in systematics," in *Ontogeny and Systematics*, C. J. Humphries, Ed., pp. 27–56, New York, NY, USA, Columbia University Press edition, 1988.

[105] F.-J. Sun and G. Caetano-Anollés, "The evolutionary history of the structure of 5S ribosomal RNA," *Journal of Molecular Evolution*, vol. 69, no. 5, pp. 430–443, 2009.

[106] X. Fang, T. Pan, and T. R. Sosnick, "Mg^{2+}-dependent folding of a large ribozyme without kinetic traps," *Nature Structural Biology*, vol. 6, no. 12, pp. 1091–1095, 1999.

[107] F.-J. Sun and G. Caetano-Anollés, "Evolutionary patterns in the sequence and structure of transfer RNA: early origins of

Archaea and viruses," *PLoS Computational Biology*, vol. 4, no. 3, Article ID e1000018, 2008.

[108] F.-J. Sun and G. Caetano-Anollés, "The origin and evolution of tRNA inferred from phylogenetic analysis of structure," *Journal of Molecular Evolution*, vol. 66, no. 1, pp. 21–35, 2008.

[109] W. R. Ashby, *An Introduction to Cybernetics*, Taylor & Francis, London, UK, 1955.

[110] J. G. Lundberg, "Wagner networks and ancestors," *Systematic Biology*, vol. 21, no. 4, pp. 398–413, 1972.

[111] P. T. Chippindale and J. J. Wiens, "Weighting, partitioning, and combining characters in phylogenetic analysis," *Systematic Biology*, vol. 43, no. 2, pp. 278–287, 1994.

[112] J. Mittenthal, D. Caetano-Anollés, and G. Caetano-Anollés, "Biphasic patterns of diversification and the emergence of modules," *Frontiers in Genetics*, vol. 3, article 147, 2012.

[113] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[114] C. A. Orengo, A. Michie, S. Jones, D. T. Jones, M. Swindells, and J. M. Thornton, "CATH—a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093–1109, 1997.

[115] G. Sawa, J. Dicks, and I. N. Roberts, "Current approaches to whole genome phylogenetic analysis," *Briefings in Bioinformatics*, vol. 4, no. 1, pp. 63–74, 2003.

[116] R. Sathyapriya, J. M. Duarte, H. Stehr, I. Filippis, and M. Lappe, "Defining an essence of structure determining residue contacts in proteins," *PLoS Computational Biology*, vol. 5, no. 12, Article ID e1000584, 2009.

[117] D. Penny and L. J. Collins, "Evolutionary genomics leads the way," in *Evolutionary Genomics and Systems Biology*, G. Caetano-Anollés, Ed., pp. 1–16, Wiley-Blackwell, Hoboken, NJ, USA, 2010.

[118] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[119] A. Andreeva, D. Howorth, J. Chandonia et al., "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Research*, vol. 36, pp. D419–D425, 2008.

[120] L. S. Yafremava, M. Wielgos, S. Thomas et al., "A general framework of persistence strategies for biological systems helps explain domains of life," *Frontiers in Genetics*, vol. 4, article 16, 2013.

[121] J. L. Friar, T. Goldman, and J. Pérez-Mercader, "Genome sizes and the Benford distribution," *PLoS ONE*, vol. 7, no. 5, Article ID e36624, 2012.

[122] A. Nasir, K. M. Kim, and G. Caetano-Anollés, "Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya," *BMC Evolutionary Biology*, vol. 12, article 156, 2012.

[123] A. Nasir and G. Caetano-Anollés, "Comparative analysis of proteomes and functionomes provides insights into origins of cellular diversification," *Archaea*, vol. 2013, Article ID 648746, 13 pages, 2013.

[124] W. Martin and M. Müller, "The hydrogen hypothesis for the first eukaryote," *Nature*, vol. 392, no. 6671, pp. 37–41, 1998.

[125] A. Nasir, K. M. Kim, and G. Caetano-Anollés, "Global patterns of protein domain gain and loss in superkingdoms," *PLoS Computational Biology*, vol. 10, no. 1, Article ID e1003452, 2014.

[126] M. Wang, L. S. Yafremava, D. Caetano-Anollés, J. E. Mittenthal, and G. Caetano-Anollés, "Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world," *Genome Research*, vol. 17, no. 11, pp. 1572–1585, 2007.

[127] L. Guy and T. J. G. Ettema, "The archaeal "TACK" superphylum and the origin of eukaryotes," *Trends in Microbiology*, vol. 19, no. 12, pp. 580–587, 2011.

[128] D. Schlieper, M. A. Oliva, J. M. Andreu, and J. Löwe, "Structure of bacterial tubulin BtubA/B: evidence for horizontal gene transfer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 26, pp. 9170–9175, 2005.

[129] M. P. Hoeppner, P. P. Gardner, and A. M. Poole, "Comparative analysis of RNA families reveals distinct repertoires for each domain of life," *PLoS Computational Biology*, vol. 8, no. 11, Article ID e1002752, 2012.

[130] H. Xue, K. Tong, C. Marck, H. Grosjean, and J. T. Wong, "Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life," *Gene*, vol. 310, no. 1-2, pp. 59–66, 2003.

[131] H. Xue, S. Ng, K. Tong, and J. T. Wong, "Congruence of evidence for a Methanopyrus-proximal root of life based on transfer RNA and aminoacyl-tRNA synthetase genes," *Gene*, vol. 360, no. 2, pp. 120–130, 2005.

[132] M. Di Giulio, "Nanoarchaeum equitans is a living fossil," *Journal of Theoretical Biology*, vol. 242, no. 1, pp. 257–260, 2006.

[133] M. Di Giulio, "The tree of life might be rooted in the branch leading to Nanoarchaeota," *Gene*, vol. 401, no. 1-2, pp. 108–113, 2007.

[134] J. T. Wong, J. Chen, W. Mat, S. Ng, and H. Xue, "Polyphasic evidence delineating the root of life and roots of biological domains," *Gene*, vol. 403, no. 1-2, pp. 39–52, 2007.

[135] F.-J. Sun and G. Caetano-Anollés, "The ancient history of the structure of ribonuclease P and the early origins of Archaea," *BMC Bioinformatics*, vol. 11, article 153, 2010.

[136] C. Marck and H. Grosjean, "tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features," *RNA*, vol. 8, no. 10, pp. 1189–1232, 2002.

[137] W. Mat, H. Xue, and J. T. Wong, "The genomics of LUCA," *Frontiers in Bioscience*, vol. 13, no. 14, pp. 5605–5613, 2008.

[138] C. Brochier, S. Gribaldo, Y. Zivanovic, F. Confalonieri, and P. Forterre, "Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales?" *Genome Biology*, vol. 6, no. 5, article R42, 2005.

[139] B. Billoud, M. Guerrucci, M. Masselot, and J. S. Deutsch, "Cirripede phylogeny using a novel approach: molecular morphometrics," *Molecular Biology and Evolution*, vol. 17, no. 10, pp. 1435–1445, 2000.

[140] L. J. Collins, V. Moulton, and D. Penny, "Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP," *Journal of Molecular Evolution*, vol. 51, no. 3, pp. 194–204, 2000.

[141] G. Caetano-Anollés, "Novel strategies to study the role of mutation and nucleic acid structure in evolution," *Plant Cell, Tissue and Organ Culture*, vol. 67, no. 2, pp. 115–132, 2001.

[142] F.-J. Sun, S. Fleurdépine, C. Bousquet-Antonelli, G. Caetano-Anollés, and J. Deragon, "Common evolutionary trends for SINE RNA structures," *Trends in Genetics*, vol. 23, no. 1, pp. 26–33, 2007.

[143] D. Caetano-Anollés, K. M. Kim, J. E. Mittenthal, and G. Caetano-Anollés, "Proteome evolution and the metabolic origins of translation and cellular life," *Journal of Molecular Evolution*, vol. 72, no. 1, pp. 14–33, 2011.

[144] M. Wang, Y. Jiang, K. M. Kim et al., "A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation," *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 567–582, 2011.

[145] S. A. Bukhari and G. Caetano-Anollés, "Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes," *PLoS Computational Biology*, vol. 9, no. 3, Article ID e1003009, 2013.

[146] M. Vesteg and J. Krajčovič, "The falsifiability of the models for the origin of eukaryotes," *Current Genetics*, vol. 57, no. 6, pp. 367–390, 2011.

[147] I. N. Berezovsky and E. I. Shakhnovich, "Physics and evolution of thermophilic adaptation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12742–12747, 2005.

[148] C. Woese, "The universal ancestor," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 12, pp. 6854–6859, 1998.

[149] M. J. Seufferheld and G. Caetano-Anollés, "Phylogenomics supports a cellularly structured urancestor," *Journal of Molecular Microbiology and Biotechnology*, vol. 23, no. 1-2, pp. 178–191, 2013.

[150] O. Kandler, "Cell wall biochemistry and three-domain concept of life," *Systematic and Applied Microbiology*, vol. 16, no. 4, pp. 501–509, 1994.

[151] C. R. Woese, "On the evolution of cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8742–8747, 2002.

[152] S. L. Baldauf, J. D. Palmer, and W. F. Doolittle, "The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 15, pp. 7749–7754, 1996.

[153] H. Leffers, J. Kjems, L. Østergaard, N. Larsen, and R. A. Garrett, "Evolutionary relationships amongst archaebacteria. A comparative study of 23 S ribosomal RNAs of a sulphur-dependent extreme thermophile, an extreme halophile and a thermophilic methanogen," *Journal of Molecular Biology*, vol. 195, no. 1, pp. 43–61, 1987.

[154] J. Martin, D. Blackburn, and E. Wiley, "Are node-based and stem-based clades equivalent? Insights from graph theory," *PLoS Currents*, vol. 2, Article ID RRN1196, 2010.

Hindawi Publishing Corporation
Archaea
Volume 2014, Article ID 898453, 18 pages
http://dx.doi.org/10.1155/2014/898453

Research Article

Towards a Computational Model of a Methane Producing Archaeum

Joseph R. Peterson,[1] Piyush Labhsetwar,[2] Jeremy R. Ellermeier,[3] Petra R. A. Kohler,[3] Ankur Jain,[2] Taekjip Ha,[2,4] William W. Metcalf,[3] and Zaida Luthey-Schulten[1,2,4]

[1] *Department of Chemistry, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA*
[2] *Center for Biophysics and Computational Biology, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA*
[3] *Department of Microbiology, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA*
[4] *Department of Physics, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA*

Correspondence should be addressed to Zaida Luthey-Schulten; zan@illinois.edu

Received 14 November 2013; Accepted 18 December 2013; Published 4 March 2014

Academic Editor: Gustavo Caetano-Anollés

Copyright © 2014 Joseph R. Peterson et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Progress towards a complete model of the methanogenic archaeum *Methanosarcina acetivorans* is reported. We characterized size distribution of the cells using differential interference contrast microscopy, finding them to be ellipsoidal with mean length and width of 2.9 μm and 2.3 μm, respectively, when grown on methanol and 30% smaller when grown on acetate. We used the single molecule pull down (SiMPull) technique to measure average copy number of the Mcr complex and ribosomes. A kinetic model for the methanogenesis pathways based on biochemical studies and recent metabolic reconstructions for several related methanogens is presented. In this model, 26 reactions in the methanogenesis pathways are coupled to a cell mass production reaction that updates enzyme concentrations. RNA expression data (RNA-seq) measured for cell cultures grown on acetate and methanol is used to estimate relative protein production per mole of ATP consumed. The model captures the experimentally observed methane production rates for cells growing on methanol and is most sensitive to the number of methyl-coenzyme-M reductase (Mcr) and methyl-tetrahydromethanopterin:coenzyme-M methyltransferase (Mtr) proteins. A draft transcriptional regulation network based on known interactions is proposed which we intend to integrate with the kinetic model to allow dynamic regulation.

1. Introduction

Molecular signatures of ribosomal rRNA evolution were used by Woese and his associates to establish the three primary groupings of living organisms: archaea, bacteria, and eukarya [1–6]. Although the ancestral or communal origins of these three domains remains a matter of debate, increasingly large amounts of data regarding the RNA phylogeny and molecular makeup of cells accumulated over the last several decades continue to support the division between the three primary domains [7]. Furthermore, comparative analysis of the sequences of proteins and RNA involved in translation provides strong evidence that the existence of highly developed translational machinery was a necessary condition for the emergence of cells as we know them [8, 9]. Molecular signatures in the ribosome—idiosyncrasies in its rRNA [7] and/or r-proteins characteristic of each domain of life—were locked in place at the time of evolutionary divergence, destined to become molecular fossils. As Woese postulated in his theory of genetic annealing, ancestors of the three primary groupings of organisms developed into a number of increasingly complex cell types. The various subsystems of the cell "crystallized," that is, became refractory to lateral gene transfer with the translation apparatus probably crystallizing first.

While the rRNA phylogeny is supported by phylogenetic analysis of concatenated protein sequences of the fundamental genes in the translational machinery, the effects of lateral gene transfer (LGT) among organisms in the three domains of life are clearly seen in the aminoacyl-tRNA synthetases, a modular subsystem that charges tRNA, and helped to establish the genetic code [10–12].

As one moves beyond the information processing systems of translation and transcription, an increasing amount of LGT also extends into other cellular networks. If the early community of cells was more like a modern bacterial consortium, the cells could have cross-fed one another not only genetically but also metabolically. Every improvement in translation that increased its accuracy would have permitted new proteins to emerge, which in turn could have further developed the metabolic pathways within cells. With metabolic functions being modular in nature, these genes could be transferred laterally. Many cases are now known in which a bacterial metabolic gene occurs in one or a few archaea or vice-versa and has prompted the search for signatures in the metabolic networks that are distinctive of the archaea [13–16].

Many theories of early life argue for a reducing environment in which anaerobic organisms would likely be the first to have evolved [17]. A phylogenetic analysis of proteins that are distinctive of archaea and its main subgroups has led to hypotheses in which methanogens—anaerobic archaeal organisms that derive all of their metabolic energy by reduction of single carbon compounds to methane—feature prominently in the early evolution of life. Methanogens are phylogenetically diverse group of strict anaerobes estimated to produce a billion tonnes of methane per year [18]. They are found in niche environments including shallow and deep hydrothermal vents [19], swamps, paddy fields, land fills [20], hot-springs, and oxygen-depleted sediments beneath kelp beds [18].

The *Methanosarcineae* are the most metabolically diverse methanogens known. Only *M. acetivorans* and other archaea in the genus *Methanosarcina* use all four known metabolic pathways for methanogenesis under different growth conditions. While systems biology studies have long used *E. coli* as a model organism in understanding the response of cellular networks to changes in various environmental conditions or gene knock-outs, computational models of methanogen metabolism are just beginning to be established [21, 22]. Based in part on our own work modeling genetic switches in *E. coli* and the effects of heterogeneity in protein expression on the metabolism of large populations of bacteria [23], we present here our progress toward a comprehensive computational model of a methanogen. In doing so we have been profoundly influenced by our association with Carl Woese, who published the first genome of a methanogen, *Methanocaldococcus janaschii*, and greatly inspired our interest in characterizing both the translational and metabolic machinery of methanogenic archaea [24].

We have focused our study on *M. acetivorans* for several reasons. First, the organism can grow on three classes of substrates demonstrating use of three methanogenic pathways: (1) methylotrophic pathway, wherein the organism grows on methyl containing substrates including methanol, tri-, di-, and mono-methylamine (TMA, DMA, MMA), and methylsulfides (DMS, MMS); (2) acetoclastic pathway, wherein the organism grows on acetate; (3) carboxidotrophic pathway, wherein carbon monoxide is oxidized to acetate, formate, and methane [25–27]. Second, the genome of *M. acetivorans* has been sequenced [25], and considerable effort has been expended towards determining the regulation of gene expression of methanogenesis proteins [28]. Third, the genome exhibits considerable homology to two other well studied members of the genus *Methanosarcina*: *M. barkeri* and *M. mazei* and, therefore, a model for one will likely be easily modified to work for the others.

Developing a model of the archaeum requires characterization of its physical and biochemical properties. To that end the physical dimensions of the cells, including their length and width, were measured. Modeling also requires estimation of protein/ribosome copy numbers in single cells; the single molecule pulldown (SiMPull) technique [29]—a marriage of the conventional pull-down assay with single molecule fluorescence microscopy—was used to measure the mean copy number of two key proteins. The first protein measured was the γ subunit (McrG) of methyl-coenzyme-M reductase (Mcr) complex as a proxy for number of Mcr complexes, which catalyze the methane producing step of methanogenesis. Second, the ribosomal protein Rpl18p in the large subunit of ribosome was counted as a proxy for the number of ribosomes. A kinetic model for methanogenesis pathways capable of representing growth on methanol and acetate was developed using RNA-seq data and kinetic parameters from the literature. This model captures several features of comparable experimental data [30, 31]. The model further allows us to probe the sensitivity of the growth (and indirectly the methane production) on the copy number of each protein, directing further experimental study. In an effort to extend the model to simulate growth on other substrates, we compile a list of all experimentally known and hypothetical transcriptional regulatory interactions. These interactions will be used to modulate protein expression as a function of growth substrate that we can marry with the kinetic model in future.

2. Experimental and Computational Methods

2.1. Strains, Media, and Growth Conditions. *M. acetivorans* C2A strains (wild-type, WWM 889 :: *SNAP-mcrG*, and WWM890 :: *rpl18p-SNAP*) were grown in single cell morphology [48] at 37°C in high-salt (HS) medium containing either 125 mM methanol or 40 mM acetate [49]. Handling and manipulation of all strains were carried out under strict anaerobic conditions in an anaerobic glove box, using sterile anaerobic media and stocks. Solid media plates (HS medium, 1.5% agar) were used for selection of SNAP integrants in two steps: puromycin (Research Products International, Mt. Prospect, IL) at a final concentration of 2 μg/mL was used for selection of strains carrying puromycin transacetylase (*pac*) and the purine analogue 8-aza-2,6-dia-minopurine (8-ADP) (Sigma, St. Louis, MO) at a final concentration of 20 μg/mL was used for selection against the *hpt* gene [50–52]. All plates were incubated in an anaerobic intrachamber incubator [53]. Standard methods were used throughout for isolation and manipulation of plasmid DNA from *E. coli*. DNA purification was performed using appropriate kits (OmegaBio-Tek, Norcross, GA). Growth was quantified by measuring the optical density at 600 nm (OD_{600}, Milton Roy

Company Spectronic 21 spectrophotometer) and generation times were calculated during exponential growth.

2.2. Genetic Constructs in Methanosarcina Acetivorans. Genetic fusions with SNAP were made by first constructing plasmids with the SNAP gene near an *aphII* cassette flanked by NheI restriction sites. pJK1048A was used as the template for making fusions to the C-terminus of genes of interest, while pJK1047B was used for fusions to the N-terminus. DNA oligonucleotides (IDT, Iowa City, IA) with homology to the template and gene of interest were used to amplify the SNAP-aphII constructs. The Lambda Red method was then used to integrate SNAP aphII construct into specific N- or C-terminal locations [54], selecting for kanamycin resistance. The *mcrG* and *rpl18p* genes are carried on cosmids created during an *M. acetivorans* cosmid library construction previously performed in the Metcalf lab (Zhang and Metcalf, unpublished). The *aphII* allele was then excised from the cosmid by NheI restriction digest, leaving an in-frame SNAP fusion to the gene of interest. The wild type copies of the genes in question were replaced by the SNAP tagged versions using homologous recombination, as previously described [51].

2.3. Cell Morphology from DIC Microscopy. Cell cultures were grown into exponential phase to an OD_{600} of 0.6 and 1 mL of cultures was removed and centrifuged at 14,000 g for 5 minutes. The cell pellet obtained was resuspended in 100 μL HS media without resazurin, and the cells were observed using the differential interference contrast (DIC) microscopy technique on a Zeiss LSM700 confocal microscope.

2.4. RNA-Seq Analysis. M. acetivorans C2A wild type was adapted to methanol and acetate for 33 generations. The total RNA was isolated from early exponential phase cultures (OD_{600} = 0.4) using TRIzole (Invitrogen, Carslbad, CA) and the Zymo Direct-zol RNA MiniPrep kits (Zymo Research, Irvine, CA). The RNA samples were depleted of the 16S- and 23S-rRNA through hybridization to complementary biotinylated oligonucleotides and subsequent removal with streptavidin-magnetic beads (modified from [55]). Construction of cDNA libraries and high throughput sequencing of RNA was carried out by the Roy J. Carver Biotechnology Center at University of Illinois at Urbana Champaign. All measurements were done in triplicate. The Rockhopper [56] bacterial RNA-seq analysis software was used to map RNA reads to the *M. acetivorans* genome using the default parameters with verbose output enabled. Reads per kilobase per million reads (RPKM) values from the three replicates were averaged and used in subsequent analysis.

2.5. SiMPull Experiments. The single molecule pulldown, or SiMPull technique [57], was used to determine mean protein counts for two proteins in *M. acetivorans*. Briefly, SiMPull is a microscopy technique wherein a fluorescently labeled protein of interest is "captured" out of cell lysate by an immobilized antibody attached to a passivated microscope slide. In these experiments, the genetic SNAP-tag system (New England Biolabs (NEB), Ipswich, MA) was used for labeling either the N- or C-terminus of each protein studied (Figure 1).

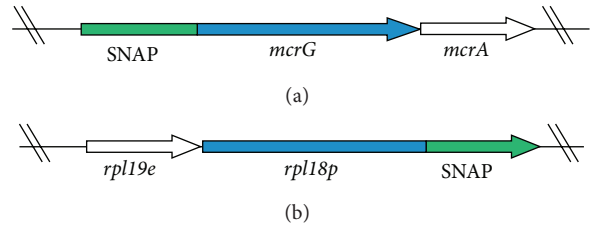


Figure 1: Genetic constructs showing position of SNAP relative to our protein of interest on chromosome of *M. acetivorans*. (a) N-terminal label on *mcrG* gene. (b) C-terminal label on *rpl18p* gene.

Labeled mutants were grown to exponential phase and harvested at an OD_{600} of 0.6. Cell density was estimated using a Petroff-Hausser counting chamber. One milliliter of cell culture was centrifuged at 14,000 g for 5 minutes to obtain cell pellets which were subsequently lysed upon re-suspending the cells in 100 μL of the recommended SNAP labeling buffer: 50 mM Tris-Hcl (pH 7.5); 100 mM NaCl; 0.1% Tween 20; 1 mM DTT (NEB) with 1 μg DNAse. The cell lysate was then incubated with AlexaFluor 488 (NEB) at a final concentration of 10 μM at room temperature for one hour. In order to remove free dye, samples were washed three times with SNAP labeling buffer and concentrated using 10 K Amicon ultra centrifugal filters. SiMPull analysis was performed as previously described [57].

Microscope slides were coated with polyethylene glycol (PEG) which minimizes nonspecific biomolecule adsorption. Surfaces were doped with 2–5% biotin-conjugated PEG during slide preparation. The bait recruiting rabbit anti-SNAP antibody (NEB) was immobilized onto the surfaces by successively flowing in NeutrAvidin (4 μM) and a biotin conjugated antirabbit antibody (20 nM), as depicted in Figure 3(a).

Lysate was washed away and the pulled-down proteins were imaged using a prism type Total Internal Reflection Microscopy (TIRF) with excitation at 488 nm. The resulting images (see Figure 3(b)) were analyzed using custom software as described previously [29], to quantify single spots in the field of view of the microscope. A single spot may correspond to more than one fluorophore which can be discerned by the observation of multiple discrete photobleaching steps (as in the case of Rpl18p, Figure 9) indicating that results are lower bounds on the actual number of proteins.

2.6. Kinetic Model. RNA-seq expression data for *M. acetivorans* growing on methanol and acetate [59] provide enough parameters for a preliminary kinetic model of the methanogenesis pathways (Table 1). The model includes reactions for the methylotrophic, acetoclastic, and electron transport pathways shown in Figure 4. An additional reaction simulating biomass growth is included to the model that converts ATP created by the methanogenesis driven proton gradient into cell mass. Because 98% of carbons that come into methanogenesis leave as CH_4 or CO_2 [60], ATP is assumed

Table 1: Kinetic model of methanogenesis. Reactions are from *iMB745* [22]. Rate constants used in the kinetic model are taken from the literature where indicated or were fit to experiments of growth on methanol. Water molecules and the intracellular protons (H^+), which are shown for completeness, are assumed to be constant and are not explicitly modeled. The "Type" column specifies the reaction mechanism: B: irreversible bimolecular Michaelis-Menten, B/C: irreversible bimolecular Michaelis-Menten for the two underlined reactants with a constant flux term for the others, which is set to the flux calculated for bimolecular reaction, M: irreversible unimolecular Michaelis-Menten, F: first order reaction.

Enzyme	Reaction	k_{cat} (s^{-1})	K_M (mM)	Type	Citation
	Acetoclastic pathway				
Ack	$ATP + Ac \rightarrow AcP + ADP$	1055	0.0713	B	[32]
Ack	$ADP + AcP \rightarrow Ac + ATP$	1260	0.098	B	[33]
Pta	$CoA + AcP \rightarrow AcCoA + P_i$	1500	0.186	B	[34]
Pta	$AcCoA + P_i \rightarrow CoA + AcP$	65.8	0.18	B	[35]
Cdh	$\underline{AcCoA} + 2Fd_{ox} + \underline{H4SPT} + H_2O \rightarrow CO + CoA + 2Fd_{red} + MeH4SPT + 2H^+$	358.5	7.1	B/C	[36]
Cdh	$\underline{CO} + CoA + 2Fd_{red} + \underline{MeH4SPT} + 2H^+ \rightarrow AcCoA + 2Fd_{ox} + H4SPT + H_2O$	1130	0.9	B/C	[37]
	Methylotrophic pathway				
MtaCBA1	$MeOH + CoM + H^+ \rightarrow MeCoM + H_2O$	17	50	M	[38, 39]
MtaCBA2	$MeOH + CoM + H^+ \rightarrow MeCoM + H_2O$	15	50	M	[38, 39]
MtaCBA3	$MeOH + CoM + H^+ \rightarrow MeCoM + H_2O$	5	50	M	[38, 39]
Mer	$MeH4SPT + F_{420} + H^+ \rightarrow MethyleneH4SPT + F_{420}H_2$	119.7	0.25	B	[40]
Mer	$MethyleneH4SPT + F_{420}H_2 \rightarrow MeH4SPT + F_{420} + H^+$	815	0.3	B	[41]
Mtd	$MethyleneH4SPT + F_{420} + 2H^+ \rightarrow MethenylH4SPT + F_{420}H_2$	2650	0.065	B	[42]
Mtd	$MethenylH4SPT + F_{420}H_2 \rightarrow MethyleneH4SPT + F_{420} + 2H^+$	408	0.065	B	—
Mch	$MethenylH4SPT + H_2O \rightarrow FormylH4SPT + H^+$	701	0.57	M	[43]
Mch	$FormylH4SPT + H^+ \rightarrow MethenylH4SPT + H_2O$	100	0.57	M	—
Ftr	$FormylH4SPT + Mfr \rightarrow H4SPT + H^+ + FormylMfr$	1787	0.1	B	[42]
Ftr	$H4SPT + H^+ + FormylMfr \rightarrow FormylH4SPT + Mfr$	262	0.1	B	—
Fmd/Fwd	$FormylMfr + 2Fd_{ox} + H_2O \rightarrow CO_2 + 2Fd_{red} + H^+ + Mfr$	1225	0.02	B/C	[44]
Fmd/Fwd	$CO_2 + 2Fd_{red} + H^+ + Mfr \rightarrow FormylMfr + 2Fd_{ox} + H_2O$	175	0.02	B/C	—
	Shared pathway				
Mtr	$H^+ + MeH4SPT + 2Na_c^+ + CoM \rightarrow H4SPT + 2Na_e^+ + MeCoM$	50	3.7	B	—
Mtr	$H4SPT + 2Na_e^+ + MeCoM \rightarrow H^+ + MeH4SPT + 2Na_c^+ + CoM$	50	3.7	B	—
Mcr	$MeCoM + CoB \rightarrow CoBCoM + CH_4$	5.0	2	B	[45]
	Electron transport pathway				
HdrDE	$CobCoM + MphenH_2 + 2H^+ \rightarrow Cob + CoM + Mphen + 2H_e^+$	74	0.092	B	[46]
Rnf	$2Fd_{red} + 3Na_c^+ + Mphen + 2H^+ \rightarrow 2Fd_{ox} + 3Na_e^+ + MphenH_2$	80	0.1	B/C	—
Fpo	$F_{420}H_2 + Mphen + H^+ \rightarrow F_{420} + MphenH_2 + 2H_e^+$	80	0.1	B	—
	Cell growth				
ATP synthase	$ADP + P_i + 4H_e^+ \rightarrow ATP + H_2O + 3H^+$	16	0.1	B/C	[47]
Cell Mass[a]	$ATP \rightarrow ADP + P_i + Cell\ mass$	0.125[b]	—	F	[22, 31]

[a]The reaction that converts ATP into ADP and cell mass, generating proteins via the stoichiometry in Table 2. [b]This rate is in units of hr^{-1} which is equivalent to an 8 hr doubling time for *M. acetivorans*.

to be a good analog for the growth of the colony. A model schematic is shown in Figure 5.

The kinetic model in Table 1 is based on the reactions from metabolic model *iMB745* [22]. The reactions are modeled as a set of coupled differential equations (ODEs) which are solved deterministically using the COPASI software [61]. Rate data for 17 of the 26 methanogenesis reactions were taken from the literature [32–38, 40–47] as reported in the BRENDA database [62]. The other 9 parameters were fit to experiments wherein a cell culture was grown on 125 mM methanol [30]. Three types of reaction mechanisms are used to model the reactions: irreversible unimolecular Michaelis-Menten, irreversible bimolecular Michaelis-Menten, and first order. In cases with more than two reactants, the two most important reactants were selected for bimolecular reaction and a constant flux reaction was added that converts the additional reactants to products at the same flux as the rate of bimolecular reaction. In bimolecular reactions, k_{cat} and

K_M for both substrates were assumed to be the same. When missing from the literature, K_M parameters for reverse reactions were assumed to be the same as that for forward reactions (e.g., Mtd, Mch, Ftr, and Fmd/Fwd). The forward and reverse rate constant are known for Mer giving a ratio of about 6.8. This ratio was assumed for Mtd, Mch, Ftr, and Fmd/Fwd as they are in the same pathway. Because Mtr is known to be nearly at equilibrium [63], we assumed the forward and reverse rates were the same. A value of 50 s^{-1} was chosen for this reaction.

Finally, Rnf and Fpo were assumed to have similar rates to the Hdr protein as they also catalyze the motion of a similar number of ions across the cell membrane. The reactions modeled and rate constants used in the model can be found in Table 1.

A value of 15.4 grams of cell mass per mole of ATP [22, 64] was used in the biomass expression to match the stationary phase mass of a culture calculated from experimental OD_{420} measurements [31]. The rate of the cell mass reaction was set to match the approximate maximal doubling time of 8 hours known for growth on methanol. The accumulation of biomass in the model leads to an accumulation of enzymes; for each gram of biomass, 63% is assumed to be proteins (in accordance with [22]) of which some are the methanogenic enzymes that themselves catalyze growth. The results of RNA-seq experiments provide estimates for the stoichiometry of methanogenic enzymes per mole ATP. A linear relationship between methanogenic proteins and mRNA was assumed. We determined the relative mass of protein as

$$0.63M_{\text{total}} = \sum_{i=1}^{N_{\text{genes}}} a_i \times m_{\text{Protein},i}, \quad (1)$$

where the coefficients a_i are the mass fraction of ith protein calculated with (2). From the value of a_i and the molecular weight of protein $m_{\text{protein},i}$, the number of moles of protein per mole of ATP was determined; these values are provided in Table 2. We have

$$a_i = \frac{m_{\text{protein},i}}{\sum_{i=1}^{N_{\text{genes}}} m_{\text{protein},i} \times \text{RPKM}_i}. \quad (2)$$

The model was solved in a 1 mL volume with an initial cell mass of 0.1 mg, calculated from the optical density at the start of growth [30, 31]. The concentrations of water and internal protons are assumed to be constant and therefore their effect on the rate constants is implicit and not explicitly modeled. The concentration of extracellular protons was initially set to physiological pH of 7 and protons are modeled explicitly in the ATP synthase reaction. This reduces the complexity of most of the reactions to either one or two substrate Michaelis-Menten kinetics. Initial concentrations of ATP, ADP, and P_i were set to physiological concentrations of 10, 1, and 10 mM, respectively [47]. Intermediate energy carriers (CoB, CoM, ferredoxin, etc.) initial concentrations were assumed to be 0.009 mM, which was calculated from the measured value of 474 nmol/g protein measured for coenzyme F_{420} in *M. barkeri* grown on methanol [65].

Table 2: A list of enzyme stoichiometries in the cell mass reaction. The moles of the indicated protein that are created from 1 mole of ATP calculated in Section 2.6.

Enzyme	Methanol (μmol/mol)	Acetate (μmol/mol)
Ack	37.5	102.0
ATP	132.0	406
Cdh	151.0	134.4
Fmd/Fwd	57.4	6.4
Fpo	27.8	3.96
Ftr	9.60	4.72
HdrDE	45.3	38.1
Mch	30.0	11.6
Mcr	321.8	398
Mer	25.5	1.26
MtaCBA1	1.97[a]	3.57
MtaCBA2	10.78	5.66
MtaCBA3	0.20[a]	141.0
Mtd	36.9	1.69
Mtr	112.4	144.4
Pta	36.2	171.0
Rnf	22.8	17.1

[a]Expression values of MtaCBA1 and MtaCBA3 were adjusted such that their ratios to MtaCBA2 were in agreement with the protein expression values measured experimentally [30].

2.7. Transcriptional Model. A putative model of transcriptional regulation was constructed using experimental data and inferred regulatory interactions based on gene annotation and sequence homology with proteins known to be regulated in other archaea. Two different models were developed: the first involving only direct interactions and the second involving indirect and hypothetical interactions. The direct interactions model was based on experimental evidence of actual binding of the activator/repressor to the promoter region causing up/downregulation of target gene. In addition, genes that showed differential expression and contained the known promoter region, were included in the direct model. The indirect interaction model includes interactions reported in the literature where proteins were differentially expressed under different growth conditions or when expression correlated with a regulator that is differentially expressed, but no direct evidence for the interaction exists. Strength of interactions in the direct and indirect models were taken from the literature; when the transcriptional regulator was overexpressed, the strength of interactions was normalized by the overexpression level. A full enumeration of the literature used to develop these transcriptional regulation models is reported in Section 3.5.

3. Results and Discussion

3.1. Cell Characterization. DIC images of methanol and acetate grown cells were obtained and analyzed in order to quantify their physical dimensions. As seen in Figure 7(a), DIC microscopy yields enhanced contrast images by taking

advantage of a gradient in optical path length between beams of light passing through adjacent points in the illuminated sample. The enhanced contrast is directional and appears strongest along the shear vector. No contrast occurs perpendicular to the shear vector, which can make the demarcation of cell boundaries difficult. A Hilbert transform has been used in the past with DIC microscopy in order to aid in image segmentation [66]. Custom Matlab scripts were developed to normalize and apply a Hilbert transform to the DIC images. The transformed image shows clearer boundaries around the imaged cells (Figure 7(b)). The CellProfiler software was used to identify cell boundaries [67] and measure the cells' dimensions. Figure 6 shows the distributions of lengths and widths obtained from approximately 10,000 identified cells. The mean length and width observed were 2.9 μm and 2.3 μm for methanol grown cells, while for acetate grown cell they were 2.3 μm and 1.7 μm, respectively. Assuming the cells to be ellipsoid in shape, volume of a methanol grown cell would be approximately 9 fl and that for acetate is approximately 4 fl. Cells have mean aspect ratios of 1.27 for methanol grown cells and 1.33 for acetate grown cells.

3.2. SiMPull Measurements. SiMPull was used to measure the mean copy numbers of two proteins integral to the growth and physiology of methanogens. The first is the γ subunit of the methyl-coenzyme M-reductase (Mcr) complex. This complex is a lynchpin in the metabolic network, catalyzing the last step of methanogenesis that produces methane. The second is Rpl18p, a ribosomal protein counted as proxy for the ribosome, the protein producing machinery of the cell. We have inserted SNAP tags at the N-terminus of the *mcrG* gene and at the C-terminus of the *rpl18p* gene (Figure 1). The fusion of SNAP at the C-terminus of Rpl18P exposes it to the outside of the ribosome enabling the immobilized anti-SNAP proteins to capture whole ribosomes during the SiMPull assay (see Figure 2). Using these calibration curves (Figure 8), along with estimates of cell density prior to lysing from cell counting experiments, copy numbers of Mcr and ribosomes per cell were obtained (Table 3). Mcr numbers agree qualitatively with a recent study where Mcr was imaged on TEM immunocytochemistry techniques [68].

3.3. RNA-Seq Experiments. RNA sequencing experiments were performed in order to elucidate the differential expression of methanogenic enzymes on different growth substrates. Comparison of the ratio of mRNA expression on methanol-grown cells to that of acetate-grown cells shows good agreement with the results of quantitative reverse transcriptase polymerase chain reaction (qrtPCR) data previously reported [69].

However, two discrepancies are observed. The first is that Hdr, which was found to be more highly expressed in methanol-grown cells than in acetate-grown cells, was previously reported to be more highly expressed under acetate growth conditions. This is likely attributable to experimental noise as in both cases the expression ratios are close to 1. The other difference is more pronounced; expression of ATP synthase was found to be 3 times higher on acetate in our experiments, whereas the previously reported results indicate only a two fold enhancement. In spite of this, expression of methanogenic proteins generally agree with previous reports [59, 69], and, importantly, our experiments were run in triplicate and, therefore, offer greater confidence in our results in addition to some means of error estimation.

Table 3: Mean protein copy numbers per cell for Mcr complex and ribosomes as estimated by dividing concentration of McrG and Rpl18p subunits in cell lysates by number of cells in the culture. All experiments were done with three technical replicates and two biological replicates grown in methanol.

Methlyl-coenzyme M-reductase			
Biol. Rep.	Conc. (nM)	Cells/mL	Count/cell
1	1.1 ± 0.13	$(5 \pm 2) \times 10^8$	1320 ± 713
2	0.37 ± 0.03	$(8 \pm 3) \times 10^8$	273 ± 124
Ribosome			
Biol. Rep.	Conc. (nM)	Cells/mL	Count/cell
1	5	$(3 \pm 1) \times 10^8$	10038 ± 3340
2	27.1	$(9 \pm 3) \times 10^8$	18135 ± 6040

Protein numbers, computed using the assumption that they are linearly proportional to mRNA number from RNA-seq, were compared with experimentally measured ones from SiMPull. From [26], we know that OD_{420} of 1 corresponds to 0.41 ± 0.07 mg dry mass per mL of culture grown on CO. Assuming 63% of this mass being protein and Mcr having 1.2% mass fraction in the proteome, we obtain 3.1 ± 0.5 μg of Mcr per mL of culture or 10.3 ± 1.6 picomoles per mL of culture (molecular weight of Mcr = 300 kDa [70]). Using cell density of 500 million cells per mL of culture at OD_{420} of 1 (data not shown), we obtain around 12,400 ± 2,000 copies of Mcr per cell grown in methanol and approximately 6,000 for cells grown on CO estimated due to size differences. SiMPull measurements for Mcr in methanol grown cells ranged from 273 to 1320 copies per cell (see Table 3).

3.4. Kinetic Model. A kinetic model of the methanogenesis pathways in *M. acetivorans* with single-reaction resolution was developed. The model was fit to cell culture growth experimental results reported previously [30] wherein the methanol consumption rate and OD_{420} of cell culture were studied over time. A comparison of the model fit, seen in Figure 10, demonstrates that the chosen rate parameters capture the methanol behavior within 10%. The cell mass in the culture was calculated from OD_{420} traces from [30] using the calibration point of 0.41 ± 0.07 mg/mL at OD_{420} 1.0. At maximum growth rate, the model predicts methane formation of 565 nmol/mL × min (slope of simulated curve in Figure 10) compared with 372 ± 69 nmol/mL × min measured experimentally [31]. The model correctly predicts the mass of the cells in culture (within 10%), and it captures the 3 : 1 methane to CO_2 efflux ratio that is necessary for the correct redox intermediate behaviors. Using a stoichiometry of 3.5 protons per ATP, as measured in some experiments for other

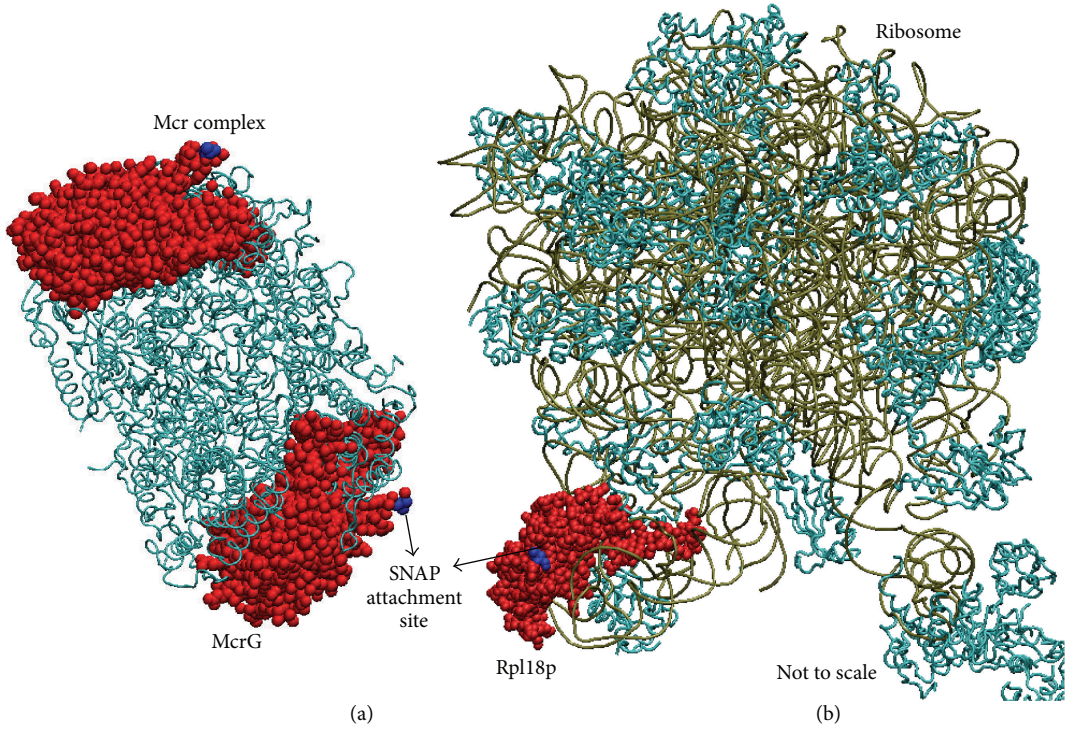


FIGURE 2: (a) The Mcr complex from *M. barkeri* (1E6Y [58]) with the McrG subunit shown in red with the SNAP attachment site shown in blue. (b) The large subunit of an archaeal ribosome (*Haloarcula marismortui*, 4HUB) showing the L18p subunit in red and the C-terminus where the SNAP is attached in blue. These suggest that the position of SNAP is on the outer part of the complexes, which enables capture by the SNAP antibody.

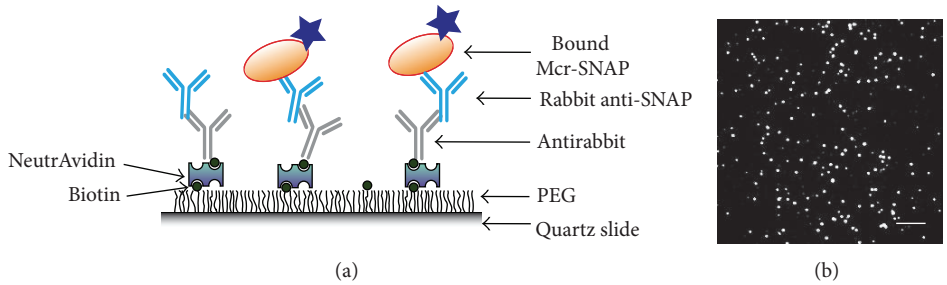


FIGURE 3: SiMPull experiments for protein count measurement. (a) Anti-SNAP antibody immobilized on a microscope slide using biotin. The antirabbit antibody captures SNAP labeled McrG. (b) A TIRF microscopy image obtained for captured McrG, where each spot corresponds to at least one immobilized protein.

organisms [71], would make the modeled cell mass growth exactly match the experiments.

Acetate growth uses a different methanogenesis pathway and is a good test for the rate constants. Using RNA expression values of proteins in acetate-grown cells and the same kinetic parameters obtained from the methanol fit, the results along with experimental results for cells grown in 120 mM acetate [31], shown in Figure 11, were obtained. Cell mass entering stationary phase is of the right level, but the rate of growth is much too high. The model predicts a significant buildup of carbon monoxide, which should be converted to CO_2 as that step produces more electrons used to drive protons across the membrane. The methane production rate was determined to be 269 nmol/mL × min, which is less than the rate of methane production on methanol but is quite a bit higher than experimental measurements of 82 ± 31 nmol/mL × min [31].

This model represents a powerful tool for its ability to be used in testing the sensitivity of cell growth to model parameters such as enzyme copy numbers and rate constants. Moreover, because the growth rate can be thought of as a proxy for the amount of methane produced, understanding its sensitivity to enzyme expression is interesting from a biofuels perspective. The relative sensitivity, s, is calculated using the standard expression:

$$s = \frac{x}{Y(x)} \frac{\partial Y(x)}{\partial x}, \tag{3}$$

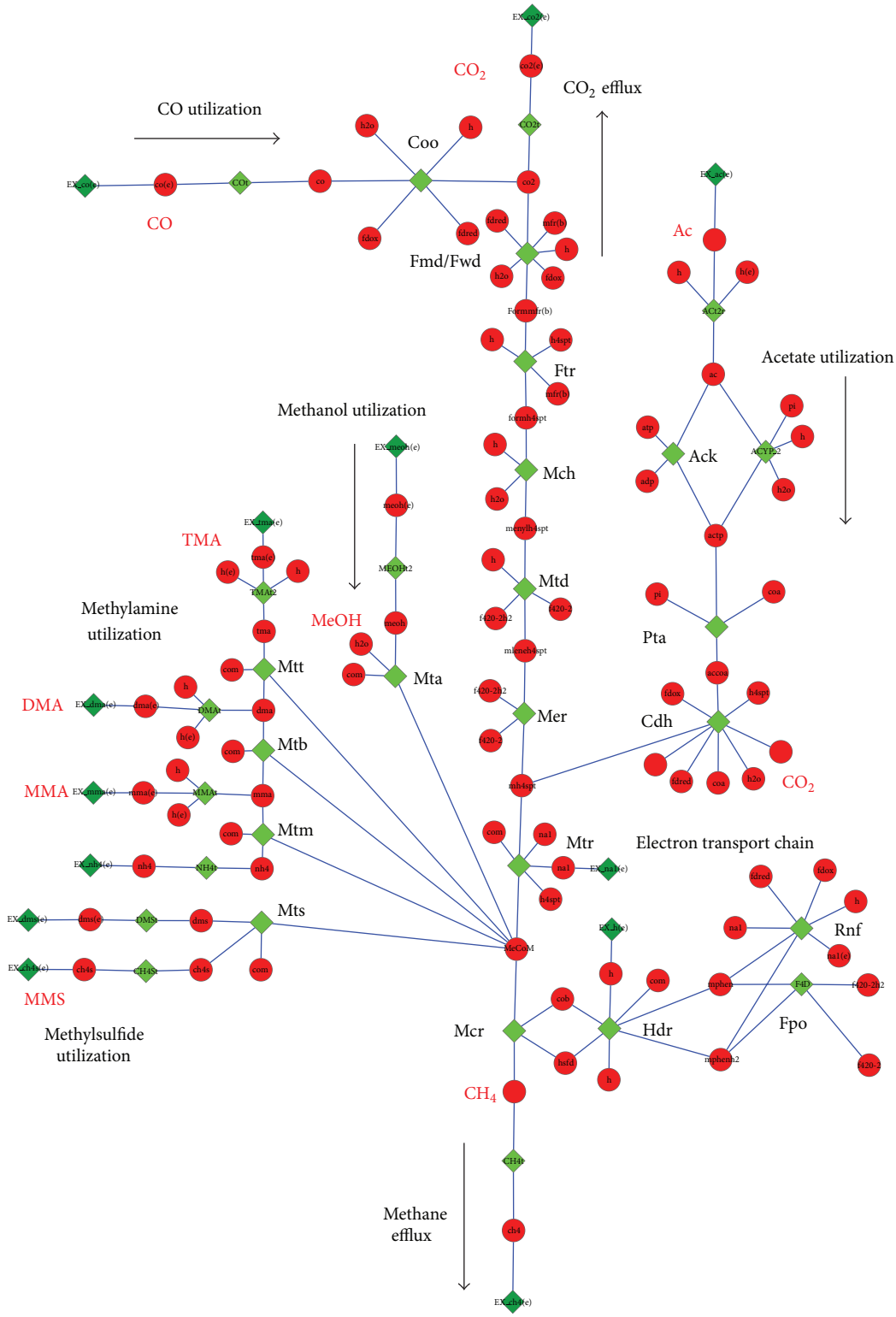


Figure 4: Methanogenesis pathways in metabolic map of *M. acetivorans* [22]. Enzymes and metabolites are depicted as nodes while reactions are depicted as edges between these nodes. Enzymes that catalyse reactions are shown as green diamonds and metabolites are shown as red circles. Enzyme names are in black and select metabolite names are in red.

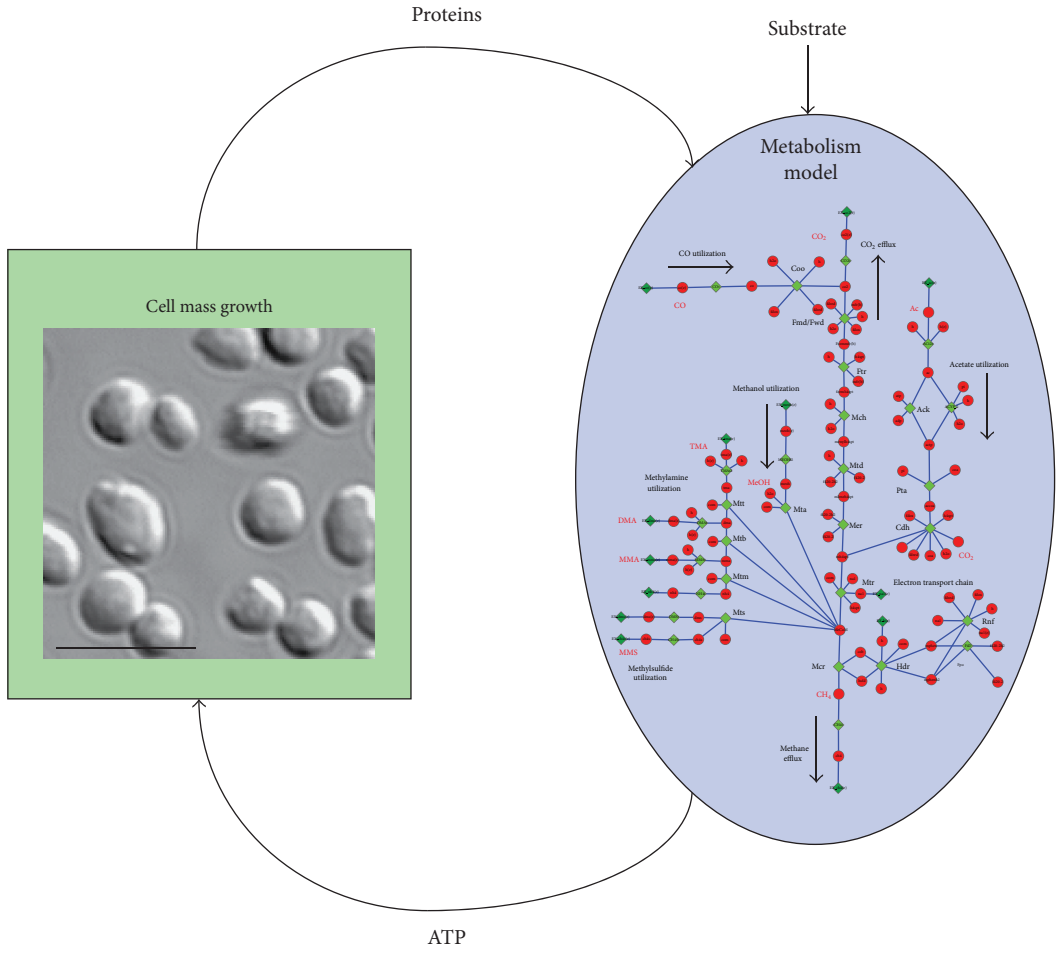


FIGURE 5: A schematic of the kinetic model. Flux of ATP from the methanogenesis pathway (see Figure 4) feeds into a cell growth term that updates protein numbers used by the kinetic model, simulating the growth when fed on a certain substrate. The inset in the cell mass growth expression is from DIC microscopy with a 5 μm scale bar.

where Y is the observable (e.g., growth rate) and x is the parameter (e.g., enzyme copy number). Performing this analysis for methanol growth shows that cellular growth rate is most sensitive to the copy number of Mcr and, in order from greatest to least sensitivity, the copy numbers of Mtr, Rnf, Fpo, MtaCBA2, HdrDE, Mer, and MtaCBA1. This suggests that growth rate is most dependent on the rate at which methyl-coenzyme M can be reduced to methane. These results also indicate that growth rate depends on the equilibrium of species at the branching point that directs substrates either to CO_2 or to methane via the Mtr reaction. This is in line with the fact that the acetoclastic pathway is highly downregulated during growth on methyl substrates, driving flux through that reaction in the reverse direction. In addition, the rate at which protons are pumped across the membrane and the intermediates regenerated (via Hdr and Rnf) effect the rate significantly. Finally, it is of no surprise to see that the rate of methane production is strongly dependent on the rate at which methanol is brought into the methanogenesis pathway as demonstrated by the dependence on Mta proteins.

Examination of the sensitivity of growth rate to various enzyme copy numbers under acetate-growth conditions yields a different trend, in order of decreasing sensitivity, Mer, Mcr, HdrDE, Rnf, and Mtr. The sensitivity to Mer is directly due to the fact that the reaction can divert flux away from methane production to CO_2 production.

Ongoing work on this model aims to test the behavior on other growth substrates such as CO, MMA, DMA and MMS, as well as mixtures of growth substrates. Future work to refine the rate parameter estimates in order to better capture growth defects with gene knockouts of nonessential methanogenesis genes such as heterodisulfide reductase is planned [31].

3.5. Transcriptional Regulation Model

3.5.1. Direct Interactions. The direct interaction map, seen in Figure 12(a), is largely made up of TATA binding proteins (TBPs) which are common across all archaea. Other direct interactions were largely identified only for methyltransferases, nitrogen fixation proteins, and oxidative stress proteins. Three TATA-box binding proteins (TBPs) were

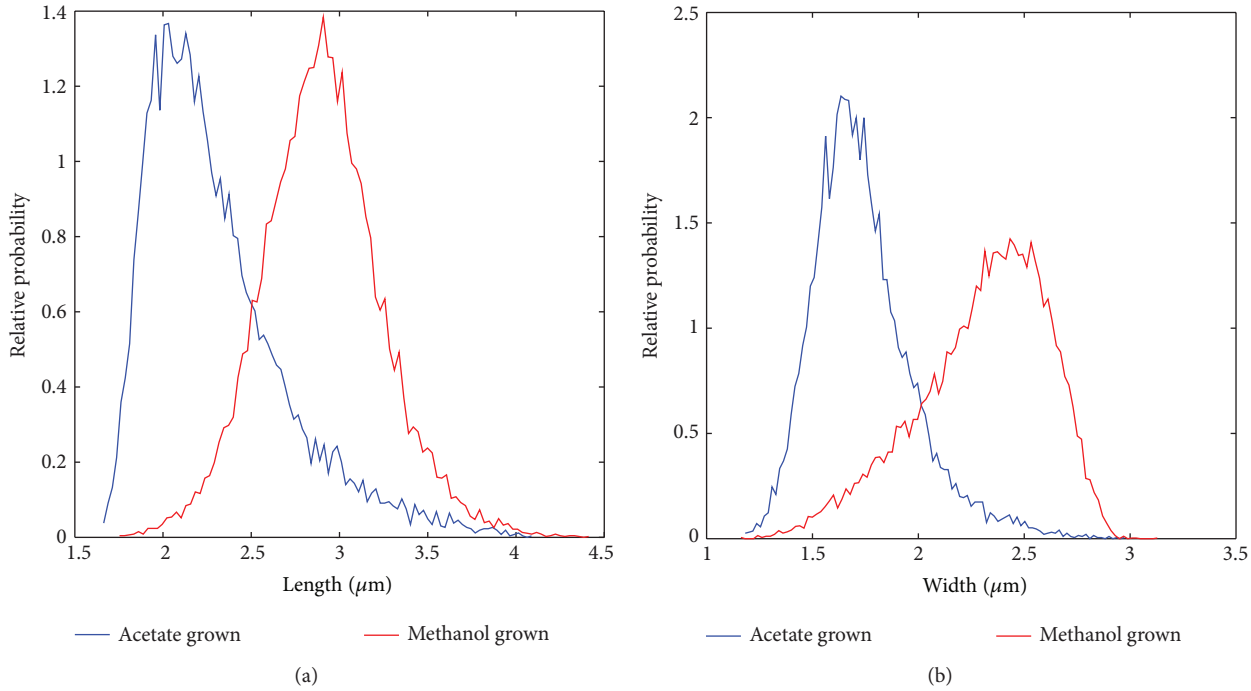


FIGURE 6: Distributions of (a) lengths and (b) widths of single *M. acetivorans* cells grown on methanol (red) and acetate (blue) as determined by DIC microscopy and image analysis. Data corresponds to approximately 10,000 cells.

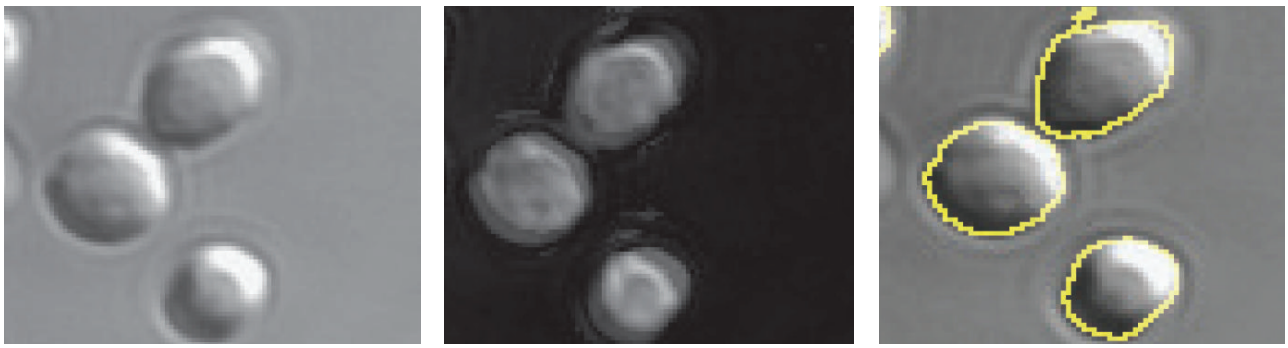


FIGURE 7: (a) DIC images of *Methanosarcina acetivorans* grown on methanol. (b) After applying Hilbert transform to DIC images and adding them to the original image, the boundaries of the cells become clearer. (c) CellProfiler is used to perform segmentation on images in (b). Here identified cell boundaries are superimposed onto original DIC image to illustrate segmentation.

identified in *Methanosarcina* spp. and one experiment characterized their role in regulation [72]. While TBP1 is required for growth, and likely the main transcription regulator, TBP2 and TBP3 are dispensable. These two differentially regulate approximately 123 genes on acetate versus methanol growth, and the authors of the study concluded that the two transcription factors optimized protein expression for low-energy substrate (e.g., acetate) growth [72]. These interactions are shown in Figure 12(a).

The second type of direct regulators—those interacting with methyltransferases—act as the mediators for methyl containing organic chemicals entering the methanogenesis pathway. There are separate methyltransferases for each substrate, including methanol, trimethylamine, dimethylamine, monomethylamine, and methylsulfide. Because methyltransferases are some of the most highly expressed genes, they are tightly regulated to preserve the energy balance in the cell [28]. Considerable experimental effort has uncovered eight methyltransferease specific regulators (Msr's). Msr's can act as both up- and downregulators. It was found that in the case of MsrA and MsrB, both proteins act in concert to upregulate expression of MtaCB1, and knockout of either can prevent expression [73]. Similarly, knockout of either *msrD* or *msrE* will prevent expression of MtaCB2 [73]. MsrD and to a lesser

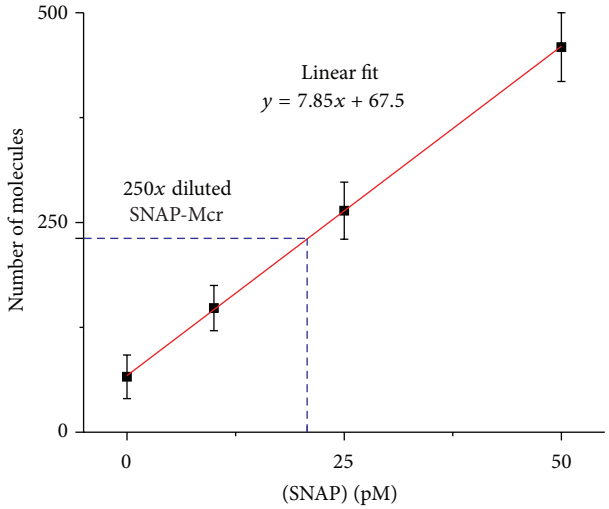


Figure 8: Calibration curve for SiMPull experiments that provides a mapping between protein concentration in cell lysate and number of spots observed on the slide.

extent MsrE also repress MtaCB3 [73]. Some Msr's upregulate one gene, while downregulating other genes; for example, MsrF enhances expression of methylsulfide methyltransferase *mtsD* on all growth substrates except methanol, while MsrC enhances expression of *mtsF* when on all three methylamines [74, 75]. The full set of interactions can be seen in Figure 12(a).

Nitrogen fixation regulation is the third direct set of regulation interactions identified. Two widely conserved nitrogen regulatory proteins named NrpRI and NrpRII have been studied in *Methanosarcinales*. In *M. mazei* Gö1 they were found to regulate 23 proteins that shared a similar DNA regulatory sequence. Overall about 5% of all of *M. mazei* Gö1 genes were regulated under nitrogen limitation, with 83 genes being upregulated [76]. Another study showed that several of the upregulated genes under nitrogen limitation were methylamine specific proteins [77]. However, support for direct regulation of all 83 genes by Nrp regulators was not established, and these connections are instead included in the indirect interaction map. The method of action for repression was shown to be that NrpRI binds the DNA and NrpRII which interacts directly with the TBPs, preventing the RNAP from binding [78]. Importantly, homologs of NrpRI and NrpRII have been identified in *M. acetivorans* that were differentially expressed under nitrogen limiting versus nitrogen sufficient growth [79], and it is likely that these highly conserved (92–94% identical amino acid sequences) regulators have similar function. An additional two small RNA (sRNA) molecules, $sRNA_{154}$ and $sRNA_{159}$ whose function is as of yet unknown, include Nrp binding sites upstream of the start codon [80].

One final set of strongly supported interactions include the repression of certain proteins involved in oxidative stress. As Isom et al. point out [81], the MsvR regulator is homologous to a well characterized variant in *Methanothermobacter thermautotrophicus* and 43 genes in addition to the *msvR* gene in *M. acetivorans* contain the two binding sequences upstream of the TATA box. Their study shows support for a homodimer with cysteine residues which likely are oxidized in an oxygen rich environment, causing the dimer to be released from the binding site.

Overall, the total number of interactions in the direct model is 248, with 10 regulators. Strengths of interaction, where known, are indicated by the width of the arrows in Figure 12.

3.5.2. Indirect/Hypothesized Interactions. Studies of nitrogen-related regulation in *M. mazei* [76] have led to the identification of 69 proteins that were differentially expressed by at least 3-fold [76, 77]. Of the proteins, 35 were involved in nitrogen and energy metabolism, 7 were transport system genes, and 10 were potential regulators. Of particular interest was the upregulation of the *mtb* and *mtm* genes used in methylamine degradation to generate energy or ammonia, the latter of which must be synthesized from N_2 under starvation conditions. Because many of the identified genes did not have the binding site for the Nrp regulators upstream of their start sites, they are likely regulated by another protein.

One of the more exciting regulations that has been discovered in archaea is an sRNA that targets both *cis*- and *trans*-encoding mRNAs called $sRNA_{162}$ [82]. Overexpression of $sRNA_{162}$ in *M. mazei* greatly upregulated many of the methylamine processing proteins. The work also implicated an ArsR family transcription factor as the mediating component in the regulation [82]. Homologs with high sequence identity (>90%) to both the sRNA and the ArsR regulators (gene MA1531) exist in *M. acetivorans*; therefore, we have included the same interactions in our hypothetical map.

MreA (*Methanosarcina* regulator of energy-converting metabolism) was recently implicated as the global regulator of methanogenesis after it was observed to be 38-fold more highly expressed on acetate than on TMA or methanol [59]. A study comparing the ratio of methanogenesis protein encoding genes in a strain containing a knockout to the wild type indicated that MreA acts to upregulate acetoclastic proteins and downregulate methylotrophic pathways [59], the latter of which is mediated by changes in expression of the Msr proteins previously discussed. Therefore, MreA could act as a switch between methanol and acetate utilization. Adding the interactions reported in the paper to the indirect graph allows MreA to have the greatest putative sphere of influence on gene expression (Figure 12(b)).

Cadmium resistance has been studied in a number of different archaea and bacteria. A well known CadC regulator represses cadmium resistance genes. It is stimulated to unbind by Cd^{2+}, Bi^{3+}, and Pb^{2+} [83]. This is a particularly interesting regulation as *M. acetivorans* growing on acetate in the presence of cadmium chloride shows between a two- and fivefold increase in methane production, likely attributed to higher levels of acetate kinase and carbonic anhydrase and lower phosphate kinase (Pta) [84]. Furthermore, it has been shown that levels of Coenzyme M increased roughly proportionally to Cd^{2+} concentration [85]. A homolog of the *cadC* gene, MA3940, likely regulates two cadmium efflux

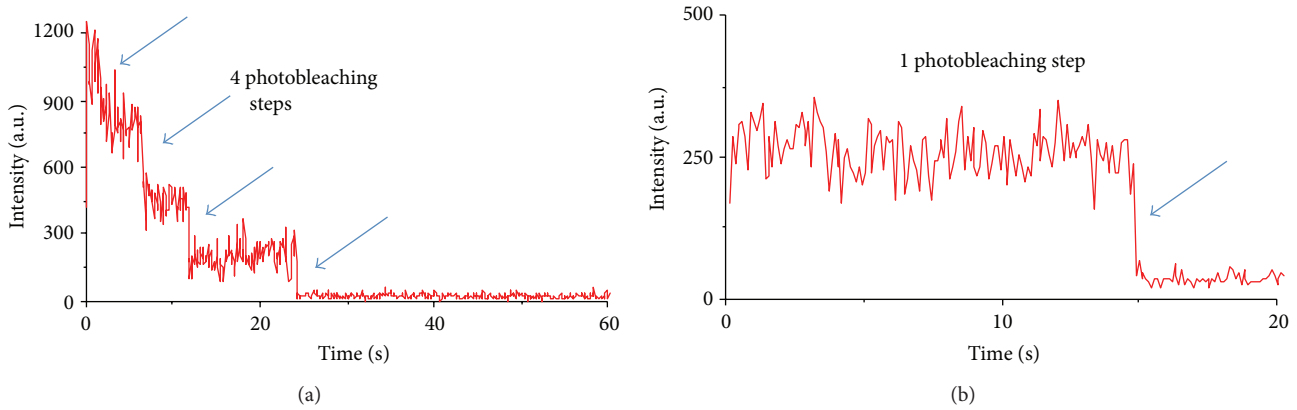


FIGURE 9: Multiple photobleaching steps in SiMPull experiments with rpl18p-SNAP strains (a) indicate multiple immobilized proteins as compared to pure SNAP protein which shows only single photobleaching step (b).

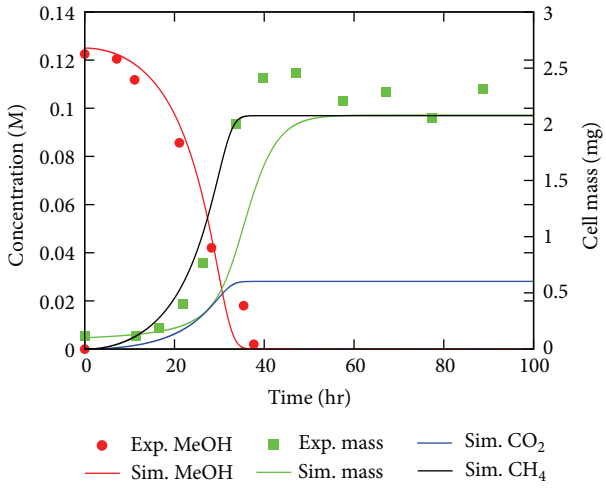


FIGURE 10: Results of the kinetic model. Comparison of the kinetic model for growth of *M. acetivorans* culture on 125 mM methanol to the experiment to which it was fit [30]. Lines indicate model results while symbols indicate experimental measurements.

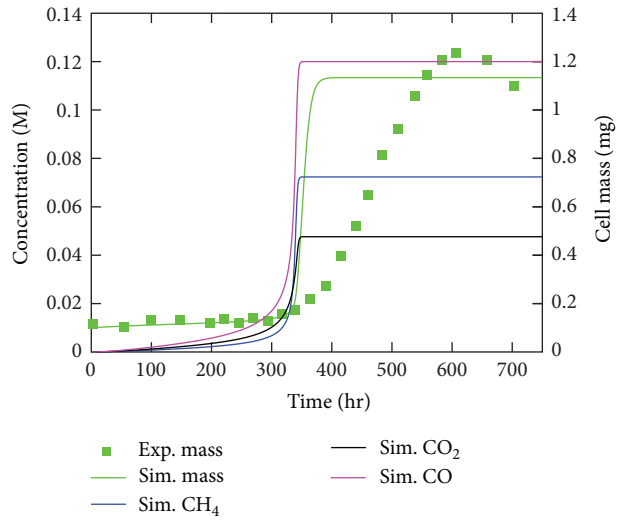
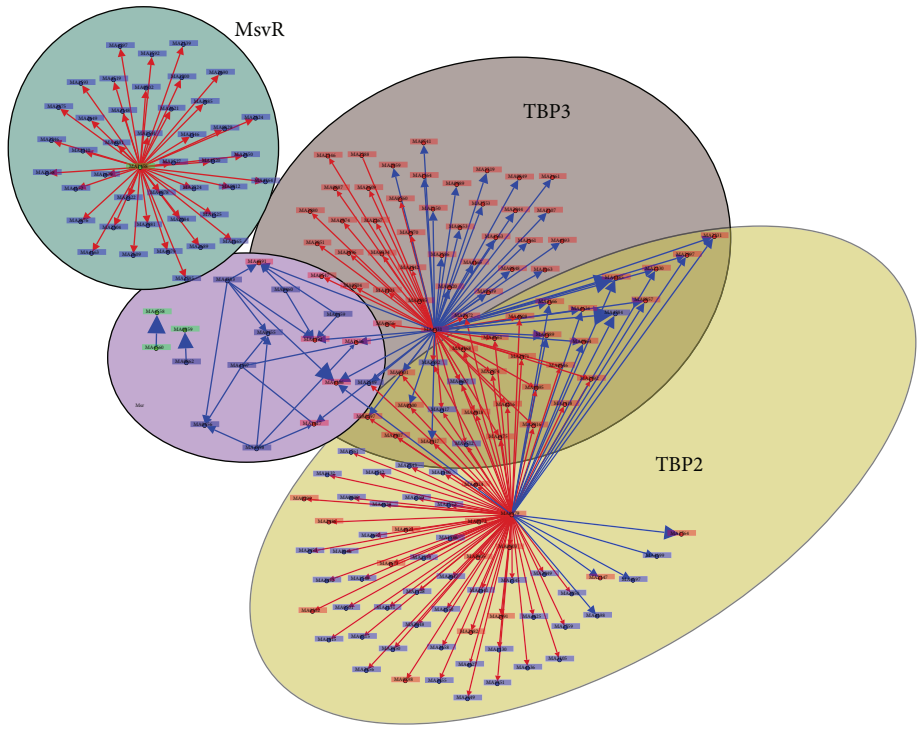


FIGURE 11: Results of the kinetic model. Comparison of the kinetic model for growth of *M. acetivorans* culture on 120 mM acetate to the experiment [31]. Lines indicate model results while symbols indicate experimental measurements.

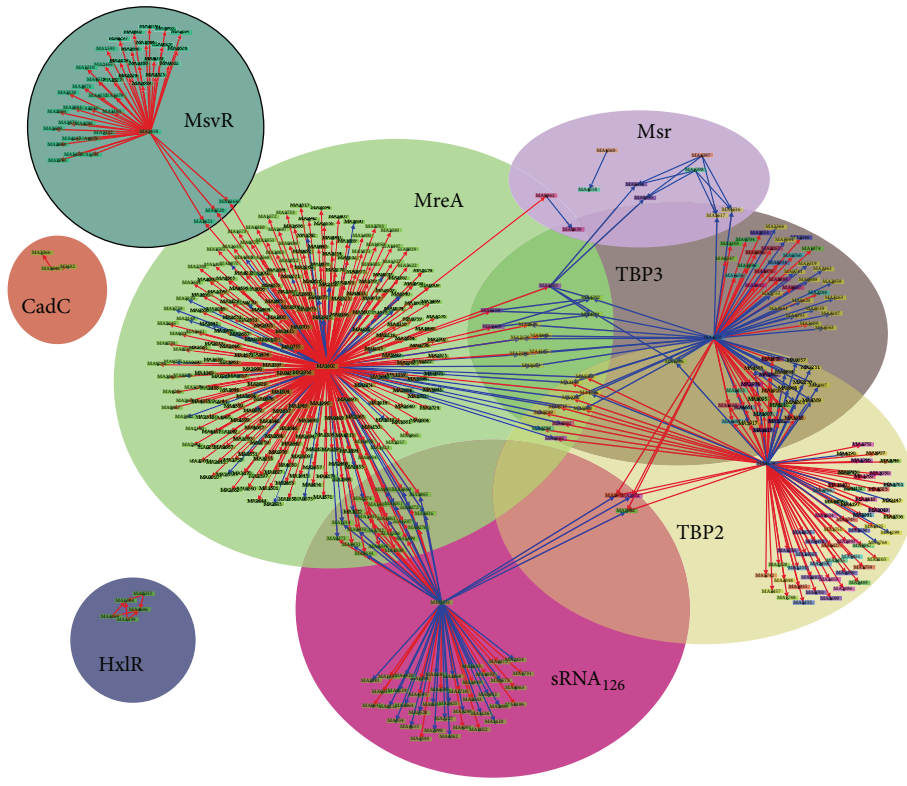
encoding genes MA3366 and MA3632 in *M. acetivorans*. In addition, evidence exists for a putative interaction between CadC or one of the genes it regulates and *ack*, *pta*, and carbonic anhydrase. The former two of these interactions are shown in Figure 12(b).

3.5.3. Methanogenesis Gene Regulation. Simplification of the two regulation maps to only those interactions that directly modulate expression of methanogenesis genes yields the map shown in Figure 13. From this map it can be seen that two regulators ($sRNA_{162}$ and MreA) interact broadly with methanogenesis genes. Correlations of regulator and methanogenesis proteins across several substrates (data not shown) indicate three broad classes of regulation: methanol, acetate, and other methyl-containing substrates (MMA, DMA, TMA, and MMS). Regulation motifs usually embody one of two modes of action in many different species: (1) a global regulator that activates/represses many genes or (2) activation of a single or handful of genes via very specific interaction [86]. Therefore, a possible energetically efficient way to regulate metabolism may be to have a few global regulators that activate/repress many genes, some of which may in turn act as specific regulators capable of fine tuning individual gene expression.

With this knowledge in mind and assuming that the MreA and $sRNA_{162}$ regulators interact in the way proposed in the literature, it can be hypothesized that between the two regulators the three classes of regulation can be covered. In this hypothesis MreA is a global regulator that facilitates the switch between methyl substrates and acetate. It does this primarily by turning off the CO_2 efflux pathway while turning on the acetate utilization pathway. Upregulation of TMA, DMA, and MMA utilizing proteins is accomplished by expression of the specific regulator $sRNA_{162}$, which turns off



(a) Direct interactions



(b) Indirect interactions

FIGURE 12: Graph representations of the direct and indirect regulations and associated spheres of influence. Red arrows indicate downregulation and blue arrows indicate up regulation. The regulator name is indicated by the large black text.

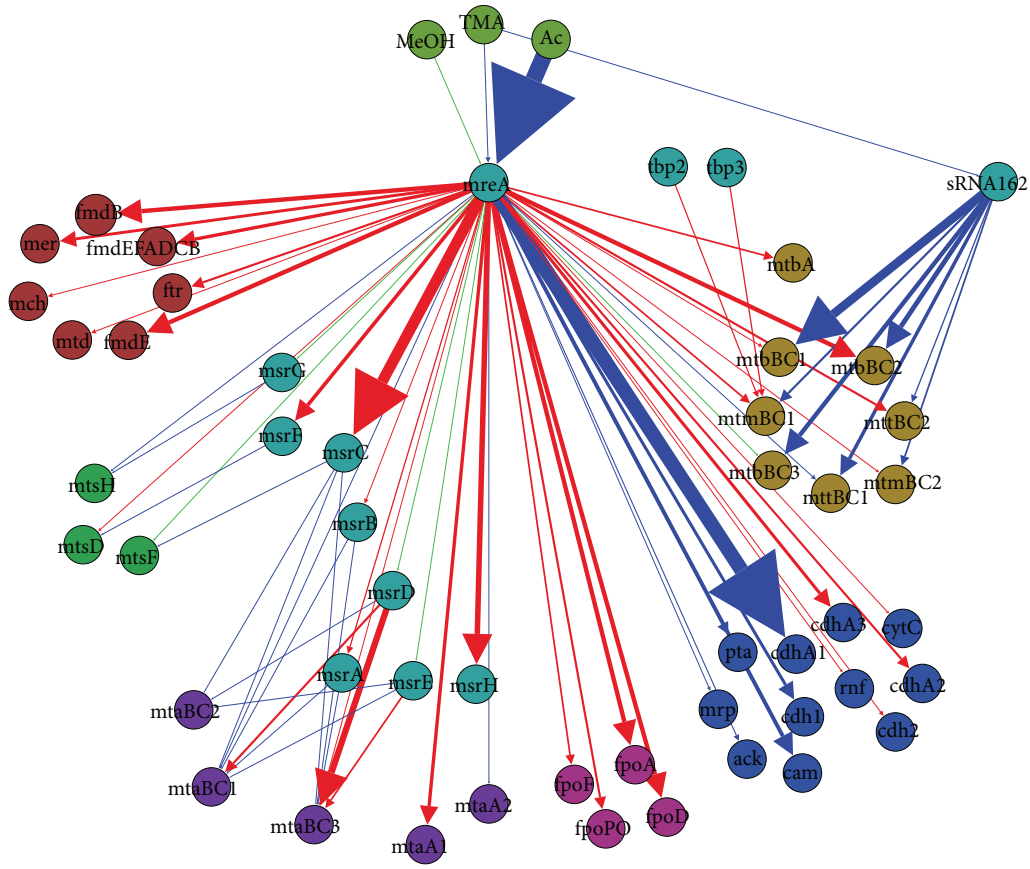


Figure 13: A graph representing the regulation of methanogenesis genes. Red arrows indicate downregulation, blue arrows indicate upregulation, and green arrows indicate weak regulation. The thickness of the arrow shows the strength of the interaction. Turquoise nodes are known regulators. Red nodes are CO_2 efflux proteins. Olive green nodes are growth substrates. Green nodes are methyl-sulfide methyltransferase proteins. Purple nodes are methanol methyltransferase proteins. Blue nodes are acetate utilization proteins and gold nodes are methylamine methyltransferase proteins.

expression MA1531, which may be a repressor of methylamine genes.

4. Conclusions

We have shown that SiMPull can be used as a method to measure the number of proteins in anaerobic organisms. This entailed genetically engineering a SNAP tag gene into the chromosome along with the gene encoding for the protein of interest, McrG and ribosomal protein Rpl18p. Using this technique we were able to estimate the number of protein complexes in single cells in their exponential growth phase which is important data for modeling. With Mcr's unique position in the methanogenic pathway, knowledge of its copy number is important for modeling metabolic dynamics. As ribosomes are known to be one of the dominant components of molecular crowding, their numbers are important to generate accurate *in silico* whole cell models of methanogens.

Using SiMPull and RNA-seq expression data from monoclonal cell cultures of the methanogens growing on acetate and methanol, we were able to estimate the number of proteins in the methanogenesis pathways. Coupling the resulting cell mass growth reaction to the methanogenesis reactions, we were able to fit unknown rate constants to experiments for growth on methanol. Applying the model to growth on acetate, we were able to capture the correct timescale for use of acetate and production of the methane; however, the cell mass growth rate in exponential phase was too high.

In order to apply the model to more complex scenarios, especially time-varying growth substrate conditions potentially found in the environment, we need a regulation mechanism for expression of the proteins. Examining correlations of protein expression across different substrates leads to the observation that there appear to be three classes of growth: methanol, acetate, and another methyl substrate (TMA, DMA, MMA, and MMS). Towards the goal of developing a regulation model, we have compiled known transcriptional regulation with putative regulation interactions to create a draft model for *M. acetivorans*. Reducing the draft regulation map to just interactions with methanogenesis protein encoding genes, two regulators arise as global regulators. MreA appears to switch between the acetoclastic pathway and the CO_2 efflux pathway and, therefore, is hypothesized as

the switch between acetoclastic growth and methylotrophic growth. $sRNA_{162}$ appears to turn on expression of genes necessary for utilizing methylamines and, therefore, optimizes the organism for methylamine growth.

The physical and stoichiometric properties and kinetic model reported here complement the metabolic reconstructions and constitutes significant progress towards a full computational model of *M. acetivorans*. Spatial heterogeneity, such as that caused by large crowders like the ribosome, is known to cause stochastic effects in similar cells from a monoclonal culture; therefore, quantifying the number and distribution is necessary. Because many reactions in methanogenesis occur in the membrane, stochasticity due to the local environment could have a large effect. Larger spatial organization, such as membrane bound protein complex locality and number, can be determined by cryoelectron tomography. Such data could be used with the kinetic and regulation models developed here to construct detailed full cell reaction-diffusion models similar to those that have been created previously [87]. Such models would allow study of stochasticity in individual organisms. Ultimately, these models could be used with hybrid reaction-diffusion master equation/flux balance analysis techniques [88] that provide full metabolic modeling with spatial effects due to cell culture organization. The utility of the computational models is that they should be easily extendable to the other *Methanosarcina* spp.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank John A. Cole for helpful review of this paper and Elijah Roberts for the code that segments cells as well as initial DIC measurements. This work was supported by Office of Science (BER), U.S. Department of Energy Grant no. DE-FG02-10ER6510, NASA Astrobiology Institute Grants nos. NNA13AA91A (ZLS, PL) and DE-FG02-02ER15296 (WWM, JRE, and PRAK) National Institutes of Health (NIH) grants U19 AI083025 and GM065367 (TH and AJ) and National Science Foundation (NSF) Physics Frontier Center Grant no. PHY-0822613 (TH and ZLS), and Molecular Biophysics Training Grant no. PHS 5 T32 GM008276 (JRP).

References

[1] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[2] S. Winker and C. R. Woese, "A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics," *Systematic and Applied Microbiology*, vol. 14, no. 4, pp. 305–310, 1991.

[3] G. E. Fox, L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese, "Classification of methanogenic bacteria by 16S ribosomal RNA characterization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 10, pp. 4537–4541, 1977.

[4] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.

[5] C. R. Woese, "Bacterial evolution," *Microbiological Reviews*, vol. 51, no. 2, pp. 221–271, 1987.

[6] R. R. Gutell and C. R. Woese, "Higher order structural elements in ribosomal RNAs: pseudo-knots and the use of noncanonical pairs," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 2, pp. 663–667, 1990.

[7] E. Roberts, A. Sethi, J. Montoya, C. R. Woese, and Z. Luthey-Schulten, "Molecular signatures of ribosomal evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 37, pp. 13953–13958, 2008.

[8] C. Woese, "The universal ancestor," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 12, pp. 6854–6859, 1998.

[9] C. R. Woese, "On the evolution of cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8742–8747, 2002.

[10] C. R. Woese, G. J. Olsen, M. Ibba, and D. Söll, "Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 1, pp. 202–236, 2000.

[11] P. O'Donoghue and Z. Luthey-Schulten, "Evolution of structure in aminoacyl-tRNA synthetases," *Microbiology and Molecular Biology Reviews*, vol. 67, no. 4, pp. 550–573, 2003.

[12] P. O'Donoghue, A. Sethi, C. R. Woese, and Z. A. Luthey-Schulten, "The evolutionary history of Cys-tRNACys formation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 52, pp. 19003–19008, 2005.

[13] J. N. Reeve, J. Nölling, R. M. Morgan, and D. R. Smith, "Methanogenesis: genes, genomes, and who's on first?" *Journal of Bacteriology*, vol. 179, no. 19, pp. 5975–5986, 1997.

[14] J. R. Brown and W. F. Doolittle, "Archaea and the prokaryote-to-eukaryote transition," *Microbiology and Molecular Biology Reviews*, vol. 61, no. 4, pp. 456–502, 1997.

[15] D. E. Graham, R. Overbeek, G. J. Olsen, and C. R. Woese, "An archaeal genomic signature," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 7, pp. 3304–3308, 2000.

[16] B. Gao and R. S. Gupta, "Phylogenomic analysis of proteins that are distinctive of archaea and its main subgroups and the origin of methanogenesis," *BMC Genomics*, vol. 8, article 86, 2007.

[17] F. Gaillard, B. Scaillet, and N. T. Arndt, "Atmospheric oxygenation caused by a change in volcanic degassing pressure," *Nature*, vol. 478, no. 7368, pp. 229–232, 2011.

[18] R. K. Thauer, A.-K. Kaster, H. Seedorf, W. Buckel, and R. Hedderich, "Methanogenic archaea: ecologically relevant differences in energy conservation," *Nature Reviews Microbiology*, vol. 6, no. 8, pp. 579–591, 2008.

[19] S. Burggraf, H. Fricke, A. Neuner et al., "*Methanococcus igneus* sp. nov., a novel hyperthermophilic methanogen from a shallow submarine hydrothermal system," *Systematic and Applied Microbiology*, vol. 13, no. 3, pp. 263–269, 1990.

[20] R. J. Cicerone and R. S. Oremland, "Biogeochemical aspects of atmospheric methane," *Global Biogeochem Cycles*, vol. 2, no. 4, pp. 299–327, 1988.

[21] A. M. Feist, J. C. M. Scholten, B. Ø. Palsson, F. J. Brockman, and T. Ideker, "Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*," *Molecular Systems Biology*, vol. 2, no. 1, 2006.

[22] M. N. Benedict, M. C. Gonnerman, W. W. Metcalf, and N. D. Price, "Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon *Methanosarcina acetivorans* C2A," *Journal of Bacteriology*, vol. 194, no. 4, pp. 855–865, 2012.

[23] P. Labhsetwar, J. Cole, E. Roberts, N. Price, and Z. Luthey-Schulten, "Heterogeneity in protein expression induces metabolic variability in a modeled *Escherichia coli* population," *Proceedings of the National Academy of Sciences*, vol. 110, no. 34, 2013.

[24] C. J. Bult, O. White, G. J. Olsen et al., "Complete genome sequence of the Methanogenic archaeon, Methanococcus jannaschii," *Science*, vol. 273, no. 5278, pp. 1058–1073, 1996.

[25] J. E. Galagan, C. Nusbaum, A. Roy et al., "The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity," *Genome Research*, vol. 12, no. 4, pp. 532–542, 2002.

[26] M. Rother and W. W. Metcalf, "Anaerobic growth of *Methanosarcina acetivorans* C2A on carbon monoxide: an unusual way of life for a *methanogenic archaeon*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 48, pp. 16929–16934, 2004.

[27] J. G. Ferry and C. H. House, "The stepwise evolution of early life driven by energy conservation," *Molecular Biology and Evolution*, vol. 23, no. 6, pp. 1286–1292, 2006.

[28] P. Browne and H. Cadillo-Quiroz, "Contribution of transcriptomics to systems-level understanding of *Methanogenic Archaea*," *Archaea*, vol. 2013, Article ID 586369, 11 pages, 2013.

[29] A. Jain, R. Liu, B. Ramani et al., "Probing cellular protein complexes using single-molecule pull-down," *Nature*, vol. 473, no. 7348, pp. 484–488, 2011.

[30] A. Bose, M. A. Pritchett, M. Rother, and W. W. Metcalf, "Differential regulation of the three methanol methyltransferase isozymes in *Methanosarcina acetivorans* C2A," *Journal of Bacteriology*, vol. 188, no. 20, pp. 7274–7283, 2006.

[31] N. R. Buan and W. W. Metcalf, "Methanogenesis by *Methanosarcina acetivorans* involves two structurally and functionally distinct classes of heterodisulfide reductase," *Molecular Microbiology*, vol. 74, no. 4, pp. 843–853, 2010.

[32] C. Ingram-Smith, A. Gorrell, S. H. Lawrence, P. Iyer, K. Smith, and J. G. Ferry, "Characterization of the acetate Binding Pocket in the *Methanosarcina thermophila* acetate Kinase," *Journal of Bacteriology*, vol. 187, no. 7, pp. 2386–2394, 2005.

[33] A. Gorrell, S. H. Lawrence, and J. G. Ferry, "Structural and kinetic analyses of arginine residues in the active site of the acetate kinase from *Methanosarcina thermophila*," *Journal of Biological Chemistry*, vol. 280, no. 11, pp. 10731–10742, 2005.

[34] S. H. Lawrence, K. B. Luther, H. Schindelin, and J. G. Ferry, "Structural and functional studies suggest a catalytic mechanism for the phosphotransacetylase from *Methanosarcina thermophila*," *Journal of Bacteriology*, vol. 188, no. 3, pp. 1143–1154, 2006.

[35] M. E. Rasche, K. S. Smith, and J. G. Ferry, "Identification of cysteine and arginine residues essential for the phosphotransacetylase from *Methanosarcina thermophila*," *Journal of Bacteriology*, vol. 179, no. 24, pp. 7712–7717, 1997.

[36] D. A. Grahame and T. C. Stadtman, "Carbon monoxide dehydrogenase from *Methanosarcina barkeri*. Disaggregation, purification, and physicochemical properties of the enzyme," *Journal of Biological Chemistry*, vol. 262, no. 8, pp. 3706–3712, 1987.

[37] R. I. L. Eggen, R. van Kranenburg, A. J. M. Vriesema et al., "Carbon monoxide dehydrogenase from *Methanosarcina frisia* Göl: characterization of the enzyme and the regulated expression of two operon-like *cdh* gene clusters," *Journal of Biological Chemistry*, vol. 271, no. 24, pp. 14256–14263, 1996.

[38] M. A. Pritchett and W. W. Metcalf, "Genetic, physiological and biochemical characterization of multiple methanol methyltransferase isozymes in *Methanosarcina acetivorans* C2A," *Molecular Microbiology*, vol. 56, no. 5, pp. 1183–1194, 2005.

[39] R. B. Opulencia, A. Bose, and W. W. Metcalf, "Physiology and posttranscriptional regulation of methanol:coenzyme M methyltransferase isozymes in *Methanosarcina acetivorans* C2A," *Journal of Bacteriology*, vol. 191, no. 22, pp. 6928–6935, 2009.

[40] K. Ma and R. K. Thauer, "Purification and properties of N5,N10-methylenetetrahydromethanopterin reductase from *Methanobacterium thermoautotrophicum* (strain Marburg)," *European Journal of Biochemistry*, vol. 191, no. 1, pp. 187–193, 1990.

[41] B. W. J. te Brommelstroet, W. J. Geerts, J. T. Keltjens, C. van der Drift, and G. D. Vogels, "Purification and properties of 5,10-methylenetetrahydromethanopterin dehydrogenase and 5,10-methylenetetrahydromethanopterin reductase, two coenzyme F420-dependent enzymes, from *Methanosarcina barkeri*," *Biochimica et Biophysica Acta*, vol. 1079, no. 3, pp. 293–302, 1991.

[42] S. Shima and R. K. Thauer, "Tetrahydromethanopterin-specific enzymes from *Methanopyrus kandleri*," *Methods in Enzymology*, vol. 331, pp. 317–353, 2001.

[43] B. W. te Brommelstroet, C. M. H. Hensgens, W. J. Geerts, J. T. Keltjens, C. van der Drift, and G. D. Vogels, "Purification and properties of 5,10-methenyltetrahydromethanopterin cyclohydrolase from *Methanosarcina barkeri*," *Journal of Bacteriology*, vol. 172, no. 2, pp. 564–571, 1990.

[44] M. Karrasch, G. Borner, M. Enssle, and R. K. Thauer, "The molybdoenzyme formylmethanofuran dehydrogenase from *Methanosarcina barkeri* contains a pterin cofactor," *European Journal of Biochemistry*, vol. 194, no. 2, pp. 367–372, 1990.

[45] P. E. Jablonski and J. G. Ferry, "Purification and properties of methyl coenzyme M methylreductase from acetate-grown *Methanosarcina thermophila*," *Journal of Bacteriology*, vol. 173, no. 8, pp. 2481–2487, 1991.

[46] E. Murakami, U. Deppenmeier, and S. W. Ragsdale, "Characterization of the intramolecular electron transfer pathway from 2-Hydroxyphenazine to the heterodisulfide reductase from *Methanosarcina thermophila*," *Journal of Biological Chemistry*, vol. 276, no. 4, pp. 2432–2439, 2001.

[47] R. Iino, R. Hasegawa, K. V. Tabata, and H. Noji, "Mechanism of inhibition by C-terminal α-helices of the ϵ subunit of *Escherichia coli* FoF1-ATP synthase," *Journal of Biological Chemistry*, vol. 284, no. 26, pp. 17457–17464, 2009.

[48] K. R. Sowers, J. E. Boone, and R. P. Gunsalus, "Disaggregation of *Methanosarcina* spp. and growth as single cells at elevated osmolarity," *Applied and Environmental Microbiology*, vol. 59, no. 11, pp. 3832–3839, 1993.

[49] W. W. Metcalf, J.-K. Zhang, X. Shi, and R. S. Wolfe, "Molecular, genetic, and biochemical characterization of the serC gene of *Methanosarcina barkeri* Fusaro," *Journal of Bacteriology*, vol. 178, no. 19, pp. 5797–5802, 1996.

[50] W. W. Metcalf, J. K. Zhang, E. Apolinario, K. R. Sowers, and R. S. Wolfe, "A genetic system for Archaea of the genus

Methanosarcina: liposome-mediated transformation and construction of shuttle vectors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 6, pp. 2626–2631, 1997.

[51] M. A. Pritchett, J. K. Zhang, and W. W. Metcalf, "Development of a markerless genetic exchange method for *Methanosarcina acetivorans* C2A and its use in construction of new genetic tools for methanogenic archaea," *Applied and Environmental Microbiology*, vol. 70, no. 3, pp. 1425–1433, 2004.

[52] A. M. Guss, M. Rother, J. K. Zhang, G. Kulkarni, and W. W. Metcalf, "New methods for tightly regulated gene expression and highly efficient chromosomal integration of cloned genes for Methanosarcina species," *Archaea*, vol. 2, no. 3, pp. 193–203, 2008.

[53] W. W. Metcalf, J. K. Zhang, and R. S. Wolfe, "An anaerobic, intrachamber incubator for growth of *Methanosarcina* spp. on methanol-containing solid media," *Applied and Environmental Microbiology*, vol. 64, no. 2, pp. 768–770, 1998.

[54] K. A. Datsenko and B. L. Wanner, "One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 12, pp. 6640–6645, 2000.

[55] F. J. Stewart, E. A. Ottesen, and E. F. Delong, "Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics," *ISME Journal*, vol. 4, no. 7, pp. 896–907, 2010.

[56] R. McClure, D. Balasubramania, Y. Sun et al., "Computational analysis of bacterial RNA-seq data," *Nucleic Acids Research*, vol. 41, no. 14, p. e140, 2013.

[57] A. Jain, R. Liu, Y. K. Xiang, and T. Ha, "Single-molecule pull-down for studying protein interactions," *Nature Protocols*, vol. 7, no. 3, pp. 445–452, 2012.

[58] W. Grabarse, F. Mahlert, S. Shima, R. K. Thauer, and U. Ermler, "Comparison of three methyl-coenzyme M reductases from phylogenetically distant organisms: unusual amino acid modification, conservation and adaptation," *Journal of Molecular Biology*, vol. 303, no. 2, pp. 329–344, 2000.

[59] M. Reichlen, V. Vepachedu, K. Murakami, and J. Ferry, "MreA functions in the global regulation of methanogenic pathways in *Methanosarcina acetivorans*," *mBio*, vol. 3, no. 4, Article ID e00189-12, 2012.

[60] S. T. Yang and M. R. Okos, "Kinetic study and mathematical modeling of methanogenesis of acetate using pure cultures of methanogens," *Biotechnology and Bioengineering*, vol. 30, no. 5, pp. 661–667, 1987.

[61] S. Hoops, S. Sahle, R. Gauges et al., "COPASI—a complex pathway sImulator," *Bioinformatics*, vol. 22, no. 24, pp. 3067–3074, 2006.

[62] I. Schomburg, A. Chang, S. Placzek et al., "BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA," *Nucleic Acids Research*, vol. 41, pp. 764–772, 2013.

[63] R. K. Thauer, "Biochemistry of methanogenesis: a tribute to Marjory Stephenson: 1998 Marjory Stephenson Prize Lecture," *Microbiology*, vol. 144, no. 9, pp. 2377–2406, 1998.

[64] R. K. Thauer, K. Jungermann, and K. Decker, "Energy conservation in chemotrophic anaerobic bacteria," *Bacteriological Reviews*, vol. 41, no. 1, pp. 100–180, 1977.

[65] M. W. Peck, "Changes in concentrations of coenzyme F420 analogs during batch growth of *Methanosarcina barkeri* and *Methanosarcina mazei*," *Applied and Environmental Microbiology*, vol. 55, no. 4, pp. 940–945, 1989.

[66] B. Obara, M. Roberts, J. Armitage, and V. Grau, "Bacterial cell identification in differential interference contrast microscopy images," *BMC Bioinformatics*, vol. 14, no. 134, 2013.

[67] A. E. Carpenter, T. R. Jones, M. R. Lamprecht et al., "CellProfiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biology*, vol. 7, no. 10, 2006.

[68] C. Wrede, U. Walbaum, A. Ducki, I. Heieren, and M. and Hoppert, "Localization of methyl-coenzyme M reductase as metabolic marker for diverse *Methanogenic archaea*," *Archaea*, vol. 2013, Article ID 920241, 7 pages, 2013.

[69] L. Rohlin and R. P. Gunsalus, "Carbon-dependent control of electron transfer and central carbon pathway genes for methane biosynthesis in the Archaean, *Methanosarcina acetivorans* strain C2A," *BMC Microbiology*, vol. 10, article 62, 2010.

[70] U. Ermler, W. Grabarse, S. Shima, M. Goubeaud, and R. K. Thauer, "Crystal structure of methyl-coenzyme M reductase: the key enzyme of biological methane formation," *Science*, vol. 278, no. 5342, pp. 1457–1462, 1997.

[71] B. E. Krenn, H. S. van Walraven, M. J. C. Scholts, and R. Kraayenhof, "Modulation of the proton-translocation stoichiometry of H^+-ATP synthases in two phototrophic prokaryotes by external pH," *Biochemical Journal*, vol. 294, no. 3, pp. 705–709, 1993.

[72] M. J. Reichlen, K. S. Murakami, and J. G. Ferry, "Functional analysis of the three TATA binding protein homologs in *Methanosarcina acetivorans*," *Journal of Bacteriology*, vol. 192, no. 6, pp. 1511–1517, 2010.

[73] A. Bose and W. W. Metcalf, "Distinct regulators control the expression of methanol methyltransferase isozymes in *Methanosarcina acetivorans* C2A," *Molecular Microbiology*, vol. 67, no. 3, pp. 649–661, 2008.

[74] A. Bose, G. Kulkarni, and W. W. Metcalf, "Regulation of putative methyl-sulphide methyltransferases in *Methanosarcina acetivorans* C2A," *Molecular Microbiology*, vol. 74, no. 1, pp. 227–238, 2009.

[75] E. Oelgeschläger and M. Rother, "*In vivo* role of three fused corrinoid/methyl transfer proteins in *Methanosarcina acetivorans*," *Molecular Microbiology*, vol. 72, no. 5, pp. 1260–1272, 2009.

[76] K. Veit, C. Ehlers, A. Ehrenreich et al., "Global transcriptional analysis of *Methanosarcina mazei* strain Gö1 under different nitrogen availabilities," *Molecular Genetics and Genomics*, vol. 276, no. 1, pp. 41–55, 2006.

[77] K. Veit, C. Ehlers, and R. A. Schmitz, "Effects of nitrogen and carbon sources on transcription of soluble methyltransferases in *Methanosarcina mazei* strain Gö1," *Journal of Bacteriology*, vol. 187, no. 17, pp. 6147–6154, 2005.

[78] K. Weidenbach, C. Ehlers, J. Kock, and R. A. Schmitz, "NrpRII mediates contacts between NrpRI and general transcription factors in the archaeon Methanosarcina mazei Gö1," *FEBS Journal*, vol. 277, no. 21, pp. 4398–4411, 2010.

[79] T. J. Lie, J. A. Dodsworth, D. C. Nickle, and J. A. Leigh, "Diverse homologues of the archaeal repressor NrpR function similarly in nitrogen regulation," *FEMS Microbiology Letters*, vol. 271, no. 2, pp. 281–288, 2007.

[80] D. Jäger, C. M. Sharma, J. Thomsen, C. Ehlers, J. Vogel, and R. A. Schmitz, "Deep sequencing analysis of the *Methanosarcina mazei* Gö 1 transcriptome in response to nitrogen availability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, pp. 21878–21882, 2009.

[81] C. Isom, J. Turner, D. Lessner, and E. Karr, "Redox-sensitive DNA binding by homodimeric *Methanosarcina acetivorans*

MsvR is modulated by cysteine residues," *BMC Micro-Biology*, vol. 13, no. 163, 2013.

[82] D. Jäger, S. Pernitzsch, A. Richter, R. Backofen, C. Sharma, and R. Schmitz, "An archaeal sRNA targeting cis- and trans-encoded mRNAs via two distinct domains," *Nucleic Acids Research*, vol. 40, no. 21, pp. 10964–10979, 2012.

[83] G. Endo and S. Silver, "CadC, the transcriptional regulatory protein of the cadmium resistance system of *Staphylococcus aureus* plasmid pI258," *Journal of Bacteriology*, vol. 177, no. 15, pp. 4437–4441, 1995.

[84] E. Lira-Silva, M. Santiago-Martinez, V. Hernández-Juárez, R. Garcia-Contreras, R. Moreno-Sánchez, and R. Jasso-Chávez, "Activation of methanogenesis by cadmium in the marine archaeon *Methanosarcina acetivorans*," *PLoS ONE*, vol. 7, no. 11, 2012.

[85] E. Lira-Silva, M. Santiago-Martinez, R. Garcia-Contreras et al., "Cd^{2+} resistance mechanisms in *Methanosarcina acetivorans* involve the increase in coenzyme M content and induction of biofilm synthesis," *Environmental Microbiology Reports*, vol. 5, no. 6, Article ID 10.1111/1758-2229.12080, pp. 799–808, 2013.

[86] L. Gerosa, K. Kochanowski, M. Heinemann, and U. Sauer, "Dissecting specific and global transcriptional regulation of bacterial gene expression," *Molecular Systems Biology*, vol. 9, p. 658, 2013.

[87] E. Roberts, A. Magis, J. O. Ortiz, W. Baumeister, and Z. Luthey-Schulten, "Noise contributions in an inducible genetic switch: a whole-cell simulation study," *PLOS Computational Biology*, vol. 7, no. 3, p. e1002010, 2011.

[88] J. Cole, M. Hallock, P. Labhsetwar, J. Peterson, J. Stone, and Z. Luthey-Schulten, *Stochastic Simulations of Cellular Processes: From Single Cells to Colonies*, Chapter 13, Academic Press, 2013.

Hindawi Publishing Corporation
Archaea
Volume 2014, Article ID 206735, 24 pages
http://dx.doi.org/10.1155/2014/206735

Review Article

Archaeal Genome Guardians Give Insights into Eukaryotic DNA Replication and Damage Response Proteins

David S. Shin,[1] Ashley J. Pratt,[1] and John A. Tainer[1,2]

[1] *Life Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 6R2100, Berkeley, CA 94720, USA*
[2] *Department of Integrative Structural and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, MB-4, La Jolla, CA 92037, USA*

Correspondence should be addressed to John A. Tainer; jat@scripps.edu

Received 21 June 2013; Revised 27 September 2013; Accepted 29 November 2013; Published 20 February 2014

Academic Editor: Celine Brochier-Armanet

Copyright © 2014 David S. Shin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As the third domain of life, archaea, like the eukarya and bacteria, must have robust DNA replication and repair complexes to ensure genome fidelity. Archaea moreover display a breadth of unique habitats and characteristics, and structural biologists increasingly appreciate these features. As archaea include extremophiles that can withstand diverse environmental stresses, they provide fundamental systems for understanding enzymes and pathways critical to genome integrity and stress responses. Such archaeal extremophiles provide critical data on the periodic table for life as well as on the biochemical, geochemical, and physical limitations to adaptive strategies allowing organisms to thrive under environmental stress relevant to determining the boundaries for life as we know it. Specifically, archaeal enzyme structures have informed the architecture and mechanisms of key DNA repair proteins and complexes. With added abilities to temperature-trap flexible complexes and reveal core domains of transient and dynamic complexes, these structures provide insights into mechanisms of maintaining genome integrity despite extreme environmental stress. The DNA damage response protein structures noted in this review therefore inform the basis for genome integrity in the face of environmental stress, with implications for all domains of life as well as for biomanufacturing, astrobiology, and medicine.

1. Introduction

From the ideas of Lamarck, to Darwin and Mendel, to Huxley and those involved in the modern evolutionary synthesis, the accepted views behind the driving force of inheritance and evolution have certainly themselves "evolved" over time. Likewise, how we categorize different life forms has similarly evolved. Before the advent of tools like the microscope, natural intuition would seem to differentiate living things as either plant or animal. In the mid-1800s, Haeckel expanded these categorical divisions of life, which, by the 1960s, eventually grew to 5 "Kingdoms," allowing the *Monera*, *Protista,* and *Fungi* to be positioned alongside the *Animalia* and *Plantae* [1, 2]. Later, the two-empire system, whose highest level encompassed prokaryotes ("before the kernel," i.e., lacking a membrane-bound nucleus) and the eukaryotes (containing a "true kernel"), gained attraction as a model to classify life alongside the 5 kingdom system [3]. Beliefs at the time were that these predominantly unicellular organisms that lacked a nucleus must be a less complex predecessor of the more complex eukaryotic cells.

Fast-forwarding to more recent times, the advent of DNA sequencing, along with its subsequent improvements in cost and speed, led to a redefining of the evolutionary divisions of life by providing a shift from phenotypic taxonomy to genotypic computationally aided phylogenetics. Previously, archaea and bacteria were combined as prokaryotes (or monera), based largely upon their similar features (or lack thereof) when compared to eukaryotes: the lack of membrane-bound nuclei and organelles, and being generally unicellular [3]. The sequencing of ribosomal RNA genes eventually led to the separation and clustering of the prokaryotes, which ultimately gave rise to the main, top-level classification systems used today, defining 3 "domains" of life: archaea and bacteria along with the eukarya [4, 5].

For a general understanding of genome integrity, archaea provide master keys to understanding enzymes and pathways critical to stress response and genome integrity, as archaea include many extremophiles that must withstand great environmental stress. At the most basic level, archaea provide critical data for understanding the "periodic table for life." Notably, all three domains have proteins containing the 21st proteinogenic amino acid selenocysteine, which is one of the few amino acids synthesized in a tRNA-dependent fashion with its specific incorporation directed by the UGA codon, as noted by Mullenbach and others [6–8]. The striking cross domain conservation for use of the otherwise toxic element selenium may reflect life's origins as well as the element's utility. For genome integrity, widely conserved domains and functional structures may likewise reveal not only the original selection for methods to preserve genome integrity but also essential aspects of stress responses for genome maintenance. Here we will highlight informative examples where archaeal genome maintenance protein structures were found to be more similar in organization to eukaryal structures than to bacterial or perhaps provide more insights into how protein structures impact human health, where the first or only structure(s) of a particular protein system was derived from archaea. In all these cases structural results on archaeal proteins are important cornerstones for understanding human homologs involved in disease. Furthermore, as human cell genetics and biological tools improve along with structural results, data from human system and archaeal systems will further complement each other to provide a deeper and more unified understanding as illustrated for FEN1-PCNA, Mre11-Rad50, Rad51, and XPD systems that are among the examples presented here.

2. Archaeal Species Speak for Structural Biology

Interestingly, one of the characteristics of many archaeal proteins led to visual support that these "simpler" organisms may be more related to eukarya at certain levels than their bacterial counterparts. The use of proteins isolated from single microbes have filled an impressive number of highly different niches in biotechnological and industrial applications [9]. Likely one of the most desirous common characteristics for these proteins to possess is long-term stability. Organisms with optimal growth temperatures above 80°C are defined as hyperthermophiles [10], the majority of which are classified as archaea [2], and as such contain proteins that are highly stable. Thus, characterization of systems requiring high stability may be facilitated by using genes or proteins isolated from archaea (or bacteria), which natively reside at high temperature and/or pressure [11]. These results may also help efforts to stabilize eukaryotic proteins by designed mutations, such as seen for superoxide dismutase [12]. Relationships of single-site changes and stability have more than academic interest as mutations that destabilize protein frameworks can cause fatal diseases, also as seen for superoxide dismutase [13–15]. Interest in archaeal species and the development of new applications that exploit their enzymes continues to increase. In part, this is due to the appreciation of just how impressive the diverse range of harsh environments in which they inhabit actually appears to be. These "extremophiles" are found in areas not only of high temperature and pressure but also high alkalinity, acidity, salinity, metal content, and even low temperature [9, 11, 16]. In the basic sciences, thermophilic enzymes have been successfully exploited for use in a variety of applications. Archaeal metalloenzymes in general have provided an improved means of understanding and predicting protein metal ion binding sites [17]. Archaea also opened structural doors for understanding the CRISPR system of genetic regulation [18]. Another readily recognized system in molecular biology includes the thermostable DNA polymerases that catalyze DNA synthesis at elevated temperatures during the polymerase chain reaction (PCR). Similarly, in structural biology, enzymes from archaeal thermophiles are frequently characterized, often recombinantly, for a variety of reasons. First, the inherent stability of proteins expressed in cultured cells [19] often yields better-behaved samples *in vitro* at mesophilic temperatures [20]. Second, structural biology such as X-ray crystallography often requires large amounts of highly purified protein, which can often be accomplished by heat denaturation of mesophilic host proteins during purification of recombinant thermophilic proteins [20, 21]. Third, many thermophilic archaeal proteins are significantly closer in amino acid sequence similarity to human proteins than are their thermophilic bacterial counterparts. This aids comparative analyses and inferences into structure-function relationships on the individual atom-, amino acid residue-, domain- and subunit interaction-levels that can be translated into human systems of interest when structures derived from human sources are unavailable. As a consequence of these features, the number of characterized structures derived from archaeal source organisms is rising. Notably, over 3200 such X-ray crystal structures have now been deposited in the protein databank (PDB). Thus, in this postgenomics era, structural biology provides a "seeing is believing" form of support that in many cases, archaea and eukarya together branch off the evolutionary tree from bacteria. Direct structural and three-dimensional computational comparisons of proteins that perform essential basic cellular functions can reveal similarities at the tertiary and quaternary levels between archaeal and eukaryotic proteins [20] and divergence from bacterial proteins. In some instances, archaeal proteins and their structures may be more similar or useful to inform on human proteins of interest than those derived from other eukaryotic model systems like yeast. For example, human and *Pyrococcus furiosus* Rad51 homologs have a similar domain organization, whereas in yeast, there is a significant N-terminal sequence or domain not shared between these homologs. Thus, this information lent to designing a truncated yeast Rad51 construct for crystallization, and due to the stability of the *P. furiosus* protein, a humanized version is being used for inhibitor design [22–26]. This paper will highlight structural insights into several central proteins, enzymes, and complexes involved in basic DNA metabolism, to illustrate key similarities and differences among the three domains.

3. DNA Replication and Repair

DNA replication is the basic fundamental process for transferring or copying the "blueprint of life" to budding or dividing cells. Fidelity is required to ensure that errors do not alter the genotype of the cell or are passed on. Death or disease, either at the cellular, organismal, or familial levels, may be a consequence of improper DNA replication. Like other fundamental cellular processes, it would be expected *a priori* that the macromolecules and mechanisms responsible for genome maintenance are conserved in all domains of life. However, early biochemical comparisons of enzymes such as the DNA-dependent RNA polymerases from archaeal *Sulfolobus acidocaldarius* with bacterial and eukaryotic homologs suggested that perhaps archaeal systems involved in nucleic acid metabolism were less similar than bacterial and more similar to and shared properties with eukaryal homologs [27–29]. However, since archaea lack nuclei and typically contain a singular circular genome, this observation appeared counterintuitive. Combining biochemistry with early sequencing efforts determined that archaeal DNA polymerases likewise were seemingly more eukaryotic-like than bacterial polymerases [30, 31]. The sequencing of the first archaeal genome along with other studies further supported the notion that structural and functional aspects of transcription and translation were often similar to those of eukaryotes [32–36].

DNA replication at its heart entails the separation of duplex DNA into two template strands for synthesis of new complementary DNA to give two identical sets of duplex DNA, whereby one set may be allocated to daughter cells. In the initiation phase, DNA is unwound by helicases to provide the template bases and may also be primed by short RNA segments. Because single-stranded DNA (ssDNA) anneals with opposite polarity to form double-stranded DNA (dsDNA), the elongation phase of replicating DNA contains two complementary processes. For the "leading strand" DNA, replication proceeds continuously in the $5'$ to $3'$ direction along with the replication fork as it is unwound by helicases. For the discontinuously synthesized "lagging strand," RNA primers are deposited on the template DNA by primase and are extended by another polymerase to generate DNA-RNA Okazaki fragments. RNA primers are later removed, and the gaps on the complementary strand are filled in by polymerases and ligases.

An important principle derived from the double helix but not originally recognized is that the double helix provides the basis not only for DNA replication but also for error-free DNA repair. DNA fidelity within the genome does not depend upon extreme stability of dsDNA but rather on robust DNA repair machinery that extends proofreading by polymerases and responds to all the different forms of DNA damage. For example, damage of DNA bases are repaired by the Base Excision Repair (BER) pathway [37–39], while larger base lesions, crosslinks, or protein-DNA adducts is repaired by the Nucleotide Excision Repair (NER) pathway [40–43]. Other forms of base lesions, such as pyrimidine dimers, apurinic/apyrimidinic (AP) sites, and 8-oxoG, may be bypassed by translesion synthesis (TLS) polymerases [44–46]. Misincorporated DNA bases, or single base insertions or deletions, are repaired by Mismatch Repair (MMR) systems [47, 48]. DNA double-strand breaks (DSBs) that may give rise to threatening gross chromosomal rearrangements are repaired with fidelity by homologous recombination (HR) when possible or by nonhomologous end-joining (NHEJ) in a pinch but with small losses of fidelity [49–51]. Thus, life has evolved such that multiple mechanisms promote fidelity of the genome.

3.1. It Starts with a Ring. As mentioned above, replicative polymerases add nucleotides to DNA in the $5'$ to $3'$ direction, and both strands of dsDNA are used as templates to generate new daughter strands from moving replication forks running in opposite directions. Again, since the two template strands are of opposite polarity, one polymerase (leading) is allowed to run continuously, while the other (lagging) synthesizes DNA discontinuously from constantly added RNA-DNA primers from the primase-Polα complex. Coordinating these efforts in archaea and eukarya is the proliferating nuclear cell antigen (PCNA) protein. PCNA is a multimeric, nonenzymatic scaffold protein that encircles DNA as a ring and is otherwise known as a DNA clamp. In replication, it enhances the activity of the leading and lagging polymerases and also plays a role in Okazaki fragment processing. Besides acting in replication, PCNA also serves as a factor in numerous DNA repair, genome maintenance, and cell cycle processes. This includes DNA repair and recombination pathways such as BER, NER, MMR, and HR [52]. Extensive lists of PCNA protein interaction partners have been noted in reviews [53, 54]. Moreover a variety of posttranslational modifications, including phosphorylation, ubiquitination, and SUMOylation regulate PCNA protein partner interactions in different species.

DNA clamp proteins and the enzymes that help them encircle DNA, the DNA "clamp-loaders," are found in all three domains of life. However, while DNA clamps act as central proteins in a relatively large number of cellular processes, their sequences are generally not conserved. Despite the lack of sequence conservation, their general shapes have preserved features, and the domain and subunit organization of the archaeal and eukaryotic proteins are similar (Figures 1(a) and 1(b)). The majority of archaeal and eukaryotic PCNA proteins are homotrimers [20, 55–57]. The first PCNA structure revealed that each subunit consists of two domains that are topologically similar yet, interestingly, do not share significant sequence identity [57]. A long interdomain connector loop (IDCL) joins the two domains, and an extended β-sheet is also formed between subunits. When condensed in head-to-tail fashion into a trimer, the organization is such that an inner ring is formed by 12 α-helices, which are flanked by a circumscribing set of 6 β-sheets. The organization of fold and assembly of both eukaryotic and archaeal PCNA proteins is shared, again despite lack of sequence similarity. In the bacterium *Escherichia coli*, DNA replicative machinery consists of the large 10-subunit DNA polymerase III (Pol III) holoenzyme. Pol III is divided into the Pol III core, the clamp-loader complex, and the DNA clamp, which is known as

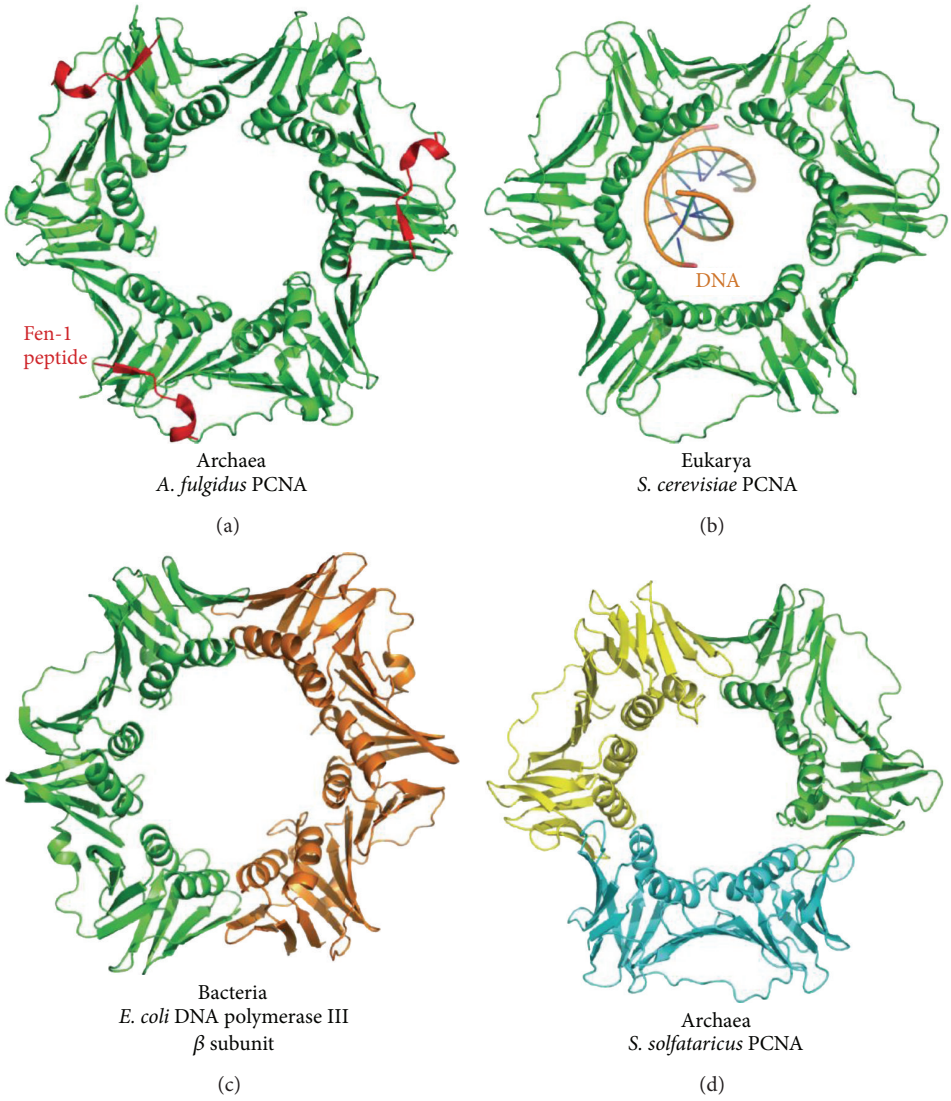


Figure 1: Comparison of PCNAs. The semblance of DNA sliding clamp proteins is made of differing general assemblies. (a) Archaeal *A. fulgidus* PCNA is a trimer composed of three identical subunits, which is the general case for PCNA proteins found in both archaea and eukaryotes. This particular structure has a FEN-1 peptide bound to each subunit (PDB code 1RXZ). PCNA proteins dock other enzymes to bring them into proximity to DNA when their functions are required. (b) The *S. cerevisiae* PCNA shares the general homotrimer assembly with other archaeal and eukaryotic PCNA proteins (PDB code 3K4X). This particular structure was engineered such that a DNA molecule was sequestered within the ring. (c) The β subunit of bacterial DNA polymerase III complexes shares the PCNA fold with its archaeal and eukaryotic counterparts. However, the assembly is formed by a homodimer (PDB code 2POL). (d) Archaeal *S. solfataricus* PCNA is unusual as it is assembled as a heterotrimer (PDB code 2HIK). It perhaps evolved this way to dock different enzymes with specificity.

the β-clamp [56, 58]. With an even greater lack of sequence similarity between the bacterial protein with the eukaryotic and archaeal proteins, coupled with stark differences in protein length, it was likely a surprise to some researchers that the bacterial β-clamp shares the overall PCNA ring-shape, including the 12 helix/6 sheet organization [57, 58]. To share the shape in the context of the extended β-clamp sequence, the domain and assembly organization differs considerably; it is a homodimer composed of 2 subunits containing 3 domains. In essence, one subunit of a bacterial β-clamp resembles one subunit of an archaeal/eukaryotic protein plus one additional domain (Figure 1(c)). Several other PCNA variants have also been discovered in archaea. Crenarcheal homologs, such as that from *Sulfolobus solfataricus*, are heterotrimers composed of subunits PCNA1, PCNA2, and PCNA3 [59, 60] (Figure 1(d)). A recent study suggested the possibility that PCNA from *Sulfolobus tokodaii* forms a heterotetramer from two PCNA2-PCNA3 complexes [61].

Studies have suggested that, in many cases, PCNA stimulates enzymatic activity of partner proteins by influencing their affinity for their respective DNA substrates [62]. To facilitate this type of function with such a large number of proteins (which in turn must be exchanged to meet the needs of reaction steps during pathway progression), some general

mechanisms that lend to regulation are expected. As revealed in the *H. sapiens* PCNA:p21 peptide complex structure, followed by others, PCNA uses a conserved binding mode to interact with a number of proteins via the PCNA-interacting protein (PIP) motif [52, 63]. The consensus sequence consists of Q-x-x-L/I/M-x-x-F/Y/W-F/Y/W, where x is a variable amino acid residue. In general, the Gln residue makes contributions to anchor binding of middle residues within this sequence. These middle residues form a short 3_{10}-helix and anchor the hydrophobic residues of the motif into a correspondingly hydrophobic pocket on PCNA's surface. The C-terminal end of the peptide then forms β-sheet interactions with the IDCL [52, 63, 64]. An additional antiparallel β-zipper, analogous to that found in Rad51 [25, 50], is found in some cases [65]. Many enzymes contain this motif at their C-terminus, allowing conformational flexibility and the possibility for more than one protein to bind PCNA at any given time [60, 66]. Again, numerous posttranslational modifications also aid the process of regulating interactions [53, 54]. Recently, advances in solution small-angle X-ray scattering [67, 68] have aided generating experimentally based structures of flexible and multiconformational examples of these systems. This includes visualization of flexibility and the opening and closing of conformational states of PCNA with protein partner Ligase 1 [60], along with regulatory element ubiquitin [69].

3.2. Did I Make a Mistake? Flap endonuclease 1 (Fen1 in archaea or FEN-1 in eukaryotes) plays roles in both DNA replication and repair, often in conjunction with PCNA. During these processes, DNA polymerases synthesize new DNA that displace RNA or damaged DNA creating 5′ single-strand flap structures [70–74]. Similar intermediates are formed during repair processes involving new DNA synthesis, such as long-patch BER (LP-BER). RNaseH may help in the removal of longer RNA stretches but cannot remove the final RNA base [75], and crystal structures and structure-based mutational analysis of RNase HII from *Archaeoglobus fulgidus*, both with and without a bound metal ion, identified it as a molecular ruler and revealed the means whereby type 2 RNase H specifically cleaves the RNA portion of an RNA-DNA/DNA hybrid duplex [76]. Furthermore for short flaps, FEN-1 is the primary structure-specific endonuclease that removes the 5′ ssDNA or RNA flap to produce a single, nicked product that can be sealed by DNA Ligase I [74, 77–79]. FEN-1 can remove 5′ single stranded DNA or RNA flaps from several types of DNA substrates *in vitro* [55, 71, 73, 80–87]. Besides sequence-independent flap endonuclease activity, FEN-1 has other nuclease activities that include 5′ exonuclease activity during recombination and gap-dependent endonuclease (GEN) activity to aid replication fork processes [88–91].

FEN-1 defects are associated with genomic instability and subsequent development of cancer [92–94] and other diseases [95, 96] in eukaryotes. Preventing PCNA-FEN-1 mutations in mice gave rise to defects in RNA primer removal, which was subsequently embryonic lethal [97]. Screening of human cancers for FEN-1 mutations revealed that defects could be identified that affect 5′ exonuclease activity and GEN activity. When one such mutation was transferred to a mouse model, progeny developed autoimmunity and chronic inflammation in addition to cancer predisposition [94]. Therefore the roles of FEN-1 as a structure-specific flap endonuclease, a 5′ exonuclease, and a gap-dependent nuclease have important implications for human health.

The discovery of the first archaeal Fen1 sequences revealed that they had significantly more sequence similarity with their eukaryotic FEN-1 counterparts than with related bacterial sequences [96]. For example, in viruses and bacteria, Fen1 homologs include T4 RNaseH, T5 5′-3′ exonuclease, and the proofreading element of bacterial Polymerase 1. They also were of comparable length to the eukaryotic proteins suggesting they were independent enzymes and not part of other machinery such as in the bacterial case. Two regions that contain elements responsible for nuclease activity, termed the N (N-terminal) and I (intermediate) domains, are predominant areas of homology between these proteins. Other proteins also have similar domains, such as Xeroderma pigmentosum complementation group G (XPG), which is involved in both Xeroderma pigmentosum and Cockayne's syndrome.

Due to their stability and ease of purification from heterologous expression systems, the archaeal *P. furiosus* and *Methanococcus jannaschii* Fen1 proteins were the first to be crystallized and structurally characterized [55, 98]. These and other [65] archaeal Fen1 structures revealed that the enzyme is a saddle-shaped, single-domain protein with a ~20 Å deep groove formed from a central seven-stranded β-sheet, an antiparallel β-ribbon, and two α-helical bundles (Figure 2(a)). The C-terminal edge of the β-sheet is identified as the substrate-binding region by the presence of catalytically important residues. The two halves of Fen1 are joined by a "helical clamp" or "helical gateway," which, depending on the set of coordinates, ranges from a flexible unstructured region to a pair of ordered α-helices. With additional information from a later DNA-bound human FEN-1 structure, it was found that the protein recognizes DNA 5′ flaps by being able to form a sharp ~100 degree bend with dsDNA on either side (Figure 2(b)). A flap or break is required to bend dsDNA to such a degree at a single phosphodiester site. Binding a 3′ flap causes a ~5 Å shift in the α2-α3 loop, which creates a "hydrophobic wedge" that packs against the terminal base pair of the DNA. A 3′ flap-binding pocket encloses a single unpaired nucleotide that ensures an eventual product suitable for ligation. FEN-1 also requires the 5′ flap to pass under a cap to enter the helical gateway and the active site. The structure of the bacterial *Thermus aquaticus* Polymerase 1 revealed a relatively conserved fold for the 5′-3′ exonuclease domain that shares homology with the flap endonuclease proteins (Figure 2(c)). The many archaeal results provided a strong foundation for a determination of DNA substrate and product complexes for the human FEN1 as well as FEN1 complexes with PCNA and its repair analogue 9-1-1 that supported and extended the results from archaeal systems [99, 100]. Indeed, the FEN-1 superfamily structure and unpairing mechanism for specificity is broadly conserved with RNA enzymes regulating transcription as well

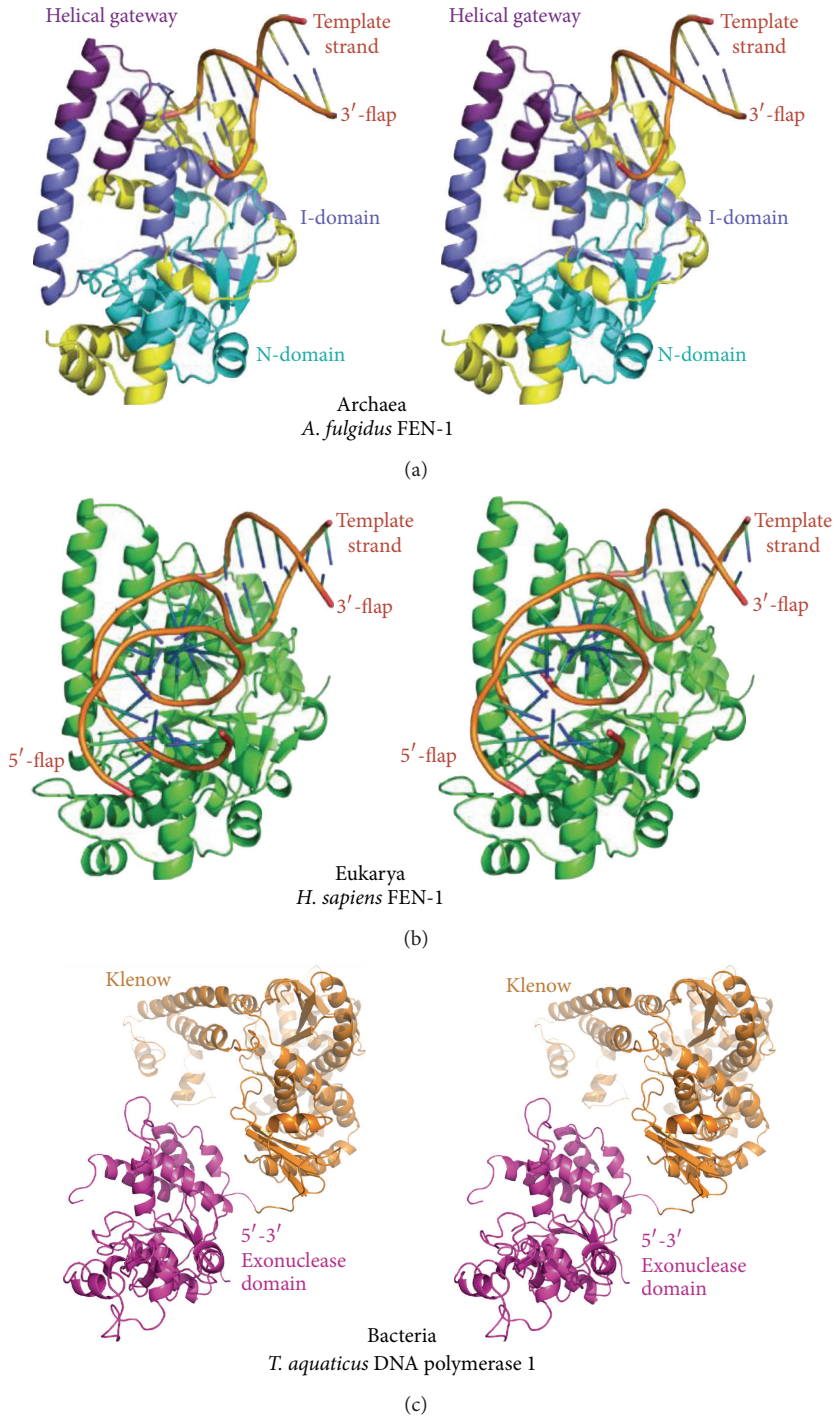


Figure 2: Comparison of FEN structures. Archaeal FEN-1 homologs share structural conservation with eukaryotic FEN1 proteins. (a) Stereoview of archaeal *A. fulgidus* FEN-1 in complex with DNA (PDB code 1RXW). The original conserved N and I regions are shown in light and dark blue, respectively. One strand of the short duplex DNA segment represents a 3′ flap substrate, while the other represents the template strand. The helical gateway segment used to guide DNA is shown in purple. (b) The stereoview of the human FEN1 structure reveals the conservation of tertiary structural fold between archaeal FEN-1 proteins and eukaryotic FEN1 proteins. The more complex double-flap substrate revealed insights into the DNA binding mode and active site chemistry (PDB code 3Q8M). (c) The structurally related "FEN-1" of bacteria is the 5′-3′ exonuclease of DNA polymerase 1 shown in the stereogram in magenta (PDB code 1TAQ). This domain is tethered to the Klenow fragment that carries the DNA polymerase activity.

as with replication and repair enzymes such as exonuclease 1, DNA repair protein XPG, endonuclease GEN1, and the 5′-3′-exoribonucleases. Together these enzymes play key roles in many cellular processes such as DNA replication and repair, recombination, transcription, RNA turnover, and RNA interference [101].

With respect to FEN-1 and PCNA functions in DNA replication and repair, questions arise in how the mechanistic steps are regulated. In particular, how are proteins exchanged during these processes, or do multiple proteins bind to PCNA simultaneously? During Okazaki fragment maturation in archaeal replication, work in the *S. solfataricus* system determined that the main proteins involved appear to be PCNA, Fen1, PolB1, and Lig1 [62]. This is analogous to that of PCNA, FEN-1, Polδ, and DNA ligase 1 being the main players for Okazaki fragment processing in eukaryotes [53, 62]. In a homotrimeric PCNA system, it seems logical that binding of one protein partner to a PCNA subunit may either influence the conformations of the unbound PCNA subunits or perhaps sterically exclude other partner proteins, until either DNA conformations or product complex conformations induce a "handoff" to the next protein in the pathway. For the heterotrimeric *S. solfataricus* PCNA, the differing subunits bind their cognate partners as follows: PCNA1:Fen1, PCNA2:PolB1, PCNA3:DNA ligase 1 [59]. Furthermore, data suggests that these interactions may occur simultaneously. *In vitro*, multiple FEN-1 proteins may bind a single eukaryotic PCNA molecule; however, crystallographic information revealed that they did so with different binding modes [66]. Therefore, it is plausible that PCNA may bind different protein partners simultaneously. Moreover, it appears that both inactive "carrier" and "active" conformations may exist in some circumstances when binding a particular protein. It does also appear however that one partner protein may also displace another, as is the case for the DNA clamp loaders and polymerases. Therefore, PCNA's binding interactions appear to be regulated at many levels.

3.3. They Dislike U. High temperatures can increase deamination of cytosine resulting in conversion into uracil [102]. Therefore, many archaea have multiple means to control the presence of uracil in DNA and thus suppress possible mutagenic or genotoxic effects. For instance, some archaeal family-B DNA polymerases have a "read-ahead" scanning mechanism employing an N-terminal pocket to detect template-strand uracil and halt the polymerase [103–107]. Additionally, the heterodimeric euryarchaeal family-D polymerases may also possess a uracil detection-response system, though likely through a distinct mechanism [108]. Archaea can also contain dUTPases [109] including DCD-DUT, which converts dUTP to dUMP to prevent misincorporation of dUTP into DNA and is exemplified by a jelly-roll fold with two helices and a β-arm [110]. Other conserved uracil repair activities include steps leading to BER. Some archaea possess enzymes such as the helix-hairpin-helix folded uracil-DNA glycosylases (UDG) similar to the MIG/EndoIII enzymes [111, 112], and UDGs from other superfamilies are also present in archaea, as discussed below.

As a first step in the BER pathway, glycosylases must target incorrect base lesions and cleave the bond between the base and deoxyribose sugar in DNA, creating an apurinic/apyrimidinic (AP) site. Following this step, the actions of conserved AP endonucleases, polymerases, and ligases finalize the repairs [113]. Of the BER-triggering glycosylases, the UDG superfamily is a well-studied example specific to removal of uracil [114]. Within this superfamily, 6 families have been classified [115]. Family 1 is comprised of the uracil DNA N-glycosylases (UDGs/UNGs) [116] and related homologs whose substrates include ssDNA and dsDNA (Figure 3). These enzymes are found in bacteria and eukaryotes, and in humans; the UNG2 protein is involved in somatic hypermutation for immunoglobulin gene diversification in the immune system [117]. Family 2 includes mismatch-specific uracil-DNA glycosylases (MUGs) [118] and thymine-DNA glycosylases (TDGs) [112, 119]; family 3 (mostly eukaryotes and some bacteria) include the single-strand-selective monofunctional uracil DNA-glycosylases (SMUGs) [120]; families 4 [121] and 5 [122] have distinct specificities [123, 124] and contain an Fe-S cluster specific to thermophiles (bacteria and archaea); family 6 are hypoxanthine-DNA glycosylases [125] found in all domains. Thus, of these, only the first 5 contain UDG activity [125]. Archaea appear to utilize UDGs from families 4, 5, and 6 and sometimes 2 [125, 126].

Structures from these UDG family members are typically characterized by a β-sheet bordered by α-helices (the $\alpha/\beta/\alpha$ sandwich) and contain a pocket that positions the uracil for cleavage (see Figure 3). Structures of family 4 UDGs also reveal a similar fold with the Fe-S cluster adjacent to the active site. Interestingly, despite a common evolutionary ancestor and fold for family 1–5 UDGs, divergence has been observed at the sequence level and manifests in part via active site differences [126, 127]. Steric features help to recognize uracil via hydrogen bonding, bending of DNA and nucleotide flipping [128]. Two active site motifs are variable between the UDGs and contribute to their subtle mechanistic distinctions [125]. Recently, crystallization of the first archaeal (family 4) UDG from *S. tokodaii* was reported [129]. When this structure is finalized, it will be fascinating to compare the structural determinants of this enzyme with those known from bacterial and eukaryotic UDGs. More generally, variation and conservation of BER enzymes from archaea to humans provide a deeper understanding of strategies to remove or reverse base damage. For example, N1-methyladenine (m1A) and N3-methylcytosine (m3C) are major toxic and mutagenic lesions induced by alkylation in single-stranded DNA. In bacteria and eukaryotes, m1A and m3C are repaired by AlkB-mediated or AlkB-like (ABH) oxidative demethylation [130, 131]. Yet, no AlkB homologues have been identified in Archaea, and m1A and m3C are repaired by the AfAlkA base excision repair glycosylase of *A. fulgidus*, suggesting a different repair mechanism for these lesions in the third domain of life [132].

3.4. Unwind or Move On. Large DNA lesions caused by chemicals or UV radiation, such as thymine dimers, threaten genomic fidelity in all three domains of life. Archaea, like

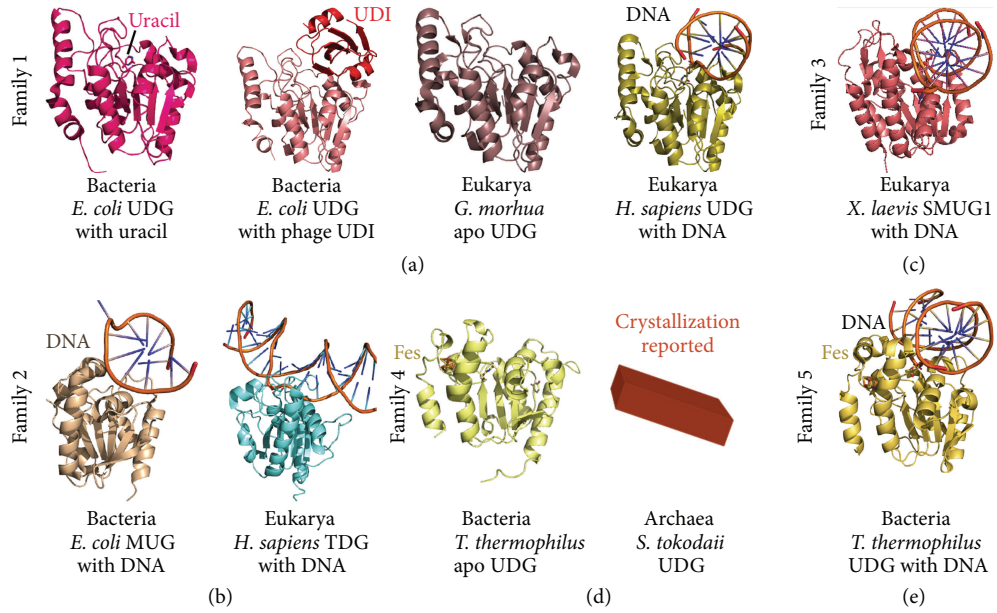


Figure 3: Representative structures of UDGs from families 1–5. (a) Family 1: bacterial UDG monomer with uracil from *E. coli* (PDB code 1FLZ) and *E. coli* UDG (PDB code 1UUG) bound to a *Bacillus* phage inhibitor (UDI). Eukaryotic UDG structures exemplified by apo *G. morhua* UDG (PDB code 1OKB) and DNA-bound human UDG (PDB code 1EMH). (b) Family 2: *E. coli* MUG in complex with DNA (PDB code 1MWJ); human TDG (PDB code 2RBA). (c) Family 3: *X. laevis* SMUG1 in complex with DNA (PDB code 1OE4). (d) Family 4: bacterial UDG (with Fe-S) from *T. thermophilus* UDG (PDB code 1 UI0). Determination of the X-ray structure from *S. tokodaii* will shed light on archaeal UDG homologs. (e) Family 5: *T. thermophilus* UDG (with Fe-S) bound to DNA (PDB code 2DEM). Archaea contain homologs from families 2, 4, and 5.

bacteria and eukarya, utilize NER to repair such DNA lesions [40]. During NER, either enacted globally or during transcription, damage must be recognized, dsDNA must be unwound and the lesion bracketed with incisions before the damaged stretch of ssDNA can be excised and resynthesized. In contrast with other DNA repair mechanisms, no significant homology exists between the NER machinery in bacteria and eukarya. Bacteria perform NER using a complex of Uvr proteins (UvrABCD) [40, 133], whereas eukaryotes use the multicomponent transcriptional and repair factor TFIIH in addition to other proteins [134]. Archaeal organisms lack many components of eukaryotic TFIIH, but sometimes encode homologs of the bacterial Uvr proteins. Exceptions to this trend are the Xeroderma pigmentosum complementation group D protein, XPD (also known as ERCC2 or Rad3), and Xeroderma pigmentosum complementation group B protein, XPB [135], which both form part of TFIIH in eukaryotes and are encoded by many archaea. XPD and XPB also help regulate general transcriptional as part of the TFIIH complex [136–138].

In both eukarya and archaea, XPD functions as an ATP-driven DNA helicase recruited to unwind dsDNA near the lesion-site for NER. XPD is superfamily 2 (SF2) helicase that contains a pair of Rad51/RecA-like helicase domains (HD1 and HD2), is characterized by the insertion of Arch and Fe-S domains [139] into HD1, and functions as a $5'$-$3'$ helicase. In eukarya, XPD also helps dictate cell cycle progression via cyclin-dependent kinase activating kinase (CAK) interactions [136] and, in humans, the Arch domain may represent a recruitment platform for CAK to TFIIH [140]. The Fe-S domain was proposed to recognize the DNA (at the dsDNA-ssDNA junction) and place the enzyme in an appropriate position for unwinding [141]. The Fe-S domain is also likely to play an important role in $5'$-$3'$ processing, since all Fe-S containing helicases characterized to date operate with $5'$-$3'$ polarity [142]. Furthermore, XPD may serve to verify DNA damage, as yeast Rad3 [143, 144] and *F. acidarmanus* XPD (FaXPD) appear to stall at damaged sites. For FaXPD, this abortion of helicase activity was accompanied by stimulation of ATPase activity [144]. Thus, the damage-specific stalling of XPD was recently described as a central decision point in the NER reaction [145].

Consistent with this critical role, dysfunctions in the NER pathway cause a UV-hypersensitive phenotype in organisms such as humans and yeast, and moreover, deletion of XPD is embryonic lethal in mice [146]. In humans, point mutations in XPD produce three different diseases: xeroderma pigmentosum (XP), Cockayne syndrome with XP (XP/CS), and trichothiodystrophy (TTC) [42]. Structural information has been crucial to elucidating the molecular determinants of XPD mutations. Whereas human XPD has not proved amenable to structural studies, a number of archaeal XPD structures have been solved (Figure 4). The structures of XPD from *S. acidocaldarius* (SaXPD) [147] and *S. tokodaii* (StXPD) [148] have been solved in the absence of DNA, while XPD structures *from T. acidophilum* (TaXPD) have been solved in the presence [149] and absence of DNA [150]. These structures have allowed investigators to rationalize how

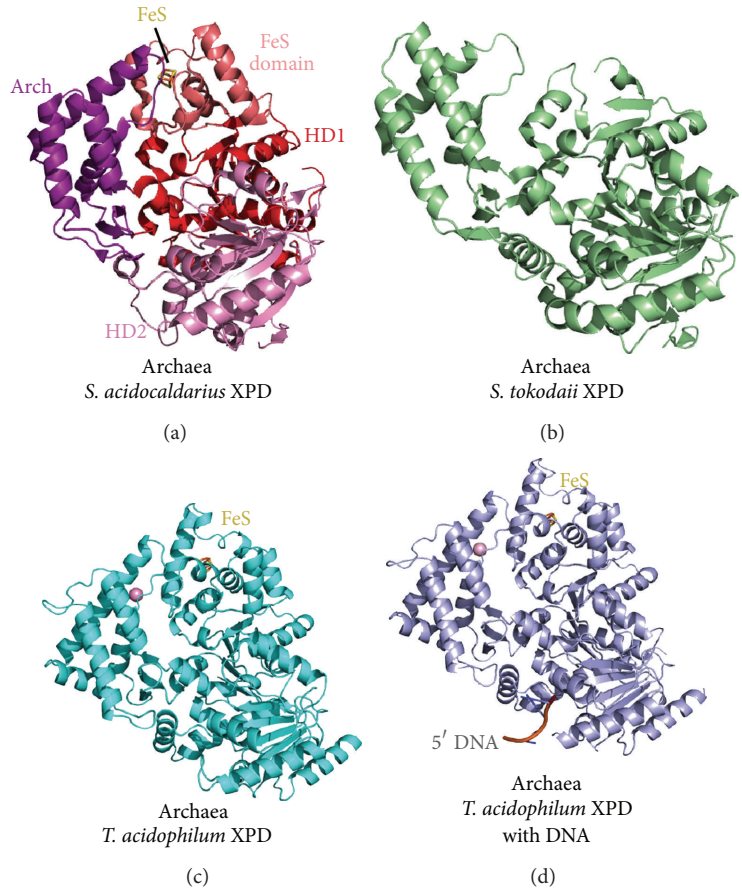


Figure 4: Insights into XPD from archaeal structures. Structures of XPDs from *S. acidocaldarius* (PDB code 3CRV), *S. tokodaii* (PDB code 2VL7), and *T. acidophilum* XPD with (PDB code 4A15) and without DNA (PDB code 2VSF) reveal Arch, Fe-S, HD1, and HD2 domains forming a seat-shaped molecule. Notably, despite similar architecture, cleft pore size between Arch and Fe-S domains varies, suggesting flexibility in this region. DNA orientation is shown for *T. acidophilum* XPD, and a calcium ion is shown as a pink sphere.

sequence-related XPD mutants can cause different diseases: ATP and DNA binding mutations give rise to XP; flexibility and conformational mutations give rise to XP/CS; and framework destabilizing mutations give rise to TTD [147, 148]. The structure of TaXPD bound to a 4 nucleotide DNA above motor domain 2 [150] plus spectroscopic results on XPD with bound DNA [151] helped define the path of the translocating DNA through the pore region of the protein created by the arch and FeS domains (Figure 4), as well as elucidating how XPD-like helicases operate in the 5′-3′ direction. Rather than reorienting the DNA, XPD helicases grip it in the same orientation as the 3′-5′ SF2 helicases but process the DNA in the opposite direction. These structures have also helped researchers develop models of DNA-damage detection based on charge transfer (or electron tunneling) between the Fe-S domain and the DNA [142, 152]. Thus, archaeal XPD structures have been critical in helping elucidate more general insights into NER. Indeed, archaeal protein structures have informed the activities of other ATP-driven motors such as the secretion ATPase superfamily [153].

Furthermore, some archaea contain alkyltransferase-like (ATL) proteins, whose protein-DNA complexes can switch base damage to the NER pathway. ATLs share functional motifs with the cancer chemotherapy target O(6)-alkylguanine-DNA alkyltransferase (AGT) and paradoxically protect cells from the biological effects of DNA alkylation damage, despite lacking the reactive cysteine and alkyltransferase activity of AGT. Structural results on the ATL from *Schizosaccharomyces pombe* without and with damaged DNA containing endogenous lesions revealed nonenzymatic DNA nucleotide flipping plus increased DNA distortion and binding pocket size compared to AGT. Analysis of lesion-binding site conservation identified ATLs in sea anemone and ancestral archaea, indicating that ATL interactions are ancestral to present-day repair pathways in all domains of life [154]. Genetic connections to mammalian XPG (also known as ERCC5) and biochemical interactions with *E. coli* UvrA and UvrC combined with structural results reveal that ATLs sculpt alkylated DNA to create a genetic and structural intersection of base damage processing with nucleotide excision repair. Such sculpting of DNA to create cross-talk among

different DNA repair pathways may prove to be a general strategy to regulate DNA damage response networks [155].

3.5. Staying Together. DNA double strand breaks (DSBs) are a particularly threatening type of DNA damage, posing a risk of genetic information loss. DSBs can occur as a result of DNA replication and repair events or from extrinsic factors or toxins. Various forms of DSB repair exist, and several key double strand break (DSB) repair players are conserved in all domains of life. The repair of DSB typically utilizes one of two major pathways: homologous recombination (HR) or nonhomologous end-joining (NHEJ), but in some instances microhomology-mediated end-joining may be used [156]. In all domains, a core complex termed MR (for Mre11 and Rad50) plays key roles in detecting and repairing DSBs. In eukaryotes a third accessory protein (Nbs1 or Xrs2) is present to form MRN or MRX complexes, respectively.

The first archaeal structures [157–159] combined with recent structural insights have been especially powerful in illuminating our understanding of MR architecture and mechanism, in particular in *P. furiosus* [160, 161], *M. jannaschii* [162], and *T. maritima* [163, 164] (Figure 5). The arch-shaped Mre11 homodimer is assembled by interactions of two manganese-containing nuclease domains, which are each flanked by nuclease capping domains controlling active site access [158] (together comprising the phosphodiesterase domain) for DNA ends [161] and is then trailed by C-terminal Rad50-binding domains often not included in crystallographic structures. Dimer formation is required for stable DNA binding in the cleft between subunits but not endonuclease activity [161]. Rad50 is a dumbbell-shaped ATP-binding cassette protein containing a conserved signature motif [165] with joined ends connected by 600–900 amino acids of coiled coil linker [159] containing a zinc hook mediating complex bridging [157]. The two termini form a bowl-shaped globular domain containing two lobes with a signature motif, N-terminal Walker A and C-terminal Walker B motifs, magnesium ions, and several key loops required for Rad50 catalysis [159]. Upon ATP and Mg^{2+} binding, the MR complex moves from a wing-shaped heterotetramer to a globular structure and as such acts as an ATP-stimulated nuclease to degrade DNA ends and bridge them in repair and recombination [157, 166]. Although a precise mechanism for ATP-coupled catalysis is currently being resolved, recent insights from archaeal MR complexes have been crucial in efforts toward this goal [160, 162] (Figure 5).

Signaling interactions by eukaryotic MRN complexes have also been informed by structural studies on eukaryotic homologs [167]. Structures of Nbs1 domains alone [161, 168, 169] and bound to a CTP1 peptide [161] or Mre11 [167, 170] have illuminated our knowledge of the eukaryotic-specific components. Moreover, human Mre11 contains a distinct orientation of the dimer heads [170]. In metazoans, these complexes are linked to the cell cycle, telomere maintenance, and activation of ATM kinase. Overall the utility of the complexes are underscored by observations that mutations in MR components lead to disease or severe phenotypes [171]. Nonetheless, archaea may also have domain-specific interacting partners, for instance, the DNA-associated MlaA (or HerA) ATPase that may work with MR in processing (or restarting) stalled replication forks or Holliday junctions in hyperthermophilic archaea [172]. The intricate functions of the MR complex await further characterization.

3.6. Protection. Homologous recombination (HR) is regarded as an ancient essential DNA metabolism system [51] that plays important roles in the repair of DNA DSBs from exogenous agents, replication associated repair of DSBs, gene rearrangement, mitosis, and meiosis [50, 51, 173]. While the basic process of HR is conserved among the three domains of life, only the central enzymes are conserved. Many other enzyme factors, usually dubbed mediators in the recombination field, aid steps of the reactions and also perform key signaling functions in specific species.

HR begins with end resection to form $3'$ ssDNA overhangs. In archaea, this process is mediated by the previously mentioned systems: the helicase MlaA/HerA, the MR complex, and the NurA nuclease [174, 175]. These overhangs are protected by single-stranded DNA binding proteins [50, 173, 176], an important process to prevent degradation of these ssDNA ends, as well as keeping them from base pairing which may inhibit recombination processes. The DNA binding domains of these molecules are composed of oligosaccharide/oligonucleotide binding (OB) folds. However, depending on the domain of life and further divisions there are differences in overall architecture and quaternary structures, and different naming systems are used. The human single-stranded DNA binding protein is termed replication protein A (RPA), where the protein exists as a heterotrimer. The first crystal structure of the protein revealed how the largest subunit RPA70 binds ssDNA via two tandem OB folds (Figure 6(a)) with sidechains stacking with the ssDNA bases to produce an irregularly shaped ssDNA chain [177]. The contorsion of the ssDNA is such that in the larger more intact RPA ssDNA complex structure, the bases are protected by generally facing inward [178]. In contrast, the domain organization of the bacterial single-stranded DNA binding protein (SSB) from *E. coli* consists of a heterotetramer, and while the individual domain structures contain OB-folds, they differ from eukaryotic RPA OB-folds (Figure 6(c)) [179–181].

Interestingly, archaea appear to have perhaps followed two paths for their ssDNA binding proteins. The euryarchaea subdivision contain more eukaryotic-like RPA proteins, whereas the crenarchaea subdivision tend to have ssDNA binding proteins that resemble bacterial SSBs in terms of having a more similar domain organization [182, 183]. Overall, the OB folds of both euryarchaeal RPAs and crenarchaeal SSBs resemble those of eukarya (Figure 6) [184]. However, the monomeric crenarchaeal SSB has an unconserved C-terminal extension reminiscent of bacterial proteins, where their C-terminal ends are involved with interactions with the exonuclease1 protein [184, 185]. The crenarchaeal SSB eventually led to the discovery of new eukaryotic ssDNA binding proteins, where in humans these have been termed hSSB1 and hSSB2. Initial characterization of hSSB1 led to the findings that it

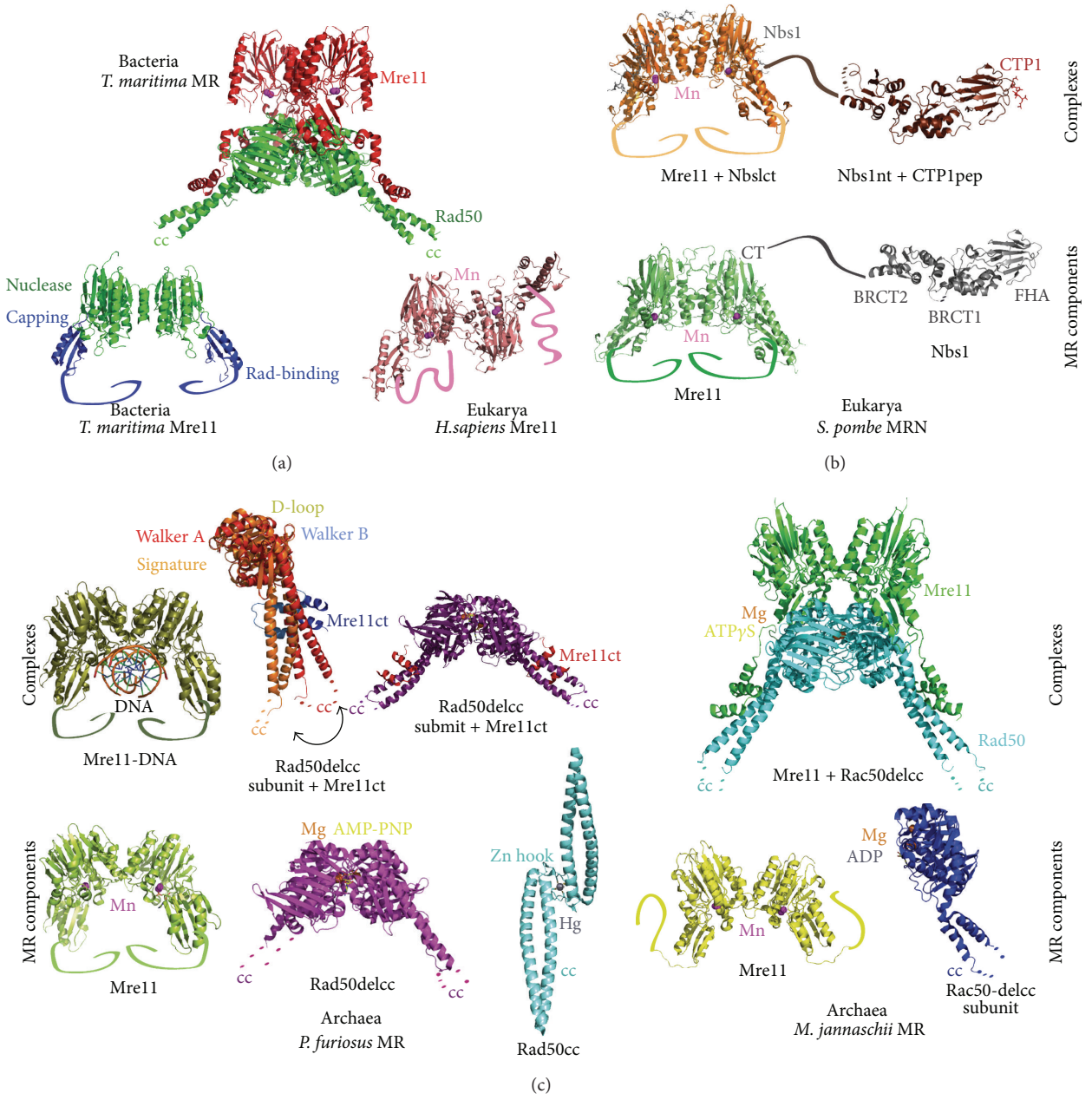


Figure 5: Conservation of MR family members in three domains. (a) Structures of bacterial homologs have elucidated domain organization of MR components. The bacterial Mre11 dimer, exemplified by the *T. maritima* homolog (PDB code 2Q8U), contains the larger, N-terminal nuclease domains (green), the adjacent capping domains, and Rad-family member binding domains (not structurally defined for TmMre11). The Rad50 portion of the Mre11-Rad50 complex (PDB code 3THO) in *T. maritima* is comprised of a curved, globular domain of two lobes with intervening coiled coils (lower regions). (b) Human and yeast MRN structures reveal eukaryote-specific features. The subunit orientation of human Mre11 dimer (PDB 3T1I) is substantially rotated, as compared to other known homolog structures. Mre11 and Nbs1 complexes exemplified by *S. pombe* (PDB codes 4FCX (SpMre11); 4FBW (SpMre11 with SpNbs1 C-terminal region); 3HUE (C-terminally truncated SpNbs1); 3HUF (SpNbs1 in complex with CTP1 peptide)) have revealed key regulatory interactions including the binding site of the Nbs1 C-terminal tail on Mre11 and the FHA domain interaction of Nbs1 with a phosphopeptide of CTP1. (c) Insights from archaeal MR components and complexes. Mre11 structures from *P. furiosus* (PDB codes I117 (PfMre11); 3DSC (PfMre11 with DNA); 3QKT (PfRad50 core); 3QKU (PfRad50 core with PfMre11 C-termini); 3QKS, 3QKR (PfRad50 subunits with PfMre11 C-termini); 1L8D (PfRad50 coiled coil and Zn hook)) have revealed DNA binding and partner interaction sites key to MR assembly. Likewise, complementary *M. jannaschii* structures have confirmed key architecture and interaction sites between MjMre11 and MjRad50 (PDB codes 3AUZ (MjMre11); 3AUX (MjRad50 core); 3AV0(MjRad50 core with MjMre11)).

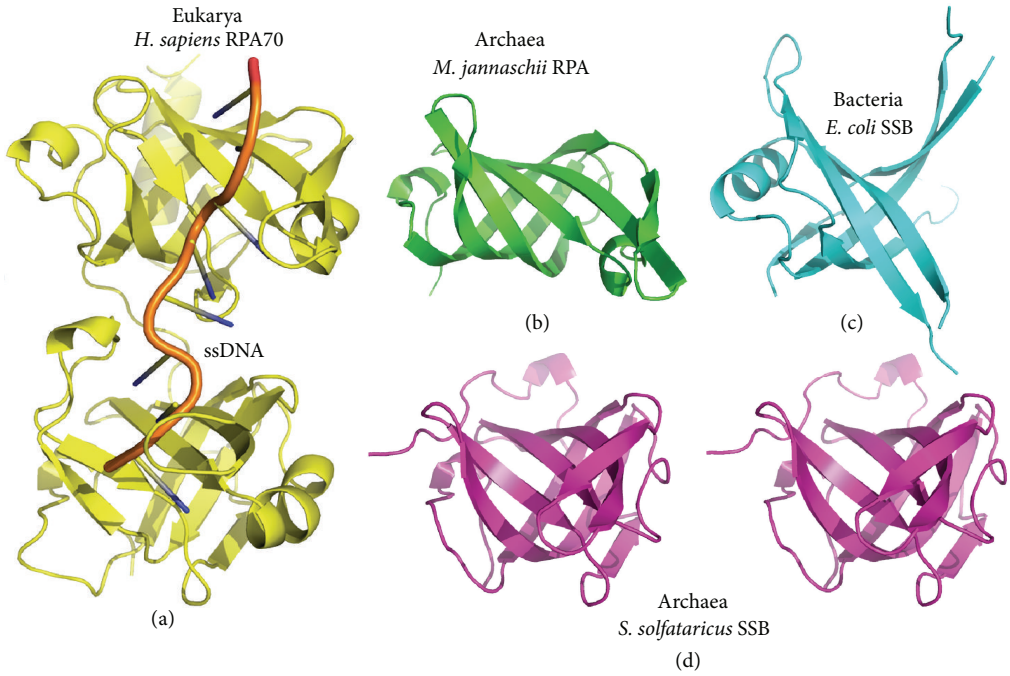


Figure 6: Single-stranded DNA binding proteins. Bacterial members of this class of protein are generally termed SSBs, which stands for single-stranded DNA binding proteins, whereas, originally, eukaryotic versions were called replication protein A (RPA). Archaeal single-stranded DNA binding proteins are generally split between the euryarchaeal RPA proteins and the crenarchaeal SSB proteins. The DNA binding elements of these single-stranded DNA binding proteins are oligosaccharide/oligonucleotide binding (OB) folds. (a) Two tandem OB folds representing residues 181–422 from the largest subunit of human replication protein A, RPA70 (PDB code 1JMC), reveal the binding mode to ssDNA. (b) Example of the OB fold from the euryarchaeal *M. jannaschii* RPA structure (PDB code 3DM3 chain A). The structure is in the same orientation as the top domain in (a). (c) An example of a bacterial SSB domain from *E. coli*. (PDB code 1SRU chain A). (d) The characterization of a crenarchaeal SSB protein from *S. solfataricus*, whose OB fold (stereoview shown in same orientation as the bottom domain in (a), PDB code 1O7I chain A) resembles that of eukaryal RPA. Interestingly, its overall domain organization is more similar to bacterial SSBs, and this led to the discovery of additional single-stranded DNA binding proteins in humans.

appeared to accumulate within the nucleus to form foci with other proteins following the induction of DSBs, where the colocalization does not seem to correlate with RPA binding at the same sites [186]. Moreover, experimental results suggested a role in HR with interactions with Rad51. Later the hSSB1 and hSSB2 proteins were found to coalesce with the INTS3 and C9orf80 proteins to form the sensor of single-stranded DNA complex 1 (SOSS1). This complex is under very active investigation and results suggest involvement with a variety of DNA damage response proteins, such as ATM, Rad51, and Exo1, through recruitment interactions, signaling, or regulation [187–189].

3.7. Infidelity and Fidelity. In the next step of homologous recombination the ssDNA binding proteins are dislodged and replaced by the central DNA strand exchange enzyme RadA (or Rad51) with the aid of mediators [25, 50, 190, 191]. Interestingly, eukaryotic mediator BRCA2 appears to use mimicry to accomplish this task as it also contains OB folds [50, 192]. The DNA bound RadA/Rad51 subunits form a nucleoprotein filament that invades a homologous segment of dsDNA, which then serves as a template for new DNA synthesis by polymerases such as pol D [193]. Following synthesis, the resulting Holliday junction DNA structures generated during recombination [173] are then rearranged by other enzymes known as resolvases. In the archaea, the Holliday junction cleavage (Hjc) and Holliday junction endonuclease (Hje) are examples of proteins that are implicated for this role [194–196].

The function of the central homologous pairing and strand exchange enzyme is conserved among the 3 domains of life. In the 1960s, the finding that bacterial resistance to radiation was correlated to the *recA* gene [197] was the first step to determine that the RecA protein performed the pairing and strand exchange function. While its archaeal (RadA/Rad51) and eukaryotic (Rad51) counterparts share the same function, their sequences were found to differ significantly. Archaeal Rad51/RadA proteins [198] generally have approximately 40% primary sequence identity with eukaryotic Rad51, and these enzymes also share similar overall domain architecture. In contrast, only about 20% sequence identity is shared between Rad51/RadA proteins and bacterial RecA proteins, and this is localized to a single domain. Early sequence alignment programs were unable to correctly align structurally similar regions of the bacterial RecA proteins with their functional equivalents from archaea and eukarya. ApoRecA usually exists as a protein filament [199, 200], while apoRad51 exists primarily

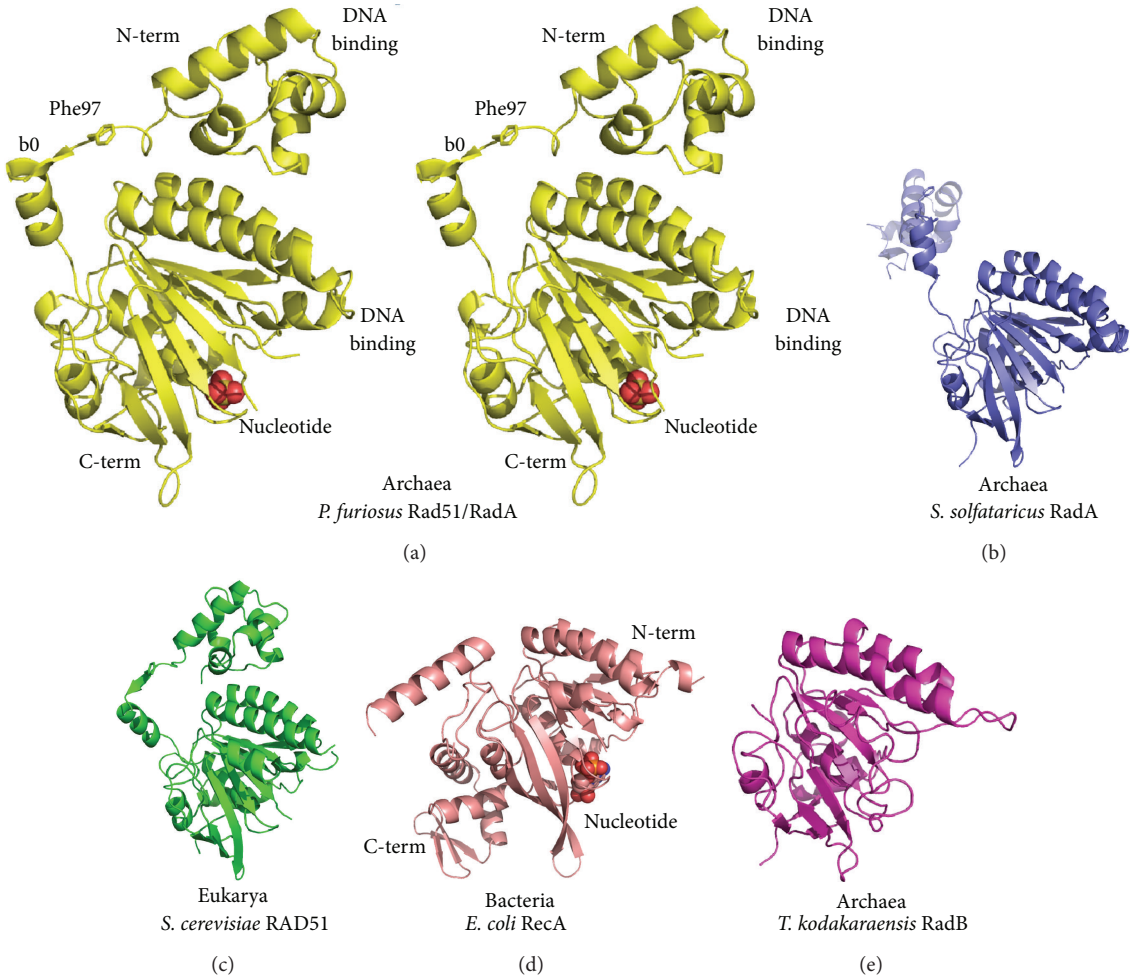


Figure 7: Structural comparisons of Rad51 family proteins. (a) Stereoview of archaeal *P. furiosus* Rad51 (PDB code 1PZN). While Rad51 generally forms larger homopolymeric assemblies, the prototypical fold of a single archaeal Rad51 or RadA protein consists of a small 4-helix bundle N-terminal DNA-binding domain, which is tethered to a larger C-terminal ATPase domain. The ATPase domain has several loops that are also implicated in binding DNA. The interdomain linker that tethers the two domains contains a polymerization motif (PM) that consists of a β-strand β0, which, upon contact with a neighboring subunit, extends the β-sheet of the ATPase domain. Conserved Phe97 forms a ball and socket to stabilize this interaction. (b) The two domains of the archaeal *S. solfataricus* structure individually superpose well with the PfRad51 domains. However, this structure reveals the flexibility of the interdomain linker, where the N-terminal domain has swung outward (PDB code 2BKE). (c) The structure of *S. cerevisiae* RAD51 reveals structural conservation with the archaeal proteins (PDB code 1SZP). (d) The bacterial RecA protein shares the ATPase domain fold (PDB code 2REB). However, in this structure the ATPase is represented in the N-terminal domain, while in archaea and eukarya the ATPase is represented in the C-terminal domain. A small N-terminal arm extends from the bacterial ATPase and serves a similar function as the Rad51 PM. (e) The archaeal *T. kodakaraensis* RadB protein structure illuminates how extensions to the ATPase likely served as critical components of primordial recombination structures, where in the archaea and eukarya the N-terminal domain became an accessory domain, whereas in bacteria the C-terminus gave rise to an accessory domain (PDB code 2CVF).

as a polymeric ring. Similar to RadA/Rad51, in the presence of DNA, RecA will form a nucleoprotein filament [201–203].

The first full-length RadA/Rad51 crystal structure solved was derived from the archaeal thermophile *P. furiosus* (PfRad51) (Figure 7(a)) [25]. A single subunit of Rad51 consists of a small N-terminal 4-helix bundle and a larger C-terminal ATPase domain. The N-terminal domain contains an HhH motif [204], which acts in Mg-coordinated DNA phosphate backbone binding [205]. The larger ATPase domain consists of a central beta-sheet surrounded between alpha-helices. The ATPase contains the Walker A and B motifs and a rare *cis*-linked glycine at position 141 within the active site. Two loops, termed L1 and L2, that are analogous to disordered loop regions found within the first RecA structure [200] and implicated to become ordered upon DNA binding, are also contained in the C-terminal ATPase domain. Comparisons with other archaeal crystal structures that followed, such as those from *Methanococcus voltae* (MvRadA) [206] and *S. solfataricus* (SsRadA) [207], revealed that the N- and C-terminal domains are highly conserved (Figure 7(b)).

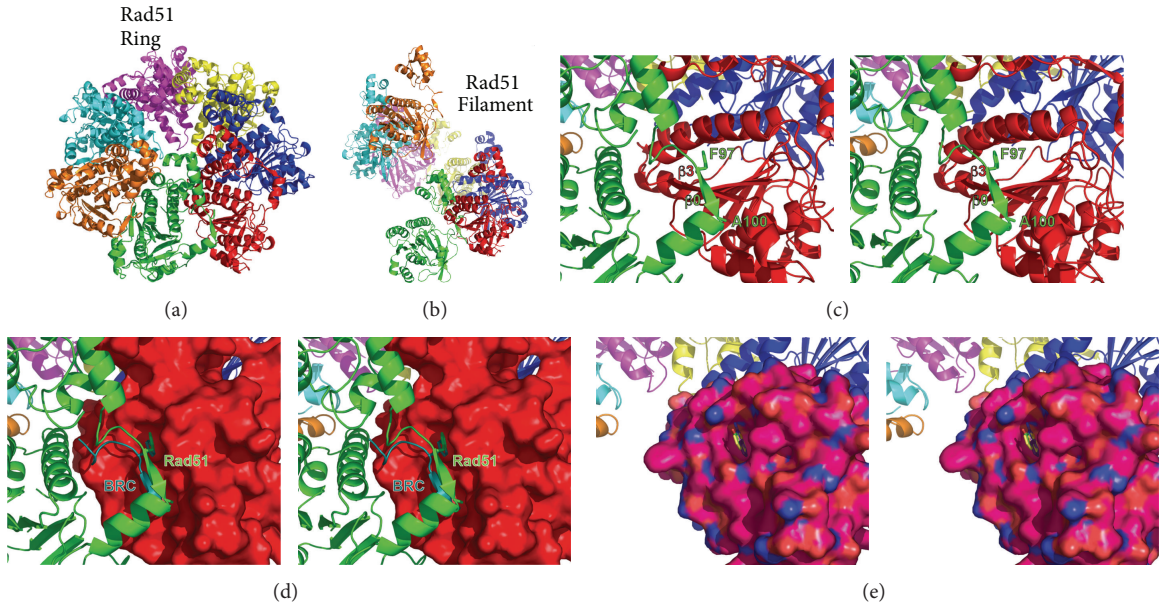


Figure 8: Rad51 assemblies and interface mimicry and exchange. (a) A Rad51 ring model derived from a *P. furiosus* crystal structure is composed of 7 identical subunits, each colored differently (PDB code 1PZN). While substantial intersubunit contacts are made on both sides of each individual subunit, the interactions made by the polymerization motif (PM) shown in (c) are responsible for the ability of the assembly to transition from a ring to a filament, as shown in (b). (b) A Rad51 filament model derived from docking in a *P. furiosus* crystal structure into *S. solfataricus* Rad51 electron microscopy 3D reconstruction density. This assembly generally forms upon binding DNA to generate a nucleoprotein filament, where DNA resides in the interior of the assembly. (c) Wall-eyed-stereoview zoom of Rad51 subunits from (a). The PM resides in the interdomain linker that tethers the N-terminal and C-terminal domains together. The PM $\beta 0$ (green) extends the β-sheet made by the adjacent subunit (red) by bonding to $\beta 3$. The conserved Phe buries itself into a pocket formed by the adjacent subunit. The conserved Ala residue also stabilizes the PM via hydrophobic interactions. (d) Same view as in (c), however the surface of the red subunit is shown. Overlay of the HsRad51 ATPase:BRC4 domain fusion structure (PDB code 1N0W) with PfRad51 reveals that BRC4 mimics interactions made by the Rad51 PM, and it is proposed that the interfaces may be exchanged to form Rad51:BRCA2 complexes in eukarya. (e) The same PM interface is now a target for small-molecule fragment-based approaches to develop ligands that work in conjunction with radiation and genotoxic drugs used to treat cancer. Due to the conservation of the *P. furiosus* and human enzymes, researchers developed a humanized PfRad51 mutant (overlaid in the position of the red subunit in other panels) and identified compounds that bind to the hydrophobic pocket that is normally occupied by the PM Phe residue (PDB codes 4B33, 4B34, and 4B3C). After optimization of a designed ligand, the interaction between Rad51 and BRCA2 may be prevented.

Likely the most unusual feature of the structure is the interdomain linker. This 19 amino acid linker between the N- and C-terminal domains appears to be highly flexible. In the PfRad51 structure the majority of N-terminal domains were disordered presumably due to motion and lack of crystal contacts, whereas solution SAXS measurements supported their presence. The quaternary structure of the PfRad51 consists of two oppositely stacked heptameric rings. The interdomain linker contains a polymerization motif (Rad51-PM) that consists of the conserved sequence G*FxxAxE (* = possible insertion, x = differing residues) and provides a key interface for the ring assembly [25, 50] (Figures 7(a), 8(a), and 8(c)). Residues of this motif form a beta-strand $\beta 0$ that extends the central beta-sheet of the neighboring subunit. For additional stability, the Phe residue also buries itself into a hydrophobic pocket formed by the adjacent subunit. These interactions allow the ring assembly to transition to a helical filament during DNA binding. Visualizing the interdomain linker in subunits from the different quaternary assemblies found within the PfRad51 (ring) and SsRadA (extended structure) crystal structures illustrates the flexibility of the linker (Figures 7(a) and 7(b)). Additionally, by also comparing these structures with the MvRadA (filament), it is also revealed that the $\alpha 5/\beta 0$ elbow-like bend usually remains rigid within the linker [206].

During HR, Rad51 binds to DNA in two steps, first binding to the 3′ ssDNA overhang of a DSB via the primary DNA binding site and then binding to homologous dsDNA at a secondary site. Electron microscopy (EM) studies show that the primary Rad51 DNA binding site likely lies at the center of the Rad51 filament, as for RecA [202]. DNA-bound RecA and Rad51 filaments have a large outer groove with one smooth side and one lobed side. Biochemical evidence indicates that the stoichiometry of the Rad51 nucleoprotein filament is 1 Rad51 monomer per 3 or 4 nucleotides [23, 25, 191, 208–210]. These filaments expand or contract in the presence of ATP, ADP, or other nucleotide analogs. The lobes within the DNA binding groove are likely to include the N- or C-terminal domains of Rad51 and RecA, respectively, and these regions also undergo significant nucleotide-induced conformational change [202, 203]. Filament expansion likely involves a change in DNA base rotamer conformation, as

observed for RecA-bound DNA by NMR [211]. *In vitro*, Rad51 filaments bind ssDNA or dsDNA in the primary DNA binding site.

Comparing the individual subunits from the 3 domains of life, it is readily apparent that archaeal RadA/Rad51 proteins are structurally similar to eukaryotic Rad51 proteins. Despite that the subunit structures of PfRad51 and *S. cerevisiae* RAD51 (ScRAD51) [23] in Figure 7 are derived from ring and filament assemblies, respectively, they are obviously quite similar in tertiary structure. Moreover, the eukaryotic proteins also possess the Rad51-PM and have also been found to exist as rings [212] or filaments [23]. On the other hand, a structure of *E. coli* RecA (EcRecA) [200] illustrates a much different prototypical subunit organization for the bacterial DNA strand exchange proteins (Figure 7(d)). For these proteins, the ATPase domain now represents the N-terminal domain, and the C-terminus consists of a small DNA binding domain that is roughly the same size as RadA/Rad51 N-terminal domains. A small N-terminal arm consisting of an alpha-helix and linker provides the interactions for polymerization. Interestingly, many archaea contain paralogs of RadA, which could perhaps perform similar roles to eukaryal Rad51 homologs [173, 213]. Some euryarchaea contain a paralog implicated in HR that is smaller than RadA called RadB [191, 213]. This protein consists of just the ATPase domain [214] (Figure 7(e)). RadB has DNA binding properties and, depending on the species from which it was characterized, has been implicated in a variety of interactions and functions. These include inhibiting and activating DNA strand exchange [173], and interacting with RadA/Rad51, the polD polymerase, and the Hjc holiday junction cleavage protein [191, 193, 214].

3.8. Utility. The properties of archaeal DNA repair enzymes, the similarity to eukaryotic homologs, and stability, are being exploited for inhibitor design. In eukaryotes, unrepaired DNA double-strand breaks (DSBs) can trigger cells to undergo programmed cell death, a process known as apoptosis. Alternatively, DSBs may lead to gross chromosomal rearrangements or loss, thus threatening genome stability. Illustrating the importance of Rad51 in metazoans, knockout mice deficient in Rad51 die during embryogenesis [215], and chicken DT40 cells lacking Rad51 show reduced viability [216]. The breast cancer susceptibility protein BRCA2 acts as a mediator for generating Rad51 nucleoprotein filaments, thus playing a role in HR. Women who carry a BRCA2 mutation have a greatly increased lifetime risk for developing breast or ovarian cancer [217]. The central region of BRCA2 contains a set of 8 noncontiguous ~30 amino acid repeat sequences. These sequences termed BRC repeats contain many tumorigenic polymorphisms, where a single mutation within a repeat can increase cancer risk [24, 218]. BRC repeats bind directly to the Rad51 filament to mediate their loading onto DNA; however BRC repeat-derived peptides prevent Rad51 polymerization into rings and nucleoprotein filaments *in vitro* [219, 220] and prevent nuclear aggregates of Rad51 *in vivo* [221]. By overlaying the PfRad51 structure with a structure of the human Rad51 ATPase domain fused to repeat BRC4 [24], it was revealed that the BRC repeat-derived peptide mimics the Rad51-PM and would disrupt Rad51:Rad51 intersubunit interactions, so that they may be loaded onto DNA as individual subunits (Figure 8) [25, 50]. While BRC repeats do not bind archaeal RadA/Rad51 proteins, the extreme structural similarity between PfRad51 and HsRad51 allowed the generation of a mutant PfRad51 that could be bound by BRCA2 and transported to nuclei in irradiated human cells that would contain DSBs. The further utility of the archaeal PfRad51 enzyme is being exploited as a platform for drug design (Figure 8(e)), as it is more stable and homogeneous in solution than the human enzyme [26]. Again, a few mutations replicate the surface properties of the human enzyme at the pharmaceutical target site, the binding site of the Rad51-PM.

4. Conclusions and Prospects

With the first views of cells under the light microscope, classifying microbes without nuclei together against cells from animals and plants with nuclei was indeed intuitive. With the advent of techniques such as X-ray crystallography and NMR, a closer look "under the hood" revealed that some of the "parts" of these cells certainly resembled those believed to be parts of more distant relatives, paralleling insights from available sequencing data. The stability of macromolecules from archaeal thermophiles often allows obtaining "the first structure" of a class of enzymes. Furthermore, the realization that the overall folds and architectures between many archaeal and eukaryal proteins are similar is a huge benefit for using structures to understand human disease. For instance, not only are the tertiary structures often conserved, but also at the primary level residues that result in disease when mutated are often conserved. This aids interpretation of mechanistic defects at the basic research level and the use of structures at the application level. As systems that inform responses to extreme environmental stress, the archaeal proteins provide biological insights along with precise structural knowledge of complexes and conformations that are often prototypical and foundational. Defining the abilities and limits of the adaptive strategies employed by extremophiles to thrive under extreme stress is also relevant to determining the chemical and physical boundaries that limit life on Earth and beyond for life as we know it. The huge efforts on human systems often provide complementary information so that the combination of archaeal and human structural and biological data provides a deep and comprehensive understanding of great value and utility.

Here we highlighted such prototypic examples of DNA replication and repair systems, and, for several of these proteins, archaeal structures predated those of human structures or served as the only representatives of their class of protein. We have noted that archaea provide deep insights into mechanisms of maintaining genome integrity in the face of extreme environmental stress, with prospects of temperature-trapping flexible complexes and revealing core domains and transient and dynamic complexes. Indeed, archaeal windows into genome integrity have proven exceptionally bright and

clear compared to other choices. In concert with archaeal systems, bacterial thermophiles such as *T. maritima* have also provided pertinent examples for X-ray crystallographic studies of proteins involved in DNA damage responses but are outside the scope of this review. However, these include recombination repair by RuvB [222], nucleotide excision repair by UvrC [223], and deaminated base excision repair by endonuclease V [224]. Likewise, many informative archaeal topoisomerase structures have added greatly to our understanding of these systems [225, 226], but these examples came after other human and *E. coli* topoisomerase family structures [227–230] and are also outside the scope of this review, which again is mainly focused upon systems where archaeal results had led the way to understanding human systems and processes.

Archaeal proteins have allowed us and other researchers to start the process of bridging structures to pathways and systems. They can provide structures that are not just "parts-lists" but include interactions and conformations that link to functional networks. The coupling of advanced archaeal genetics and advanced structural methods to combine MX and SAXS promises to provide an integrated and predictive knowledge of the dynamic structural machines critical to cell biology [231, 232]. For instance, the development of genetics for prototypic archaeal systems such as *Sulfolobus* [233, 234] and *Pyrococcus* [67, 235, 236] coupled to advanced small-angle X-ray scattering methods [237–239] are but a few powerful applications of these advances. Although many archaea are anaerobic, the development of anaerobic iLOV as well as the aerobic green fluorescent protein GFP now allows fluorescent labeling of archaeal proteins in anaerobic and aerobic systems [240, 241]. Going forward, a structural and mechanistic understanding of critical networks, such as those that respond to environmental stress and change, will enable applications such as rewiring bugs for synthetic biology and biomanufacturing. Archaea furthermore provide insights into responses to environmental stresses, such as heavy metal ions, that pose challenges for DNA integrity and repair [242]. Currently, archaeal DNA enzymes are already widely used in biotechnology for PCR [243] and detection assays [244]. Furthermore experiments with SAXS show archaeal thermophiles allow the direct testing and visualization of dynamic ATP-driven conformational changes that control different biological outcomes [245]. Thus, we can expect archaeal structural biology to remain both important and vibrant in the next decade with both medical and industrial impacts.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to acknowledge the pioneering work of Carl Woese. We thank Gareth J. Williams for comments and suggestions. Researcher efforts on archaeal systems are supported in part by the National Institute of Health funding from CA081967, AI22160, GM046312, CA112093, CA117638, and GM105404. Ashley J. Pratt was supported by an NIH/NIA T32AG000266 postdoctoral training grant. Research content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] N. R. Pace, J. Sapp, and N. Goldenfeld, "Phylogeny and beyond: scientific, historical, and conceptual significance of the first tree of life," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 4, pp. 1011–1018, 2012.

[2] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[3] N. R. Pace, "Problems with 'procaryote'," *Journal of Bacteriology*, vol. 191, no. 7, pp. 2008–2010, 2009.

[4] G. E. Fox, L. J. Magrum, and W. E. Balch, "Classification of methanogenic bacteria by 16S ribosomal RNA characterization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 10, pp. 4537–4541, 1977.

[5] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.

[6] T. Stock, M. Selzer, S. Connery, D. Seyhan, A. Resch, and M. Rother, "Disruption and complementation of the selenocysteine biosynthesis pathway reveals a hierarchy of selenoprotein gene expression in the archaeon Methanococcus maripaludis," *Molecular Microbiology*, vol. 82, no. 3, pp. 734–747, 2011.

[7] T. C. Stadtman, "Selenocysteine," *Annual Review of Biochemistry*, vol. 65, pp. 83–100, 1996.

[8] G. T. Mullenbach, A. Tabrizi, B. D. Irvine, G. I. Bell, J. A. Tainer, and R. A. Hallewell, "Selenocysteine's mechanism of incorporation and evolution revealed in cDNAs of three glutathione peroxidases," *Protein Engineering*, vol. 2, no. 3, pp. 239–246, 1988.

[9] K. Egorova and G. Antranikian, "Industrial relevance of thermophilic archaea," *Current Opinion in Microbiology*, vol. 8, no. 6, pp. 649–655, 2005.

[10] K. O. Stetter, "A brief history of the discovery of hyperthermophilic life," *Biochemical Society Transactions*, vol. 41, no. 1, pp. 416–420, 2013.

[11] J. K. Yano and T. L. Poulos, "New understandings of thermostable and peizostable enzymes," *Current Opinion in Biotechnology*, vol. 14, no. 4, pp. 360–365, 2003.

[12] D. E. McRee, S. M. Redford, E. D. Getzoff, J. R. Lepock, R. A. Hallewell, and J. A. Tainer, "Changes in crystallographic structure and thermostability of a Cu,Zn superoxide dismutase mutant resulting from the removal of a buried cysteine," *Journal of Biological Chemistry*, vol. 265, no. 24, pp. 14234–14241, 1990.

[13] M. DiDonato, L. Craig, M. E. Huff et al., "ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization," *Journal of Molecular Biology*, vol. 334, no. 1, pp. 175–175, 2003.

[14] D. S. Shin, M. DiDonato, D. P. Barondeau et al., "Superoxide dismutase from the eukaryotic thermophile Alvinella pompejana:

structures, stability, mechanism, and insights into amyotrophic lateral sclerosis," *Journal of Molecular Biology*, vol. 385, no. 5, pp. 1534–1555, 2009.

[15] J. J. P. Perry, D. S. Shin, E. D. Getzoff, and J. A. Tainer, "The structural biochemistry of the superoxide dismutases," *Biochimica et Biophysica Acta*, vol. 1804, no. 2, pp. 245–262, 2010.

[16] S. M. Yannone, S. Hartung, A. L. Menon, M. W. W. Adams, and J. A. Tainer, "Metals in biology: defining metalloproteomes," *Current Opinion in Biotechnology*, vol. 23, no. 1, pp. 89–95, 2012.

[17] W. A. Lancaster, J. L. Praissman, F. L. Poole II et al., "A computational framework for proteome-wide pursuit and prediction of metalloproteins using ICP-MS and MS/MS data," *BMC Bioinformatics*, vol. 12, article 64, 2011.

[18] N. G. Lintner, K. A. Frankel, S. E. Tsutakawa et al., "The structure of the CRISPR-associated protein csa3 provides insight into the regulation of the CRISPR/Cas system," *Journal of Molecular Biology*, vol. 405, no. 4, pp. 939–955, 2011.

[19] M. Robinson-Rechavi, A. Alibés, and A. Godzik, "Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritima," *Journal of Molecular Biology*, vol. 356, no. 2, pp. 547–557, 2006.

[20] L. Fan, R. S. Williams, D. S. Shin et al., "Master keys to DNA replication, repair, and recombination from the structural biology of enzymes from thermophiles," in *Thermophiles*, F. T. Robb, G. Antranikian, D. Grogan, and A. Driessen, Eds., pp. 239–263, CRC Press, New York, NY, USA, 2008.

[21] M. DiDonato, A. M. Deacon, H. E. Klock, D. McMullan, and S. A. Lesley, "A scaleable and integrated crystallization pipeline applied to mining the Thermotoga maritima proteome," *Journal of Structural and Functional Genomics*, vol. 5, no. 1-2, pp. 133–146, 2004.

[22] H. Aihara, Y. Ito, H. Kurumizaka, S. Yokoyama, and T. Shibata, "The N-terminal domain of the human Rad51 protein binds DNA: structure and a DNA binding surface as revealed by NMR," *Journal of Molecular Biology*, vol. 290, no. 2, pp. 495–504, 1999.

[23] A. B. Conway, T. W. Lynch, Y. Zhang et al., "Crystal structure of a Rad51 filament," *Nature Structural and Molecular Biology*, vol. 11, no. 8, pp. 791–796, 2004.

[24] L. Pellegrini, D. S. Yu, T. Lo et al., "Insights into DNA recombination from the structure of a RAD51-BRCA2 complex," *Nature*, vol. 420, no. 6913, pp. 287–293, 2002.

[25] D. S. Shin, L. Pellegrini, D. S. Daniels et al., "Full-length archaeal Rad51 structure and mutants: mechanisms for RAD51 assembly and control by BRCA2," *The EMBO Journal*, vol. 22, no. 17, pp. 4566–4576, 2003.

[26] D. E. Scott, M. T. Ehebauer, T. Pukala et al., "Using a fragment-based approach to target protein-protein interactions," *Chembiochem*, vol. 14, no. 3, pp. 332–342, 2013.

[27] W. Zillig, "Comparative biochemistry of archaea and bacteria," *Current Opinion in Genetics and Development*, vol. 1, no. 4, pp. 544–551, 1991.

[28] W. Zillig, K. O. Stetter, and D. Janekovic, "DNA-dependent RNA polymerase from the archaebacterium Sulfolobus acidocaldarius," *European Journal of Biochemistry*, vol. 96, no. 3, pp. 597–604, 1979.

[29] W. Zillig, K. O. Stetter, and M. Tobien, "DNA-dependent RNA polymerase from Halobacterium halobium," *European Journal of Biochemistry*, vol. 91, no. 1, pp. 193–199, 1978.

[30] P. Forterre, C. Elie, and M. Kohiyama, "Aphidicolin inhibits growth and DNA synthesis in halophilic archaebacteria," *Journal of Bacteriology*, vol. 159, no. 2, pp. 800–802, 1984.

[31] F. M. Pisani, C. De Martino, and M. Rossi, "A DNA polymerase from the archaeon Sulfolobus solfataricus shows sequence similarity to family B DNA polymerases," *Nucleic Acids Research*, vol. 20, no. 11, pp. 2711–2716, 1992.

[32] C. J. Bult, O. White, G. J. Olsen et al., "Complete genome sequence of the Methanogenic archaeon, Methanococcus jannaschii," *Science*, vol. 273, no. 5278, pp. 1058–1073, 1996.

[33] J. N. Reeve, K. Sandman, and C. J. Daniels, "Archaeal histones, nucleosomes, and transcription initiation," *Cell*, vol. 89, no. 7, pp. 999–1002, 1997.

[34] P. P. Dennis, "Ancient ciphers: translation in archaea," *Cell*, vol. 89, no. 7, pp. 1007–1010, 1997.

[35] D. R. Edgell and W. F. Doolittle, "Archaea and the origin(s) of DNA rplication poteins," *Cell*, vol. 89, no. 7, pp. 995–998, 1997.

[36] G. J. Olsen and C. R. Woese, "Archaeal genomics: an overview," *Cell*, vol. 89, no. 7, pp. 991–994, 1997.

[37] K. Hitomi, S. Iwai, and J. A. Tainer, "The intricate structural chemistry of base excision repair machinery: implications for DNA damage recognition, removal, and repair," *DNA Repair*, vol. 6, no. 4, pp. 410–428, 2007.

[38] D. J. Hosfield, D. S. Daniels, C. D. Mol, C. D. Putnam, S. S. Parikh, and J. A. Tainer, "DNA damage recognition and repair pathway coordination revealed by the structural biochemistry of DNA repair enzymes," *Progress in Nucleic Acid Research and Molecular Biology*, vol. 68, pp. 315–347, 2001.

[39] S. S. Wallace, D. L. Murphy, and J. B. Sweasy, "Base excision repair and cancer," *Cancer Letters*, vol. 327, no. 1-2, pp. 73–89, 2012.

[40] J. O. Fuss and J. A. Tainer, "XPB and XPD helicases in TFIIH orchestrate DNA duplex opening and damage verification to coordinate repair with transcription and cell cycle via CAK kinase," *DNA Repair*, vol. 10, no. 7, pp. 697–713, 2011.

[41] T. Iyama and D. M. Wilson III, "DNA repair mechanisms in dividing and non-dividing cells," *DNA Repair*, vol. 12, no. 8, pp. 620–636, 2013.

[42] A. R. Lehmann, "The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases," *Genes and Development*, vol. 15, no. 1, pp. 15–23, 2001.

[43] A. R. Lehmann, "DNA polymerases and repair synthesis in NER in human cells," *DNA Repair*, vol. 10, no. 7, pp. 730–733, 2011.

[44] S. Kashiwagi, I. Kuraoka, Y. Fujiwara et al., "Characterization of a Y-family DNA polymerase eta from the eukaryotic thermophile Alvinella pompejana," *Journal of Nucleic Acids*, vol. 2010, Article ID 701472, 13 pages, 2010.

[45] A. R. Lehmann, "New functions for Y family polymerases," *Molecular Cell*, vol. 24, no. 4, pp. 493–495, 2006.

[46] A. R. Lehmann, "Translesion synthesis in mammalian cells," *Experimental Cell Research*, vol. 312, no. 14, pp. 2673–2676, 2006.

[47] M. A. Edelbrock, S. Kaliyaperumal, and K. J. Williams, "Structural, molecular and cellular functions of MSH2 and MSH6 during DNA mismatch repair, damage signaling and other noncanonical activities," *Mutation Research*, vol. 743, pp. 53–74466, 2013.

[48] A. G. Schroering, M. A. Edelbrock, T. J. Richards, and K. J. Williams, "The cell cycle and DNA mismatch repair," *Experimental Cell Research*, vol. 313, no. 2, pp. 292–304, 2007.

[49] E. Mladenov and G. Iliakis, "Induction and repair of DNA double strand breaks: the increasing spectrum of non-homologous

end joining pathways," *Mutation Research*, vol. 711, no. 1-2, pp. 61–72, 2011.

[50] D. S. Shin, C. Chahwan, J. L. Huffman, and J. A. Tainer, "Structure and function of the double-strand break repair machinery," *DNA Repair*, vol. 3, no. 8-9, pp. 863–873, 2004.

[51] L. H. Thompson, "Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: the molecular choreography," *Mutation Research*, vol. 751, no. 2, pp. 158–246, 2012.

[52] E. Warbrick, "The puzzle of PCNA's many partners," *Bioessays*, vol. 22, no. 11, pp. 997–1006, 2000.

[53] G. L. Moldovan, B. Pfander, and S. Jentsch, "PCNA, the maestro of the replication fork," *Cell*, vol. 129, no. 4, pp. 665–679, 2007.

[54] H. D. Ulrich and T. Takahashi, "Readers of PCNA modifications," *Chromosoma*, vol. 122, no. 4, pp. 259–274, 2013.

[55] D. J. Hosfield, C. D. Mol, B. Shen, and J. A. Tainer, "Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: coupling DNA and PCNA binding to FEN-1 activity," *Cell*, vol. 95, no. 1, pp. 135–146, 1998.

[56] C. Indiani and M. O'Donnell, "The replication clamp-loading machine at work in the three domains of life," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 10, pp. 751–761, 2006.

[57] T. S. Krishna, X. P. Kong, S. Gary, P. M. Burgers, and J. Kuriyan, "Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA," *Cell*, vol. 79, no. 7, pp. 1233–1243, 1994.

[58] X. P. Kong, R. Onrust, M. O'Donnell, and J. Kuriyan, "Three-dimensional structure of the β subunit of E. coli DNA polymerase III holoenzyme: a sliding DNA clamp," *Cell*, vol. 69, no. 3, pp. 425–437, 1992.

[59] I. Dionne, R. K. Nookala, S. P. Jackson, A. J. Doherty, and S. D. Bell, "A heterotrimeric PCNA in the hyperthermophilic archaeon Sulfolobus solfataricus," *Molecular Cell*, vol. 11, no. 1, pp. 275–282, 2003.

[60] J. M. Pascal, O. V. Tsodikov, G. L. Hura et al., "A flexible interface between DNA ligase and PCNA supports conformational switching and efficient ligation of DNA," *Molecular Cell*, vol. 24, no. 2, pp. 279–291, 2006.

[61] A. Kawai, H. Hashimoto, S. Higuchi et al., "A novel heterotetrameric structure of the crenarchaeal PCNA2-PCNA3 complex," *Journal of Structural Biology*, vol. 174, no. 3, pp. 443–450, 2011.

[62] T. R. Beattie and S. D. Bell, "The role of the DNA sliding clamp in Okazaki fragment maturation in archaea and eukaryotes," *Biochemical Society Transactions*, vol. 39, no. 1, pp. 70–76, 2011.

[63] J. M. Gulbis, Z. Kelman, J. Hurwitz, M. O'Donnell, and J. Kuriyan, "Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA," *Cell*, vol. 87, no. 2, pp. 297–306, 1996.

[64] J. A. Winter and K. A. Bunting, "Rings in the extreme: PCNA interactions and adaptations in the archaea," *Archaea*, vol. 2012, Article ID 951010, 10 pages, 2012.

[65] B. R. Chapados, D. J. Hosfield, S. Han et al., "Structural basis for FEN-1 substrate specificity and PCNA-mediated activation in DNA replication and repair," *Cell*, vol. 116, no. 1, pp. 39–50, 2004.

[66] S. Sakurai, K. Kitano, H. Yamaguchi et al., "Structural basis for recruitment of human flap endonuclease 1 to PCNA," *The EMBO Journal*, vol. 24, no. 4, pp. 683–693, 2005.

[67] G. L. Hura, A. L. Menon, M. Hammel et al., "Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS)," *Nature Methods*, vol. 6, no. 8, pp. 606–612, 2009.

[68] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer, "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution," *Quarterly Reviews of Biophysics*, vol. 40, no. 3, pp. 191–285, 2007.

[69] S. E. Tsutakawa, A. W. van Wynsberghe, B. D. Freudenthal et al., "Solution X-ray scattering combined with computational modeling reveals multiple conformations of covalently bound ubiquitin on PCNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 43, pp. 17672–17677, 2011.

[70] J. J. Harrington and M. R. Lieber, "The characterization of a mammalian DNA structure-specific endonuclease," *The EMBO Journal*, vol. 13, no. 5, pp. 1235–1246, 1994.

[71] J. J. Harrington and M. R. Lieber, "Functional domains within FEN-1 and RAD2 define a family of structure- specific endonucleases: implications for nucleotide excision repair," *Genes and Development*, vol. 8, no. 11, pp. 1344–1355, 1994.

[72] M. S. DeMott, B. Shen, M. S. Park, R. A. Bambarat, and S. Zigman, "Human RAD2 homolog 1 5′- to 3′-exo/endonuclease can efficiently excise a displaced DNA fragment containing a 5′-terminal abasic lesion by endonuclease activity," *Journal of Biological Chemistry*, vol. 271, no. 47, pp. 30068–30076, 1996.

[73] R. S. Murante, J. A. Rumbaugh, C. J. Barnes, J. R. Norton, and R. A. Bambara, "Calf RTH-1 nuclease can remove the initiator RNAs of Okazaki fragments by endonuclease activity," *Journal of Biological Chemistry*, vol. 271, no. 42, pp. 25888–25897, 1996.

[74] R. A. Bambara, R. S. Murante, and L. A. Henricksen, "Enzymes and reactions at the eukaryotic DNA replication fork," *Journal of Biological Chemistry*, vol. 272, no. 8, pp. 4647–4650, 1997.

[75] Q. Chai, J. Qiu, B. R. Chapados, and B. Shen, "Archaeoglobus fulgidus RNase HII in DNA replication: enzymological functions and activity regulation via metal cofactors," *Biochemical and Biophysical Research Communications*, vol. 286, no. 5, pp. 1073–1081, 2001.

[76] B. R. Chapados, Q. Chai, D. J. Hosfield, J. Qiu, B. Shen, and J. A. Tainer, "Structural biochemistry of a type 2 RNase H: RNA primer recognition and removal during DNA replication," *Journal of Molecular Biology*, vol. 307, no. 2, pp. 541–556, 2001.

[77] L. A. Henricksen and R. A. Bambara, "Multiprotein reactions in mammalian DNA replication," *Leukemia Research*, vol. 22, no. 1, pp. 1–5, 1998.

[78] Y. Matsumoto, "Molecular mechanism of PCNA-dependent base excision repair," *Progress in Nucleic Acid Research and Molecular Biology*, vol. 68, pp. 129–138, 2001.

[79] E. Dogliotti, P. Fortini, B. Pascucci, and E. Parlanti, "The mechanism of switching among multiple BER pathways," *Progress in Nucleic Acid Research and Molecular Biology*, vol. 68, pp. 3–27, 2001.

[80] C. J. Bornarth, T. A. Ranalli, L. A. Henricksen, A. F. Wahl, and R. A. Bambara, "Effect of flap modifications on human FEN1 cleavage," *Biochemistry*, vol. 38, no. 40, pp. 13347–13354, 1999.

[81] S. J. Garforth, D. Patel, M. Feng, and J. R. Sayers, "Unusually wide co-factor tolerance in a metalloenzyme; divalent metal ions modulate edno-exonuclease activity in T5 exonuclease," *Nucleic Acids Research*, vol. 29, no. 13, pp. 2772–2779, 2001.

[82] M. W. Kaiser, N. Lyamicheva, W. Ma et al., "A comparison of eubacterial and archaeal structure-specific 5′- exonucleases," *Journal of Biological Chemistry*, vol. 274, no. 30, pp. 21387–21394, 1999.

[83] H.-I. Kao, L. A. Henricksen, Y. Liu, and R. A. Bambara, "Cleavage specificity of Saccharomyces cerevisiae flap endonuclease 1 suggests a double-flap structure as the cellular substrate," *Journal of Biological Chemistry*, vol. 277, no. 17, pp. 14379–14389, 2002.

[84] S. Kimura, T. Ueda, M. Hatanaka, M. Takenouchi, J. Hashimoto, and K. Sakaguchi, "Plant homologue of flap endonuclease-1: molecular cloning, characterization, and evidence of expression in meristematic tissues," *Plant Molecular Biology*, vol. 42, no. 3, pp. 415–427, 2000.

[85] V. Lyamichev, M. A. D. Brow, V. E. Varvel, and J. E. Dahlberg, "Comparison of the 5′ nuclease activities of Taq DNA polymerase and its isolated nuclease domain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 11, pp. 6143–6148, 1999.

[86] Y. Xie, Y. Liu, J. L. Argueso et al., "Identification of rad27 mutations that confer differential defects in mutation avoidance, repeat tract instability, and flap cleavage," *Molecular and Cellular Biology*, vol. 21, no. 15, pp. 4889–4899, 2001.

[87] Y. Xu, O. Potapova, A. E. Leschziner, N. D. F. Grindley, and C. M. Joyce, "Contacts between the 5′ nuclease of DNA polymerase I and its DNA substrate," *Journal of Biological Chemistry*, vol. 276, no. 32, pp. 30167–30177, 2001.

[88] K. Kikuchi, Y. Taniguchi, A. Hatanaka et al., "Fen-1 facilitates homologous recombination by removing divergent sequences at DNA break ends," *Molecular and Cellular Biology*, vol. 25, no. 16, pp. 6948–6955, 2005.

[89] J. Z. Parrish, C. Yang, B. Shen, and D. Xue, "CRN-1, a Caenorhabditis elegans FEN-1 homologue, cooperates with CPS-6/EndoG to promote apoptotic DNA degradation," *The EMBO Journal*, vol. 22, no. 13, pp. 3451–3460, 2003.

[90] B. Shen, P. Singh, R. Liu et al., "Multiple but dissectible functions of FEN-1 nucleases in nucleic acid processing, genome stability and diseases," *Bioessays*, vol. 27, no. 7, pp. 717–729, 2005.

[91] L. Zheng, M. Zhou, Q. Chai et al., "Novel function of the flap endonuclease 1 complex in processing stalled DNA replication forks," *The EMBO Reports*, vol. 6, no. 1, pp. 83–89, 2005.

[92] M. Kucherlapati, A. Nguyen, M. Kuraguchi et al., "Tumor progression in Apc1638N mice with Exo1 and Fen1 deficiencies," *Oncogene*, vol. 26, no. 43, pp. 6297–6306, 2007.

[93] M. Kucherlapati, K. Yang, M. Kuraguchi et al., "Haploinsufficiency of flap endonuclease (Fen1) leads to rapid tumor progression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 15, pp. 9924–9929, 2002.

[94] L. Zheng, H. Dai, M. Zhou et al., "Fen1 mutations result in autoimmunity, chronic inflammation and cancers," *Nature Medicine*, vol. 13, no. 7, pp. 812–819, 2007.

[95] G. Henneke, E. Friedrich-Heineken, and U. Hübscher, "Flap endonuclease 1: a novel tumour suppresser protein," *Trends in Biochemical Sciences*, vol. 28, no. 7, pp. 384–390, 2003.

[96] D. J. Hosfield, G. Frank, Y. Weng, J. A. Tainer, and B. Shen, "Newly discovered archaebacterial flap endonucleases show a structure- specific mechanism for DNA substrate binding and catalysis resembling human flap endonuclease-1," *Journal of Biological Chemistry*, vol. 273, no. 42, pp. 27154–27161, 1998.

[97] L. Zheng, H. Dai, J. Qiu, Q. Huang, and B. Shen, "Disruption of the FEN-1/PCNA interaction results in DNA replication defects, pulmonary hypoplasia, pancytopenia, and newborn lethality in mice," *Molecular and Cellular Biology*, vol. 27, no. 8, pp. 3176–3186, 2007.

[98] K. Y. Hwang, K. Baek, H.-Y. Kim, and Y. Cho, "The crystal structure of flap endonuclease-1 from Methanococcus jannaschii," *Nature Structural Biology*, vol. 5, no. 8, pp. 707–713, 1998.

[99] J. Querol-Audi, C. Yan, X. Xu et al., "Repair complexes of FEN1 endonuclease, DNA, and Rad9-Hus1-Rad1 are distinguished from their PCNA counterparts by functionally important stability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 22, pp. 8528–8533, 2012.

[100] S. E. Tsutakawa, S. Classen, B. R. Chapados et al., "Human flap endonuclease structures, DNA double-base flipping, and a unified understanding of the FEN1 superfamily," *Cell*, vol. 145, no. 2, pp. 198–211, 2011.

[101] J. A. Grasby, L. D. Finger, S. E. Tsutakawa, J. M. Atack, and J. A. Tainer, "Unpairing and gating: sequence-independent substrate recognition by FEN superfamily nucleases," *Trends in Biochemical Sciences*, vol. 37, no. 2, pp. 74–84, 2012.

[102] T. Lindahl and B. Nyberg, "Heat-induced deamination of cytosine residues in deoxyribonucleic acid," *Biochemistry*, vol. 13, no. 16, pp. 3405–3410, 1974.

[103] M. A. Greagg, M. J. Fogg, G. Panayotou, S. J. Evans, B. A. Connolly, and L. H. Pearl, "A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 16, pp. 9045–9050, 1999.

[104] M. J. Fogg, L. H. Pearl, and B. A. Connolly, "Structural basis for uracil recognition by archaeal family B DNA polymerases," *Nature Structural Biology*, vol. 9, no. 12, pp. 922–927, 2002.

[105] G. Shuttleworth, M. J. Fogg, M. R. Kurpiewski, L. Jen-Jacobson, and B. A. Connolly, "Recognition of the pro-mutagenic base uracil by family B DNA polymerases from archaea," *Journal of Molecular Biology*, vol. 337, no. 3, pp. 621–634, 2004.

[106] S. J. Firbank, J. Wardle, P. Heslop, R. J. Lewis, and B. A. Connolly, "Uracil recognition in archaeal DNA polymerases captured by X-ray crystallography," *Journal of Molecular Biology*, vol. 381, no. 3, pp. 529–539, 2008.

[107] T. T. Richardson, X. H. Wu, B. J. Keith et al., "Unwinding of primer-templates by archaeal family-B DNA polymerases in response to template-strand uracil," *Nucleic Acids Research*, vol. 41, no. 4, pp. 2466–2478, 2013.

[108] T. T. Richardson, L. Gilroy, Y. Ishino et al., "Novel inhibition of archaeal family-D DNA polymerase by uracil," *Nucleic Acids Research*, vol. 41, no. 7, pp. 4207–4218, 2013.

[109] S. S. Cho, Y. Sun, M. Yu et al., "Characterization and PCR applications of dUTPase from the hyperthermophilic euryarchaeon Thermococcus pacificus," *Enzyme and Microbial Technology*, vol. 51, no. 6-7, pp. 342–347, 2012.

[110] J. L. Huffman, H. Li, R. H. White, and J. A. Tainer, "Structural basis for recognition and catalysis by the bifunctional dCTP deaminase and dUTPase from Methanococcus jannaschii," *Journal of Molecular Biology*, vol. 331, no. 4, pp. 885–896, 2003.

[111] J. H. Chung, E. K. Im, H.-Y. Park et al., "A novel uracil-DNA glycosylase family related to the helix-hairpin-helix DNA glycosylase superfamily," *Nucleic Acids Research*, vol. 31, no. 8, pp. 2045–2055, 2003.

[112] C. D. Mol, A. S. Arvai, T. J. Begley, R. P. Cunningham, and J. A. Tainer, "Structure and activity of a thermostable thymine-DNA glycosylase: evidence for base twisting to remove mismatched normal DNA bases," *Journal of Molecular Biology*, vol. 315, no. 3, pp. 373–384, 2002.

[113] S. E. Tsutakawa, D. S. Shin, C. D. Mol et al., "Conserved structural chemistry for incision activity in structurally non-homologous apurinic/apyrimidinic endonuclease APE1 and

endonuclease IV DNA repair enzymes," *The Journal of Biological Chemistry*, vol. 288, no. 12, pp. 8445–8455, 2013.

[114] T. Lindahl, "An N glycosidase from Escherichia coli that releases free uracil from DNA containing deaminated cytosine residues," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 71, no. 9, pp. 3649–3653, 1974.

[115] L. H. Pearl, "Structure and function in the uracil-DNA glycosylase superfamily," *Mutation Research*, vol. 460, no. 3-4, pp. 165–181, 2000.

[116] C. D. Putnam, M. J. N. Shroyer, A. J. Lundquist et al., "Protein mimicry of DNA from crystal structures of the uracil-DNA glycosylase inhibitor protein and its complex with Escherichia coli uracil-DNA glycosylase," *Journal of Molecular Biology*, vol. 287, no. 2, pp. 331–346, 1999.

[117] B. Kavli, M. Otterlei, G. Slupphaug, and H. E. Krokan, "Uracil in DNA-General mutagen, but normal intermediate in acquired immunity," *DNA Repair*, vol. 6, no. 4, pp. 505–516, 2007.

[118] T. E. Barrett, R. Savva, G. Panayotou et al., "Crystal structure of a G:T/U mismatch-specific DNA glycosylase: mismatch recognition by complementary-strand interactions," *Cell*, vol. 92, no. 1, pp. 117–129, 1998.

[119] A. Maiti, M. T. Morgan, E. Pozharski, and A. C. Drohat, "Crystal structure of human thymine DNA glycosylase bound to DNA elucidates sequence-specific mismatch recognition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 26, pp. 8890–8895, 2008.

[120] J. E. A. Wibley, T. R. Waters, K. Haushalter, G. L. Verdine, and L. H. Pearl, "Structure and specificity of the vertebrate anti-mutator uracil-DNA glycosylase SMUG1," *Molecular Cell*, vol. 11, no. 6, pp. 1647–1659, 2003.

[121] J. Hoseki, A. Okamoto, R. Masui et al., "Crystal structure of a family 4 uracil-DNA glycosylase from Thermus thermophilus HB8," *Journal of Molecular Biology*, vol. 333, no. 3, pp. 515–526, 2003.

[122] H. Kosaka, J. Hoseki, N. Nakagawa, S. Kuramitsu, and R. Masui, "Crystal structure of family 5 uracil-DNA glycosylase bound to DNA," *Journal of Molecular Biology*, vol. 373, no. 4, pp. 839–850, 2007.

[123] A. A. Sartori, P. Schär, S. Fitz-Gibbon, J. H. Miller, and J. Jiricny, "Biochemical characterization of uracil processing activities in the hyperthermophilic archaeon pyrobaculum aerophilum," *Journal of Biological Chemistry*, vol. 276, no. 32, pp. 29979–29986, 2001.

[124] A. A. Sartori, S. Fitz-Gibbon, H. Yang, J. H. Miller, and J. Jiricny, "A novel uracil-DNA glycosylase with broad substrate specificity and an unusual active site," *The EMBO Journal*, vol. 21, no. 12, pp. 3182–3191, 2002.

[125] H.-W. Lee, B. N. Dominy, and W. Cao, "New family of deamination repair enzymes in uracil-DNA glycosylase superfamily," *Journal of Biological Chemistry*, vol. 286, no. 36, pp. 31282–31287, 2011.

[126] J. I. Lucas-Lledó, R. Maddamsetti, and M. Lynch, "Phylogenomic analysis of the uracil-DNA glycosylase superfamily," *Molecular Biology and Evolution*, vol. 28, no. 3, pp. 1307–1317, 2011.

[127] L. Aravind and E. V. Koonin, "The alpha/beta fold uracil DNA glycosylases: a common origin with diverse fates," *Genome biology*, vol. 1, no. 4, p. RESEARCH0007, 2000.

[128] G. Slupphaug, C. D. Mol, B. Kavli, A. S. Arvai, H. E. Krokan, and J. A. Tainer, "A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA," *Nature*, vol. 384, no. 6604, pp. 87–92, 1996.

[129] A. Kawai, S. Higuchi, M. Tsunoda, K. T. Nakamura, and S. Miyamoto, "Purification, crystallization and preliminary X-ray analysis of the PCNA2-PCNA3 complex from Sulfolobus tokodaii strain 7," *Acta Crystallographica F*, vol. 65, no. 12, pp. 1282–1284, 2009.

[130] O. Sundheim, C. B. Vågbø, M. Bjørås et al., "Human ABH3 structure and key residues for oxidative demethylation to reverse DNA/RNA damage," *The EMBO Journal*, vol. 25, no. 14, pp. 3389–3397, 2006.

[131] C. Yi, C.-G. Yang, and C. He, "A non-heme iron-mediated chemical demethylation in DNA and RNA," *Accounts of Chemical Research*, vol. 42, no. 4, pp. 519–529, 2009.

[132] I. Leiros, M. P. Nabong, K. Grøsvik et al., "Structural basis for enzymatic excision of N1-methyladenine and N3-methylcytosine from DNA," *The EMBO Journal*, vol. 26, no. 8, pp. 2206–2217, 2007.

[133] A. Sancar and W. D. Rupp, "A novel repair enzyme: UVRABC excision nuclease of Escherichia coli cuts a DNA strand on both sides of the damaged region," *Cell*, vol. 33, no. 1, pp. 249–260, 1983.

[134] E. Compe and J. M. Egly, "TFIIH: when transcription met DNA repair," *Nature Reviews Molecular Cell Biology*, vol. 13, no. 6, pp. 343–354, 2012.

[135] L. Fan, A. S. Arvai, P. K. Cooper, S. Iwai, F. Hanaoka, and J. A. Tainer, "Conserved XPB core structure and motifs for DNA unwinding: implications for pathway selection of transcription or excision repair," *Molecular Cell*, vol. 22, no. 1, pp. 27–37, 2006.

[136] E. Cameroni, K. Stettler, and B. Suter, "On the traces of XPD: cell cycle matters—untangling the genotype-phenotype relationship of XPD mutations," *Cell Division*, vol. 5, article 24, 2010.

[137] Y. He, J. Fang, D. J. Taatjes et al., "Structural visualization of key steps in human transcription initiation," *Nature*, vol. 495, no. 7442, pp. 481–486, 2013.

[138] A. H. Sarker, S. E. Tsutakawa, S. Kostek et al., "Recognition of RNA polymerase II and transcription bubbles by XPG, CSB, and TFIIH: insights for transcription-coupled repair and Cockayne syndrome," *Molecular Cell*, vol. 20, no. 2, pp. 187–198, 2005.

[139] J. Rudolf, V. Makrantoni, W. J. Ingledew, M. J. R. Stark, and M. F. White, "The DNA repair helicases XPD and FancJ have essential iron-sulfur domains," *Molecular Cell*, vol. 23, no. 6, pp. 801–808, 2006.

[140] W. Abdulrahman, I. Iltis, L. Radu et al., "ARCH domain of XPD, an anchoring platform for CAK that conditions TFIIH DNA repair and transcription activities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 8, pp. E633–E642, 2013.

[141] R. A. Pugh, M. Honda, H. Leesley et al., "The iron-containing domain is essential in Rad3 helicases for coupling of ATP hydrolysis to DNA translocation and for targeting the helicase to the single-stranded DNA-double-stranded DNA junction," *Journal of Biological Chemistry*, vol. 283, no. 3, pp. 1732–1743, 2008.

[142] Y. L. Wu and R. M. Brosh, "DNA helicase and helicase-nuclease enzymes with a conserved iron-sulfur cluster," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4247–4260, 2012.

[143] H. Naegeli, L. Bardwell, and E. C. Friedberg, "Inhibition of Rad3 DNA helicase activity by DNA adducts and abasic sites: implications for the role of a DNA helicase in damage-specific incision of DNA," *Biochemistry*, vol. 32, no. 2, pp. 613–621, 1993.

[144] N. Mathieu, N. Kaczmarek, and H. Naegeli, "Strand- and site-specific DNA lesion demarcation by the xeroderma pigmentosum group D helicase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 41, pp. 17545–17550, 2010.

[145] N. Mathieu, N. Kaczmarek, P. Ruthemann et al., "DNA quality control by a lesion sensor pocket of the xeroderma pigmentosum group D helicase subunit of TFIIH," *Current Biology*, vol. 23, no. 3, pp. 204–212, 2013.

[146] J. de Boer, I. Donker, J. de Wit, J. H. J. Hoeijmakers, and G. Weeda, "Disruption of the mouse xeroderma pigmentosum group D DNA repair/basal transcription gene results in preimplantation lethality," *Cancer Research*, vol. 58, no. 1, pp. 89–94, 1998.

[147] L. Fan, J. O. Fuss, Q. J. Cheng et al., "XPD helicase structures and activities: insights into the cancer and aging phenotypes from XPD mutations," *Cell*, vol. 133, no. 5, pp. 789–800, 2008.

[148] H. Liu, J. Rudolf, K. A. Johnson et al., "Structure of the DNA Repair Helicase XPD," *Cell*, vol. 133, no. 5, pp. 801–812, 2008.

[149] S. C. Wolski, J. Kuper, P. Hänzelmann et al., "Crystal structure of the FeS cluster-containing nucleotide excision repair helicase XPD," *PLoS Biology*, vol. 6, no. 6, article e149, pp. 1332–1342, 2008.

[150] J. Kuper, S. C. Wolski, G. Michels, and C. Kisker, "Functional and structural studies of the nucleotide excision repair helicase XPD suggest a polarity for DNA translocation," *The EMBO Journal*, vol. 31, no. 2, pp. 494–502, 2012.

[151] R. A. Pugh, M. Honda, H. Leesley et al., "The iron-containing domain is essential in Rad3 helicases for coupling of ATP hydrolysis to DNA translocation and for targeting the helicase to the single-stranded DNA-double-stranded DNA junction," *Journal of Biological Chemistry*, vol. 283, no. 3, pp. 1732–1743, 2008.

[152] T. P. Mui, J. O. Fuss, J. P. Ishida, J. A. Tainer, and J. K. Barton, "ATP-stimulated, DNA-mediated redox signaling by XPD, a DNA repair and transcription helicase," *Journal of the American Chemical Society*, vol. 133, no. 41, pp. 16378–16381, 2011.

[153] A. Yamagata and J. A. Tainer, "Hexameric structures of the archaeal secretion ATPase GspE and implications for a universal secretion mechanism," *The EMBO Journal*, vol. 26, no. 3, pp. 878–890, 2007.

[154] J. L. Tubbs, V. Latypov, S. Kanugula et al., "Flipping of alkylated DNA damage bridges base and nucleotide excision repair," *Nature*, vol. 459, no. 7248, pp. 808–813, 2009.

[155] B. Dalhus, L. Nilsen, H. Korvald et al., "Sculpting of DNA at abasic sites by DNA glycosylase homolog mag2," *Structure*, vol. 21, no. 1, pp. 154–166, 2013.

[156] E. A. Rahal, L. A. Henricksen, Y. Li, R. S. Williams, J. A. Tainer, and K. Dixon, "ATM regulates Mre11-dependent DNA end-degradation and microhomology-mediated end joining," *Cell Cycle*, vol. 9, no. 14, pp. 2866–2877, 2010.

[157] K. P. Hopfner, L. Craig, G. Moncalian et al., "The Rad50 zinc-hook is a structure joining Mre11 complexes in DNA recombination and repair," *Nature*, vol. 418, no. 6897, pp. 562–566, 2002.

[158] K. P. Hopfner, A. Karcher, L. Craig, T. T. Woo, J. P. Carney, and J. A. Tainer, "Structural biochemistry and interaction architecture of the DNA double-strand break repair Mre11 nuclease and Rad50-ATPase," *Cell*, vol. 105, no. 4, pp. 473–485, 2001.

[159] K. P. Hopfner, A. Karcher, D. S. Shin et al., "Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily," *Cell*, vol. 101, no. 7, pp. 789–800, 2000.

[160] G. J. Williams, R. S. Williams, J. S. Williams et al., "ABC ATPase signature helices in Rad50 link nucleotide state to Mre11 interface for DNA repair," *Nature Structural and Molecular Biology*, vol. 18, no. 4, pp. 423–431, 2011.

[161] R. S. Williams, G. E. Dodson, O. Limbo et al., "Nbs1 flexibly tethers Ctp1 and Mre11-Rad50 to coordinate DNA double-strand break processing and repair," *Cell*, vol. 139, no. 1, pp. 87–99, 2009.

[162] H. S. Lim, J. S. Kim, Y. B. Park, G. H. Gwon, and Y. Cho, "Crystal structure of the Mre11-Rad50-ATPγS complex: understanding the interplay between Mre11 and Rad50," *Genes and Development*, vol. 25, no. 10, pp. 1091–1104, 2011.

[163] C. Möckel, K. Lammens, A. Schele, and K.-P. Hopfner, "ATP driven structural changes of the bacterial Mre11:Rad50 catalytic head complex," *Nucleic Acids Research*, vol. 40, no. 2, pp. 914–927, 2012.

[164] D. Das, D. Moiani, H. L. Axelrod et al., "Crystal structure of the first eubacterial Mre11 nuclease reveals novel features that may discriminate substrates during DNA repair," *Journal of Molecular Biology*, vol. 397, no. 3, pp. 647–663, 2010.

[165] G. Moncalian, B. Lengsfeld, V. Bhaskara et al., "The Rad50 signature motif: essential to ATP binding and biological function," *Journal of Molecular Biology*, vol. 335, no. 4, pp. 937–951, 2004.

[166] M. de Jager, J. van Noort, D. C. van Gent, C. Dekker, R. Kanaar, and C. Wyman, "Human Rad50/Mre11 is a flexible complex that can tether DNA ends," *Molecular Cell*, vol. 8, no. 5, pp. 1129–1135, 2001.

[167] C. B. Schiller, K. Lammens, I. Guerini et al., "Structure of Mre11-Nbs1 complex yields insights into ataxia-telangiectasia-like disease mutations and DNA damage signaling," *Nature Structural and Molecular Biology*, vol. 19, no. 7, pp. 693–700, 2012.

[168] J. Lloyd, J. R. Chapman, J. A. Clapperton et al., "A supramodular FHA/BRCT-repeat architecture mediates Nbs1 adaptor function in response to DNA damage," *Cell*, vol. 139, no. 1, pp. 100–111, 2009.

[169] C. Xu, L. Wu, G. Cui, M. V. Botuyan, J. Chen, and G. Mer, "Structure of a second BRCT domain identified in the nijmegen breakage syndrome protein Nbs1 and its function in an MDC1-dependent localization of Nbs1 to DNA damage sites," *Journal of Molecular Biology*, vol. 381, no. 2, pp. 361–372, 2008.

[170] Y. B. Park, J. Chae, Y. C. Kim, and Y. Cho, "Crystal structure of human Mre11: understanding tumorigenic mutations," *Structure*, vol. 19, no. 11, pp. 1591–1602, 2011.

[171] O. Limbo, D. Moiani, A. Kertokalio et al., "Mre11 ATLD17/18 mutation retains Tel1/ATM activity but blocks DNA double-strand break repair," *Nucleic Acids Research*, vol. 40, no. 22, pp. 11435–11449, 2012.

[172] A. Manzan, G. Pfeiffer, M. L. Hefferin, C. E. Lang, J. P. Carney, and K.-P. Hopfner, "MlaA, a hexameric ATPase linked to the Mre11 complex in archaeal genomes," *The EMBO Reports*, vol. 5, no. 1, pp. 54–59, 2004.

[173] M. F. White, "Homologous recombination in the archaea: the means justify the ends," *Biochemical Society Transactions*, vol. 39, no. 1, pp. 15–19, 2011.

[174] J. K. Blackwood, N. J. Rzechorzek, A. S. Abrams, J. D. Maman, L. Pellegrini, and N. P. Robinson, "Structural and functional insights into DNA-end processing by the archaeal HerA helicase-NurA nuclease complex," *Nucleic Acids Research*, vol. 40, no. 7, pp. 3183–3196, 2012.

[175] B. B. Hopkins and T. T. Paull, "The P. furiosus Mre11/Rad50 Complex Promotes 5′ Strand Resection at a DNA Double-Strand Break," *Cell*, vol. 135, no. 2, pp. 250–260, 2008.

[176] T. Sugiyama, E. M. Zaitseva, and S. C. Kowalczykowski, "A single-stranded DNA-binding protein is needed for efficient presynaptic complex formation by the Saccharomyces cerevisiae Rad51 protein," *Journal of Biological Chemistry*, vol. 272, no. 12, pp. 7940–7945, 1997.

[177] A. Bochkarev, R. A. Pfuetzner, A. M. Edwards, and L. Frappier, "Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA," *Nature*, vol. 385, no. 6612, pp. 176–181, 1997.

[178] J. Fan and N. P. Pavletich, "Structure and conformational change of a replication protein A heterotrimer bound to ssDNA," *Genes and Development*, vol. 26, no. 20, pp. 2337–2347, 2012.

[179] T. Matsumoto, Y. Morimoto, N. Shibata et al., "Roles of functional loops and the C-terminal segment of a single- stranded DNA binding protein elucidated by X-ray structure analysis," *Journal of Biochemistry*, vol. 127, no. 2, pp. 329–335, 2000.

[180] S. Raghunathan, C. S. Ricard, T. M. Lohman, and G. Waksman, "Crystal structure of the homo-tetrameric DNA binding domain of Escherichia coli single-stranded DNA-binding protein determined by multiwavelength x-ray diffraction on the selenomethionyl protein at 2.9-Å resolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 13, pp. 6652–6657, 1997.

[181] S. N. Savvides, S. Raghunathan, K. Fütterer, A. G. Kozlov, T. M. Lohman, and G. Waksman, "The C-terminal domain of full-length E. coli SSB is disordered even when bound to DNA," *Protein Science*, vol. 13, no. 7, pp. 1942–1947, 2004.

[182] F. Chédin, E. M. Seitz, and S. C. Kowalczykowski, "Novel homologs of replication protein A in archaea: implications for the evolution of ssDNA-binding proteins," *Trends in Biochemical Sciences*, vol. 23, no. 8, pp. 273–277, 1998.

[183] R. I. Wadsworth and M. F. White, "Identification and properties of the crenarchaeal single-stranded DNA binding protein from Sulfolobus solfataricus," *Nucleic Acids Research*, vol. 29, no. 4, pp. 914–920, 2001.

[184] I. D. Kerr, R. I. M. Wadsworth, L. Cubeddu, W. Blankenfeldt, J. H. Naismith, and M. F. White, "Insights into ssDNA recognition by the OB fold from a structural and thermodynamic study of Sulfolobus SSB protein," *The EMBO Journal*, vol. 22, no. 11, pp. 2561–2570, 2003.

[185] J. Genschel, U. Curth, and C. Urbanke, "Interaction of E. coli-single-stranded DNA binding protein (SSB) with exonuclease I. The carboxy-terminus of SSB is the recognition site for the nuclease," *Biological Chemistry*, vol. 381, no. 3, pp. 183–192, 2000.

[186] D. J. Richard, E. Bolderson, L. Cubeddu et al., "Single-stranded DNA-binding protein hSSB1 is critical for genomic stability," *Nature*, vol. 453, no. 7195, pp. 677–681, 2008.

[187] J. Huang, Z. Gong, G. Ghosal, and J. Chen, "SOSS complexes participate in the maintenance of genomic stability," *Molecular Cell*, vol. 35, no. 3, pp. 384–393, 2009.

[188] J. R. Skaar, D. J. Richard, A. Saraf et al., "INTS3 controls the hSSB1-mediated DNA damage response," *Journal of Cell Biology*, vol. 187, no. 1, pp. 25–32, 2009.

[189] S. H. Yang, R. Zhou, J. Campbell et al., "The SOSS1 single-stranded DNA binding complex promotes DNA end resection in concert with Exo1," *The EMBO Journal*, vol. 32, no. 1, pp. 126–139, 2013.

[190] K. Komori and Y. Ishino, "Replication protein A in Pyrococcus furiosus is involved in homologous DNA recombination," *Journal of Biological Chemistry*, vol. 276, no. 28, pp. 25654–25660, 2001.

[191] K. Komori, T. Miyata, J. DiRuggiero et al., "Both RadA and RadB are involved in homologous recombination in Pyrococcus furiosus," *Journal of Biological Chemistry*, vol. 275, no. 43, pp. 33782–33790, 2000.

[192] H. Yang, P. D. Jeffrey, J. Miller et al., "BRCA2 function in DNA binding and recombination from a BRCA2-DSS1-ssDNA structure," *Science*, vol. 297, no. 5588, pp. 1837–1848, 2002.

[193] I. Hayashi, K. Morikawa, and Y. Ishino, "Specific interaction between DNA polymerase II (PolD) and RadB, a Rad51/Dmc1 homolog, in Pyrococcus furiosus," *Nucleic Acids Research*, vol. 27, no. 24, pp. 4695–4702, 1999.

[194] C. L. Middleton, J. L. Parker, D. J. Richard, M. F. White, and C. S. Bond, "Crystallization and preliminary X-ray diffraction studies of Hje, a Holliday junction resolving enzyme from Sulfolobus solfataricus," *Acta Crystallographica D*, vol. 59, no. 1, pp. 171–173, 2003.

[195] T. Nishino, K. Komori, D. Tsuchiya, Y. Ishino, and K. Morikawa, "Crystal structure of the archaeal Holliday junction resolvase Hjc and implications for DNA recognition," *Structure*, vol. 9, no. 3, pp. 197–204, 2001.

[196] T. Nishino, K. Komori, Y. Ishino, and K. Morikawa, "Dissection of the regional roles of the archaeal Holliday Junction resolvase Hjc by structural and mutational analyses," *Journal of Biological Chemistry*, vol. 276, no. 38, pp. 35735–35740, 2001.

[197] A. J. Clark and A. D. Margulies, "Isolation and characterization of recombination-deficient mutants of E. coli K-12," *Proceedings of the National Academy of Sciences of the United States of*, vol. 53, pp. 451–459, 1965.

[198] S. J. Sandler, L. H. Satin, H. S. Samra, and A. J. Clark, "RecA-like genes from three archaean species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast Saccharomyces cerevisiae," *Nucleic Acids Research*, vol. 24, no. 11, pp. 2125–2132, 1996.

[199] S. Datta, M. M. Prabu, M. B. Vaze et al., "Crystal structures of Mycobacterium tuberculosis RecA and its complex with ADP-AIF4: implications for decreased ATPase activity and molecular aggregation," *Nucleic Acids Research*, vol. 28, no. 24, pp. 4964–4973, 2000.

[200] R. M. Story, I. T. Weber, and T. A. Steitz, "The structure of the E. coli RecA protein monomer and polymer," *Nature*, vol. 355, no. 6358, pp. 318–325, 1992.

[201] S. Yang, M. S. VanLoock, X. Yu, and E. H. Egelman, "Comparison of bacteriophage T4 UvsX and human Rad51 filaments suggests that RecA-like polymers may have evolved independently," *Journal of Molecular Biology*, vol. 312, no. 5, pp. 999–1009, 2001.

[202] X. Yu, S. A. Jacobs, S. C. West, T. Ogawa, and E. H. Egelman, "Domain structure and dynamics in the helical filaments formed by RecA and Rad51 on DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 15, pp. 8419–8424, 2001.

[203] M. S. VanLoock, X. Yu, S. Yang et al., "ATP-mediated conformational changes in the RecA filament," *Structure*, vol. 11, no. 2, pp. 187–196, 2003.

[204] M. M. Thayer, H. Ahern, D. Xing, R. P. Cunningham, and J. A. Tainer, "Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure," *The EMBO Journal*, vol. 14, no. 16, pp. 4108–4120, 1995.

[205] H. Pelletier, M. R. Sawaya, W. Wolfle, S. H. Wilson, and J. Kraut, "Crystal structures of human DNA polymerase β complexed

with DNA: implications for catalytic mechanism, processivity, and fidelity," *Biochemistry*, vol. 35, no. 39, pp. 12742–12761, 1996.

[206] Y. Wu, Y. He, I. A. Moya, X. Qian, and Y. Luo, "Crystal structure of archaeal recombinase RadA: a snapshot of its extended conformation," *Molecular Cell*, vol. 15, no. 3, pp. 423–435, 2004.

[207] A. Ariza, D. J. Richard, M. F. White, and C. S. Bond, "Conformational flexibility revealed by the crystal structure of a crenarchaeal RadA," *Nucleic Acids Research*, vol. 33, no. 5, pp. 1465–1473, 2005.

[208] P. Sung and D. L. Robberson, "DNA strand exchange mediated by a RAD51-ssDNA nucleoprotein filament with polarity opposite to that of RecA," *Cell*, vol. 82, no. 3, pp. 453–461, 1995.

[209] F. E. Benson, A. Stasiak, and S. C. West, "Purification and characterization of the human Rad51 protein, an analogue of E.coli RecA," *The EMBO Journal*, vol. 13, no. 23, pp. 5764–5771, 1994.

[210] G. Tombline and R. Fishel, "Biochemical characterization of the human RAD51 protein. I. ATP hydrolysis," *Journal of Biological Chemistry*, vol. 277, no. 17, pp. 14417–14425, 2002.

[211] T. Nishinaka, Y. Ito, S. Yokoyama, and T. Shibata, "An extended DNA structure through deoxyribose-base stacking induced by RecA protein," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 13, pp. 6623–6628, 1997.

[212] T. Kinebuchi, W. Kagawa, R. Enomoto et al., "Structural basis for octameric ring formation and DNA interaction of the human homologous-pairing protein Dmc1," *Molecular Cell*, vol. 14, no. 3, pp. 363–374, 2004.

[213] S. Haldenby, M. F. White, and T. Allers, "RecA family proteins in archaea: RadA and its cousins," *Biochemical Society Transactions*, vol. 37, no. 1, pp. 102–107, 2009.

[214] T. Akiba, N. Ishii, N. Rashid, M. Morikawa, T. Imanaka, and K. Harata, "Structure of RadB recombinase from a hyperthermophilic archaeon, Thermococcus kodakaraensis KOD1: an implication for the formation of a near-7-fold helical assembly," *Nucleic Acids Research*, vol. 33, no. 10, pp. 3412–3423, 2005.

[215] D. S. Lim and P. Hasty, "A mutation in mouse rad51 results in an early embryonic lethal that is suppressed by a mutation in p53," *Molecular and Cellular Biology*, vol. 16, no. 12, pp. 7133–7143, 1996.

[216] E. Sonoda, M. S. Sasaki, J. M. Buerstedde et al., "Rad51-deficient vertebrate cells accumulate chromosomal breaks prior to cell death," *The EMBO Journal*, vol. 17, no. 2, pp. 598–608, 1998.

[217] K. N. Nathanson, R. Wooster, and B. L. Weber, "Breast cancer genetics: what we know and what we need," *Nature Medicine*, vol. 7, no. 5, pp. 552–556, 2001.

[218] A. R. Venkitaraman, "Cancer susceptibility and the functions of BRCA1 and BRCA2," *Cell*, vol. 108, no. 2, pp. 171–182, 2002.

[219] A. A. Davies, J. Y. Masson, M. J. McIlwraith et al., "Role of BRCA2 in control of the RAD51 recombination and DNA repair protein," *Molecular Cell*, vol. 7, no. 2, pp. 273–282, 2001.

[220] A. K. Wong, R. Pero, P. A. Ormonde, S. V. Tavtigian, and P. L. Bartel, "RAD51 interacts with the evolutionarily conserved BRC motifs in the human breast cancer susceptibility gene brca2," *Journal of Biological Chemistry*, vol. 272, no. 51, pp. 31941–31944, 1997.

[221] P. L. Chen, C. F. Chen, Y. Chen, J. Xiao, Z. D. Sharp, and W.-H. Lee, "The BRC repeats in BRCA2 are critical for RAD51 binding and resistance to methyl methanesulfonate treatment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 9, pp. 5287–5292, 1998.

[222] C. D. Putnam, S. B. Clancy, H. Tsuruta, S. Gonzalez, J. G. Wetmur, and J. A. Tainer, "Structure and mechanism of the RuvB holliday junction branch migration motor," *Journal of Molecular Biology*, vol. 311, no. 2, pp. 297–310, 2001.

[223] J. J. Truglio, B. Rhau, D. L. Croteau et al., "Structural insights into the first incision reaction during nucleotide excision repair," *The EMBO Journal*, vol. 24, no. 5, pp. 885–894, 2005.

[224] B. Dalhus, A. S. Arvai, I. Rosnes et al., "Structures of endonuclease V with DNA reveal initiation of deaminated adenine repair," *Nature Structural and Molecular Biology*, vol. 16, no. 2, pp. 138–143, 2009.

[225] M. D. Nichols, K. DeAngelis, J. L. Keck, and J. M. Berger, "Structure and function of an archaeal topoisomerase VI subunit with homology to the meiotic recombination factor Spo11," *The EMBO Journal*, vol. 18, no. 21, pp. 6177–6188, 1999.

[226] A. Chapin Rodríguez and D. Stock, "Crystal structure of reverse gyrase: insights into the positive supercoiling of DNA," *The EMBO Journal*, vol. 21, no. 3, pp. 418–426, 2002.

[227] C. D. Lima, J. C. Wang, and A. Mondragón, "Three-dimensional structure of the 67K N-terminal fragment of E. coli DNA topoisomerase I," *Nature*, vol. 367, no. 6459, pp. 138–146, 1994.

[228] J. H. Morais Cabral, A. P. Jackson, C. V. Smith, N. Shikotra, A. Maxwell, and R. C. Liddington, "Crystal structure of the breakage-reunion domain of DNA gyrase," *Nature*, vol. 388, no. 6645, pp. 903–906, 1997.

[229] M. R. Redinbo, L. Stewart, P. Kuhn, J. J. Champoux, and W. G. J. Hol, "Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA," *Science*, vol. 279, no. 5356, pp. 1504–1513, 1998.

[230] L. Stewart, M. R. Redinbo, X. Qiu, W. G. J. Hol, and J. J. Champoux, "A model for the mechanism of human topoisomerase I," *Science*, vol. 279, no. 5356, pp. 1534–1541, 1998.

[231] R. P. Rambo and J. A. Tainer, "Super-resolution in solution x-ray scattering and its applications to structural systems biology," *Annual Review of Biophysics*, vol. 42, pp. 415–441, 2013.

[232] S. Classen, G. L. Hura, J. M. Holton et al., "Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the advanced light source," *Journal of Applied Crystallography*, vol. 46, no. 1, pp. 1–13, 2013.

[233] S. Reindl, A. Ghosh, G. J. Williams et al., "Insights into flaI functions in archaeal motor assembly and motility from structures, conformations, and genetics," *Molecular Cell*, vol. 49, no. 6, pp. 1069–1082, 2013.

[234] A. Ghosh, S. Hartung, C. van der Does, J. A. Tainer, and S.-V. Albers, "Archaeal flagellar ATPase motor shows ATP-dependent hexameric assembly and activity stimulation by specific lipid binding," *Biochemical Journal*, vol. 437, no. 1, pp. 43–52, 2011.

[235] S. L. Bridger, W. A. Lancaster, F. L. Poole et al., "Genome sequencing of a genetically tractable Pyrococcus furiosus strain reveals a highly dynamic genome," *Journal of Bacteriology*, vol. 194, no. 15, pp. 4097–4106, 2012.

[236] A. Cvetkovic, A. L. Menon, M. P. Thorgersen et al., "Microbial metalloproteomes are largely uncharacterized," *Nature*, vol. 466, no. 7307, pp. 779–782, 2010.

[237] R. P. Rambo and J. A. Tainer, "Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law," *Biopolymers*, vol. 95, no. 8, pp. 559–571, 2011.

[238] R. P. Rambo and J. A. Tainer, "Accurate assessment of mass, models and resolution by small-angle scattering," *Nature*, vol. 496, no. 7446, pp. 477–481, 2013.

[239] G. L. Hura, H. Budworth, K. N. Dyer et al., "Comprehensive macromolecular conformations mapped by quantitative SAXS analyses," *Nature Methods*, vol. 10, no. 6, pp. 453–454, 2013.

[240] J. M. Christie, K. Hitomi, A. S. Arvai et al., "Structural tuning of the fluorescent protein iLOV for improved photostability," *The Journal of Biological Chemistry*, vol. 287, no. 26, pp. 22295–22304, 2012.

[241] D. P. Barondeau, C. D. Putnam, C. J. Kassmann, J. A. Tainer, and E. D. Getzoff, "Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12111–12116, 2003.

[242] C. T. McMurray and J. A. Tainer, "Cancer, cadmium and genome integrity," *Nature Genetics*, vol. 34, no. 3, pp. 239–241, 2003.

[243] A. I. Majerník, E. R. Jenkinson, and J. P. Chong, "DNA replication in thermophiles," *Biochemical Society Transactions*, vol. 32, no. 2, pp. 236–239, 2004.

[244] R. W. Kwiatkowski, V. Lyamichev, M. de Arruda et al., "Clinical, genetic, and pharmacogenetic applications of the Invader assay," *Molecular Diagnosis*, vol. 4, no. 4, pp. 353–364, 1999.

[245] R. A. Deshpande, G. J. Williams, O. Limbo et al., "ATP-driven Rad50 conformations regulate DNA tethering, end resection, and ATM checkpoint signaling," *EMBO Journal*, vol. 33, no. 3, pp. 173–276, 2014.

Hindawi Publishing Corporation
Archaea
Volume 2014, Article ID 374146, 11 pages
http://dx.doi.org/10.1155/2014/374146

Research Article

Unique Characteristics of the Pyrrolysine System in the 7th Order of Methanogens: Implications for the Evolution of a Genetic Code Expansion Cassette

Guillaume Borrel,[1,2] Nadia Gaci,[1] Pierre Peyret,[1] Paul W. O'Toole,[2] Simonetta Gribaldo,[3] and Jean-François Brugère[1]

[1] *1EA-4678 CIDAM, Clermont Université, Université d'Auvergne, Place Henri Dunant, 63001 Clermont-Ferrand, France*
[2] *Department of Microbiology and Alimentary Pharmabiotic Centre, University College Cork, Western Road, Cork, Ireland*
[3] *Institut Pasteur, Department of Microbiology, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, 28 rue du Dr. Roux, 75015 Paris, France*

Correspondence should be addressed to Jean-François Brugère; jf.brugere@udamail.fr

Received 30 August 2013; Accepted 19 October 2013; Published 27 January 2014

Academic Editor: Kyung Mo Kim

Copyright © 2014 Guillaume Borrel et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Pyrrolysine (Pyl), the 22nd proteogenic amino acid, was restricted until recently to few organisms. Its translational use necessitates the presence of enzymes for synthesizing it from lysine, a dedicated amber stop codon suppressor tRNA, and a specific aminoacyl tRNA synthetase. The three genomes of the recently proposed Thermoplasmata-related 7th order of methanogens contain the complete genetic set for Pyl synthesis and its translational use. Here, we have analyzed the genomic features of the Pyl-coding system in these three genomes with those previously known from *Bacteria* and *Archaea* and analyzed the phylogeny of each component. This shows unique peculiarities, notably an *amber* $tRNA^{Pyl}$ with an imperfect anticodon stem and a shortened $tRNA^{Pyl}$ synthetase. Phylogenetic analysis indicates that a Pyl-coding system was present in the ancestor of the seventh order of methanogens and appears more closely related to Bacteria than to Methanosarcinaceae, suggesting the involvement of lateral gene transfer in the spreading of pyrrolysine between the two prokaryotic domains. We propose that the Pyl-coding system likely emerged once in Archaea, in a hydrogenotrophic and methanol-H_2-dependent methylotrophic methanogen. The close relationship between methanogenesis and the Pyl system provides a possible example of expansion of a still evolving genetic code, shaped by metabolic requirements.

1. Introduction

Protein synthesis relies on 20 canonical amino acids encoded accordingly to a genetic code, each codon being recognized by an aminoacyl-tRNA. The molecular basis of the genotype to phenotype correspondence relies on the conjunction of tRNAs and of aminoacyl-tRNA synthetases (aaRS). Post-translational modifications of amino acids extend the chemical nature of proteins with functional implications in cellular processes [1, 2]. Another naturally occurring mechanism expands the genetic code to 22 amino acids by adding Selenocysteine (Sec, 2-selenoalanine) [3, 4] and Pyrrolysine (Pyl, 4-methyl-pyrroline-5-carboxylate linked to the $^{\varepsilon}$N of lysine through an amide linkage) [5, 6]. Sec is present in many organisms from the three domains of life [7] with potentially almost a quarter of sequenced bacteria synthesizing it [8]. It is at least present in two orders of Euryarchaeota, the methanogens Methanococcales and Methanopyrales [9, 10]. Its synthesis and incorporation differ from other amino acids, as it is synthesized after a serine has been branched into $tRNA^{Sec}_{UCA}$ (recognizing the *opal* codon UGA) that is next modified into a Sec-$tRNA^{Sec}_{UCA}$.

In contrast, Pyl is restricted to a very small number of organisms and proteins. It necessitates a complex system with specialized enzymes for biosynthesis of Pyl, a dedicated tRNA and an associated unique aaRS [11, 12]: Pyl is first synthesized

as a free amino acid in the cell by the products of the *pylBCD* genes [13, 14] from two lysines, one being methylated into 3-methylornithine (catalyzed by PylB, a lysine mutase-proline-2 methylase) and condensed to the second lysine to form 3-methylornithinyl-N^6-lysine (PylC, Pyrrolysine synthetase) [15]. The pyrrole ring is then formed by oxidation with an atypical dehydrogenase, the Proline reductase (also called Pyl synthase, PylD) [16], with concomitant release of an amino group during the cycle formation. The gene product of *pylT* forms a dedicated tRNA used for the translational incorporation of Pyl. It is an amber decoding $tRNA_{CUA}$ [6], showing some unusual features compared to other tRNAs [11, 17–19]: the anticodon stem is longer (6 nucleotides instead of 5), while other parts are shorter: D-loop with only 5 bases, acceptor, and D-stems separated by one base instead of 2 and a variable loop of only 3 bases. Moreover, differing from almost all tRNAs, the D-loop does not contain the $G_{18}G_{19}$ bases nor the usual $T_{54}\Psi_{55}C_{56}$ bases in the T-loop. Finally, the aaRS encoded by *pylS* catalyzes the ligation of Pyl to its cognate tRNA species warranting the right correspondence between DNA and proteins. The anticodon of the tRNA seems not to be recognized by the PylRS, at least in *D. hafniense* [20]. Whereas the archaeal PylRS is encoded by a single gene (*pylS*), the bacterial PylRS is encoded by two genes, *pylSn* for the first N-terminal 140 amino acids and *pylSc* for the remaining sequence. This pyrrolysyl-tRNA synthetase belongs to the class II aaRS and its structure has been resolved, with and without its substrates/analogs [17, 21–24]. It has likely a homodimeric or homotetrameric quaternary structure.

To date, Pyl-containing proteins have been found only in the methanogens of the family Methanosarcinaceae and in a few bacteria belonging to Firmicutes and two members of deltaproteobacteria, *Bilophila wadsworthia* and an endosymbiont of the worm *Olavius algarvensis* [25]. Pyl is present almost specifically in the methyltransferases (MT) involved in the methanogenesis pathway from monomethyl-, dimethyl-, and trimethylamine in methanogens, respectively, encoded by the genes *mtmB*, *mtbB*, and *mttB* [26, 27] and in their bacterial homologs, whose function is unclear. However, many bacteria harbor *mttB* homologs that lack the in-frame *amber* codon and the Methanosarcinales *Methanococcoides burtonii* has *mttB* genes with and without an in-frame *amber* codon [11, 25]. In contrast, the *mtbB* gene is almost exclusively present in Methanosarcinales and harbors an in-frame *amber* codon. The *mtmB* gene represents an intermediary case, with a few Pyl-containing bacteria possessing it, leading either to a MT with Pyl (*Acetohalobium arabaticum*, *Desulfotomaculum acetoxidans*) or not (*Bilophila wadsworthia*) [25].

The existence of a 7th order of methanogens related to Thermoplasmatales was proposed from molecular data of 16S rRNA and *mcrA* (methanogenesis marker) sequences retrieved from human stools [28, 29]. This is strengthened by growing molecular data from various environments [30, 31] and by phylogenomics studies [32] established from the first genomes of members of this clade [33, 34]. Moreover, a first member of this order, *Methanomassiliicoccus luminyensis*, has been isolated and grows on methanol + H_2 [35]. We identified the presence of genes for methanogenesis from methanol and also from dimethylsulfide and methylamines in the two other sequenced genomes ("*Candidatus* Methanomethylophilus alvus" [33], "*Candidatus* Methanomassiliicoccus intestinalis" [36]), together with the Pyl-coding genes and in-frame *amber* codon in the *mtmB*, *mtbB*, and *mttB* genes of these three species. It has now been shown that the H_2-dependent methanogenesis from trimethylamine is effective for *M. luminyensis* [37]. Also, ruminal methanogens of the same lineage as "*Ca*. M. alvus" (Rumen Cluster C, or RCC), when cultured in consortia with trimethylamine (TMA), show an increase of methanogenesis and mRNAs for MTs, with an *amber* in-frame codon, are detected [38]. Altogether, this greatly suggests the presence of Pyl in these MTs.

The recent uncovered lineage with all the features needed for Pyl encoding and use, representing a new archaeal methanogenic Order, provides an opportunity to better understand the origin, distribution, and diversity of Pyrrolysine-coding systems.

2. Materials and Methods

Genomic sequence data were obtained through GenBank. For the 7th order, Thermoplasmata-related methanogens, the accession numbers are CP004049.1 ("*Ca*. Methanomethylophilus alvus"), CP005934.1 ("*Ca*. Methanomassiliicoccus intestinalis"), and NZ_CAJE0100001 to NZ_CAJE0100026 (*Methanomassiliicoccus luminyensis*). For the genomic organization and comparison of *pyl* genes, genomic sequences were either treated with the RAST annotation server [39] or a local Artemis platform [40]. Blast searches [41] were either performed directly on the RAST server, on the nr database at NCBI or locally. Sequence alignment (DNA, RNA, and proteins) was performed using ClustalW [42], T-coffee [43], and MUSCLE [44]. For RNAs, the dedicated R-Coffee program was also used [45], accessible at http://www.tcoffee.org/, which generates a multiple sequence alignment using structural information. RNAfold [46] (available at http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) and RNAstructure [47] (available at http://rna.urmc.rochester.edu/) were used to determine the secondary structure of tRNA. The structure of the $tRNA^{Pyl}$ was also manually drawn by comparison with the structure of the *D. hafniense* and *M. acetivorans* $tRNA^{Pyl}$ [11, 20]. Phylogenies were both inferred from maximum likelihood and Bayesian procedures. Datasets of protein homologues were aligned by MUSCLE [44] with default parameters, and unambiguously aligned positions were automatically selected by using the BMGE software for multiple-alignment trimming [48] with a BLOSUM30 substitution matrix. Maximum likelihood trees were calculated by PhyML [49] and the LG amino acid substitution model [50] with 4 rate categories as suggested by the AIC criterion implemented in Treefinder [51]. Trees were also calculated by Bayesian analysis with PhyloBayes [52], with the LG model (for single gene trees) or the CAT model (for the concatenated dataset) and 4 categories of evolutionary rates. In this case, two MCMC chains were run in parallel until convergence and the consensus tree was calculated by removing the first 25% of trees as burnin.

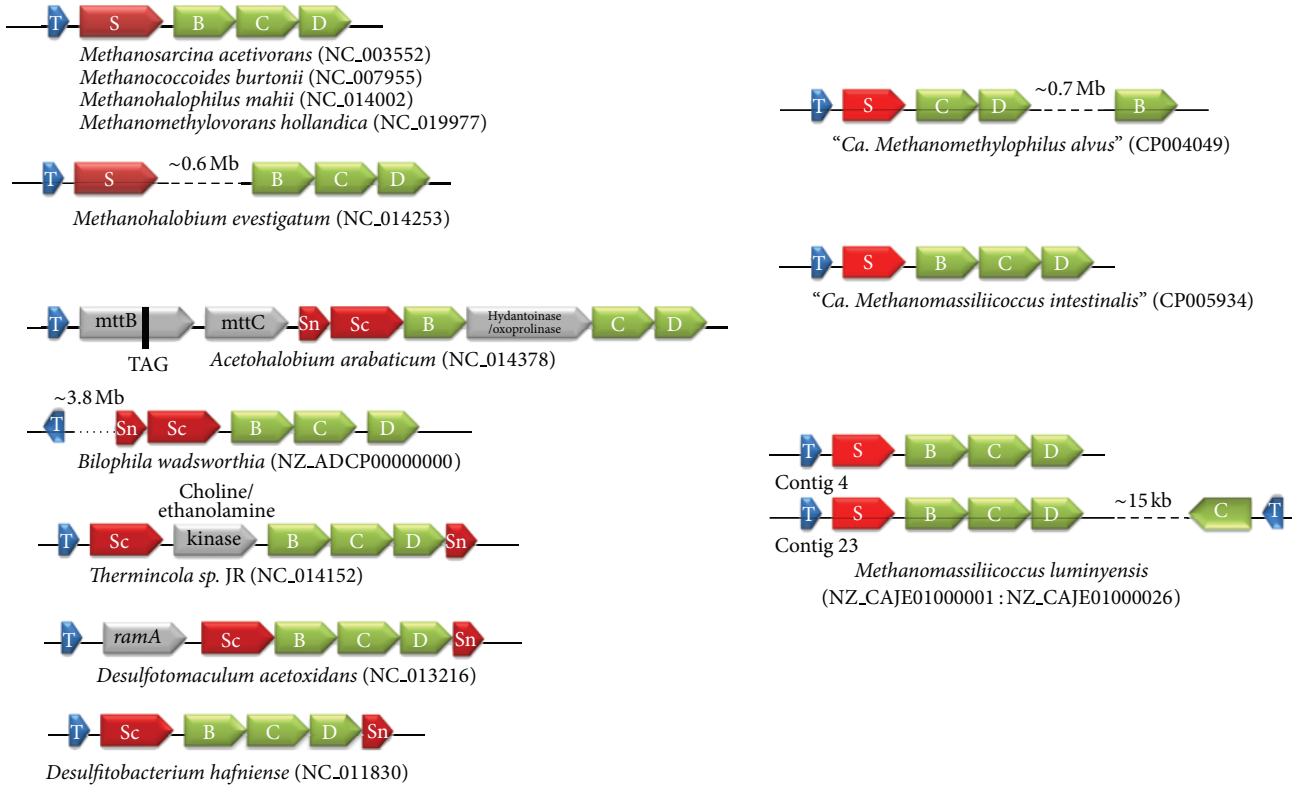


FIGURE 1: Gene organization of the Pyl system. On the left, the gene organization of the *pylBCD* and *pylS* (*pylSc* and *pylSn* in bacteria) is shown for Methanosarcinaceae and some representative bacteria (adapted and updated from [11, 13, 53]). On the right is shown the organization of these genes in the 7th order of methanogens.

3. Results

3.1. Genomic Organization of the pyl Genes. The *pyl* genes usually occur in close association in archaeal and bacterial genomes [11, 53]. In archaeal genomes, they form a cluster *pylTSBCD* that is not interrupted by other genes, except for *Methanohalobium evestigatum* (Figure 1). In bacterial genomes, the *pyl* genes are generally organized as *pylTScBCDSn* or *pylSnScBCD*. Moreover, the cluster can be interrupted by one or several genes. Each of the three genomes affiliated to 7th order of methanogen displays a distinct organization of the *pyl* genes. "*Ca.* M. intestinalis" has a single *pyl* cluster akin to the general organization observed in most of the Methanosarcinales. Two copies of an identically organized cluster are also found in *M. luminyensis* (contigs 4 and 23), together with a third isolated copy of *pylC* and *pylT* present 15 kb away on the complementary strand (contig 23). In the "Ca. M. alvus" genome, the pyl genes occur in single copy and the pylB gene is ~0.7 Mb distant from the pylTSCD cluster (Figure 1).

Usually, these genes are closely associated with the methylamines MT genes (*mtmB*, *mtbB*, and *mttB*) and other genes involved in these pathways (which do not encode Pyl-containing proteins). A similar gathering of the *pyl* genes cluster with the genes involved in methylotrophic methanogenesis, including *mtmB*, *mtbB*, and *mttB* is observed in the genomes of the 7th order of methanogens (data not shown).

3.2. tRNA^{Pyl}. The tRNA^{Pyl} homologues were retrieved from the three 7th order genomes by using the tRNA^{Pyl} of *Methanosarcina barkeri* strain MS-DSM 800 (Accession number AY064401.1) as seed. As mentioned above the "*Ca.* M. alvus" and "*Ca.* M. intestinalis" genomes harbor one *pylT* gene, while three are present in the *M. luminyensis* genome (Figure 1). The third tRNA^{Pyl} of *M. luminyensis* shows no typical stem-loop structures of tRNAs using dedicated bioinformatics tools [54] and is therefore likely a pseudogene. On the contrary, the remaining tRNAs from the three genomes all have a similar shape, different from previously known tRNA^{Pyl} and with stabilities comprised between −28.5 and −23.5 kJ·mol^{-1} (Figure 2). The D-loop already shortened to 5 bases in other tRNA^{Pyl} is here even shorter, with 4 bases in "*Ca.* M. alvus" and in one of the *M. luminyensis* sequences (contig 4, tRNA^{Pyl} no. 1) and even with 3 bases in "*Ca.* M. intestinalis" and in one of the *M. luminyensis* sequences (contig 23, tRNA^{Pyl} no. 2). The acceptor- and D-stems are either not separated ("*Ca.* M. alvus") or separated by one ("*Ca.* M. intestinalis"; tRNA^{Pyl} no. 2 of *M. luminyensis*) or two bases (tRNA^{Pyl} no. 2 of *M. luminyensis*). The variable loop is conserved in all sequences and is equivalent to that of *D. hafniense*, formed of the three bases CAG. In the anticodon loop, the adjacent base of the anticodon CUA is A in "*Ca.* M. alvus" as observed in *D. hafniense* and *M. barkeri* and C in the two other species. However, the most striking feature

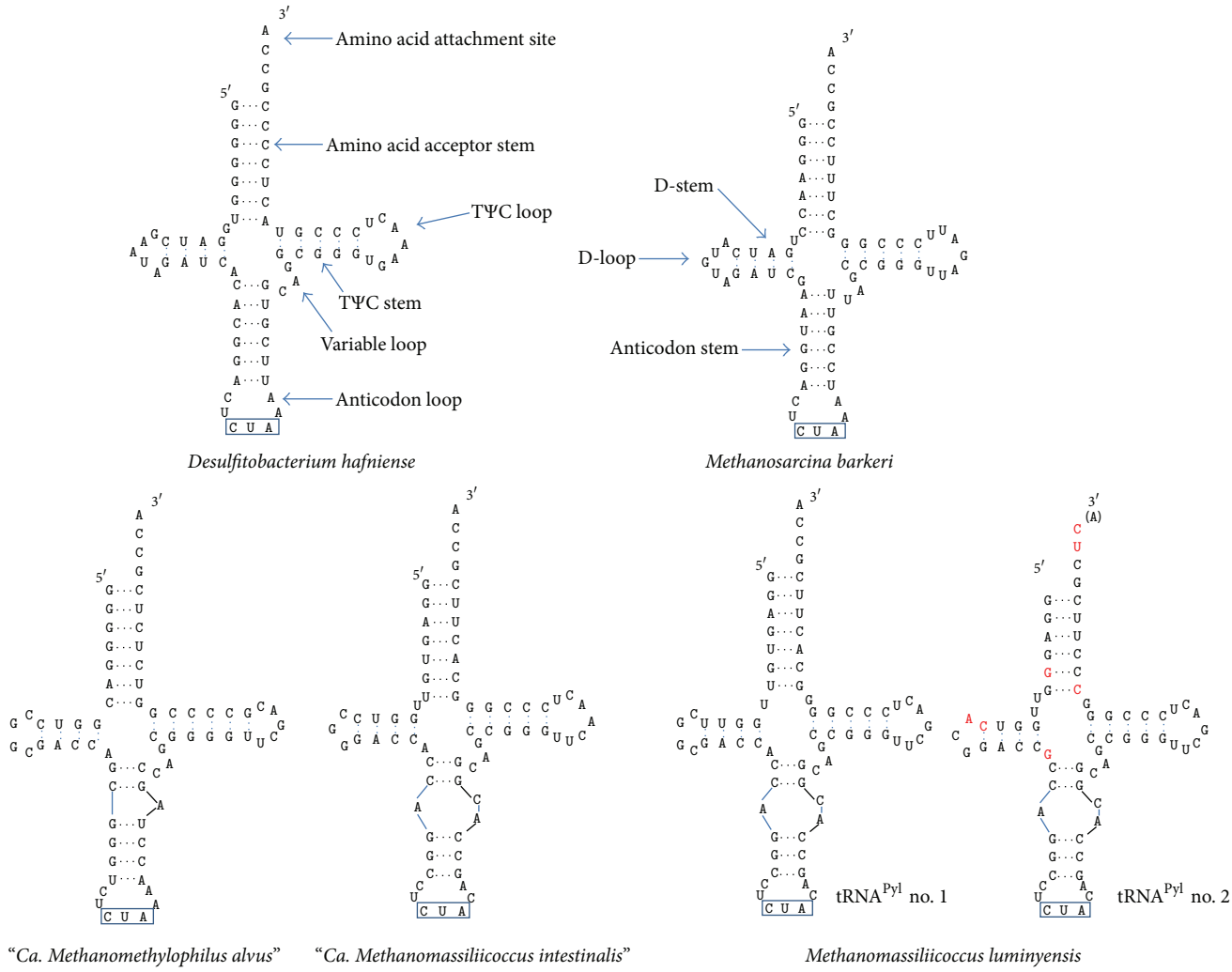


FIGURE 2: Secondary structure of the $tRNA^{Pyl}$ in the Thermoplasmata-related 7th order. The stem-loop structure of the $tRNA^{Pyl}$ in the 7th methanogenic order is shown, in comparison with the structure in bacteria (*Desulfitobacterium hafniense*, left) and in Methanosarcinaceae (*Methanosarcina barkeri*, right). The name of each region of the tRNA is indicated. The anticodon CUA (corresponding to the *amber* codon) is outlined in blue. For the two $tRNA^{Pyl}$ in *M. luminyensis*, the modified bases in no. 2 compared to no. 1 are written in red. Adapted and updated from [11, 13].

is the anticodon stem which is broken in all the $tRNA^{Pyl}$ observed in the 7th order of methanogens. This forms a small loop with a different shape in "*Ca*. M. alvus" and in "*Ca*. M. intestinalis"/*M. luminyensis* (Figure 2).

3.3. pyl-tRNA Synthetase. The *pylS* gene encoding the Pyl-tRNA synthetase PylRS is a class II-aaRS (subclass IIc) [21]. The four homologues present in the three genomes are shortened in their 5′ end compared to their Methanosarcinales counterparts (usually around 420–460 AA) and encode, respectively, 275 amino acids in "*Ca*. M. alvus", "*Ca*. M. intestinalis," and for one of the two PylRS of *M. luminyensis*, and 271 amino acids (or possibly 308 due to uncertainty of the start codon) for the second. This should lead to an N-ter truncated protein with ~140 residues less, similar to the PylSc proteins that are present in bacteria. We could not identify in any of the three complete genomes a gene encoding a protein similar to these ~140 N-terminal residues present in the Methanosarcinaceae PylS or the bacterial PylSn. Moreover, we found no homologue in these genomes of the particular domain TIGR03912 that is present in archaeal PylS and bacterial PylSn proteins and that enhances the interaction of the tRNA synthetase to its specific tRNA [55]. These particular features of the PylRS may be linked to the peculiar characteristics of the $tRNA^{Pyl}$ and suggest a different kind of interaction.

3.4. Phylogenetic Analysis. Phylogenetic analysis was carried out with each of the Pyl gene products by recovering all homologues available in current sequence databases. It appears that bacterial members belong to related Firmicutes species, and a single deltaproteobacterium, *Bilophila wadsworthia*, a common resident of the gut. For PylC, PylD, and PylS, no evident closely related homologues

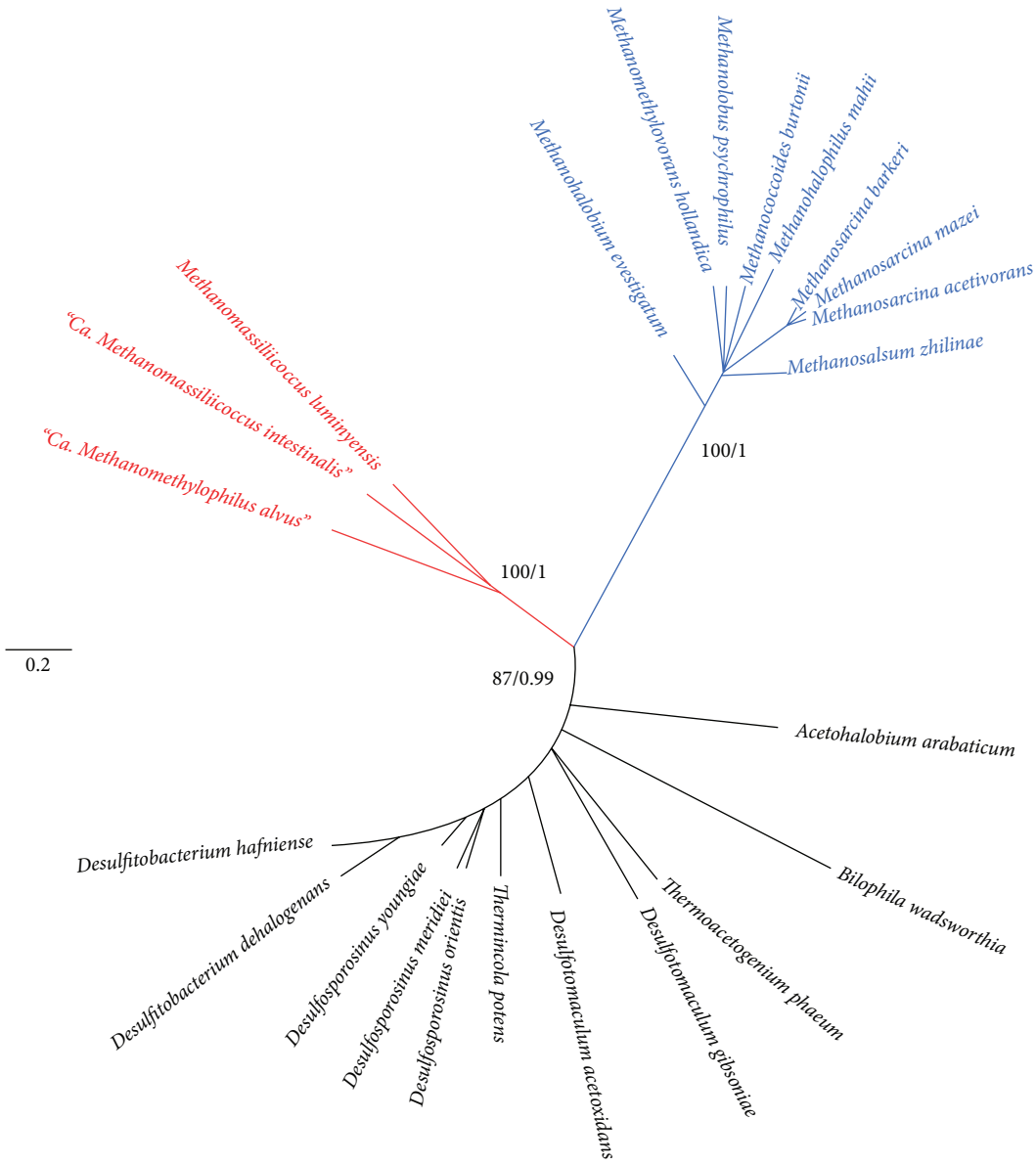


Figure 3: Phylogenetic trees of the *pylSCD* geneproducts. The phylogenetical analysis (PhyML) was performed from a concatenation (672 positions). See Sections 2 and 3 for more details. Individual trees are also reported in the additional file (supplementary Figures S1 to S3 with indication of the corresponding accession numbers). The 7th methanogenic order members are indicated in red, the Methanosarcinaceae in blue, and the bacteria in black.

could be found outside of the known Pyl-containing organisms, whereas PylB had a few distant homologues in various bacteria (not shown). Pyl proteins are very well conserved at the sequence level among *Archaea* and *Bacteria* (alignments are available on request from the authors). The PylC, PylD, and PylS datasets gave consistent results (additional files, Figures S1 to S3 available online at http://dx.doi.org/10.1155/2014/374146) and were therefore concatenated to provide increased phylogenetic signal (Figure 3). "Ca. M. alvus", "Ca. M. intestinalis," and "*M. luminyensis*" form a robustly supported monophyletic cluster, indicating a common origin of the Pyl system in the 7th methanogenic order. The products of the two *pyl* gene clusters of *M. luminyensis* (including the third extra PylC) are more closely related among them than to the other sequences, indicating that they originated from a specific duplication. Methanosarcinaceae and bacterial sequences also form two distinct and robustly supported monophyletic clusters. Moreover, the sequences from the 7th methanogenic order appear to be more closely related to their bacterial counterparts than to Methanosarcinaceae (Figure 3).

Concerning PylB, phylogenetic analyses gave more ambiguous results, with a poorly supported branching of the 7th methanogenic order within Bacteria, and different inferred evolutionary relationships according to exclusion or inclusion of the distant bacterial outgroup (data not shown).

An unrooted PhyML phylogenetic tree inferred exclusively from PylB (without the distant bacterial outgroup) is given in the additional file Figure S4.

4. Discussion

We have shown here that all the genetic components for Pyl synthesis and use are present in a new order of Archaea performing methylotrophic methanogenesis from methanol and methylamines, therefore enlarging the taxonomic distribution of this genetic code expansion cassette. Moreover, this Pyl system has unique features and appears more closely related to the bacterial one than to those present in Methanosarcinaceae. Whether or not the Pyl system is active in these archaeons is beyond the scope of this paper. However, there are many arguments supporting a functional system. These include, for the three genomes available and analyzed to date, at least the global presence of all the genes for Pyl synthesis, charging and cotranslational incorporation, in cooccurrence with all the genes for methylotrophic methanogenesis from methylamines, that is, those coding for methylamines-corrinoid protein methyltransferases (MT) genes (*mtmB, mtbB*, and *mttB*) all bearing an in-frame *amber* codon. It is interesting to note that, as a nonsense (stop) codon, the *amber* codon is used with a low frequency, this value being maximal in the *M. luminyensis* genome, with only 11% of stop codon being an amber one. It is likely an adaptive strategy to the presence of the Pyl system in this order, similarly to what is observed in Methanosarcinaceae, but different from Bacteria: these lasts deal with the presence of Pyl through regulation rather than codon avoidance [25]. Ability to metabolize TMA into methane has been shown for *M. luminyensis* [37]. Moreover, it has also been shown that methylotrophic methanogenesis from methylamines is active in the rumen, very likely carried out by Rumen Cluster C, members of the 7th order of methanogens which are close neighbors or similar to "*Ca*. M. alvus" [38]. This is therefore conceivable that methylotrophic methanogenesis is constitutive of this whole order by dedicated MTs containing Pyl. The presence of unique features in the Pyl system in these archaeons raises the exciting question of how it functions. PylRS harbors a shorter form, more closely related to the product of the bacterial pylSc gene than to the C-terminus of the PylS present in Methanosarcinaceae, and lacks the ~140 amino acids equivalent to the bacterial *pylSn* or to the N-terminal of the archaeal PylS. Bacterial PylSn and the Nter part of Methanosarcinaceae PylRS contain the protein domain TIGR03912 [55]. In the three genomes of the 7th order of methanogens, no CDS containing this peculiar domain was found by an InterProScan analysis [56]. The lack of these elements in the $tRNA^{Pyl}$ synthetase (PylSn-TIGR03912 domain) indicate that PylRS must solely rely on a homomeric structure and might be able to fulfill this role without the domain corresponding to PylSn. Indeed, it has been shown that the *D. hafniense* PylSc is alone sufficient to charge a cognate $tRNA^{Pyl}$ with an analogue of pyrrolysine in a recombinant system [20]. Alternatively, it may rely on the presence of a yet unknown enhancing component. Also exciting are the sequence and inferred structure of the $tRNA^{Pyl}$: its extremely condensed nature, with a peculiar broken anticodon stem is of concern for future functional and evolutionary studies, together with the molecular mechanisms which sustain this orthogonal code between the $tRNA^{Pyl}$ and the PylRS. The peculiar structure of the $tRNA^{Pyl}$ may strengthen the interaction with its cognate PylRS and be an adaptation following the loss of the PylSn domain. Alternatively, mutations in the sequence of the $tRNA^{Pyl}$ may have led to the loss of the PylSn. However, it is also possible that this domain was never present in the Pyl system of these archaeons. In any case, active or not, the well-supported monophyly of the Pyl encoding genes in the three genomes argue for their presence in their common ancestor. Because these represent at least two distinct clades [32], this indicates the likely presence of a Pyl system early in the 7th order of methanogens or even in the ancestor of the whole order. The same conclusion stands also for the Methanosarcinaceae.

How the 22nd proteogenic amino acid Pyrrolysine was added to the genetic code is an exciting question. Based on experimental data, it has been proposed that the universal codon catalog was anterior to the aminoacylation systems [57]. 3D structure-based phylogeny of PylRS has led to the suggestion that it was already present in the last universal common ancestor (LUCA) of all present-day life beings and has a common origin with the phenylalanine tRNA synthetase PheRS [21]. If true, this would mean that Pyl was inherited from the LUCA and retained only in the few present-day methylamines-utilizing bacterial and archaeal species. Similarly, Fournier and colleagues proposed that Pyl originated in a pre-LUCA lineage representing a yet unidentified or more probably extinct 4th domain of life [58–60]. In this hypothesis, Pyl would have been acquired in a few *Bacteria* and *Archaea* via several independent lateral gene transfers (LGT) from members of this hypothetical 4th domain of life. However, its highly specific function and restricted distribution make it more likely that the addition of Pyl is as a more recent event.

The strong link of the Pyl system with methanogenesis, especially the methylotrophic one, argues for a common evolutionary history. Pyl is now essential for methylotrophic methanogenesis from methylamines, while methylotrophic methanogenesis from methanol is independent of Pyl. Pyl has been found in *Archaea* almost exclusively in the methyltransferases involved in methanogenesis from mono-, di-, and tri-methylamines (MtmB, MtbB, and MttB). Moreover, *pyl* genes and *mtm-mtb-mtt* genes are clustered in the genomes of methylamines-utilizing archaeal *species*; therefore, a stimulating hypothesis is that genes necessary for Pyl synthesis/incorporation and for methylotrophic methanogenesis from methylamines have coevolved, whereas this was not the case for methylotrophic methanogenesis from methanol, which is independent of pyrrolysine-containing enzymes.

Methanogenesis is believed to have arisen early in the evolution of Euryarchaeota, likely after the divergence of Thermococcales, and to have been lost several times independently in various lineages (e.g., Halobacteriales,

Archaeoglobales, and Thermoplasmatales) [32, 61, 62]. The first form of methanogenesis might have been the hydrogenotrophic one (present in the class I methanogens composed of Methanopyrales, Methanobacteriales, and Methanococcales), while other types of methanogenesis (methylotrophic non-H_2-dependent and acetoclastic, involving cytochromes) would have emerged later. Based on phylogenomic studies, we have recently shown that the pathway for methylotrophic methanogenesis, at least from methanol, was already present in the ancestor of Methanobacteriales, in the ancestor of the 7th order, and in the ancestor of Methanosarcinales, which is the common ancestor of most euryarchaeal lineages excluding Thermococcales [32]. Therefore, H_2-dependent methylotrophic methanogenesis from methanol may also be an ancient type of methanogenesis. The fact that it does not involve cytochromes and needs fewer genes could favor such hypothesis. The question is, can the same conclusions be made for the methylotrophic methanogenesis from methylamines? The fact that this methanogenesis is presently restricted to Methanosarcinaceae and the 7th order of methanogens, which emerged after class I methanogens, argues for a more recent event. It has also to be stressed that no *mta* gene (involved in methylotrophic methanogenesis from methanol) harbors an *amber*-Pyl codon. Moreover, the last MT enzyme of the methanogenesis pathway (encoded by *mtaA* gene for methanol) that precedes the steps corresponding to core methanogenesis (methyl-coM reduction to generate methane) is involved in methanogenesis from both methanol and some methylamines in Methanosarcinaceae and likely also in the 7th order (data not shown). Therefore, it is tempting to speculate that the Pyl-independent, H_2-dependent methanogenesis from methanol preceded the Pyl-dependent, H_2-dependent methylotrophic methanogenesis from methylamines.

Taking together these pieces of evidence, it is likely that, in the archaeal domain, the Pyl system arose in a methanogenic euryarchaeote, in correspondence to the emergence of the genes coding for methylamines-corrinoid protein methyltransferases (MT) (*mtmB*, *mtbB*, and *mttB*) that all bear an in-frame *amber* codon, allowing the use of new substrates (mono-, di-, and trimethylamine, resp.). We will refer to the archaeon where the system arose as the archaeal Pyrrolysine ancestor (APA). The APA may correspond to the ancestor of Methanosarcinaceae, the ancestor of the 7th order of methanogens, or the common ancestor of both (depicted on Figure 4). Now, it can be asked how the Pyl system arose in this hypothetical euryarchaeote. Two possibilities can be envisaged: (i) it was acquired *via* a lateral gene transfer (LGT) or (ii) it emerged autogenously.

Concerning the hypothesis of an LGT from a 4th domain of life by Fournier and colleagues [58, 59], in the light of our results it may be reformulated by positing a single LGT from this extinct lineage into Archaea after the divergence of Thermococcales to give birth to the APA (Figure 4, dotted green arrow). The system would have then been retained only in present-day pyl-containing organisms, for example, the 7th order and Methanosarcinaceae, whereas it would have been lost multiple times independently in most euryarchaeal lineages, including methanogenic ones (Figure 4, red dots). However, the hypothetical existence of a pre-LUCA lineage that gave rise to the Pyl system remains questionable because of the necessary coexistence of the donor and the APA and because the LGT would have taken place in a euryarchaeon, relatively recent and distant from LUCA, the 4th domain would have been composed at that time of many different lineages that would have subsequently all disappeared (broken dotted lines in the hypothetical 4th Domain, Figure 4).

Our data also make the hypothesis that the Pyl system arose in Bacteria and was introduced in the APA via LGT less likely (Figure 4). In fact, its presence in the 7th order now makes its taxonomic distribution much larger in Archaeathan the few pyl-containing bacteria representing a sublineage of Firmicutes and two deltaproteobacteria and its link to methylotrophic methanogenesis stronger.

Therefore, our preferred hypothesis is that the Pyl system is an archaeal invention (green box, Figure 4). PylRS might have emerged by gene duplication followed by fast evolutionary rates from another class II RS gene, while a mutation in the anticodon of a duplicated tRNA could have led to an *amber* decoding tRNA. This would have had less detrimental effects for the cell than the recoding of another codon, affecting potentially all the proteins of the cell. Interestingly, PylRS does not need to recognize the anticodon of the $tRNA^{Pyl}$ [17, 55], and this may be a remnant of the birth of the orthogonal pair $tRNA^{Pyl}$/PylRS. Moreover, the recent discovery of the whole synthesis pathway of Pyl, with PylB being a lysine mutase [13] shows that Pyl is entirely a derivative of a proteogenic amino acid (two lysines) and this could make sense in the light of the coevolution theory, such are the cases of Asp/Asn and Glu/Gln in Archaea[63]. APA may have been an early diverging euryarchaeote performing H_2-dependent methanogenesis from methanol, perhaps the ancestor of Methanosarcinales and the 7th order of methanogens and therefore after the emergence of Thermococcales and methanogens class I (Methanococcales, Methanobacteriales, and Methanopyrales). Vertical inheritance of the Pyl system would have paralleled that of H_2-dependent methanogenesis from methylamines whereas multiple independent losses would have occurred over time in most euryarchaeal lineages, including methanogenic ones (Figure 4, red dots), but retained only in the Methanosarcinaceae and the 7th order. Alternatively, the system would have emerged later, in either the ancestor of Methanosarcinaceae or in the ancestor of the 7th order and then spread among them via LGT (not indicated in the Figure 4 for clarity). Because their Pyl systems are different (notably the absence of the N-terminus of PylS and the unique peculiarity of the $tRNA^{Pyl}$), it may be asked which on the two is the ancestral one. However, it is likely that the unique system of the 7th order is derived from a more "classical" one such as those present in bacteria and Methanosarcinaceae.

Under the hypothesis of an archaeal invention of the Pyl system, the bacterial one may have arisen via LGT from the 7th order of methanogens, considering their closer evolutionary relations, before the loss of the N-terminus of PylS (Figure 4, PylS structure indicated above the Pyl-containing groups). It is unclear if one or several LGTs

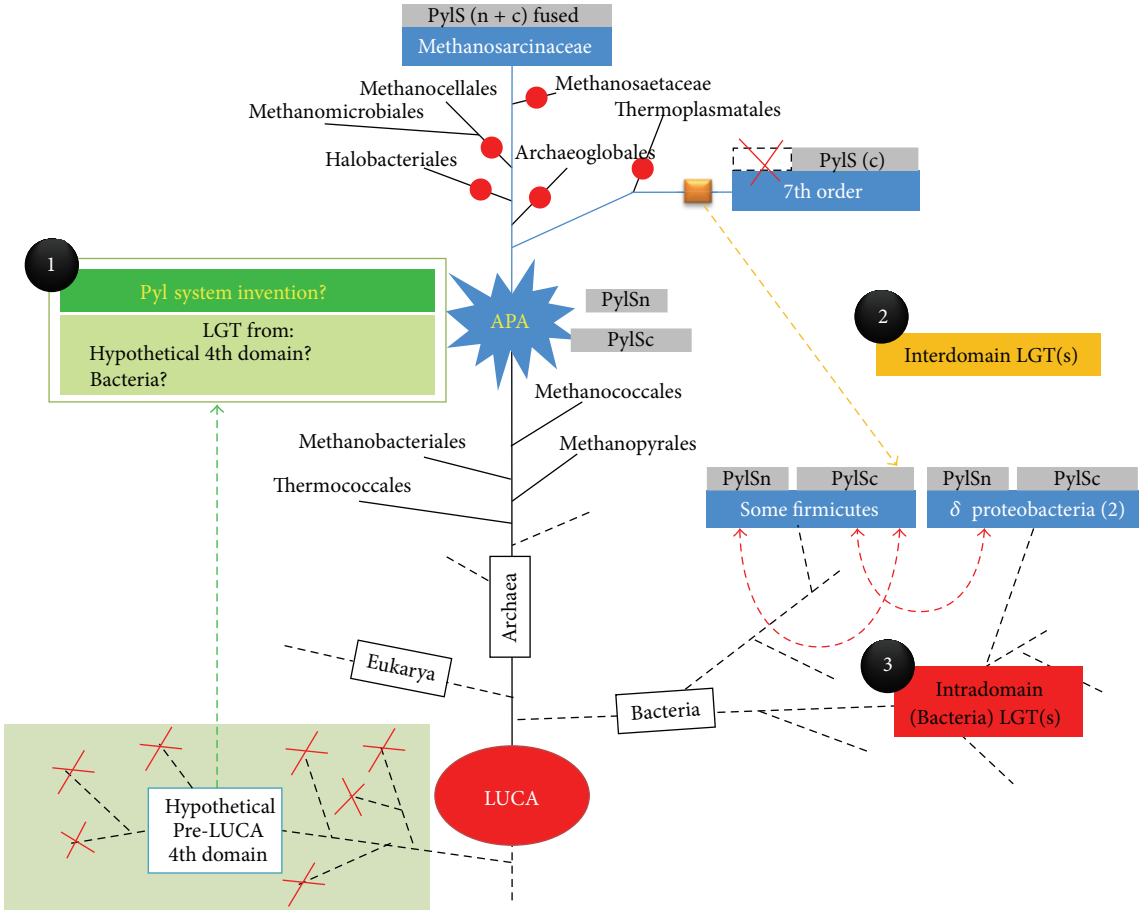


Figure 4: Proposed scenarios for the origin and evolution of the pyrrolysine system. One of the various possible models leading to the genesis of APA, the ancestral Pyl-coding archaeon, is schematized. This model supports less LGTs than previously proposed ones. APA (blue star) emerged in the Euryarchaeota (black circle no. 1) after the birth of the methanogenesis and was an ancestor of the Methanosarcinalesand/or of the Thermoplasmata-related 7th order of methanogens. Only the hypothesis of a common ancestor of both Methanosarcinalesand Thermoplasmata-related 7th order of methanogens is depicted here (see text for an alternative model with APA as a more recent ancestor, of the Methanosarcinalesor of the Thermoplasmata-related 7th order of methanogens, but not both). The putative genesis of APA (green boxes) relies either from an archaeal invention (Box 1), from a bacterial contribution, or from a whole functional system acquired from a hypothetical pre-LUCA 4th domain of life. APA was therefore likely a methanogenic archaeon performing both a methanol-H_2-dependent methylotrophic methanogenesis and a hydrogenotrophic one. The Pyl system was vertically inherited (symbolized by large blue lines) to Methanosarcinaceae and to the 7th methanogenic order and was lost several times independently on various branches, including methanogenic ones (red dots). Due to the closest relationship of the Pyl system (orange square) between the 7th order and bacteria, at least one LGT is supposed to have occurred from a 7th order ancestor (before the loss of the *pylSc* gene in this lineage), likely to a Firmicute(no. 2, orange arrow). Other LGTs from the 7th order to Bacteria (i.e., a deltaproteobacterium) are also conceivable. In a third period, intradomain LGT(s) could have originated next putatively among Firmicute and/or into a deltaproteobacterium (no. 3, red arrows). Pyl-coding organisms are symbolized in blue boxes. The nature of the PylRS is depicted with each Pyl-containing group in grey boxes and is either formed by a unique PylS (Methanosarcinaceae), a split complete version (PylSc//PylSn in bacteria), or a unique PylSc form (7th methanogenic order).

occurred from the 7th order onto bacteria (Firmicutes and deltaproteobacteria). The group of Firmicutes is likely the recipient of this first transfer and would have then given the system to a deltaproteobacterium. Alternatively, due to their close environmental relationship (the gut), a distinct LGT could have arisen from a 7th order member of the gut into *Bilophila sp*. However, it cannot be excluded that bacteria took it from Methanosarcinaceae and split the PylS gene into PylSn and PylSc and then they later passed only PylSc to the 7th order. The tree of PylB may suggest such hypothesis, albeit statistical support is not very strong and no similar pattern is observed for the other components of the system. Unfortunately, phylogenetic analysis prevents concluding the direction of these transfers, because of the absence of outgroup sequences. In every case, that these inter- and/or intradomain LGTs were not rejected is a fascinating event. Introducing a stop codon suppressor is in fact likely deleterious and has to be compensated by a strong selective advantage for the organism, such as the use of new metabolic substrates (e.g., methylamines). Introducing a Pyl-coding cassette into a heterologous genome has been successfully experimentally realized [14, 64]. It is possible that

the acquired genes remained silent or with a low expression level and would have been activated progressively, leading concomitantly to the negative selection of UAG nonsense codons. A more likely hypothesis is an LGT leading to an inducible expression of the Pyl components, such as dependence of the presence of substrates like methylamines. The natural expansion of the genetic code in the Firmicute *Acetohalobium arabaticum* able to genetically encode the 20 usual amino acids when grown on pyruvate, and to expand its repertoire to 21 by adding pyrrolysine when grown on TMA [25] provides the paradigm.

In conclusion, when considering at least the archaeal domain, there has been a Pyl-coding ancestor. It appeared likely relatively recently, as an ancestor of the Methanosarcinales, an ancestor of the 7th order of methanogens, or the common ancestor of both. It was likely a methanogen performing a methanol/H_2-dependent methanogenesis and, considering the probable coevolution history between methylotrophic methanogenesis and Pyl, with their strong interdependence, it was concomitant or rapidly followed by the emergence of methanogenesis from methylamines compounds. This has led nowadays in the archaeal lineage to conservation of a Pyl system only in methylamines-utilizing/Pyl-dependent methanogens. Therefore, this provides an example that the genetic code may be still under evolution with a conceivable expansion shaped by metabolic requirements.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was funded by PhD Scholarship supports from the French "Ministère de l'Enseignement Supérieur et de la Recherche" to Nadia Gaci Simonetta Gribaldo was supported by the French Agence Nationale de la Recherche (ANR) through Grant ANR-10-BINF-01-01 "Ancestrome." Paul W. O'Toole was supported by Science Foundation Ireland through a Principal Investigator award and by an FHRI award to the ELDERMET project by the Department of Agriculture, Fisheries and Marine of the Government of Ireland.

References

[1] A. P. Lothrop, M. P. Torres, and S. M. Fuchs, "Deciphering post-translational modification codes," *FEBS Letters*, vol. 587, pp. 1247–1257, 2013.

[2] F. Wold, "In vivo chemical modification of proteins (post-translational modification)," *Annual Review of Biochemistry*, vol. 50, pp. 783–814, 1981.

[3] A. Bock and T. C. Stadtman, "Selenocysteine, a highly specific component of certain enzymes, is incorporated by a UGA-directed co-translational mechanism," *BioFactors*, vol. 1, no. 3, pp. 245–250, 1988.

[4] T. C. Stadtman, "Selenocysteine," *Annual Review of Biochemistry*, vol. 65, pp. 83–100, 1996.

[5] M. Ibba and D. Söll, "Genetic code: introducing pyrrolysine," *Current Biology*, vol. 12, no. 13, pp. R464–R466, 2002.

[6] G. Srinivasan, C. M. James, and J. A. Krzycki, "Pyrrolysine encoded by UAG in archaea: charging of a UAG-decoding specialized tRNA," *Science*, vol. 296, no. 5572, pp. 1459–1462, 2002.

[7] A. Bock, K. Forchhammer, J. Heider et al., "Selenocysteine: the 21st amino acid," *Molecular Microbiology*, vol. 5, no. 3, pp. 515–520, 1991.

[8] Y. Zhang, H. Romero, G. Salinas, and V. N. Gladyshev, "Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues," *Genome Biology*, vol. 7, no. 10, article R94, 2006.

[9] M. Rother, A. Resch, R. Wilting, and A. Böck, "Selenoprotein synthesis in archaea," *BioFactors*, vol. 14, no. 1–4, pp. 75–83, 2001.

[10] T. Stock and M. Rother, "Selenoproteins in Archaea and Gram-positive bacteria," *Biochimica et Biophysica Acta*, vol. 1790, no. 11, pp. 1520–1532, 2009.

[11] M. A. Gaston, R. Jiang, and J. A. Krzycki, "Functional context, biosynthesis, and genetic encoding of pyrrolysine," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 342–349, 2011.

[12] M. Rother and J. A. Krzycki, "Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea," *Archaea*, vol. 2010, Article ID 453642, 14 pages, 2010.

[13] M. A. Gaston, L. Zhang, K. B. Green-Church, and J. A. Krzycki, "The complete biosynthesis of the genetically encoded amino acid pyrrolysine from lysine," *Nature*, vol. 471, no. 7340, pp. 647–650, 2011.

[14] D. G. Longstaff, R. C. Larue, J. E. Faust et al., "A natural genetic code expansion cassette enables transmissible biosynthesis and genetic encoding of pyrrolysine," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 3, pp. 1021–1026, 2007.

[15] J. A. Krzycki, "The path of lysine to pyrrolysine," *Current Opinion in Chemical Biology*, vol. 17, no. 4, pp. 619–625, 2013.

[16] F. Quitterer, P. Beck, A. Bacher, and M. Groll, "Structure and reaction mechanism of pyrrolysine synthase (PylD)," *Angewandte Chemie International Edition*, vol. 52, no. 27, pp. 7033–7037, 2013.

[17] K. Nozawa, P. O'Donoghue, S. Gundllapalli et al., "Pyrrolysyl-tRNA synthetase-tRNAPyl structure reveals the molecular basis of orthogonality," *Nature*, vol. 457, no. 7233, pp. 1163–1167, 2009.

[18] C. Polycarpo, A. Ambrogelly, A. Bérubé et al., "An aminoacyl-tRNA synthetase that specifically activates pyrrolysine," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 34, pp. 12450–12454, 2004.

[19] A. Théobald-Dietrich, M. Frugier, R. Giegé, and J. Rudinger-Thirion, "Atypical archaeal tRNA pyrrolysine transcript behaves towards EF-Tu as a typical elongator tRNA," *Nucleic Acids Research*, vol. 32, no. 3, pp. 1091–1096, 2004.

[20] S. Herring, A. Ambrogelly, C. R. Polycarpo, and D. Söll, "Recognition of pyrrolysine tRNA by the Desulfitobacterium hafniense pyrrolysyl-tRNA synthetase," *Nucleic Acids Research*, vol. 35, no. 4, pp. 1270–1278, 2007.

[21] J. M. Kavran, S. Gundllapalli, P. O'Donoghue, M. Englert, D. Söll, and T. A. Steitz, "Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 27, pp. 11268–11273, 2007.

[22] M. M. Lee, R. Jiang, R. Jain, R. C. Larue, J. Krzycki, and M. K. Chan, "Structure of Desulfitobacterium hafniense PylSc, a pyrrolysyl-tRNA synthetase," *Biochemical and Biophysical Research Communications*, vol. 374, no. 3, pp. 470–474, 2008.

[23] T. Yanagisawa, R. Ishii, R. Fukunaga, T. Kobayashi, K. Sakamoto, and S. Yokoyama, "Crystallographic studies on multiple conformational states of active-site loops in pyrrolysyl-tRNA synthetase," *Journal of Molecular Biology*, vol. 378, no. 3, pp. 634–652, 2008.

[24] T. Yanagisawa, T. Sumida, R. Ishii, and S. Yokoyama, "A novel crystal form of pyrrolysyl-tRNA synthetase reveals the pre- and post-aminoacyl-tRNA synthesis conformational states of the adenylate and aminoacyl moieties and an asparagine residue in the catalytic site," *Acta Crystallographica D*, vol. 69, pp. 5–15, 2013.

[25] L. Prat, I. U. Heinemann, H. R. Aerni, J. Rinehart, P. O'Donoghue, and D. Söll, "Carbon source-dependent expansion of the genetic code in bacteria," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 21070–21075, 2012.

[26] L. Paul, D. J. Ferguson Jr., and J. A. Krzycki, "The trimethylamine methyltransferase gene and multiple dimethylamine methyltransferase genes of Methanosarcina barkeri contain in-frame and read- through amber codons," *Journal of Bacteriology*, vol. 182, no. 9, pp. 2520–2529, 2000.

[27] J. A. Soares, L. Zhang, R. L. Pitsch et al., "The residue mass of L-pyrrolysine in three distinct methylamine methyltransferases," *The Journal of Biological Chemistry*, vol. 280, no. 44, pp. 36962–36969, 2005.

[28] A. Mihajlovski, M. Alric, and J.-F. Brugère, "A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of the mcrA gene," *Research in Microbiology*, vol. 159, no. 7-8, pp. 516–521, 2008.

[29] A. Mihajlovski, J. Doré, F. Levenez, M. Alric, and J.-F. Brugère, "Molecular evaluation of the human gut methanogenic archaeal microbiota reveals an age-associated increase of the diversity," *Environmental Microbiology Reports*, vol. 2, no. 2, pp. 272–280, 2010.

[30] T. Iino, H. Tamaki, S. Tamazawa et al., "Candidatus Methanogranum caenicola: a novel methanogen from the anaerobic digested sludge, and proposal of Methanomassiliicoccaceae fam. nov. and Methanomassiliicoccales ord. nov., for a methanogenic lineage of the class Thermoplasmata," *Microbes and Environments*, vol. 28, pp. 244–250, 2013.

[31] K. Paul, J. O. Nonoh, L. Mikulski, and A. Brune, "'Methanoplasmatales,' Thermoplasmatales-related archaea in termite guts and other environments, are the seventh order of methanogens," *Applied and Environmental Microbiology*, vol. 78, pp. 8245–8253, 2012.

[32] G. Borrel, P. W. O'Toole, P. Peyret, J. F. Brugère, and S. Gribaldo, "Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis," *Genome Biology and Evolution*, vol. 5, pp. 1769–1780, 2013.

[33] G. Borrel, H. M. Harris, W. Tottey et al., "Genome sequence of, "Candidatus Methanomethylophilus alvus" Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of methanogens," *Journal of Bacteriology*, vol. 194, pp. 6944–6945, 2012.

[34] A. Gorlas, C. Robert, G. Gimenez, M. Drancourt, and D. Raoult, "Complete genome sequence of Methanomassiliicoccus luminyensis, the largest genome of a human-associated Archaea species," *Journal of Bacteriology*, vol. 194, no. 17, p. 4745, 2012.

[35] B. Dridi, M. L. Fardeau, B. Ollivier, D. Raoult, and M. Drancourt, "Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces," *International Journal of Systematic and Evolutionary Microbiology*, vol. 62, pp. 1902–1907, 2012.

[36] G. Borrel, H. M. Harris, N. Parisot et al., "Genome sequence of "*Candidatus* methanomassiliicoccus intestinalis" Issoire-Mx1, a third *Thermoplasmatales*-related methanogenic archaeon from human feces," *Genome Announcements*, vol. 1, no. 4, Article ID e00453, 13 pages, 2013.

[37] J. F. Brugère, G. Borrel, N. Gaci, W. Tottey, P. W. O. 'Toole, and C. Malpuech-Brugère, "Archaebiotics: proposed therapeutic use of archaea to prevent trimethylaminuria and cardiovascular disease," *Gut Microbes*, vol. 5, no. 1, 2013.

[38] M. Poulsen, C. Schwab, B. B. Jensen et al., "Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen," *Nature Communications*, vol. 4, article 1428, 2013.

[39] R. K. Aziz, D. Bartels, A. Best et al., "The RAST server: rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, article 75, 2008.

[40] K. Rutherford, J. Parkhill, J. Crook et al., "Artemis: sequence visualization and annotation," *Bioinformatics*, vol. 16, no. 10, pp. 944–945, 2000.

[41] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[42] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[43] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

[44] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[45] A. Wilm, D. G. Higgins, and C. Notredame, "R-Coffee: a method for multiple alignment of non-coding RNA," *Nucleic Acids Research*, vol. 36, no. 9, article e52, 2008.

[46] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker, "The Vienna RNA websuite," *Nucleic Acids Research*, vol. 36, pp. W70–W74, 2008.

[47] J. S. Reuter and D. H. Mathews, "RNAstructure: software for RNA secondary structure prediction and analysis," *BMC Bioinformatics*, vol. 11, article 129, 2010.

[48] A. Criscuolo and S. Gribaldo, "BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments," *BMC Evolutionary Biology*, vol. 10, no. 1, article 210, 2010.

[49] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0," *Systematic Biology*, vol. 59, no. 3, pp. 307–321, 2010.

[50] S. Q. Le and O. Gascuel, "An improved general amino acid replacement matrix," *Molecular Biology and Evolution*, vol. 25, no. 7, pp. 1307–1320, 2008.

[51] G. Jobb, A. von Haeseler, and K. Strimmer, "TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics," *BMC Evolutionary Biology*, vol. 4, article 18, 2004.

[52] N. Lartillot, T. Lepage, and S. Blanquart, "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating," *Bioinformatics*, vol. 25, no. 17, pp. 2286–2288, 2009.

[53] J. Yuan, P. O'Donoghue, A. Ambrogelly et al., "Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems," *FEBS Letters*, vol. 584, no. 2, pp. 342–349, 2010.

[54] I. L. Hofacker, "RNA secondary structure analysis using the Vienna RNA package," *Current Protocols in Bioinformatics*, 2004, Chapter 12, Unit 12.2.

[55] R. Jiang and J. A. Krzycki, "PylSn and the homologous N-terminal domain of pyrrolysyl-tRNA synthetase bind the tRNA that is essential for the genetic encoding of pyrrolysine," *The Journal of Biological Chemistry*, vol. 287, pp. 32738–32746, 2012.

[56] E. M. Zdobnov and R. Apweiler, "InterProScan—an integration platform for the signature-recognition methods in InterPro," *Bioinformatics*, vol. 17, no. 9, pp. 847–848, 2001.

[57] M. J. Hohn, H.-S. Park, P. O'Donoghue, M. Schnitzbauer, and D. Söll, "Emergence of the universal genetic code imprinted in an RNA record," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 48, pp. 18095–18100, 2006.

[58] G. Fournier, "Horizontal gene transfer and the evolution of methanogenic pathways," *Methods in Molecular Biology*, vol. 532, pp. 163–179, 2009.

[59] G. P. Fournier, J. Huang, and J. Peter Gogarten, "Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life," *Philosophical Transactions of the Royal Society B*, vol. 364, no. 1527, pp. 2229–2239, 2009.

[60] A. V. Lobanov, A. A. Turanov, D. L. Hatfield, and V. N. Gladyshev, "Dual functions of codons in the genetic code," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 45, no. 4, pp. 257–265, 2010.

[61] É. Bapteste, C. Brochier, and Y. Boucher, "Higher-level classification of the Archaea: evolution of methanogenesis and methanogens," *Archaea*, vol. 1, no. 5, pp. 353–363, 2005.

[62] S. Gribaldo and C. Brochier-Armanet, "The origin and evolution of Archaea: a state of the art," *Philosophical Transactions of the Royal Society B*, vol. 361, no. 1470, pp. 1007–1022, 2006.

[63] J. T.-F. Wong, "Coevolutlon theory of genetic code at age thirty," *BioEssays*, vol. 27, no. 4, pp. 416–425, 2005.

[64] D. G. Longstaff, S. K. Blight, L. Zhang, K. B. Green-Church, and J. A. Krzycki, "In vivo contextual requirements for UAG translation as pyrrolysine," *Molecular Microbiology*, vol. 63, no. 1, pp. 229–241, 2007.

Hindawi Publishing Corporation
Archaea
Volume 2013, Article ID 648746, 13 pages
http://dx.doi.org/10.1155/2013/648746

Research Article

Comparative Analysis of Proteomes and Functionomes Provides Insights into Origins of Cellular Diversification

Arshan Nasir and Gustavo Caetano-Anollés

Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, and Illinois Informatics Institute, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA

Correspondence should be addressed to Gustavo Caetano-Anollés; gca@illinois.edu

Received 30 September 2013; Revised 22 November 2013; Accepted 25 November 2013

Academic Editor: Kyung Mo Kim

Copyright © 2013 A. Nasir and G. Caetano-Anollés. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Reconstructing the evolutionary history of modern species is a difficult problem complicated by the conceptual and technical limitations of phylogenetic tree building methods. Here, we propose a comparative proteomic and functionomic inferential framework for genome evolution that allows resolving the tripartite division of cells and sketching their history. Evolutionary inferences were derived from the spread of conserved molecular features, such as molecular structures and functions, in the proteomes and functionomes of contemporary organisms. Patterns of use and reuse of these traits yielded significant insights into the origins of cellular diversification. Results uncovered an unprecedented strong evolutionary association between Bacteria and Eukarya while revealing marked evolutionary reductive tendencies in the archaeal genomic repertoires. The effects of nonvertical evolutionary processes (e.g., HGT, convergent evolution) were found to be limited while reductive evolution and molecular innovation appeared to be prevalent during the evolution of cells. Our study revealed a strong vertical trace in the history of proteins and associated molecular functions, which was reliably recovered using the comparative genomics approach. The trace supported the existence of a stem line of descent and the very early appearance of Archaea as a diversified superkingdom, but failed to uncover a hidden canonical pattern in which Bacteria was the first superkingdom to deploy superkingdom-specific structures and functions.

1. Introduction

Tracing the evolution of extant organisms to a common universal cellular ancestor of life is of fundamental biological importance. Modern organisms can be classified into three primary cellular superkingdoms, Archaea, Bacteria, and Eukarya [1]. Molecular, biochemical, and morphological lines of evidence support this trichotomous division. While the three-superkingdom system is well accepted, establishing which of the three is the most ancient remains problematic. Initial construction of unrooted phylogenies based on the joint evolution of genes linked by an ancient gene duplication event revealed that, for each set of paralogous genes, Archaea and Eukarya were sister groups and diverged from a last archaeal-eukaryal common ancestor [2, 3]. This "canonical" rooting that places Bacteria at the base of the "Tree of Life" (ToL) is still widely accepted despite the fact that many other paralogous gene couples produced discordant topologies and despite known technical artifacts associated with these sequence-based evolutionarily deep phylogenies [4, 5]. As a result, reconstructing a truly "universal" ToL portraying the evolutionary relationships of all existing species remains one of the most controversial issues in evolutionary biology. This in part owes to the shortcomings of available phylogenetic characters and tree optimization methods that suffer from important technical and conceptual limitations [6, 7] and have failed to generate a consensus. It is further complicated by the fact that genetic material can be readily exchanged between species, especially akaryotes (i.e., Archaea and Bacteria that lack a nucleus) via horizontal gene transfer (HGT) [8–10]. Nonvertical evolutionary processes coupled with uncertainties regarding evolutionary assumptions greatly complicate the problem of reconstructing the evolutionary past. Recently, ToLs reconstructed using conserved structural

information of protein domains [11, 12], their annotated functions (Kim et al., ms. resubmitted), and universal RNA families [13–18] provided new ways to root phylogenies. These studies identified thermophilic archaeal species to be the most closely related to the primordial cells. Findings not only challenge the bacterial rooting of the ToL but also highlight the importance of employing reliable phylogenetic methods and assumptions when reconstructing deep evolutionary history [7].

Here, we advance the structural and functional approach by providing a simple solution to the problem of phylogenetic reconstruction. We argue that basic quantitative and comparative genomic analyses that do not invoke phylogenetic reconstruction are sufficient to resolve the tripartite division of cells and sketch their history. Our comparative approach involves the analysis of how superkingdoms, and their organismal constituents, relate to each other in terms of global sharing of genomic features. The genomic features we selected are entire repertoires of molecular structures and functions (collectively referred to as traits from hereinafter). They define two specific genomic datasets. The *structure* dataset encompasses the occurrence and abundance of 1,733 fold superfamily (FSF) domains in 981 completely sequenced proteomes. FSF domains were delimited using the Structural Classification of Proteins (SCOP ver. 1.75), which is a manually curated database of structural and evolutionary information of protein domains [19, 20]. The FSF level of the SCOP hierarchy includes domains that have diverged from a common ancestor and are evolutionarily conserved [21, 22]. In comparison, the *function* dataset describes the occurrence and abundance of 1,924 gene ontology (GO) terms [23, 24] in 249 functionomes. We note that the global set of FSFs portrays the entire structural repertoire of organisms and that the repertoire of GO terms portrays their true physiology. Both provide useful information about species diversification. We restricted our analyses to include only structures and functions as they are more conserved than gene sequences [25–27] and permit deep evolutionary comparisons. In contrast, nucleotide sequences are susceptible to higher mutation rates and are continuously rearranged in genomes to yield novel domain combinations and molecular functions [6]. In other words, loss of an FSF domain structure or molecular function is much more costly for cells as it sometimes involves loss of hundreds of genes that have accumulated over long periods of evolutionary time. This is compounded especially for traits that are very ancient as they had more time to multiply in genomes and increase their genomic abundance [28, 29]. Thus molecular structure and function remain preserved in cells for relatively longer periods and make reliable candidates for inferring deep evolutionary relationships.

Here, we show that an analysis of trait distribution between superkingdoms, distributions between genomic repertoires of superkingdoms, and abundance counts allow dissection of historical (ideographic) patterns using a comparative ahistorical (nomothetic) method (Figure 1). Inspired by a comparative analysis of RNA families [30], we measured the strength of evolutionary association between superkingdoms as a function of patterns of sharing of individual traits (Figure 1). We note that our approach is sufficiently

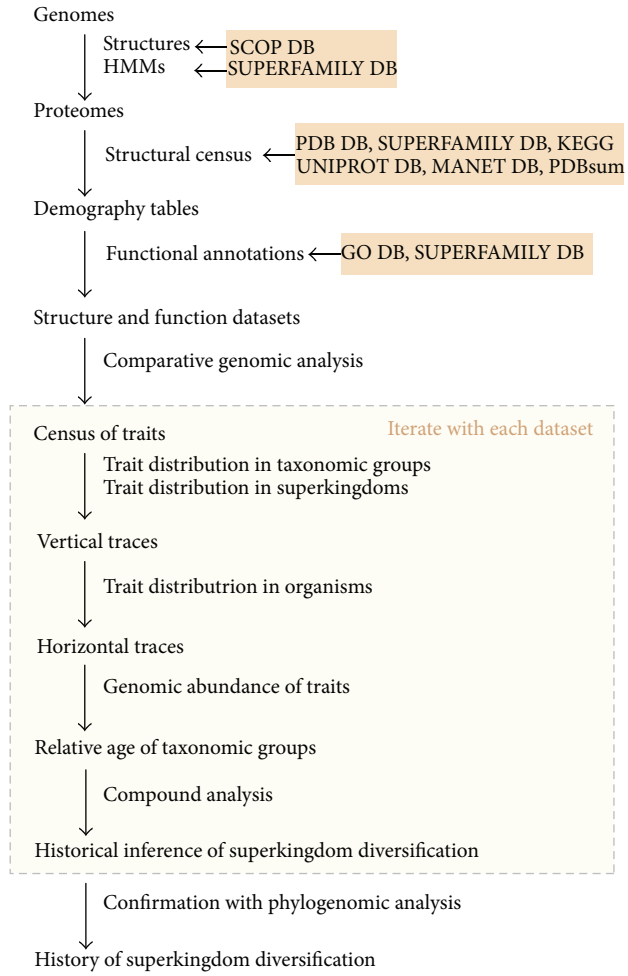


Figure 1: Overview of the comparative proteomics and functionomics methodology. Proteomes and functionomes were scanned for the occurrence and abundance of FSFs and GO terms (i.e., traits). This information was represented in data matrices that were analyzed for trends of trait sharing and traces of vertical and horizontal inheritance. Inferences were drawn regarding superkingdom diversification and were confirmed with previously published phylogenetic studies.

informative to make reliable inferences regarding different evolutionary scenarios of diversification adopted by the three superkingdoms. This approach revisits widely accepted theories regarding the origin of diversified life [31, 32] and falsifies the fusion [33] and hydrogen hypotheses [34] of eukaryotic origins, more than supporting any. This exercise then prompts validation by phylogenetic tree reconstruction, which we have reported previously (see [26, 28, 29, 35]). In light of these considerations, the comparative exercise provides an easy-to use and reliable alternative to otherwise complicated phylogenetic tree reconstruction methods. These analyses carry the potential to yield significant insights into the evolution of cells and, if carefully interpreted, provide strong arguments in favor of the rooting of the ToL in Archaea and embedded canonical pattern of FSF and GO innovation.

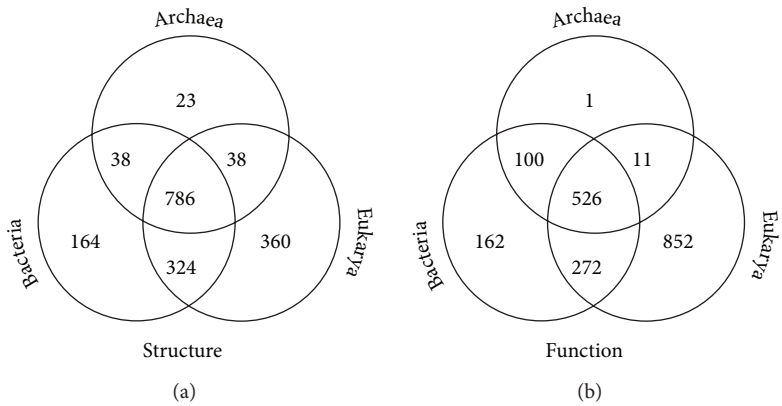


Figure 2: Global trends of trait sharing in Venn taxonomic groups. (a) Venn diagram displaying the distribution of 1,733 FSF domains in 981 completely sequenced proteomes sampled from 652 Bacteria, 70 Archaea, and 259 Eukarya. This constituted the *structure* dataset. (b) Venn diagram displaying the distribution of 1,924 terminal-level GOs in 249 free-living functionomes corresponding to 183 Bacteria, 45 Archaea, and 21 Eukarya. This constituted the *function* dataset.

2. Materials and Methods

2.1. Data Retrieval and Manipulation. FSF domain assignments for 981 completely sequenced proteomes were extracted from local MySQL installation of SUPERFAMILY ver. 1.75 database [36] using a stringent E-value cutoff of 10^{-4} [37]. The SUPERFAMILY database assigns structures to protein sequences using profile hidden Markov models (HMMs) searches that are superior in detecting remote homologies [38]. The dataset included 652 bacterial, 70 archaeal, and 259 eukaryal proteomes encoding a total repertoire of 1,733 significant FSF domains. In this study, FSFs were identified using SCOP alphanumeric identifiers (e.g., c.37.1, where c represent the class of domain structure (α, β, $\alpha + \beta$, α/β, etc.), 37 the fold, and 1 the FSF). This constituted the *structure* dataset.

To prepare the *function* dataset, we downloaded the Gene Ontology Association (GOA) files for 1,595 organisms from the European Bioinformatics Institute (http://www.ebi.ac.uk/GOA/proteomes). These files were filtered to exclude strain-level and parasitic organisms. They were subjected to a 50% GO coverage threshold (number of gene products annotated to GO terms divided by the total number of gene products) to ensure high quality annotations. In this study, we only sampled terminal-level GO terms from the GO molecular function hierarchy (simply referred to as GOs or functions from hereinafter), as they represent the highly-specialized functional annotations and approximate the molecular activities of cells (which are evolutionarily informative) [25]. We further excluded GOs that were likely candidates of HGT by scanning the total set of 2,039 terminal GOs in our dataset against proteins listed in the horizontal gene transfer database (HGT-DB) [39]. This allowed the exclusion of 115 potentially HGT-derived GOs. The final *function* dataset included 249 free-living functionomes from 183 Bacteria, 45 Archaea, and 21 Eukarya encoding a total set of 1,924 GOs.

2.2. Genomic Census of Traits. We conducted a genomic census for both *structure* and *function* datasets by counting the occurrence (presence/absence) and abundance (redundant counts) of traits in all proteomes and functionomes. These data matrices were then scanned to generate Venn diagrams and boxplots displaying patterns of trait sharing both between and within proteomes and functionomes of superkingdom groups.

2.3. Calculating the Spread of Traits in Proteomes and Functionomes. The spread of each trait in a superkingdom was calculated by an f-value indicating the number of proteomes/functionomes harboring a trait divided by the total number of proteomes/functionomes in that organismal group. The f-value approaches one for ubiquitous traits but is lower for those that are less widely distributed.

2.4. Estimating the Evolutionary Age of Traits. We used a relative time scale to pinpoint the origin of FSFs in molecular evolution. This scale was defined by node distance (*nd*) as calculated from a phylogenetic tree of FSF domains (see [26, 28, 35] for practical details). Technically, *nd* is the distance of a particular trait from its position on the phylogenetic tree to the root node. It is given on a scale from 0 (the most ancient or root node) to 1 (highly derived or terminal node). Biologically, it reflects the evolutionary age of an FSF relative to other FSFs. *nd* has been successfully used in the past to describe important events in the evolution of cells (e.g., [26, 28]) and could be considered a reliable proxy to estimate the origin of molecular traits in organisms.

3. Results

3.1. Identifying Vertical Traces. Venn diagrams demonstrate the evolutionary sorting of FSF and GO traits in the seven possible and mutually exclusive Venn taxonomic groups, ABE (i.e., present in all three superkingdoms), AB (present only in akaryotes), BE (present only in Bacteria and Eukarya), AE (present only in Archaea and Eukarya), and the three superkingdom-specific groups, A, B, and E (Figure 2).

Remarkably, a majority of the traits (45% of total structures and 27% of functions) were present in all three superkingdoms, supporting the hypothesis of common ancestry (Figure 2). Since a ToL by definition is a nested hierarchy of taxonomies, we propose that elevated sharing of traits by a taxonomic group points towards an ancient "vertical trace" indicative of divergence from a common ancestor. In turn, low numbers in a taxonomic group are indicative of other evolutionary processes besides lineage splitting, including reductive evolution, HGT, convergent evolution, differential loss, and secondary evolutionary adaptations.

The two-superkingdom taxonomic groups were most informative as each embodied a possible vertical trace and an evolutionary hypothesis of superkingdom origin. The number of traits in the AB, AE, and BE taxonomic groups is therefore indicative of the strength of evolutionary association between akaryotes, Archaea and Eukarya, and Bacteria and Eukarya, respectively. Remarkably, and against intuition, the size of the AB and AE taxonomic groups was ~9 folds smaller than that of BE in the *structure* dataset (38 and 38 versus 324) (Figure 2(a)). This trend was also recovered in the *function* dataset where BE significantly outnumbered both AB and AE (272 versus 100 and 11) (Figure 2(b)). These important biases suggest an intriguing ancestral evolutionary link between Bacteria and Eukarya, very much as the large number of ABE traits suggests an ancestral link between all organisms. While simultaneous gains of traits in both bacterial and eukaryal proteomes would be possible, the high sharing of structures and functions by the BE taxonomic group makes it parsimoniously unlikely and points instead to an evolutionary scenario in which the two superkingdoms diverged from a common ancestor. This is particularly supported by the findings that convergent evolution of structures is rare [40] and seems unlikely to occur at such high levels. We note that bacterial organisms are more intimately associated with eukaryotes, establishing many coevolving bacterial parasitic/symbiotic interactions with eukaryotic hosts; this is in marked contrast with organismal interactions involving Archaea [41]. These interactions could foster the exchange of protein and functional repertoires between the organisms. However, the *function* dataset included only free-living GO-annotated organisms with the exclusion of HGT-acquired GOs and consequently was free from adaptive effects of either parasitic or symbiotic lifestyles. The dataset still showed the high representation of the BE group relative to the AB and AE groups (Figure 2(b)). In short, the very large size difference of BE compared to the AB and AE groups is an evolutionarily significant outcome that cannot be explained merely by parasitic/symbiotic processes.

Finally, the Venn diagrams show that Eukarya-specific traits always outnumbered Bacteria-specific and Archaea-specific counterparts, suggesting either an expansive mode of evolutionary growth of eukaryotic repertoires or a reductive mode in akaryotic counterparts, or both (Figure 2). This is an expected result as eukaryotes encode a highly diverse and complex genome and are capable of carrying out many advanced molecular activities, especially those related to development and immunological responses. Based on our initial comparative genomic exercise, we put forth three preliminary conclusions: (i) all extant cells are related by common descent, (ii) Bacteria and Eukarya diverged from a mutual ancestor, and (iii) eukaryotes are significantly more complex than akaryotes in terms of numbers of unique traits.

3.2. Identifying Horizontal Traces. Venn diagrams simply describe global patterns of sharing in superkingdoms and cannot dissect how popular are traits in the organisms of each superkingdom. In other words, the presence of a trait in a superkingdom does not necessarily imply that it was vertically inherited; this trait might only be present in few of its members. In such cases, acquisition of traits by nonvertical (e.g., HGT fluxes, convergent evolution) or confounding (e.g., differential loss that mimics HGT) evolutionary processes becomes more likely. To fully explore the extent to which these real or virtual "horizontal traces" contribute to the development of the proteomes of organisms in superkingdoms and to further test the preliminary conclusions drawn from the Venn diagrams of Figure 2, we calculated the spread or popularity of FSF and GO traits in the organisms of superkingdoms, which we term f-value. The f-value is simply the number of organisms in a Venn taxonomic group harboring a trait divided by the total number of organisms in that taxonomic group and in that superkingdom. It is given on a relative scale from 0 (absent) to 1 (omnipresent). Using this simplistic approach, we first identified 17 FSFs (Table 1) and 26 GOs (Table 2) that were present in all proteomes and functionomes, respectively. This cohort of traits truly represents the "universal" core of traits that was present in the common ancestor of life, the urancestor, and was strongly retained by all of its descendants. These traits perform crucial and central metabolic and informational roles in cells such as ATP hydrolysis and ion binding, make up structural components of ribosomal proteins, and are involved in DNA replication and protein translational processes (Tables 1 and 2). Moreover, a total of 245 FSFs and 95 GOs had an $f > 0.90$ implying near-universal presence and suggesting reductive losses in the remaining 10% of the proteomes and functionomes (data not shown). This global analysis based on popularity of traits in proteomes and functionomes suggests that the urancestor was especially enriched (structurally and functionally) in metabolic functions [29] and illustrates the power of f-value in dissecting traces of vertical versus horizontal inheritance. Therefore, we extended this analysis to the proteomes and functionomes of members of each of the seven taxonomic groups.

We first compared the spread of FSFs in the *structure* dataset using boxplot representations of f-value distributions (Figure 3(a)). Our assumptions are straightforward: high f-values and balanced f-distributions reflect vertical traces while low f-values and biased f-distributions echo horizontal (flux-loss) traces, respectively. The 786 ABE structures were distributed with the highest f-values and the medians increased in the order, Archaea (median f = 0.6), Bacteria (0.74), and Eukarya (0.90) (Figure 3(a), ABE taxonomic group). The large number of ABE structures that was widespread in all three superkingdoms strengthens the hypothesis of life's common ancestry. The relatively lower median f-values in akaryotes (0.6 for Archaea and 0.74 for

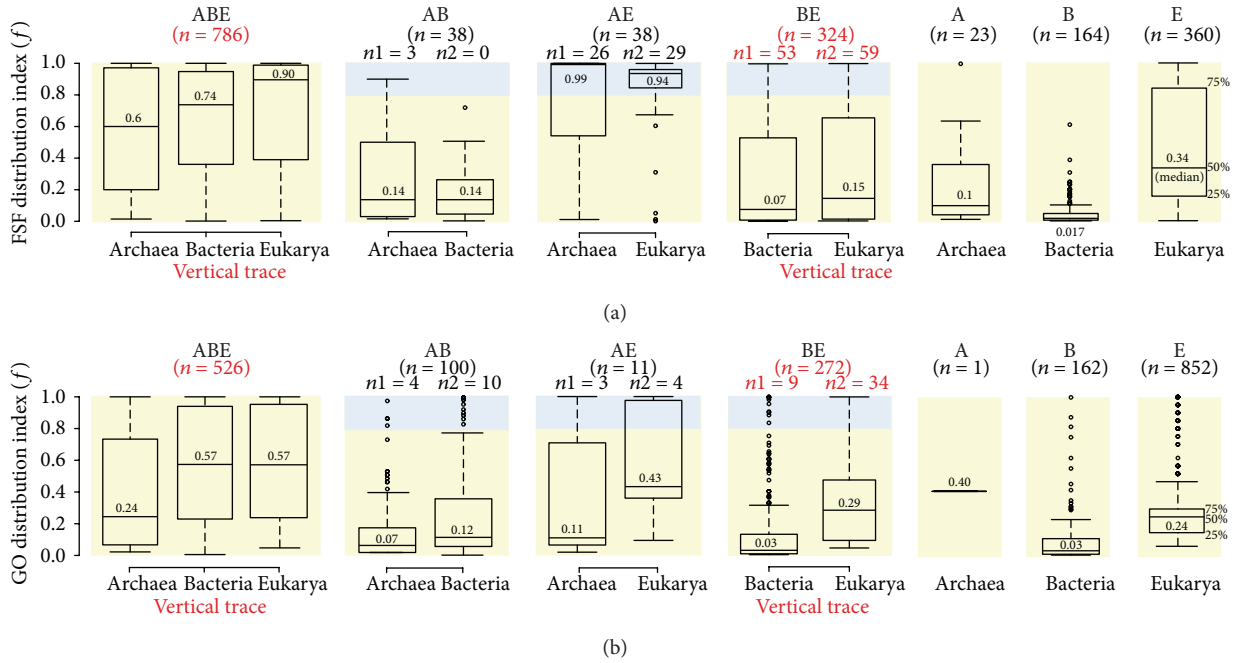


Figure 3: The spread of FSF domain structures (a) and GO terminal terms (b) in the proteomes and functionomes of each member of the superkingdom in the seven Venn taxonomic groups (panels ABE, AB, AE, BE, A, B, and E). Shaded regions indicate that FSFs or GOs were present in >80% of the proteomes ($f > 0.8$), and their numbers, n_1 and n_2. Numbers in boxplots of each distribution indicate group medians. Numbers in red suggest the strongest vertical evolutionary trace.

Table 1: List of universal FSFs that were present in all proteomes of the *structure* dataset.

No.	SCOP Id	FSF Id	FSF description
1	52540	c.37.1	P-loop containing nucleoside triphosphate hydrolases
2	50249	b.40.4	Nucleic acid-binding proteins
3	53067	c.55.1	Actin-like ATPase domain
4	51905	c.3.1	FAD/NAD(P)-binding domain
5	53098	c.55.3	Ribonuclease H-like
6	54211	d.14.1	Ribosomal protein S5 domain 2-like
7	55681	d.104.1	Class II aaRS and biotin synthetases
8	50447	b.43.3	Translation proteins
9	54980	d.58.11	EF-G C-terminal domain-like
10	50104	b.34.5	Translation proteins SH3-like domain
11	50465	b.44.1	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain
12	55174	d.66.1	Alpha-L RNA-binding motif
13	54768	d.50.1	dsRNA-binding domain-like
14	55257	d.74.3	RBP11-like subunits of RNA polymerase
15	52080	c.12.1	Ribosomal proteins L15p and L18e
16	54686	d.41.4	Ribosomal protein L16p/L10e
17	54843	d.55.1	Ribosomal protein L22

Bacteria versus 0.90 in Eukarya) can be explained by genome reduction events that are known to occur with relatively high frequency in akaryotic microbes [26, 42], and also manifest in the numbers of superkingdom-specific traits (Figure 2). The 38 AB structures were poorly but similarly distributed (median f-values = 0.14) in archaeal and bacterial proteomes, with archaeal structures exhibiting a tendency to become more widespread (longer tail) (Figure 3(a), AB). This pattern supports the existence of a horizontal trace between akaryotes, with a weak bias in flux-loss between superkingdoms (note however that no common outliers could be detected). In contrast, the 38 AE structures were highly represented (median f-values > 0.94) in the organisms of corresponding superkingdoms (Figure 3(a), AE). Again, archaeal structures appeared more widely shared but also showed a longer tail indicative of possible flux-loss episodes. At first glance, this

Table 2: List of universal GOs that were present in all functionomes of the *function* dataset.

No.	GO Id	GO description
1	GO:0005524	ATP binding
2	GO:0008270	zinc ion binding
3	GO:0000287	magnesium ion binding
4	GO:0005525	GTP binding
5	GO:0004222	metalloendopeptidase activity
6	GO:0010181	FMN binding
7	GO:0030145	manganese ion binding
8	GO:0003924	GTPase activity
9	GO:0003887	DNA-directed DNA polymerase activity
10	GO:0004252	serine-type endopeptidase activity
11	GO:0003746	translation elongation factor activity
12	GO:0009982	pseudouridine synthase activity
13	GO:0004523	ribonuclease H activity
14	GO:0004826	phenylalanine-tRNA ligase activity
15	GO:0004821	histidine-tRNA ligase activity
16	GO:0004820	glycine-tRNA ligase activity
17	GO:0004824	lysine-tRNA ligase activity
18	GO:0004831	tyrosine-tRNA ligase activity
19	GO:0004618	phosphoglycerate kinase activity
20	GO:0004634	phosphopyruvate hydratase activity
21	GO:0004749	ribose phosphate diphosphokinase activity
22	GO:0003952	NAD+ synthase (glutamine-hydrolyzing) activity
23	GO:0004815	aspartate-tRNA ligase activity
24	GO:0004807	triose-phosphate isomerase activity
25	GO:0004813	alanine-tRNA ligase activity
26	GO:0003917	DNA topoisomerase type I activity

chimes for a strong vertical trace of the AE group that could rival that of the BE group. However, this may not be the case. The 324 BE structures were on average poorly represented in bacterial and eukaryal proteomes (median f-values < 0.15) (Figure 3(a), BE). Their overall spread was relatively uniform, with a weak bias towards higher representation in Eukarya. However, 53 and 59 structures were widespread in the proteomes of Bacteria and Eukarya ($f > 0.8$), respectively (shaded region in Figure 3(a), BE). This subset of BE structures was numerically double that of the total set of the highly represented AE structures. Thus, the stronger vertical trace for BE structures continues to support a sister-group relationship between Bacteria and Eukarya and the early diversification of Archaea. We note that this inference is strengthened by the fact that we had 652 bacterial and 259 eukaryal proteomes in comparison to only 70 archaeal proteomes. Existence of any structure in such large number of genomes implies strong selective pressure and conservation of that trait. Finally, the sharing of superkingdom-specific structures was low in each superkingdom (median f-values = 0.01–0.34), with minimum average f-values for Bacteria and maximum for Eukarya (Figure 3(a), A, B, and E). Remarkably, out of the 164 Bacteria-specific structures, none, but one, was present in >50% of the proteomes (Figure 3(a), B). The absence of an expected homogenous distribution strongly suggests that the role of HGT and other homogenizing processes may be quite limited in shaping the evolution of bacterial proteomes. Eukaryal-specific structures were distributed with higher f-values (Figure 3(a), E). The relatively low spread of superkingdom-specific structures suggests that these structures were acquired independently and after divergence from the last common ancestors of each superkingdom.

Inferences drawn from boxplots of the *function* dataset (Figure 3(b)) supported the general conclusions derived from the *structure* dataset. The ABE distributions had high f-values, with those of Archaea (median $f = 0.24$) being considerably lower than those of Bacteria (0.57) and Eukarya (0.57) (Figure 3(b), ABE). Bacterial and eukaryal distributions were remarkably balanced, providing additional support to their recent divergence from a mutual ancestor. The median f-value in Archaea was lowest and could be explained by either high genome reduction events [26] or biases in the number of GO annotations for archaeal genomes. GOs are more reliably and extensively curated for Bacteria and Eukarya, and this factor could reduce the number of overall detections in archaeal genomes. However, comparing distributions of the *function* and *structure* datasets show that supporting results were consistent and suggest a limited impact of this possible shortcoming. Here, ABE distributions followed the pattern observed for FSFs and were therefore considered reliable. None of the AB, AE, and BE taxonomic groups showed balanced distributions (Figure 3(b), AB, AE, and BE). The AB taxonomic group harbored 100 GOs (~3-fold greater than corresponding structures) that were distributed with low popularity (Figure 3(b), AB). In general, these functions were more abundant in Bacteria compared to Archaea and thus suggested that some molecular activities were laterally transferred from Bacteria to Archaea (confirmed below). The AE taxonomic group failed to strongly support AE distributions in the *structure* dataset. This group included only 11 GOs that were relatively more abundant in eukaryal proteomes (Figure 3(b), AE). Finally, the BE taxonomic group also supported the increased prevalence of BE functions in eukaryal genomes compared to bacterial genomes (0.39 median versus 0.03), indicating either horizontal trace effects or biases introduced by GO annotation schemes (Figure 3(b), BE). However, the numbers of traits of the BE group were considerably greater than those of either the AB or AE groups and included a significantly large number of functions that were relatively widespread ($f > 0.8$) (Figure 3(b), BE). This was in sharp contrast with patterns in either AB or AE taxonomic groups. The subset of highly represented BE functions is therefore the most likely trace of an ancient vertical signature that unifies Bacteria and Eukarya as sister-groups in the ToL. This trace is remarkably consistent with the patterns obtained in the *structure* dataset (Figures 2(a) and 3(a)).

Finally, the superkingdom-specific functions were again distributed with low f-values. Archaea had only one unique

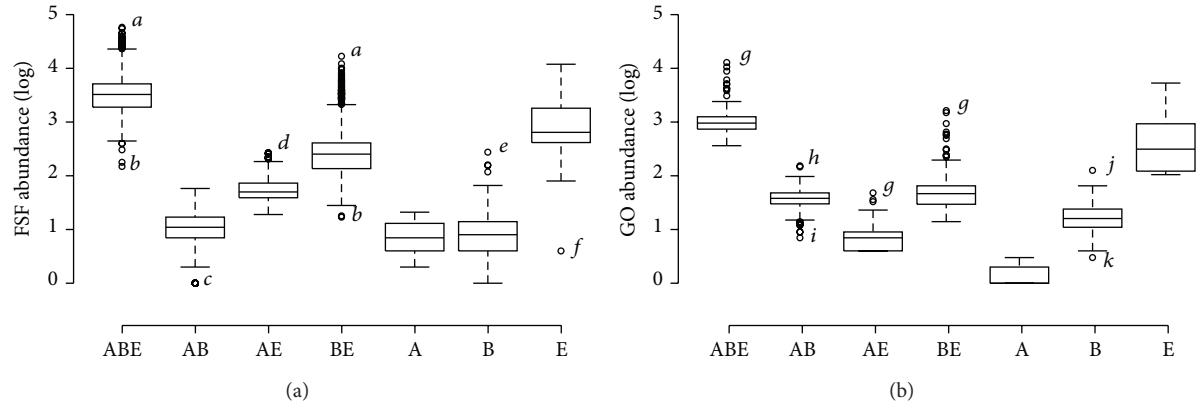


Figure 4: Boxplots comparing the log-transformed abundance values of structural (a) and functional (b) traits in the proteomes and functionomes of the seven Venn taxonomic groups. Italicized characters identify outliers with maximum and minimum abundance of traits in each group: *a, Takifugu rubripes; b, Cand.* Hodgkinia cicadicola Dsem; *c, Mycoplasma genitalium G37; d, Zea mays; e, Mycobacterium marinum; f, Guillardia theta; g, Homo sapiens; h, Rhodospirillum rubrum; i, Desulfurococcus kamchatkensis; j, Ralstonia eutropha; k, Thermosipho africanus.*

GO that was present in 40% of the archaeal genomes (Figure 3(b), A). In sharp contrast, there were 162 bacterial and 852 eukaryal-specific GOs. Bacterial functions again showed evidence of very limited spread in organisms (Figure 3(b), B) challenging claims of widespread bacterial HGT. In turn, eukaryal functions were moderately widespread (Figure 3(b), E). These results are in line with earlier inferences regarding late and independent acquisition of superkingdom-specific traits.

3.3. Identifying Patterns of Horizontal Flux. Boxplot distributions provided useful clues regarding the divergence patterns of superkingdoms. However, they did not allow us to quantify the extent of horizontal versus vertical inheritance. Therefore, we calculated a difference in the f-value for all traits in the AB, AE, and BE taxonomic groups. If the difference between f-values was >0.6, the presence of the trait in both superkingdoms was considered the result of a probable HGT event. This threshold was set arbitrarily to include only those traits that were considerably more abundant in one superkingdom but scarcely present in the other. For example, the "t-snare proteins" superfamily [SCOP id: a.47.2], which is abundantly found in yeast and mammalian cells and forms bridges to mediate intracellular trafficking [43], had an f-value of 0.996 in eukaryotes implying that it was ubiquitous. However, it was only present in one of the 652 bacterial proteomes examined ($f = 0.001$) (Table S1, Supplementary Materials available online at http://dx.doi.org/10.1155/2013/648746). This most likely is an example of structure gain via HGT that occurred in the direction from Eukarya to Bacteria. Using this criterion, only one structure ("tRNA-intron endonuclease N-terminal domain-like" [d.75.1]) was acquired horizontally in Eukarya from Archaea in the AE taxonomic group, while 6 were transferred from Eukarya to Archaea (Table S1). Similarly, only one FSF was laterally transferred to Bacteria from Archaea ("Sulfolobus fructose-1,6-bisphosphatase-like" [d.280.1]) while none were acquired in reciprocity. Finally, Bacteria likely transferred 35 structures to eukaryotes while gained 52 in return (Table S1). The rest 237 structures did not show significant deviations in terms of spread in these taxonomic groups and were possibly acquired vertically or gained independently in evolution.

In terms of *function*, none of the GO traits were likely transferred to Bacteria from Archaea. However, 9 GOs were transfer candidates from Bacteria to Archaea (Table S2). Perhaps the most interesting among these was the lateral acquisition of "penicillin binding molecular activity" [GO:0008658] that was universally present in Bacteria but also present in 11% of the archaeal proteomes (Table S2). Similarly, no molecular function was transferred to Eukarya from Archaea, while only one GO ("dolichyl-diphosphooligosaccharide-protein glycotransferase activity" [GO:0004579]) was gained. Finally, 4 molecular functions were likely transferred from Bacteria to Eukarya and 28 were gained in return (Table S2). Overall, the inferred impact of horizontal transfer processes appeared to be quite limited and did not seriously invalidate our inferences. Moreover, horizontal contributions from Archaea to either Bacteria or Eukarya were minimal, which is consistent with the minimal sharing of traits described above (Figures 2 and 3). In comparison, both Bacteria and Eukarya exhibited higher levels of vertical and horizontal inheritance of traits and indicated a much stronger evolutionary association, a conclusion intimated by likely ancient endosymbiotic events.

3.4. Identifying Ancestral Traits Using Abundance Counts. Traits that are of ancient origin are expected to be present in greater abundance than those acquired recently. This is true because traits appearing earlier have more time to accumulate in genomes and to increase their representation [6]. Thus, high abundance of traits in a particular Venn taxonomic group is indicative of the presence of relatively more ancient traits and an ancient origin. Therefore, genomic abundance can be used as one proxy to estimate the age of taxonomic groups. We calculated the abundance of traits present in each proteome and functionome and represented these values in boxplot distributions (Figure 4). The median

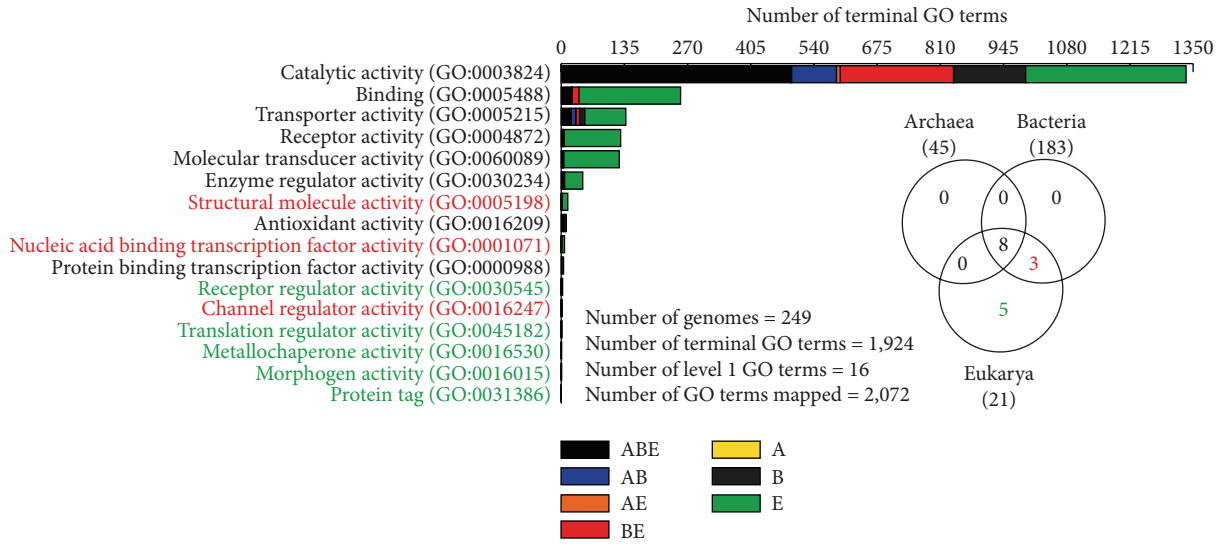


Figure 5: Bar plots illustrating the breakdown of terminal GOs in the seven taxonomic groups for level 1 GO terms. A total of 1,871 out of 1,924 GOs (97.24%) could be reliably mapped to their parents. Level 1 GOs that could not be mapped include "D-alanyl carrier activity [GO:0036370]," "electron carrier activity [GO:0009055]," "chemoattractant activity [GO:0042056]," "chemorepellent activity [GO:0045499]," and "nutrient reservoir activity [GO:0045735]." Note that terminal GOs may have more than one parent. The Venn diagram shows that none of the A, B, AB, and AE taxonomic groups uniquely code for any level 1 GO term.

abundance value was highest for the ABE taxonomic group in both the *structure* (Figure 4(a)) and *function* (Figure 4(b)) datasets, again supporting that this group retains most of the urancestral traits that have relished maximum time to multiply and become abundant in modern proteomes and functionomes. The BE group always harbored traits in much greater abundance compared to the AB and AE groups (Figure 4). Finally, Eukarya-specific traits were significantly enriched in the eukaryal proteomes and functionomes and were detected in much greater abundance compared to the genomic abundance of either Archaea-specific or Bacteria-specific traits (Figure 4). This result confirms the existence of a strong vertical trace in modern cells in the direction from ABE to BE and to E. It is likely that eukaryotes retained the majority of the most ancient traits that were progressively lost in akaryal organisms, beginning in Archaea and manifesting much later in Bacteria. Previous phylogenomic analyses have confirmed strong reductive trends in the akaryal proteomes [26, 28, 35]. Evolution of Archaea has also been linked to genome reduction events that started very early in evolution and before the appearance of the BE taxonomic group [28, 35]. However, the relatively late loss of traits in Bacteria is intriguing. Several bacterial species are known to have adapted a parasitic lifestyle following genome reduction [44]. Thus, gene loss in Bacteria is likely an ongoing evolutionary process hinting towards a major secondary evolutionary transition. This was also manifested in the very poor spread of Bacteria-specific traits (Figure 3).

We provide evidence for late loss in Bacteria by closely examining the AE traits. The majority of the 38 AE FSFs and 11 GOs are enriched in informational functions (e.g., translation initiation, ribosomal proteins, DNA binding proteins, and proteins involved in DNA replication; Tables S3 and S4). This result is consistent with existing knowledge. Indeed, Archaea and Eukarya are more related to each other in terms of informational processes, while Bacteria and Eukarya resemble each other metabolically [45]. Thus, the high popularity of AE FSFs could be due to biases attributed to late differential loss of structures in these functional categories. For example, the 11 AE GOs include crucial molecular functions such as "DNA polymerase processivity factor activity [GO:0030337]" and "tRNA-intron endonuclease activity [GO:0000213]." The former is a regulator of the replication fork [46, 47] while the latter is involved in processing tRNA introns [48]. Both of these activities could be linked to late losses in Bacteria, as they seem centrally important functions in cells. Therefore, while HGT, convergent evolution and coevolution of BE traits seems less likely, we cannot rule out the possibility of extensive genome reduction in akaryal species.

3.5. Tracking the Vertical Trace. To further dissect the evolution of Venn taxonomic groups, we mapped the 1,924 terminal GOs to 16 level 1 parent GO terms. Figure 5 shows the distribution of terminal GOs, indexed by taxonomic group, in each of the 16 parent categories. This exercise confirmed the inferences drawn from earlier experiments and highlighted the direction of the vertical trace. Remarkably, only ABE, BE, and E were enriched in level 1 molecular functions while the majority of the terminal GO terms could be identified as either "catalytic activity [GO:0003824]" or "binding [GO:0005488]" (Figure 5). This is an interesting result. A previous analysis by Kim and Caetano-Anollés [25] confirmed that these two molecular activities appeared first in evolution and were shared by all organisms. In comparison, the more derived molecular activities first appeared in the BE taxonomic group (e.g., "structural molecule activity [GO:0005198]," "nucleic acid binding transcription factor activity [GO:0001071]," and "channel regulator activity

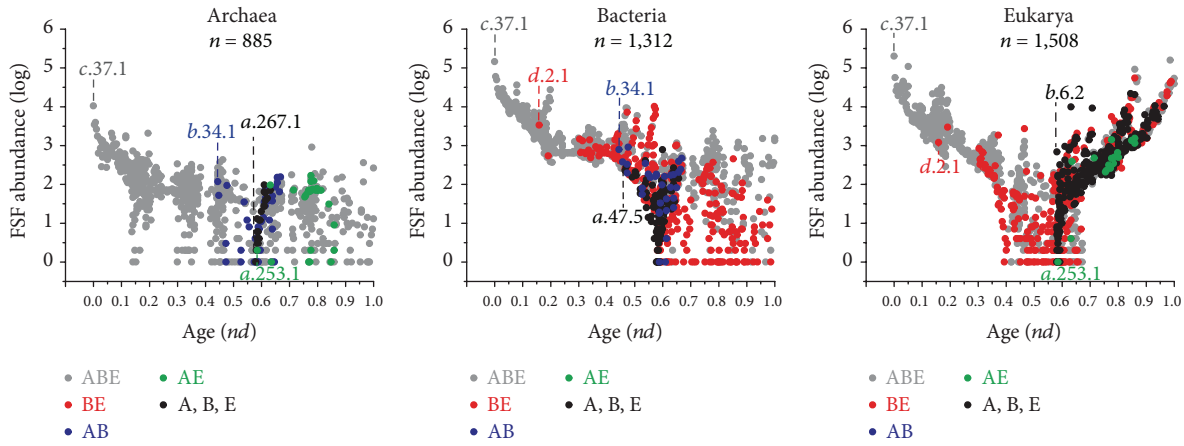


FIGURE 6: Evolutionary timelines highlighting the abundance of FSFs in superkingdom taxonomic groups. Evolutionary age (*nd*) was calculated from a phylogenetic tree of protein domains describing the evolution of 1,733 FSFs (taxa) in 981 organisms (characters) (see [26, 28, 35] for technical details). SCOP alphanumeric identifiers were used to identify the most ancient FSF in each taxonomic group. In case of multiple FSFs of same age, only the FSF with maximum abundance was labeled. *c.37.1* is the P-loop containing NTP hydrolase FSF; *b.34.1* is the C-terminal domain of transcriptional repressors FSF; *a.267.1* is the topoisomerase V catalytic domain-like FSF; *a.253.1* is the AF0941-like FSF; *d.2.1* is the Lysozyme-like FSF; *a.47.5* is the FlgN-like FSF; *b.6.2* is the major surface antigen p30, SAG1.

[GO:0016247])," while the recent innovations occurred uniquely in Eukarya (e.g., "receptor regulator activity [GO:0030545], "translation regulator activity [GO:0045182]," "metallochaperone activity [GO:0016530]," "morphogen activity [GO:0016015]," and "protein tag [GO:0031386]"). In contrast, none of the AB, AE, A, and B taxonomic groups uniquely harbored a level 1 molecular function (Figure 5). Remarkably, a significant proportion of the BE terminal GOs was devoted to the most ancient catalytic and binding activities (Figure S1). In comparison, "transporter activity [GO:0005215]" was found to be over-represented in the AB group while AE was numerically much smaller (Figure S1). These findings confirm the existence of a vertical trace from ABE to BE and finally to E (also supported by the *structure* dataset). Akaryal ancestors likely diverged from this trace by following paths towards genome reductions while eukaryotes enriched their repertoires by engaging in gene duplication events and exploring novel domain combinations [12, 49].

3.6. Validating Inferences with Evolutionary Timelines. To validate our ahistorical comparative approach, we unfolded the appearance of FSF and GO traits in evolutionary time (*nd*), while plotting their genomic abundance in each superkingdom. The historical analyses of FSF evolution (Figure 6) and GO terminal terms (data not shown) were congruent and revealed two clear patterns: (1) a pattern of ancient genomic loss embodying the early rise of the BE taxonomic group (red circles), which generally involved traits with abundance levels that were at least an order of magnitude higher than the levels of other taxonomic groups (e.g., AE and AB); and (2) a canonical pattern of appearance of superkingdom-specific traits that revealed the rise of early bacterial novelties followed by the joint appearance of unique novelties in Archaea and Eukarya. This historical analysis therefore supports the ancient vertical trace identified by comparative analysis that flows from the ABE group to the BE and E groups. These three groups were distributed with maximum abundance values in timelines indicating retention of large number of traits from the common ancestor. This vertical trace defines an ancient stem line of descent responsible for the early origination of archaeal lineages and bacterial novelties, which reconciles the canonical and archaeal rooting of the ToL. The ahistorical analysis however was unable to predict the canonical pattern, since the comparative analysis of trait distribution in Venn taxonomic groups, superkingdoms, and organisms cannot accommodate competing hypotheses of rooting that manifest at different times in evolution. The plots of Figure 6 also revealed a marked increase in the abundance of FSFs late in eukaryal evolution, which can be explained by the remarkable development of multidomain protein structures and their associated functions [12, 49]. The combinatorics of domains and functions are the likely culprit of the biphasic patterns we observed when we focused on Eukarya.

4. Discussion

Our approach is simple (Figure 1). It does not involve computation of a sequence alignment or use of complex data matrices for phylogenetic reconstruction. Instead, it focuses on the census of molecular (structural and functional) traits in the genomes of modern cells. The fundamental principle of analysis is the use of trait distributions in Venn taxonomic groups to explain vertical evolutionary traces, the use of f-values to explain horizontal traces, and the use of trait abundance as a proxy for age. The sequential combination of these approaches dissects the most likely scenario of diversification of superkingdoms, without invoking a phylogenetic framework of analysis.

Our comparative genomic exercise shows evidence in favor of a common ancestry for cells and establishes the deep branching patterns of the ToL. The genetic complexity

of Bacteria and Eukarya hints towards a strong and ancient evolutionary association between the two superkingdoms. This association is stronger than the associations of other superkingdoms. Our findings are also compatible with an evolutionary scenario in which Archaea emerged as the first superkingdom of life by diverging from a primordial stem line of descent that originated in the urancestor [26, 28]. This line likely encountered extreme temperatures that affected its proteomic growth, hampering the acquisition of new molecular traits in those environments. Under such hostile conditions, the persistence strategy of the emerging archaeal cells was most likely survival rather than enrichment [50]. This explains why we observed the lowest number of traits in extant archaeal species. In contrast, both Bacteria and Eukarya shared a protracted coevolutionary history. Their diversification occurred well after the primordial split of Archaea from the urancestral line. Bacteria followed a path towards exploring a diverse range of habitats, which enabled high rates of gene discovery. This explains the high numbers of unique bacterial traits that are unequally distributed among bacterial species. Bacterial species also engaged in genome reductive processes and simplified their trait representations. This probably occurred well after their divergence from the primordial stem line. Finally, eukaryotes evolved by (i) increasing the abundance of ancient traits (via gene duplications and domain rearrangements), (ii) discovering novel traits, or (iii) both. These findings falsify an evolutionary scenario of first appearance of bacterial cells [2, 3] or the fusion hypotheses linked to the origin of eukaryotes (e.g., [33]), as none seems compatible with our data. However, we did not consider the roles that viruses may have played during cellular evolution. Viruses are known to contribute to the genetic diversity of cells and are believed to be very ancient [35, 51–53]. We will accomplish this task in the near future.

Genome reduction is an ongoing evolutionary process that often triggers lifestyle transitions in cells (e.g., from free-living to intracellular parasites [44]). We propose that genome streamlining played a key role in the evolution of akaryotes, especially Archaea. Our data show that the BE taxonomic group was enriched in molecular traits compared to the relatively poor representations of FSFs and GOs in the AB and AE groups (Figure 2). In fact, evolutionary timelines revealed that the BE group appeared very early in evolution and was correlated with high abundance levels of BE FSFs in the bacterial and eukaryal proteomes (Figure 6). These findings were taken as an indication of loss of traits in Archaea that occurred very early in evolution. While it can be argued that such losses could have occurred much later in archaeal lineages and after their diversification from Bacteria, our comparative and evolutionary data indicate that this may not be very likely. The loss of ancient traits late in evolution is phylogenetically costly as it implies loss of many genes and proteins that have accumulated during the course of evolution to perform a particular molecular task. In comparison, loss of ancient traits early in evolution is more parsimonious and complies with the principle of spatiotemporal continuity. An alternative explanation, however, could be the confounding effects of HGT processes. However, it was shown recently that a large number of ribosomal proteins were unevenly distributed in archaeal species [54, 55]. Because ribosomal proteins are generally refractory to HGT, their patchy and uneven distribution in archaeal lineages is better explained by differential loss from a more complex archaeal ancestor. Taken together, these findings strongly suggest that primordial reductive evolutionary processes have tailored archaeal evolution.

When placed along evolutionary timelines of trait innovation (Figure 6), Venn taxonomic groups uncovered a remarkable pattern that could not be dissected with the comparative genomic approach. This hidden pattern embodies the primordial rise of Bacteria-specific traits followed much later by the concurrent appearance of Archaea-specific and Eukarya-specific innovations. This important succession supports the "canonical" rotting of the ToL in which Bacteria occupy the most basal positions while Archaea and Eukarya emerge as derived sister-groups [2, 3]. From a cladistics perspective, traits unique to a superkingdom are autapomorphies, derived features that are unique to terminal groups. These autapomorphies cannot be used to reconstruct trees in phylogenetic analysis or dissect the alternative evolutionary scenarios of our comparative genomic approach. In comparison, FSFs and GOs that are shared by any two superkingdoms reflect synapomorphies (shared and derived features) that allow both historical (phylogenetic) and ahistorical (comparative) inferences. We note that traits uniquely shared by any two superkingdoms can arise either by the gain of the feature in two superkingdoms or by the loss in one. Abundance levels and f-distribution patterns support the latter scenario, especially if the loss involves an ancient trait. Thus, an early primordial loss of FSFs and GO synapomorphies in Archaea embeds later on the early gain of autapomorphies in Bacteria.

The hidden canonical pattern of Figure 6 was already reported in an exhaustive structural phylogenomic exploration of domain evolution at fold and FSF levels of structural abstraction [26], which prompted the definition of three epochs in the evolution of proteins and the organismal world and a number of hypotheses of origin. In the first *"architectural diversification"* epoch, the emerging organismal community accumulated a rich toolkit of protein structures and functions. This communal world resembled the ancient world of multiphenotypical precells proposed by Kandler [56] that inspired Woese's more advanced scenarios of early cellular evolution [57]. However, and in contrast with the simple cellular systems sought by Kandler and Woese, the precell molecular make up that was inferred from our phylogenomic analysis was extremely rich in complex structures and functions [29]. This richness is expressed today in the sizable number of structures and functions that are shared by all superkingdoms and are revealed by our comparative exploration. Towards the end of the architectural diversification epoch, the pervasive loss of domain structures in subgroups of the urancestral precell population resulted in primordial archaeal grades, groups of diversifying organisms in active transition that were at first unified by the physiological complexity of the urancestral community but later on gained the cellular cohesiveness needed to establish lineages and true patterns

of organismal diversification. While it may prove difficult to establish the time when these "thresholds" (sensu [57]) were crossed by the primordial archaeal grades as these were stemming from the urancestral stem line, the early process of reductive evolution left deep historical signatures in the make up of the archaeal organisms that are embedded in the timelines of domain structures [26]. The second *"superkingdom specification"* epoch brought the first Bacteria-specific domain structures and later on the concurrent appearance of Archaea-specific and Eukarya-specific structures. This canonical pattern of appearance of superkingdom-specific structures, which unfolded in the absence of early and major reductive evolutionary tendencies, signals a time in which the emerging superkingdoms were being molded by innovation. During this epoch, grades turned into clades and the precell "swap shop" strategy was gradually replaced by organismal cohesiveness. Marked decreases in f-values during this time suggested that lineage sorting occurred more frequently in the growing number of lineages. Finally, in the *"organismal diversification"* epoch, commitment to strategies and lifestyles enhanced even further the divide between superkingdoms and weakened the contribution of the stem line of descent. Two forces of particular significance play crucial roles during this final epoch, the combinatorial use of domains as modules in multidomain proteins of Eukarya [12, 49] that is responsible for the high abundance levels and the biphasic patterns of Figure 6 and the HGT-driven combinatorial exchange of protein repertoires in lineages of Bacteria [26] that minimizes trait distribution in Figure 3.

We end by emphasizing that our comparative genomic inferences have been ratified previously by phylogenetic tree reconstructions (e.g., [11–13, 17, 22, 26, 28]) and thus establish the power of our methodology. However, our analysis depends upon the accuracy and sampling of structures and functions and the reliability of the datasets. The *function* dataset, in particular, is dependent upon the stability of GO annotations and is biased towards eukaryal organisms that are more carefully annotated. To minimize this factor, we sampled 183 bacterial and 45 archaeal functionomes in comparison to only 21 eukaryotes. Despite the huge number of akaryal functionomes in our dataset, we were still able to highlight the incredible enrichment of eukaryal repertoires. Moreover, inferences drawn from *function* were in agreement with *structure* and both should be considered reliable.

While tracing back evolutionary history from the present to the first cell is a complex problem, inferring the patterns of species diversification by comparing the use and reuse of molecular traits in extant cells must be considered a robust inferencial approach that is free from many of the external assumptions and technical problems faced when reconstructing phylogenetic trees. The only shortcoming may be one of interpretation, which we here showcase with the scenarios of origin we have discussed. However, we have tried to restrict our statements to scenarios that seem most compatible with the given data. An example is using a threshold of 60% difference in the popularity of traits to detect HGT-derived structures and functions. This criterion was set arbitrarily to identify only the most likely HGT-transfers but may have resulted in failure to detect some of the true HGT-acquired traits, especially for those where both intersuperkingdom and intrasuperkingdom transfers occurred rapidly. Although such events are less likely, they may still be occurring. However, detection of such transfers is a hard problem and cannot be reliably confirmed without experimental evidence. Given the conservation levels of structural and functional traits and the relatively poor repertoire of likely HGT-acquired features (Tables S1 and S2), we safely assume that this factor did not seriously compromise our inferences. Finally, our approach is a systematic application of morphological analyses that were initially used to classify higher-order organisms. Future work should be focused on advanced applications of our approach for reaching a consensus regarding the evolution of cells.

5. Conclusions

We inferred evolutionary patterns by examining the spread of molecular features in contemporary organisms. The analysis revealed a common origin for all cells, the early divergence of Archaea, and a sister relationship between Bacteria and Eukarya. Archaeal evolution was primarily influenced by genome reduction while that of Bacteria by two contrasting phases: (i) a period of early innovation that coincides with the rise and diversification of the bacterial superkingdom, and (ii) a postdivergence period of this lineage exhibiting relatively late genome reduction events. The branch leading to modern eukaryotes was minimally affected by reductive pressure and retained the majority of the ancestral traits. Eukaryotes further enriched the genomic abundance of these traits by engaging in gene duplication and domain rearrangement processes and by discovering novel structures and molecular activities. Traces of all of these events could be reliably detected in modern proteomes and functionomes. In particular, a strong vertical trace from the urancestor to the stem line unifying Bacteria and Eukarya and the ancestor of Eukarya could be inferred. This strong vertical trace strongly supports the existence of a stem line of descent, from which all three superkingdoms emerged, very much in line with Kandler's ideas of an aboriginal precellular line of early biochemical evolution that was undergoing cellularization [56]. Finally, nonvertical evolutionary processes seemed to have played only limited roles during defining steps of cellular evolution. The comparative framework enables exploration of deep evolutionary histories without invoking tree reconstruction algorithms and external hypotheses of evolution. This approach is in line with various published phylogenetic analyses and provides strong support to theories favoring an archaeal origin of diversified life.

Acknowledgments

The authors thank Jay E. Mittenthal for valuable suggestions and Richard Egel for his review and for providing extensive thought-provoking discussion. Research was supported by grants from the National Science Foundation (MCB-0749836 and OISE-1132791) and the United States Department of Agriculture (ILLU-802-909 and ILLU-483-625) to GCA. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the paper.

References

[1] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[2] J. P. Gogarten, H. Kibak, P. Dittrich et al., "Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 17, pp. 6661–6665, 1989.

[3] N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata, "Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 23, pp. 9355–9359, 1989.

[4] S. Gribaldo and H. Philippe, "Ancient phylogenetic relationships," *Theoretical Population Biology*, vol. 61, no. 4, pp. 391–408, 2002.

[5] H. Philippe and P. Forterre, "The rooting of the universal tree of life is not reliable," *Journal of Molecular Evolution*, vol. 49, no. 4, pp. 509–523, 1999.

[6] G. Caetano-Anollés and A. Nasir, "Benefits of using molecular structure and abundance in phylogenomic analysis," *Frontiers in Genetics*, vol. 3, article 172, 2012.

[7] F. Delsuc, H. Brinkmann, and H. Philippe, "Phylogenomics and the reconstruction of the tree of life," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 361–375, 2005.

[8] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal gene transfer in prokaryotes: quantification and classification," *Annual Review of Microbiology*, vol. 55, pp. 709–742, 2001.

[9] O. Popa and T. Dagan, "Trends and barriers to lateral gene transfer in prokaryotes," *Current Opinion in Microbiology*, vol. 14, no. 5, pp. 615–623, 2011.

[10] R. Jain, M. C. Rivera, and J. A. Lake, "Horizontal gene transfer among genomes: the complexity hypothesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 7, pp. 3801–3806, 1999.

[11] G. Caetano-Anollés and D. Caetano-Anollés, "An evolutionarily structural universe of protein architecture," *Genome Research*, vol. 13, no. 7, pp. 1563–1571, 2003.

[12] M. Wang and G. Caetano-Anollés, "Global phylogeny determined by the combination of protein domains in proteomes," *Molecular Biology and Evolution*, vol. 23, no. 12, pp. 2444–2454, 2006.

[13] F. J. Sun and G. Caetano-Anollés, "Evolutionary patterns in the sequence and structure of transfer RNA: early origins of Archaea and viruses," *PLoS Computational Biology*, vol. 4, no. 3, Article ID e1000018, 2008.

[14] F. J. Sun and G. Caetano-Anollés, "The origin and evolution of tRNA inferred from phylogenetic analysis of structure," *Journal of Molecular Evolution*, vol. 66, no. 1, pp. 21–35, 2008.

[15] F. J. Sun and G. Caetano-Anollés, "The evolutionary history of the structure of 5S ribosomal RNA," *Journal of Molecular Evolution*, vol. 69, no. 5, pp. 430–443, 2009.

[16] F. J. Sun and G. Caetano-Anollés, "The ancient history of the structure of ribonuclease P and the early origins of Archaea," *BMC Bioinformatics*, vol. 11, article 153, 2010.

[17] H. Xue, K. L. Tong, C. Marck, H. Grosjean, and J. T. F. Wong, "Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life," *Gene*, vol. 310, no. 1-2, pp. 59–66, 2003.

[18] M. di Giulio, "The tree of life might be rooted in the branch leading to Nanoarchaeota," *Gene*, vol. 401, no. 1-2, pp. 108–113, 2007.

[19] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[20] A. Andreeva, D. Howorth, J. M. Chandonia et al., "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Research*, vol. 36, no. 1, pp. D419–D425, 2008.

[21] D. Caetano-Anollés, K. M. Kim, J. E. Mittenthal, and G. Caetano-Anollés, "Proteome evolution and the metabolic origins of translation and cellular life," *Journal of Molecular Evolution*, vol. 72, no. 1, pp. 14–33, 2011.

[22] G. Caetano-Anollés, M. Wang, D. Caetano-Anollés, and J. E. Mittenthal, "The origin, evolution and structure of the protein world," *Biochemical Journal*, vol. 417, no. 3, pp. 621–637, 2009.

[23] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[24] M. Harris, J. Clark, A. Ireland et al., "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, pp. D258–D261, 2004.

[25] K. M. Kim and G. Caetano-Anollés, "Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data," *Molecular Biology and Evolution*, vol. 27, no. 7, pp. 1710–1733, 2010.

[26] M. Wang, L. S. Yafremava, D. Caetano-Anollés, J. E. Mittenthal, and G. Caetano-Anollés, "Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world," *Genome Research*, vol. 17, no. 11, pp. 1572–1585, 2007.

[27] K. Illergård, D. H. Ardell, and A. Elofsson, "Structure is three to ten times more conserved than sequence—a study of structural response in protein cores," *Proteins*, vol. 77, no. 3, pp. 499–508, 2009.

[28] K. M. Kim and G. Caetano-Anollés, "The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms," *BMC Evolutionary Biology*, vol. 12, no. 1, article 13, 2012.

[29] K. M. Kim and G. Caetano-Anollés, "The proteomic complexity and rise of the primordial ancestor of diversified life," *BMC Evolutionary Biology*, vol. 11, no. 1, article 140, 2011.

[30] M. P. Hoeppner, P. P. Gardner, and A. M. Poole, "Comparative analysis of RNA families reveals distinct repertoires for each domain of life," *PLoS Computational Biology*, vol. 8, no. 11, article e1002752, 2012.

[31] G. J. Olsen, C. R. Woese, and R. Overbeek, "The winds of (evolutionary) change: breathing new life into microbiology," *Journal of Bacteriology*, vol. 176, no. 1, pp. 1–6, 1994.

[32] C. R. Woese, "Bacterial evolution," *Microbiological Reviews*, vol. 51, no. 2, pp. 221–271, 1987.

[33] M. C. Rivera and J. A. Lake, "The ring of life provides evidence for a genome fusion origin of eukaryotes," *Nature*, vol. 431, no. 7005, pp. 152–155, 2004.

[34] W. Martin and M. Müller, "The hydrogen hypothesis for the first eukaryote," *Nature*, vol. 392, no. 6671, pp. 37–41, 1998.

[35] A. Nasir, K. M. Kim, and G. Caetano-Anolles, "Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea Bacteria and Eukarya," *BMC Evolutionary Biology*, vol. 12, article 156, 2012.

[36] J. Gough and C. Chothia, "SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments," *Nucleic Acids Research*, vol. 30, no. 1, pp. 268–272, 2002.

[37] D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough, "The SUPERFAMILY database in 2007: families and functions," *Nucleic Acids Research*, vol. 35, no. 1, pp. D308–D313, 2007.

[38] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure," *Journal of Molecular Biology*, vol. 313, no. 4, pp. 903–919, 2001.

[39] S. Garcia-Vallve, E. Guzman, M. A. Montero, and A. Romeu, "HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes," *Nucleic Acids Research*, vol. 31, no. 1, pp. 187–189, 2003.

[40] J. Gough, "Convergent evolution of domain architectures (is rare)," *Bioinformatics*, vol. 21, no. 8, pp. 1464–1471, 2005.

[41] C. Moissl-Eichinger and H. Huber, "Archaeal symbionts and parasites," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 364–370, 2011.

[42] M. Wang, C. G. Kurland, and G. Caetano-Anollés, "Reductive evolution of proteomes and protein structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 29, pp. 11954–11958, 2011.

[43] D. Ungar and F. M. Hughson, "SNARE protein structure and function," *Annual Review of Cell and Developmental Biology*, vol. 19, pp. 493–517, 2003.

[44] K. Georgiades, V. Merhej, K. El Karkouri, D. Raoult, and P. Pontarotti, "Gene gain and loss events in Rickettsia and Orientia species," *Biology Direct*, vol. 6, article 6, 2011.

[45] S. Gribaldo, A. M. Poole, V. Daubin, P. Forterre, and C. Brochier-Armanet, "The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse?" *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 743–752, 2010.

[46] J. Kuriyan and M. O'Donnell, "Sliding clamps of DNA polymerases," *Journal of Molecular Biology*, vol. 234, no. 4, pp. 915–925, 1993.

[47] B. Stillman, "Smart machines at the DNA replication fork," *Cell*, vol. 78, no. 5, pp. 725–728, 1994.

[48] K. Kleman-Leyer, D. W. Armbruster, and C. J. Daniels, "Properties of H. volcanii tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems," *Cell*, vol. 89, no. 6, pp. 839–847, 1997.

[49] M. Wang and G. Caetano-Anollés, "The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world," *Structure*, vol. 17, no. 1, pp. 66–78, 2009.

[50] L. S. Yafremava, M. Wielgos, S. Thomas et al., "A general framework of persistence strategies for biological systems helps explain domains of life," *Frontiers in Genetics*, vol. 4, article 16, 2013.

[51] E. V. Koonin, T. G. Senkevich, and V. V. Dolja, "Compelling reasons why viruses are relevant for the origin of cells," *Nature Reviews Microbiology*, vol. 7, no. 8, article 615, 2009.

[52] P. Forterre, "The origin of viruses and their possible roles in major evolutionary transitions," *Virus Research*, vol. 117, no. 1, pp. 5–16, 2006.

[53] A. Nasir, K. M. Kim, and G. Caetano-Anollés, "Viral evolution: primordial cellular origins and late adaptation to parasitism," *Mobile Genetic Elements*, vol. 2, no. 5, pp. 247–252, 2012.

[54] C. Brochier-Armanet, P. Forterre, and S. Gribaldo, "Phylogeny and evolution of the Archaea: one hundred genomes later," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 274–281, 2011.

[55] E. Desmond, C. Brochier-Armanet, P. Forterre, and S. Gribaldo, "On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes," *Research in Microbiology*, vol. 162, no. 1, pp. 53–70, 2011.

[56] O. Kandler, "Cell wall biochemistry and three-domain concept of life," *Systematic and Applied Microbiology*, vol. 16, no. 4, pp. 501–509, 1994.

[57] C. R. Woese, "On the evolution of cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8742–8747, 2002.

Hindawi Publishing Corporation
Archaea
Volume 2013, Article ID 202358, 12 pages
http://dx.doi.org/10.1155/2013/202358

Review Article

Close Encounters of the Third Domain: The Emerging Genomic View of Archaeal Diversity and Evolution

Anja Spang, Joran Martijn, Jimmy H. Saw, Anders E. Lind, Lionel Guy, and Thijs J. G. Ettema

Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, P.O. Box 596, 75123 Uppsala, Sweden

Correspondence should be addressed to Thijs J. G. Ettema; thijs.ettema@icm.uu.se

Received 24 July 2013; Accepted 21 September 2013

Academic Editor: Gustavo Caetano-Anollés

Copyright © 2013 Anja Spang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Archaea represent the so-called Third Domain of life, which has evolved in parallel with the Bacteria and which is implicated to have played a pivotal role in the emergence of the eukaryotic domain of life. Recent progress in genomic sequencing technologies and cultivation-independent methods has started to unearth a plethora of data of novel, uncultivated archaeal lineages. Here, we review how the availability of such genomic data has revealed several important insights into the diversity, ecological relevance, metabolic capacity, and the origin and evolution of the archaeal domain of life.

1. Introduction

The description of the three (cellular) domains of life—Eukarya, Bacteria, and Archaea—by Carl Woese and George Fox [1] represents a milestone in the modern era of microbiology. In particular, using phylogenetic reconstructions of the small-subunit (16S or 18S) ribosomal RNA gene, Woese discovered that microscopically indistinguishable prokaryotes are not a homogeneous assemblage but are comprised of two fundamentally different groups of organisms: Eubacteria (later Bacteria) on one side and an additional life form referred to as Archaebacteria (later Archaea) on the other side [1]. Though not immediately accepted by the scientific community, this finding was early on supported by Wolfram Zillig through his studies on DNA-dependent RNA polymerases, as well as by Otto Kandler investigating "bacterial" cell walls [2]. Indeed, a subset of prokaryotic organisms subsequently assigned to Archaea was found to harbor DNA-dependent RNA polymerases that bore more similarity to those of eukaryotes, and to contain proteinaceous cell walls that lack peptidoglycan as well as cell membranes composed of L-glycerol ether lipids with isoprenoid chains instead of D-glycerol ester lipids with fatty acid chains [3–6]. Since then, further investigation of cellular characteristics of archaea has revealed that this domain of life contains eukaryotic-like information-processing machineries [7–14]. These findings were later supported by genome sequences and comparative analyses of genes coding for replication, transcription, and translation machineries as well as by protein crystal structures [15–21]. Additionally, some archaeal lineages were shown to contain homologs of eukaryotic cell division and cytoskeleton genes as well as histones and seem to express a chromatin architecture similar to eukaryotes [22–28]. In contrast to information-processing and cell division genes, archaeal operational systems (energy metabolism, biosynthesis pathways, and regulation) often appear to be more closely related to bacteria [29].

Based on phylogenetic reconstructions of the evolutionary history of 16S rRNA genes, the domain Archaea was originally divided into two major phyla: the Euryarchaeota and Crenarchaeota [30], which were separated by a deep split and thought to comprise only extremophilic (thermophilic, halophilic, and acidophilic) as well as methanogenic organisms. However, novel culture-independent and high-throughput sequencing techniques have recently uncovered a huge diversity of so far uncharacterized microorganisms on Earth as well as the ubiquitous occurrence of archaeal species [31–33]. Many of these novel archaeal groups are responsible for important ecological

processes and are only distantly related to established lineages within Cren- and Euryarchaeota [31, 32, 34–39]. For example, the acquisition of genome sequences from novel archaeal representatives has led to the proposal of several additional archaeal phyla (including Nanoarchaeota, Korarchaeota, Thaumarchaeota, Aigarchaeota, and Geoarchaeota) [40–46] and the investigation of uncultivated archaea using single cell genomics has already started to add new insights into the phylogenetic diversity of the Third Domain of life and necessitates the definition of additional lineages of high taxonomic rank including novel potential phyla and superphyla [33, 39] (see also below). Furthermore, the investigation of the metabolic potential of these novel organisms has provided fundamentally new insights into major biogeochemical nutrient cycles. Indeed, archaea are now recognized as key players in various biogeochemical processes [47]. For example, the perception of the global nitrogen cycle has been deeply altered by discovering that the ability to gain energy solely from ammonia was not limited to a few bacteria but also included the ammonia-oxidizing Thaumarchaeota [48, 49]. Archaea also appear to play a significant role in the carbon cycle, since, in addition to all known methanogenic organisms on Earth, they also encompass anaerobic methane oxidizing archaea (ANME lineages 1–3) [50].

The study of archaeal genomes and diversity is also of considerable importance for a better understanding of eukaryotic evolution. Indeed, the discovery of eukaryotic features in archaea [10] has initiated a new basis for addressing the origin of eukaryotes [51–54]. Interestingly, recent phylogenetic analyses of universal proteins have suggested that eukaryotes might have evolved from a *bona fide* archaeal lineage that forms a sister-lineage of or a lineage emerging from within the TACK-superphylum comprised of Thaum-, Aig-, Cren-, and Korarchaeota [55–58].

Below we give a contemporary overview of how recent developments in archaeal genomic research have contributed to revealing new insights into the diversity, ecological relevance, metabolic capacity, and the origin and evolution of the archaeal domain of life.

2. The Methanogenic Nature of Archaea

The scientific community that addressed questions about prokaryotic energy metabolism on the early Earth or in hydrothermal vent systems [59] has proposed that methanogenesis and/or acetogenesis most likely represent ancient metabolic pathways [60–62]. Evidence for the biological production of methane as early as 3.46 Gyr ago supports these scenarios [63]. However, phylogenetic evidence placing methanogens at the base of the archaeal tree is limited and disputed. Depending on the outgroup and phylogenetic methods used, many recent analyses find either members of the Nanohaloarchaea, Nanoarchaeota, ARMAN-lineages, and/or *Thermococcales* as earliest (eury-)archaeal branches [55, 64, 65]. The latter observation is consistent with results from a base and amino acid composition analysis, which indicated that last archaeal common ancestor (LACA) was a hyperthermophilic organism [66]. The placement of *Methanopyrus kandleri* as the most basal branch of archaea in some of the earliest phylogenetic analyses can most likely be attributed to long-branch attraction (LBA) artifacts [67]. Notably, in recent phylogenetic analyses that include novel archaeal single cell genomes, Euryarchaeota form a sister group to other archaeal phyla rather than representing an early diverging lineage (Figure 1) [33]. Furthermore, gene content comparisons of extant archaeal lineages and reconstruction of the putative genetic repertoire of the LACA do not support methanogenesis as the earliest archaeal metabolism [57, 68]. In contrast, only one study has so far placed the root of archaea within a methanogenic order [69] and thus favors a methanogenic origin of the Third Domain of life. Gene content comparisons and network analyses that include novel archaeal single-cell amplified genomes (SAGs) Could potentially help to further investigate the metabolic gene repertoire of the archaeal ancestor.

Whereas the origin of methanogenic pathways that include a multitude of specific genes and cofactors is not fully resolved yet [72], it appears that several later emerging euryarchaeal lineages have lost their methanogenic lifestyles. Thus, as already noted more than a decade ago, methanogens comprise a paraphyletic group separated by nonmethanogenic euryarchaeal lineages such as the *Thermoplasmatales*, Haloarchaeota, and *Archaeoglobales* [73]. Interestingly, a novel methanogenic archaeal lineage has been described recently that is distantly affiliated with cultivated *Thermoplasmatales* including *Aciduliprofundum* sp. [74, 75]. This suggests that the last common ancestor of *Thermoplasmatales* was a methanogen and the capability to reduce methane has been independently lost several times along some branches within this group [76] or, albeit less likely, that some lineages within the *Thermoplasmatales* have regained genes for methane production.

A single acquisition of a plethora of genes (>1000) from a bacterial donor has recently been put forward as explanation for the transition from a methanogenic ancestor to aerobic heterotrophic Haloarchaeota [77]. A possible driving force for this massive gene transfer might have been a syntrophic relationship between a methanogenic recipient and a bacterial donor. However, the exact donor lineage could not be determined: the acquired genes bear conflicting phylogenetic signals, supposedly due to prevalent gene transfers between different bacterial species. So far, the origin of alternative energy metabolisms in other non-methanogenic euryarchaeal lineages that evolved from methanogenic ancestors has not been addressed properly. However, comparative genomics suggests that several of these lineages have retained specific genes that trace back to the methanogenic nature of their ancestor (e.g., *Archaeoglobus*) [78] and might point to a rather transient transition.

3. Phylogeny of New Archaeal Phyla and Lineages

In recent years, several new archaeal lineages have been identified and subjected to whole genome or metagenomic

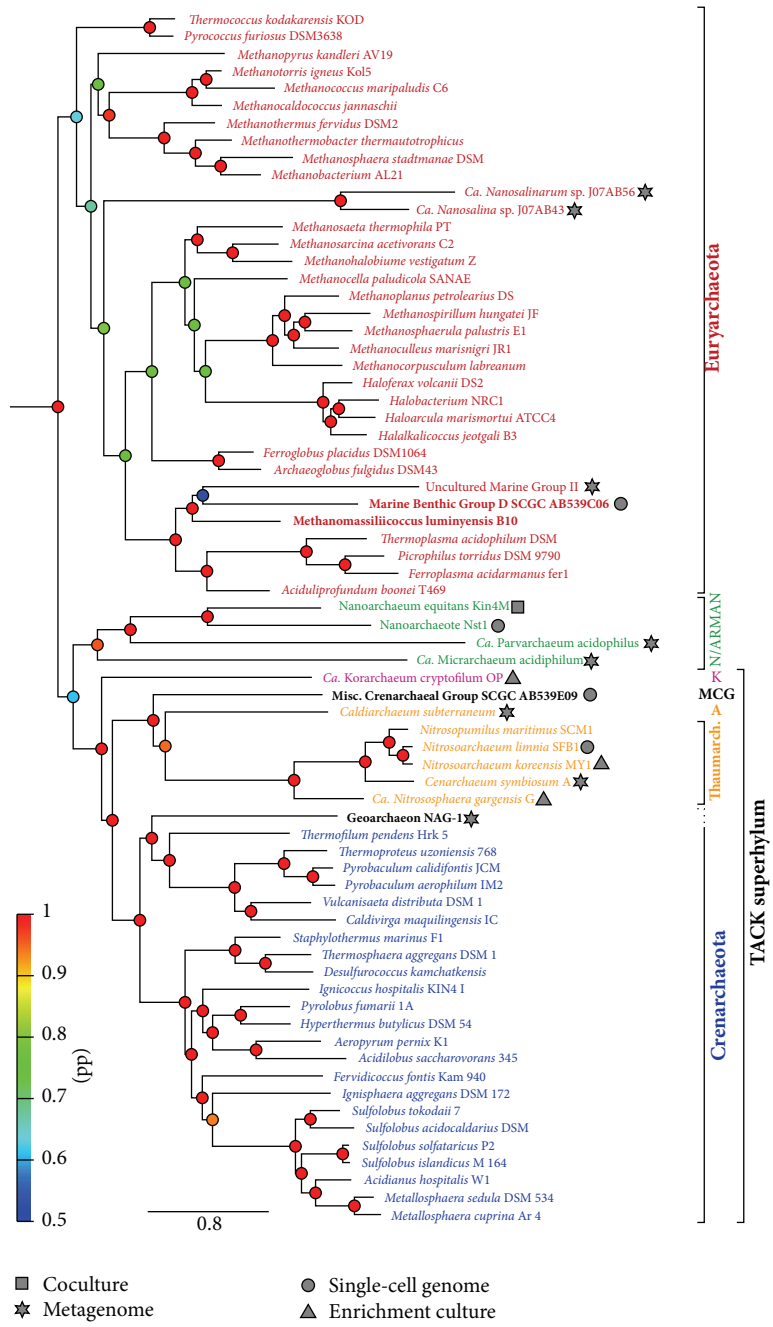


Figure 1: Bayesian phylogeny of 80 representative archaeal species. BLAST databases containing the proteome of 6 new archaeal genomes were retrieved from NCBI (in bold font on the tree): *Methanomassiliicoccus luminyensis* B10 (acc. no. CAJE01), MCG SCGC AB539E09 (acc. no. ALXK01), Marine Benthic Group D (MBGD) SCGC AB539N05, AB539C06, and AB540F20 (acc. no. ALXL01, AOSH01, and AOSI01, resp.). Protein sequence alignment from the 57 clusters in the discFilter 15 p dataset from [70] for which eukaryotes were removed were used as an input to psi-blast, with the six new proteomes as a database. Orthologs were retrieved as in [70]. For the three MBGD strains, one composite set of orthologs was constituted by using the most complete one (AB539C06) whenever possible and complementing with sequences from the other two if available. Orthologous genes selection, alignment, trimming, and concatenation were performed as in [70] resulting in a 15,069 amino-acid alignment. Four chains of Bayesian phylogenies were run with Phylobayes [71], under the CAT-Poisson model, running for approximately 10000 generations and discarding half as a burn-in. The tree was rooted with bacteria. Posterior probabilities (pp) are represented by colored dots on the nodes, with support values coloured according to the depicted heat-map colour scheme. The scale represents the number of substitutions per site. Species are colored according to the following: red, Euryarchaeota; green, Nanoarchaeota (N) and ARMAN; pink, Korarchaeota (K); black, Misc. Crenarchaeal Group (MCG); orange, Thaumarchaeota and Aigarchaeota (A); blue, Crenarchaeota. The DNA collection method, if different from pure culture, is indicated by a symbol next to the organism name: square, coculture; star, metagenome; circle, single-cell genome; triangle, enrichment culture.

sequencing. Based on phylogenetic analyses of available genomic data, some of these lineages have been proposed to represent novel archaeal phyla. Yet, some of these claims have been challenged or falsified in follow-up studies. Below, we give an overview of several such examples.

The candidate phylum Nanoarchaeota has initially been proposed on basis of the extremely divergent 16S rRNA sequence of the small parasitic cells of *Nanoarchaeum equitans* growing attached to the cell surface of *Ignicococcus hospitalis* [41]. Several subsequent and more comprehensive phylogenetic analyses as well as the finding of potentially ancestral genomic features (e.g., split tRNA genes) have provided support for the initial assignment of this tiny archaeal cells to a separate ancient archaeal phylum [65, 79, 80]. Yet, in contrast, other phylogenetic and comparative analyses testing the taxonomic position of *N. equitans* have suggested that Nanoarchaeota might rather represent a fast-evolving euryarchaeal lineage related to *Thermococcales* [81]. Genomic data from additional "nanosized" archaea (*Ca.* Parvarchaeum acidophilus ARMAN-4 and *Ca.* Micrarchaeum acidiphilum ARMAN-2) [82] as well as of a novel deep-branching member of Nanoarchaeota (Nst1) [83] have enabled a revision of phylogenetic reconstructions and genome comparisons. Although some of these analyses suggest that Nanoarchaeota and *Ca.* Parvarchaeum acidophilus are monophyletic, the placement of these groups in the archaeal tree remains unclear and is strongly dependent on dataset and phylogenetic methods used [64, 83]. For example, in our phylogenetic reconstructions Nanoarchaeota (including ARMAN-lineages) represent a sister clade of the TACK superphylum (Figure 1), although the support for this clade is low. In a recent study by Rinke et al. [33], the Nanoarchaeota (including all ARMAN strains) were grouped together in the newly proposed superphylum DPANN with two novel groups, DSEG and pMC2A384 (designated "Aenigmarchaeota" and "Diapherotrites", resp.), as well as the Nanohaloarchaea (see also below). Given that the phylogenetic methods employed by Rinke and coworkers do not accommodate rate heterogeneity across taxa, the proposed grouping of Nanoarchaeota with these archaeal clades has to be taken with care and the exact position of Nanoarchaeota still remains an unresolved question.

The Nanohaloarchaea represent yet another archaeal lineage comprised of small cells and with unresolved phylogenetic position. Based on both 16S rRNA gene and concatenated ribosomal protein phylogenies, this group was suggested to comprise a deep lineage of Haloarchaeota [84]. However, only euryarchaeal sequences were included in these maximum-likelihood (ML) analyses. Depending on the phylogenetic method and evolutionary model used, we obtained contradictory results for the phylogenetic position of this group. Whereas ML analyses tend to recover Nanohaloarchaea as earliest branching archaeal lineage (e.g., see above), a phylogenetic reconstruction using Bayesian methods (and the CAT model [71]) place this lineage within Euryarchaeota, but the exact position could not be resolved with high confidence (Figure 1). Results obtained with Bayesian methods using the CAT model might provide a better approximation of the position of Nanohaloarchaea, as this model accounts for rate variations across sites. As such, the early divergence of Nanohaloarchaea that is observed in ML-based methods is likely caused by LBA artifacts. However, novel phylogenetic analyses including the improved archaeal taxon sampling of Rinke et al. suggest that Nanohaloarchaea form a distinct lineage within the proposed superphylum DPANN and are not closely related to Euryarchaeota [33].

It will be interesting to further address the position of these organisms in the archaeal tree to be able to elucidate whether the adaptation to halophily has evolved only once in archaea or is due to convergence in Halo- and Nanohaloarchaea. The latter has received initial support from comparative genome analyses, which have revealed that each of these two archaeal groups seems to harbor diverse unique features including distinctive amino acid compositions to accommodate high salt conditions [84]. It might also be of value to address the effect of these novel genome sequences on the results obtained in the analysis of Nelson-Sathi et al. studying the origin of Haloarchaeota from a methanogenic ancestor [77].

Another novel archaeal phylum comprises the abundant and ecologically important ammonia-oxidizing archaea (AOA). On the basis of comparative genomics and phylogenetic analyses based on concatenated ribosomal proteins that were rooted with eukaryotes, Brochier-Armanet and coworkers proposed that "mesophilic crenarchaeota" constitute the novel deep branching archaeal phylum Thaumarchaeota [42, 85]. Additional comprehensive phylogenetic analyses including additional members of this group, as well as the discovery of a distinctive set of informational processing genes involved in replication, transcription, and translation as well as DNA repair and cell division machineries, have provided further support for the independent status of the Thaumarchaeota [44]. For example, in contrast to Crenarchaeota, Thaumarchaeota share several characteristics with Euryarchaeota and Korarchaeota including the presence of DNA polymerase D, histones, and cell division protein FtsZ. Furthermore, they contain putative "ancestral" features absent from Cren- or Euryarchaeota but common in Bacteria and eukaryotes (e.g., presence of an unsplit gene encoding DNA polymerase subunit A, and toposimerase IB as well as the absence of ribosomal protein LXa) [20, 44, 85]. The distinct nature of Thaumarchaeota has been accepted by many authors [45, 46, 86, 87] although the taxonomic borders of this phylum are still difficult to delineate and might only be resolved when genomes of uncultivated early branching lineages are made available. The early emergence of Thaumarchaeota in these phylogenetic reconstructions using eukaryotes as outgroup was initially assumed to indicate the ancient nature of this phylum [42, 44]. However, several recent phylogenetic analyses have recovered a monophyletic group of Thaum-, Aig-, Cren, Korarchaeota, and eukaryotes (with varying relationships in between these groups) to the exclusion of Euryarchaeota, which indicates that eukaryotes emerge from within the Archaea [55, 56, 88]. Thus, eukaryotes cannot be used as valid outgroup for the rooting of archaeal phylogenies [54].

Another lineage that emerges as a separate branch in the archaeal tree is comprised of the so-called Hot Water

Crenarchaeotic Group I (HWCG I), members of which have been detected in diverse hydrothermal environments but have not yet been cultivated [89, 90]. Until recently, the sole representative with a sequenced genome in this group was *Ca.* Caldiarchaeum subterraneum, whose composite genome has been obtained from a metagenomic library of a microbial mat in a subsurface geothermal water stream [45]. The investigation of its genome sequence has revealed the presence of components of the eukaryotic ubiquitin-like protein modifier system previously not detected in archaea or bacteria. This unique trait, as well as comparative genomics and phylogenetic analyses of concatenated protein sequences, suggested that this organism and other members of HWCG I might constitute a novel phylum (Aigarchaeota), distinct from both Thaum- and Crenarchaeota [45]. However, due to the presence of a set of informational processing genes most similar to Thaumarchaeota [45] and the highly supported monophyletic grouping of these two lineages in diverse phylogenetic analyses (e.g., see Figure 1), the separation of Thaum- and Aigarchaeota into two distinct phyla is still debated [45, 55, 56, 64, 91, 92].

Uncultivated archaea belonging to the so-called Miscellaneous Crenarchaeotal Group (MCG) (e.g., [39]) have been suggested to represent additional members of Aigarchaeota [55]. Recently, the first single-cell genome of a member of this group has been obtained and phylogenetic analyses of concatenated conserved single copy genes placed the MCG-archaeon as a lineage in between Thaum- and Aigarchaeota [97]. However, our analyses rather suggest that MCG emerges prior to the Thaum/Aigarchaeota (Figure 1). The availability of additional genome sequences of members of this group as well as the comparison of informational processing marker genes [44] of MCG-archaea with other available archaeal genomes might help both to resolve their phylogenetic placement and to determine whether MCG-archaea comprise a separate archaeal phylum [39].

Geoarchaeota represents yet another recently proposed archaeal phylum, which is proposed to emerge as a basal lineage of Crenarchaeota and includes the so-called novel archaeal group I (NAG-1) detected in acidic ferric iron mats from Yellowstone National Park [46, 98]. NAG-1 organisms thrive in hot (60–78°C) acidic mats rich in iron and are suggested to grow heterotrophically from simple carbon compounds. Though not yet enriched in culture, nearly full-length genome sequences of members of this group have been obtained from a *de novo* metagenome assembly. The description of this lineage as a separate phylum was based on phylogenetic analyses of concatenated ribosomal proteins and 16S/23S rRNA genes as well as on its specific set of informational processing genes with features in common with either Crenarchaeota or Thaum- and Aigarchaeota [46]. However, our analyses, based on a larger dataset, place Geoarchaeota as an early branching lineage of the crenarchaeal order Thermoproteales (Figure 1). This observation is confirmed by Rinke et al., who sequenced six additional NAG-1-related strains [33]. Indeed, detailed phylogenetic analyses, as well as comparative assessment of the NAG-1 composite genome, seem to refute the phylum-level status of NAG-1 (Guy, L., Spang, A., Saw, J.H. and Ettema, T.J.G., unpublished observation).

4. Archaea and the Origin of Eukaryotes

The origin of the eukaryotes remains one of the major unanswered questions in modern biology, and archaea have recently reclaimed the spotlights in heated discussions entailing this enigmatic event. A central issue in this discussion entails the placement of the root within the Tree of Life, as it has a fundamental effect on any hypothesis on the origin of eukaryotes. Whereas diverse competing hypotheses have been put forward in the past, no consensus has been reached on this topic so far. For instance, several studies, including a recent network analysis, place the root between Archaea and Bacteria [99–104]. This view is in agreement with both the observed fundamental differences distinguishing the bacterial and archaeal domains as well as with the geological record. In contrast, studies that were based on transmission analyses or the distribution of indels in protein sequences suggested a rooting within the bacterial domain [105–107], whereas a root in the archaeal domain has been proposed based on analyses of protein folds or the evolution of the tRNA molecules [108, 109]. Yet other hypotheses state that LUCA was a eukaryotic-like organism [110, 111]. Certainly, in order to reach a consensus on this controversial discussion, additional data and analyses are needed. Bearing the uncertainty of the placement of the root in the Tree of Life in mind, we will present current hypotheses on the origin of eukaryotes below, by providing a short review on the most commonly proposed scenarios.

Even though a wide variety of incompatible theories have been suggested regarding the origin of the eukaryotic cell, three aspects are now largely accepted: (i) the last eukaryotic common ancestor (LECA) contained mitochondria, (ii) eukaryotic genomes are chimeric; whereas informational genes are of archaeal descent, many metabolic genes are derived from Bacteria, and (iii) eukaryotes complement a set of proteins not found in either Archaea or Bacteria, the eukaryotic signature proteins (ESPs). Beyond this, the picture becomes blurry. Currently, two major questions are of interest. What was the nature of the cell that was host in the mitochondrial endosymbiosis and when did cellular complexity evolve, before (complexity-first) or after (mitochondria-first) mitochondrial endosymbiosis? From this perspective, theories on eukaryogenesis can be divided into two categories. In the first scenario, the host was a protoeukaryote and complexity evolved first. This theory, often referred to as the "archezoa hypothesis" [112, 113], fits with the three domains tree of life model in which eukaryotes vertically evolved from the archaea-eukaryote common ancestor (Figure 2(a)). In the second scenario, the host was a prokaryote and the acquisition of the mitochondria likely triggered the evolution of cellular complexity. The latter are often referred to as "fusion" hypotheses [51, 58, 93, 95, 96] and these are generally incompatible with the classical three domains model. Rather, in these models, Bacteria and Archaea are the primary domains of life and eukaryotes a secondary, or derived, domain of life (Figure 2(b)). Theories that

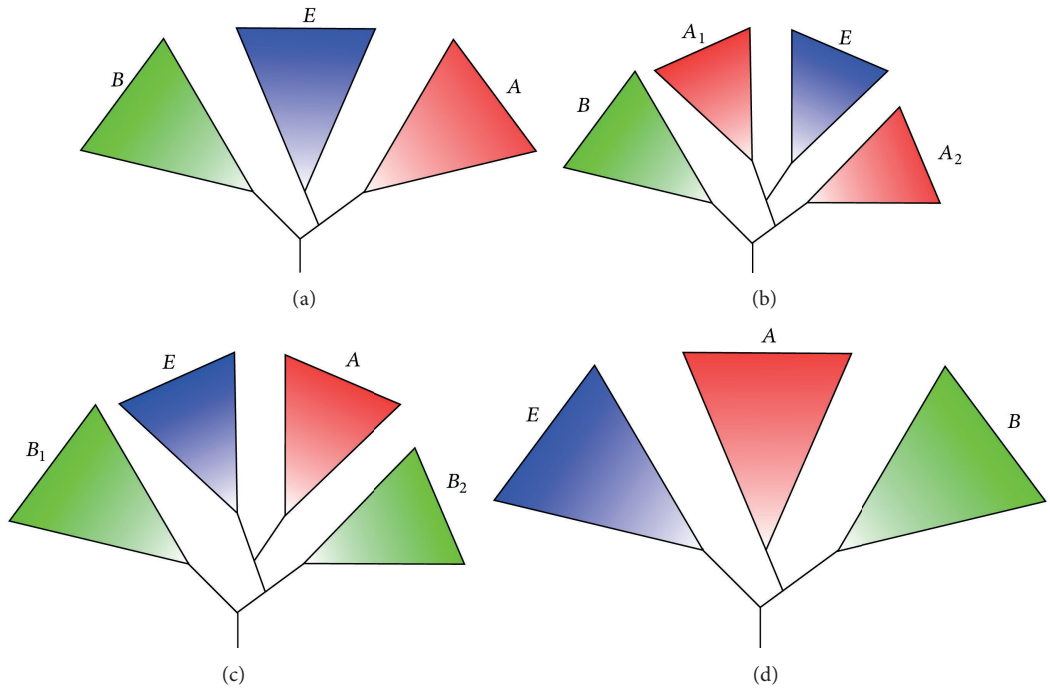


Figure 2: Overview of theories regarding the origin of the eukaryotic nuclear lineage. (a) The classical, Woesean three domains of life tree in which the nuclear lineage vertically evolved from the archaea-eukaryote common ancestor. (b) The fusion tree in which the nuclear lineage originated from the archaeal partner in the fusion. Depending on which fusion model, the archaeal parent's lineage (A_1) was either part of the euryarchaeota [SET [93], original syntrophy hypothesis [94], hydrogen hypothesis [95] or alternative syntrophy hypothesis [51]], the Crenarchaeota (eocyte hypothesis) [96], or the TACK superphylum (PhAT) [58]. "A_2" represents all archaea not directly affiliated with "A_1." (c) The neomuran tree in which the eukaryotic and archaeal lineage (combined referred to as "neomurans"), evolved vertically from ancestor shared with actinobacteria (B_2) as a result of the loss of bacterial-type cell wall (the neomuran revolution). B_1 represents all bacteria not directly affiliated with B_2. (d) The eukaryote-early tree, which suggests that the last common universal ancestor was more eukaryote-like than prokaryote-like.

fit neither of these categories exist as well. These include the neomuran hypothesis [114] (Figure 2(c)), the PVC hypothesis [115–118], virus-assisted eukaryogenesis [119–122], and a hypothesis suggesting a eukaryote-like universal common ancestor [110] (Figure 2(d)). In order to choose the correct category with high confidence, evidence is needed in the form of protoeukaryote intermediate lineage's descendants ("missing links"). Unfortunately, for either category, none has been found so far. Whereas the archezoa theory has lost much support ever since remnants of mitochondria were found in the previously thought archezoa (for review, see [123]), the fusion theory has slowly been gaining favor. Initially lightly supported by ribosomal structural features [124] and an 11-amino acid insertion in EF-1α/EF-Tu [125, 126] shared between eocytes (Crenarchaeota) and eukaryotes to the exclusion of other prokaryotes, it has now received strong support from phylogenomic [55, 56, 70, 88, 127, 128] and gene similarity network analyses [129]. In addition, a large number of ESPs has been found in Archaea, in particular within the recently proposed TACK superphylum [55]. Examples include actin [53, 130], tubulin [28], H3/H4-type histones [55], ESCRT-III [24, 25, 131], and components of the ubiquitin modifier system [45]. Fusion models can be subdivided based upon the nature of the end-product of the "fusion". In amitochondriate models the symbiosis results in a eukaryotic progenitor lacking mitochondria. They are similar to the archezoa theory in the sense that the origin of eukaryotes and the origin of mitochondria are separate events. These include the serial endosymbiosis theory (SET) [93], the original syntrophy hypothesis [94], and the eocyte hypothesis [96]. In mitochondriate models, the end product is a eukaryotic progenitor containing mitochondria. Here, the origin of eukaryotes and mitochondria are one and the same. These include the hydrogen hypothesis [95], the alternative syntrophy hypothesis [51] and the recently proposed phagocytosing archaeon theory [53, 58]. With exception of the eocyte hypothesis, all fusion theories suggest an archaeal host. Based on extensive, in-depth phylogenomic studies, the archaeal host most likely emerged from within the TACK superphylum [55, 56, 70]. Interestingly, out of all TACK phyla, a sister relationship between the Korarchaeota and eukaryotes was retrieved with significant phylogenetic support [56, 70]. Even though this placement could be a taxon sampling artifact (Korarchaeota are represented by a single, deep rooting taxon), it could also indicate that eukaryotes are affiliated with an unidentified lineage distantly related to Korarchaeota. Genomically unexplored lineages such as DSAG (Deep Sea Archaea Group), MHVG (Marine

Hydrothermal Vent Group), and AAG (Ancient Archaea Group) are likely candidates [55, 70].

5. Genomic Assessment and Taxonomic Classification of Archaeal Diversity

Recent progress in genomic sequencing technologies and cultivation-independent methods has started to unearth a plethora of novel, uncultivated archaeal lineages. The availability of such genomic data has revealed several important insights into the diversity, ecological relevance, metabolic capacity, and the origin and evolution of the archaeal domain of life. Several new archaeal lineages have been obtained by means of metagenomics approaches, such as sequencing of enrichment cultures or environmental samples. Examples of the former include the first korarchaeal genome [43] and several of the available thaumarchaeal genomes (e.g. [132, 133]). Archaeal genomes that have been retrieved from metagenomic datasets include the first thaumarchaeal genome (*Ca.* Cenarchaeum symbiosum [134]), the genome of the proposed Aigarchaeon *Ca.* C. subterraneum [45], the proposed Geoarchaeon NAG-1 [46], representatives of the Nanohaloarchaea [84], several ARMAN lineages that were part of an acid mine drainage microbial community [82], and a genome derived from a representative of the uncultivated marine group II euryarchaeota [135] (Figure 1). More recently, a number of studies have employed single cell genomic approaches to probe the genetic diversity of uncultivated archaea. For example, Lloyd and coworkers have reported the first genomic data of a representative of the Miscellaneous Crenarchaeal Group (MCG) and of members of the Marine Benthic Group D that were isolated from marine sediments and speculate that these lineages are involved in the degradation of detrital proteins [97] (Figure 1). Another large scale study that aimed at uncovering the coding potential of so-called "microbial dark matter" using single cell genomics approaches reported several genome sequences of cells that potentially represented novel phylum-level archaeal lineages, including the members of the uncultured DSEG and pMC2A384 clades, designated Aenigmarchaeota and Diapherotrites, respectively [33]. A combination of single cell genomics and metagenomics has been used to sequence the genome of the thaumarchaeon *Ca.* Nitrosoarchaeum limnia SFB1 [136].

Obviously, single cell and metagenomics-oriented projects will continue to probe the existing archaeal diversity during the coming years, and most likely, the availability of genomic data will reveal interesting insights into novel characteristics and the diversity within the Third Domain of life. In addition, the availability of such genomic data is likely to trigger discussions regarding the higher-order taxonomic classification of the major archaeal lineages. To many (micro-)biologists, it would appear that the archaeal domain is far less diverse than the bacterial domain. A reason for this could be, for instance, the discrepancy in assigned or proposed phyla, which ranges from a handful in Archaea, to well over a hundred in Bacteria. But is it really fair to say that the bacterial domain of life is more diverse than that of the Archaea? Whereas bacterial phyla generally have been assigned based on the diversity of the 16S rDNA gene sequence, archaeal taxonomy is largely founded on historic grounds, that is, adhering to the classical Cren-Euryarchaeota dichotomy (sensu Woese [30]). Only during the past decade, a handful of additional archaeal phyla have been proposed based on genome sequencing, such as the Nano-, Kor-, and Thaumarchaeota and a few other lineages that may or may not represent phylum-level archaeal clades (also see above). Yet, the majority of archaeal species that have been sequenced in recent years have been assigned to the phyla Cren- or Euryarchaeota, each of which now comprise genetically distinct groups, which differ in terms of metabolic capacity, lifestyle, and environmental distribution. In light of this and of the abovementioned "superficial" imbalance in bacterial versus archaeal diversity, one could argue that a revision of archaeal higher-order taxonomy is in place. The suggestion to bring order into archaeal systematics was recently put forward [92], but thus far, a framework as to how novel phyla and/or superphyla should be defined is debated. Nevertheless, to be able to fully appreciate the overall archaeal diversity and compare it to the diversity observed within the bacterial domain of life, a reappraisal of the archaeal taxonomy, whether it will be at the level of rRNA genes, large datasets of concatenated protein sequences, genome content, or gene networks analyses, seems to be a *conditio sine qua non*.

Abbreviations

ANME: Anaerobic methanotrophic archaea
LACA: Last archaeal common ancestor
LBA: Long-branch attraction
MCG: Miscellaneous Crenarchaeotal Group
ML: Maximum-likelihood
SAGs: Single-cell amplified genomes.

Conflicts of Interests

The authors declare that they do not have a direct financial relation with the trademarks mentioned in the paper that might lead to a conflict of interests for the authors.

Acknowledgments

The work in Ettema laboratory is supported by the Swedish Research Council (Grant no. 621-2009-4813), by the European Research Council (ERC) (Grant no. 310039-PUZZLE_CELL), by a Marie Curie European Reintegration Grant (ERG) (Grant no. 268259-RICKOCHET), and by the Carl Tryggers Stiftelse (Grant no. CTS11:127).

References

[1] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.

[2] C. R. Woese, "The birth of the Archaea: a personal retrospective," in *Archaea*, R. A. Garrett and H.-P. Klenk, Eds., Blackwell Publishing, 2007.

[3] T. A. Langworthy, P. F. Smith, and W. R. Mayberry, "Lipids of Thermoplasma acidophilum," *Journal of Bacteriology*, vol. 112, no. 3, pp. 1193–1200, 1972.

[4] T. A. Langworthy, W. R. Mayberry, and P. F. Smith, "Long chain glycerol diether and polyol dialkyl glycerol triether lipids of Sulfolobus acidocaldarius," *Journal of Bacteriology*, vol. 119, no. 1, pp. 106–116, 1974.

[5] O. Kandler and H. Hippe, "Lack of peptidoglycan in the cell walls of Methanosarcina barkeri," *Archives of Microbiology*, vol. 113, no. 1-2, pp. 57–60, 1977.

[6] W. Zillig, K. O. Stetter, and D. Janekovic, "DNA-dependent RNA polymerase from the archaebacterium *Sulfolobus acidocaldarius*," *European Journal of Biochemistry*, vol. 96, no. 3, pp. 597–604, 1979.

[7] J. Huet, R. Schnabel, A. Sentenac, and W. Zillig, "Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type," *EMBO Journal*, vol. 2, no. 8, pp. 1291–1294, 1983.

[8] J. A. Lake, "Ribosome evolution: the structural bases of protein synthesis in archaebacteria, eubacteria, and eukaryotes," *Progress in Nucleic Acid Research and Molecular Biology*, vol. 30, pp. 163–194, 1983.

[9] E. Henderson, M. Oakes, M. W. Clark, J. A. Lake, A. T. Matheson, and W. Zillig, "A new ribosome structure," *Science*, vol. 225, no. 4661, pp. 510–512, 1984.

[10] W. Zillig, R. Schnabel, and K. O. Stetter, "Archaebacteria and the origin of the eukaryotic cytoplasm," *Current Topics in Microbiology and Immunology*, vol. 114, pp. 1–18, 1985.

[11] G. Puhler, H. Leffers, F. Gropp et al., "Archaebacterial DNA-dependent RNA polymerase testify to the evolution of the eukaryotic nuclear genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 12, pp. 4569–4573, 1989.

[12] K. Sandman, J. A. Krzycki, B. Dobrinski, B. Lurz, and J. N. Reeve, "HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon Methanothermus fervidus, is most closely related to histones," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 15, pp. 5788–5791, 1990.

[13] D. Langer, J. Hain, P. Thuriaux, and W. Zillig, "Transcription in archaea: similarity to that in eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 13, pp. 5768–5772, 1995.

[14] O. Kandler and H. König, "Cell wall polymers in Archaea (Archaebacteria)," *Cellular and Molecular Life Sciences*, vol. 54, no. 4, pp. 305–308, 1998.

[15] G. J. Olsen and C. R. Woese, "Lessons from an Archaeal genome: what are we learning from Methanococcus jannaschii?" *Trends in Genetics*, vol. 12, no. 10, pp. 377–379, 1996.

[16] O. Lecompte, R. Ripp, J.-C. Thierry, D. Moras, and O. Poch, "Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale," *Nucleic Acids Research*, vol. 30, no. 24, pp. 5382–5390, 2002.

[17] E. R. Barry and S. D. Bell, "DNA replication in the archaea," *Microbiology and Molecular Biology Reviews*, vol. 70, no. 4, pp. 876–887, 2006.

[18] S. Gribaldo and C. Brochier-Armanet, "The origin and evolution of Archaea: a state of the art," *Philosophical Transactions of the Royal Society B*, vol. 361, no. 1470, pp. 1007–1022, 2006.

[19] L. L. Grochowski, H. Xu, and R. H. White, "Methanocaldococcus jannaschii uses a modified mevalonate pathway for biosynthesis of isopentenyl diphosphate," *Journal of Bacteriology*, vol. 188, no. 9, pp. 3192–3198, 2006.

[20] M. Kwapisz, F. Beckouët, and P. Thuriaux, "Early evolution of eukaryotic DNA-dependent RNA polymerases," *Trends in Genetics*, vol. 24, no. 5, pp. 211–215, 2008.

[21] F. Werner and D. Grohmann, "Evolution of multisubunit RNA polymerases in the three domains of life," *Nature Reviews Microbiology*, vol. 9, no. 2, pp. 85–98, 2011.

[22] S. L. Pereira and J. N. Reeve, "Histones and nucleosomes in Archaea and Eukarya: a comparative analysis," *Extremophiles*, vol. 2, no. 3, pp. 141–148, 1998.

[23] J. N. Reeve, K. A. Bailey, W.-T. Li, F. Marc, K. Sandman, and D. J. Soares, "Archaeal histones: structures, stability and DNA binding," *Biochemical Society Transactions*, vol. 32, no. 2, pp. 227–230, 2004.

[24] A.-C. Lindås, E. A. Karlsson, M. T. Lindgren, T. J. G. Ettema, and R. Bernander, "A unique cell division machinery in the Archaea," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 48, pp. 18942–18946, 2008.

[25] R. Y. Samson, T. Obita, S. M. Freund, R. L. Williams, and S. D. Bell, "A role for the ESCRT system in cell division in archaea," *Science*, vol. 322, no. 5908, pp. 1710–1713, 2008.

[26] R. Bernander, A. E. Lind, and T. J. G. Ettema, "An archaeal origin for the actin cytoskeleton," *Communicative & Integrative Biology*, vol. 4, pp. 664–667, 2011.

[27] R. Ammar, D. Torti, K. Tsui et al., "Chromatin is an ancient innovation conserved between Archaea and Eukarya," *Elife*, vol. 1, Article ID e00078, 2012.

[28] N. Yutin and E. V. Koonin, "Archaeal origin of tubulin," *Biology Direct*, vol. 7, article 10, 2012.

[29] M. C. Rivera, R. Jain, J. E. Moore, and J. A. Lake, "Genomic evidence for two functionally distinct gene classes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 11, pp. 6239–6244, 1998.

[30] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[31] C. Schleper, G. Jurgens, and M. Jonuscheit, "Genomic studies of uncultivated archaea," *Nature Reviews Microbiology*, vol. 3, no. 6, pp. 479–488, 2005.

[32] A. Teske and K. B. Sørensen, "Uncultured archaea in deep marine subsurface sediments: have we caught them all?" *The ISME Journal*, vol. 2, no. 1, pp. 3–18, 2008.

[33] C. Rinke, P. Schwientek, A. Sczyrba et al., "Insights into the phylogeny and coding potential of microbial dark matter," *Nature*, 2013.

[34] J. O. McInerney, M. Wilkinson, J. W. Patching, T. M. Embley, and R. Powell, "Recovery and phylogenetic analysis of novel archaeal rRNA sequences from a deep-sea deposit feeder," *Applied and Environmental Microbiology*, vol. 61, no. 4, pp. 1646–1648, 1995.

[35] E. F. Delong, "Everything in moderation: Archaea as 'non-extremophiles'," *Current Opinion in Genetics & Development*, vol. 8, no. 6, pp. 649–654, 1998.

[36] K. Takai and K. Horikoshi, "Genetic diversity of archaea in deep-sea hydrothermal vent environments," *Genetics*, vol. 152, no. 4, pp. 1285–1297, 1999.

[37] B. J. Baker, G. W. Tyson, R. I. Webb et al., "Lineages of acidophilic archaea revealed by community genomic analysis," *Science*, vol. 314, no. 5807, pp. 1933–1935, 2006.

[38] B. Chaban, S. Y. M. Ng, and K. F. Jarrell, "Archaeal habitats—from the extreme to the ordinary," *Canadian Journal of Microbiology*, vol. 52, no. 2, pp. 73–116, 2006.

[39] K. Kubo, K. G. Lloyd, J. F Biddle, R. Amann, A. Teske, and K. Knittel, "Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments," *The ISME Journal*, vol. 6, pp. 1949–1965, 2012.

[40] S. M. Barns, C. F. Delwiche, J. D. Palmer, and N. R. Pace, "Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 17, pp. 9188–9193, 1996.

[41] H. Huber, M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter, "A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont," *Nature*, vol. 417, no. 6884, pp. 63–67, 2002.

[42] C. Brochier-Armanet, B. Boussau, S. Gribaldo, and P. Forterre, "Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota," *Nature Reviews Microbiology*, vol. 6, no. 3, pp. 245–252, 2008.

[43] J. G. Elkins, M. Podar, D. E. Graham et al., "A korarchaeal genome reveals insights into the evolution of the Archaea," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 23, pp. 8102–8107, 2008.

[44] A. Spang, R. Hatzenpichler, C. Brochier-Armanet et al., "Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota," *Trends in Microbiology*, vol. 18, no. 8, pp. 331–340, 2010.

[45] T. Nunoura, Y. Takaki, J. Kakuta et al., "Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group," *Nucleic Acids Research*, vol. 39, no. 8, pp. 3204–3223, 2011.

[46] M. A. Kozubal, M. Romine, R. Jennings et al., "Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park," *The ISME Journal*, vol. 7, pp. 622–634, 2013.

[47] P. Offre, A. Spang, and C. Schleper, "Archaea in biogeochemical cycles," *Annual Review of Microbiology*, vol. 67, pp. 437–457, 2013.

[48] M. Könneke, A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl, "Isolation of an autotrophic ammonia-oxidizing marine archaeon," *Nature*, vol. 437, no. 7058, pp. 543–546, 2005.

[49] A. H. Treusch, S. Leininger, A. Kietzin, S. C. Schuster, H.-P. Klenk, and C. Schleper, "Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling," *Environmental Microbiology*, vol. 7, no. 12, pp. 1985–1995, 2005.

[50] K. Knittel and A. Boetius, "Anaerobic oxidation of methane: progress with an unknown process," *Annual Review of Microbiology*, vol. 63, pp. 311–334, 2009.

[51] P. López-Garćia and D. Moreira, "Metabolic symbiosis at the origin of eukaryotes," *Trends in Biochemical Sciences*, vol. 24, pp. 88–93, 1999.

[52] T. M. Embley and W. Martin, "Eukaryotic evolution, changes and challenges," *Nature*, vol. 440, no. 7084, pp. 623–630, 2006.

[53] N. Yutin, M. Y. Wolf, Y. I. Wolf, and E. V. Koonin, "The origins of phagocytosis and eukaryogenesis," *Biology Direct*, vol. 4, article 9, 2009.

[54] S. Gribaldo, A. M. Poole, V. Daubin, P. Forterre, and C. Brochier-Armanet, "The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse?" *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 743–752, 2010.

[55] L. Guy and T. J. G. Ettema, "The archaeal "TACK" superphylum and the origin of eukaryotes," *Trends in Microbiology*, vol. 19, no. 12, pp. 580–587, 2011.

[56] T. A. Williams, P. G. Foster, T. M. W. Nye, C. J. Cox, and T. M. Embley, "A congruent phylogenomic signal places eukaryotes within the Archaea," *Proceedings of the Royal Society B*, vol. 279, no. 1749, pp. 4870–4879, 2012.

[57] Y. I. Wolf, K. S. Makarova, N. Yutin, and E. V. Koonin, "Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer," *Biology Direct*, vol. 7, article 46, 2012.

[58] J. Martijn and J. G. Ettema Thijs, "From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell," *Biochemical Society Transactions*, vol. 41, pp. 451–457, 2013.

[59] W. Martin, J. Baross, D. Kelley, and M. J. Russell, "Hydrothermal vents and the origin of life," *Nature Reviews Microbiology*, vol. 6, no. 11, pp. 805–814, 2008.

[60] N. Lane and W. F. Martin, "The origin of membrane bioenergetics," *Cell*, vol. 151, pp. 1406–1416, 2012.

[61] Y. Liu, L. L. Beer, and W. B. Whitman, "Methanogens: a window into ancient sulfur metabolism," *Trends in Microbiology*, vol. 20, no. 5, pp. 251–258, 2012.

[62] A. Poehlein, S. Schmidt, A.-K. Kaster et al., "An ancient pathway combining carbon dioxide fixation with the generation and utilization of a sodium ion gradient for ATP synthesis," *PLoS ONE*, vol. 7, no. 3, Article ID e33439, 2012.

[63] Y. Ueno, K. Yamada, N. Yoshida, S. Maruyama, and Y. Isozaki, "Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era," *Nature*, vol. 440, no. 7083, pp. 516–519, 2006.

[64] C. Brochier-Armanet, P. Forterre, and S. Gribaldo, "Phylogeny and evolution of the Archaea: one hundred genomes later," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 274–281, 2011.

[65] N. Yutin, P. Puigbò, E. V. Koonin, and Y. I. Wolf, "Phylogenomics of prokaryotic ribosomal proteins," *PLoS ONE*, vol. 7, no. 5, Article ID e36972, 2012.

[66] M. Groussin and M. Gouy, "Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea," *Molecular Biology and Evolution*, vol. 28, no. 9, pp. 2661–2674, 2011.

[67] C. Brochier, P. Forterre, and S. Gribaldo, "Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox," *Genome Biology*, vol. 5, no. 3, p. R17, 2004.

[68] K. S. Makarova, A. V. Sorokin, P. S. Novichkov, Y. I. Wolf, and E. V. Koonin, "Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea," *Biology Direct*, vol. 2, article 33, 2007.

[69] S. Kelly, B. Wickstead, and K. Gull, "Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes," *Proceedings of the Royal Society B*, vol. 278, no. 1708, pp. 1009–1018, 2011.

[70] L. Guy, J. H. Saw, and T. J. G. Ettema, "The archaeal legacy of eukaryotes: a phylogenomic perspective," *Cold Spring Harbor Perspectives in Biology*. In press.

[71] N. Lartillot, T. Lepage, and S. Blanquart, "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating," *Bioinformatics*, vol. 25, no. 17, pp. 2286–2288, 2009.

[72] J. G. Ferry, "How to make a living by exhaling methane," *Annual Review of Microbiology*, vol. 64, pp. 453–473, 2010.

[73] O. Matte-Tailliez, C. Brochier, P. Forterre, and H. Philippe, "Archaeal phylogeny based on ribosomal proteins," *Molecular Biology and Evolution*, vol. 19, no. 5, pp. 631–639, 2002.

[74] B. Dridi, M.-L. Fardeau, B. Ollivier, D. Raoult, and M. Drancourt, "Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces," *International Journal of Systematic and Evolutionary Microbiology*, vol. 62, pp. 1902–1907, 2012.

[75] K. Paul, J. O. Nonoh, L. Mikulski, and A. Brune, "'Methanoplasmatales,' Thermoplasmatales-related archaea in termite guts and other environments, are the seventh order of methanogens," *Applied and Environmental Microbiology*, vol. 78, pp. 8245–8253, 2012.

[76] G. Borrel, P. W. O'Toole, H. M. B. Harris, P. Peyret, J.-F. Brugère, and S. Gribaldo, "Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis," *Genome Biology and Evolution*, 2013.

[77] S. Nelson-Sathi, T. Dagan, G. Landan et al., "Acquisition of 1, 000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 20537–20542, 2012.

[78] B. Schworer, J. Breitung, A. R. Klein, K. O. Stetter, and R. K. Thauer, "Formylmethanofuran: tetrahydromethanopterin formyltransferase and N5,N10-methylenetetrahydromethanopterin dehydrogenase from the sulfate-reducing Archaeoglobus fulgidus: similarities with the enzymes from methanogenic Archaea," *Archives of Microbiology*, vol. 159, no. 3, pp. 225–232, 1993.

[79] E. Waters, M. J. Hohn, I. Ahel et al., "The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 22, pp. 12984–12988, 2003.

[80] M. Di Giulio, "Formal proof that the split genes of tRNAs of nanoarchaeum equitans are an ancestral character," *Journal of Molecular Evolution*, vol. 69, no. 5, pp. 505–511, 2009.

[81] C. Brochier, S. Gribaldo, Y. Zivanovic, F. Confalonieri, and P. Forterre, "Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales?" *Genome Biology*, vol. 6, no. 5, p. R42, 2005.

[82] B. J. Baker, L. R. Comolli, G. J. Dick et al., "Enigmatic, ultrasmall, uncultivated Archaea," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 19, pp. 8806–8811, 2010.

[83] M. Podar, K. S. Makarova, D. E. Graham, Y. I. Wolf, E. V. Koonin, and A.-L. Reysenbach, "Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park," *Biology Direct*, vol. 8, article 9, 2013.

[84] P. Narasingarao, S. Podell, J. A. Ugalde et al., "De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities," *The ISME Journal*, vol. 6, no. 1, pp. 81–93, 2012.

[85] C. Brochier-Armanet, S. Gribaldo, and P. Forterre, "A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya," *Biology Direct*, vol. 3, article 54, 2008.

[86] R. S. Gupta and A. Shami, "Molecular signatures for the Crenarchaeota and the Thaumarchaeota," *Antonie van Leeuwenhoek*, vol. 99, no. 2, pp. 133–157, 2011.

[87] M. Pester, C. Schleper, and M. Wagner, "The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 300–306, 2011.

[88] C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, and T. M. Embley, "The archaebacterial origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 51, pp. 20356–20361, 2008.

[89] T. Nunoura, H. Hirayama, H. Takami et al., "Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments," *Environmental Microbiology*, vol. 7, no. 12, pp. 1967–1984, 2005.

[90] T. Nunoura, H. Oida, M. Nakaseama et al., "Archaeal diversity and distribution along thermal and geochemical gradients in hydrothermal sediments at the yonaguni knoll IV hydrothermal field in the Southern Okinawa Trough," *Applied and Environmental Microbiology*, vol. 76, no. 4, pp. 1198–1211, 2010.

[91] C. Brochier-Armanet, S. Gribaldo, and P. Forterre, "Spotlight on the Thaumarchaeota," *The ISME Journal*, vol. 6, no. 2, pp. 227–230, 2012.

[92] S. Gribaldo and C. Brochier-Armanet, "Time for order in microbial systematics," *Trends in Microbiology*, vol. 20, no. 5, pp. 209–210, 2012.

[93] L. Margulis, M. Chapman, R. Guerrero, and J. Hall, "The last eukaryotic common ancestor (LECA): acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 35, pp. 13080–13085, 2006.

[94] D. Moreira and P. López-García, "Symbiosis between methanogenic archaea and δ-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis," *Journal of Molecular Evolution*, vol. 47, no. 5, pp. 517–530, 1998.

[95] W. Martin and M. Müller, "The hydrogen hypothesis for the first eukaryote," *Nature*, vol. 392, no. 6671, pp. 37–41, 1998.

[96] J. A. Lake, "Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences," *Nature*, vol. 331, no. 6152, pp. 184–186, 1988.

[97] K. G. Lloyd, L. Schreiber, D. G. Petersen et al., "Predominant archaea in marine sediments degrade detrital proteins," *Nature*, vol. 496, pp. 215–218, 2013.

[98] M. A. Kozubal, R. E. Macur, Z. J. Jay et al., "Microbial iron cycling in acidic geothermal springs of yellowstone national park: integrating molecular surveys, geochemical processes, and isolation of novel fe-active microorganisms," *Frontiers in Microbiology*, vol. 3, article 109, 2012.

[99] J. P. Gogarten, H. Kibak, P. Dittrich et al., "Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 17, pp. 6661–6665, 1989.

[100] N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata, "Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 23, pp. 9355–9359, 1989.

[101] S. Gribaldo and P. Cammarano, "The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery," *Journal of Molecular Evolution*, vol. 47, no. 5, pp. 508–516, 1998.

[102] O. Zhaxybayeva, P. Lapierre, and J. P. Gogarten, "Ancient gene duplications and the root(s) of the tree of life," *Protoplasma*, vol. 227, no. 1, pp. 53–64, 2005.

[103] B. Boussau, S. Blanquart, A. Necsulea, N. Lartillot, and M. Gouy, "Parallel adaptations to high temperatures in the Archaean eon," *Nature*, vol. 456, no. 7224, pp. 942–945, 2008.

[104] T. Dagan, M. Roettger, D. Bryant, and W. Martin, "Genome networks root the tree of life between prokaryotic domains," *Genome Biology and Evolution*, vol. 2, no. 1, pp. 379–392, 2010.

[105] T. Cavalier-Smith, "Rooting the tree of life by transition analyses," *Biology Direct*, vol. 1, article 19, 2006.

[106] J. A. Lake, R. G. Skophammer, C. W. Herbold, and J. A. Servin, "Genome beginnings: rooting the tree of life," *Philosophical Transactions of the Royal Society B*, vol. 364, no. 1527, pp. 2177–2185, 2009.

[107] T. Cavalier-Smith, "Deep phylogeny, ancestral groups and the four ages of life," *Philosophical Transactions of the Royal Society B*, vol. 365, pp. 111–132, 2010.

[108] M. Di Giulio, "The tree of life might be rooted in the branch leading to Nanoarchaeota," *Gene*, vol. 401, no. 1-2, pp. 108–113, 2007.

[109] K. M. Kim and G. Caetano-Anollés, "The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms," *BMC Evolutionary Biology*, vol. 12, no. 1, article 13, 2012.

[110] C. G. Kurland, L. J. Collins, and D. Penny, "Genomics and the irreducible nature of eukaryote cells," *Science*, vol. 312, no. 5776, pp. 1011–1014, 2006.

[111] N. Glansdorff, Y. Xu, and B. Labedan, "The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner," *Biology Direct*, vol. 3, article 29, 2008.

[112] T. Cavalier-Smith, "Molecular phylogeny. Archaebacteria and Archezoa," *Nature*, vol. 339, no. 6220, pp. l00–01, 1989.

[113] A. M. Poole and D. Penny, "Evaluating hypotheses for the origin of eukaryotes," *BioEssays*, vol. 29, no. 1, pp. 74–84, 2007.

[114] T. Cavalier-Smith, "The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification," *International Journal of Systematic and Evolutionary Microbiology*, vol. 52, no. 1, pp. 7–76, 2002.

[115] D. P. Devos and E. G. Reynaud, "Evolution. Intermediate steps," *Science*, vol. 330, no. 6008, pp. 1187–1188, 2010.

[116] P. Forterre and S. Gribaldo, "Bacteria with a eukaryotic touch: a glimpse of ancient evolution?" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 29, pp. 12739–12740, 2010.

[117] P. Forterre, "A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong," *Research in Microbiology*, vol. 162, no. 1, pp. 77–91, 2011.

[118] E. G. Reynaud and D. P. Devos, "Transitional forms between the three domains of life and evolutionary implications," *Proceedings of the Royal Society B*, vol. 278, no. 1723, pp. 3321–3328, 2011.

[119] P. J. Livingstone Bell, "Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus?" *Journal of Molecular Evolution*, vol. 53, no. 3, pp. 251–256, 2001.

[120] M. Takemura, "Poxviruses and the origin of the eukaryotic nucleus," *Journal of Molecular Evolution*, vol. 52, no. 5, pp. 419–425, 2001.

[121] P. Forterre, "The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells," *Biochimie*, vol. 87, no. 9-10, pp. 793–803, 2005.

[122] P. Forterre, "Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 10, pp. 3669–3674, 2006.

[123] T. M. Embley, "Multiple secondary origins of the anaerobic lifestyle in eukaryotes," *Philosophical Transactions of the Royal Society B*, vol. 361, no. 1470, pp. 1055–1067, 2006.

[124] J. A. Lake, E. Henderson, M. Oakes, and M. W. Clark, "Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, no. 12, pp. 3786–3790, 1984.

[125] S. L. Baldauf, J. D. Palmer, and W. F. Doolittle, "The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 15, pp. 7749–7754, 1996.

[126] T. Hashimoto and M. Hasegawa, "Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G," *Advances in Biophysics*, vol. 32, pp. 73–120, 1996.

[127] E. Lasek-Nesselquist and J. P. Gogarten, "The effects of model choice and mitigating bias on the ribosomal tree of life," *Molecular Phylogenetics and Evolution*, vol. 69, pp. 17–38, 2013.

[128] P. G. Foster, C. J. Cox, and T. Martin Embley, "The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods," *Philosophical Transactions of the Royal Society B*, vol. 364, no. 1527, pp. 2197–2207, 2009.

[129] D. Alvarez-Ponce, P. Lopez, E. Bapteste, and J. O. McInerney, "Gene similarity networks provide tools for understanding eukaryote origins and evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 1594–1603, 2013.

[130] T. J. G. Ettema, A.-C. Lindås, and R. Bernander, "An actin-based cytoskeleton in archaea," *Molecular Microbiology*, vol. 80, no. 4, pp. 1052–1061, 2011.

[131] T. J. G. Ettema and R. Bernander, "Cell division and the ESCRT complex: a surprise from the archaea," *Communicative and Integrative Biology*, vol. 2, no. 2, pp. 86–88, 2009.

[132] B. K. Kim, M.-Y. Jung, D. S. Yu et al., "Genome sequence of an ammonia-oxidizing soil archaeon, "Candidatus Nitrosoarchaeum koreensis" MY1," *Journal of Bacteriology*, vol. 193, no. 19, pp. 5539–5540, 2011.

[133] A. Spang, A. Poehlein, P. Offre et al., "The genome of the ammonia-oxidizing Candidatus Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations," *Environmental Microbiology*, vol. 14, pp. 3122–3145, 2012.

[134] S. J. Hallam, K. T. Konstantinidis, N. Putnam et al., "Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 48, pp. 18296–18301, 2006.

[135] V. Iverson, R. M. Morris, C. D. Frazar, C. T. Berthiaume, R. L. Morales, and E. V. Armbrust, "Untangling genomes

from metagenomes: revealing an uncultured class of marine euryarchaeota," *Science*, vol. 335, no. 6068, pp. 587–590, 2012.

[136] P. C. Blainey, A. C. Mosier, A. Potanina, C. A. Francis, and S. R. Quake, "Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis," *PLoS ONE*, vol. 6, no. 2, Article ID e16626, 2011.

Hindawi Publishing Corporation
Archaea
Volume 2013, Article ID 372396, 18 pages
http://dx.doi.org/10.1155/2013/372396

Review Article

The Common Ancestor of Archaea and Eukarya Was Not an Archaeon

Patrick Forterre[1,2]

[1] *Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France*
[2] *Université Paris-Sud, Institut de Génétique et Microbiologie, CNRS UMR 8621, 91405 Orsay Cedex, France*

Correspondence should be addressed to Patrick Forterre; forterre@pasteur.fr

Received 22 July 2013; Accepted 24 September 2013

Academic Editor: Gustavo Caetano-Anollés

Copyright © 2013 Patrick Forterre. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

It is often assumed that eukarya originated from archaea. This view has been recently supported by phylogenetic analyses in which eukarya are nested within archaea. Here, I argue that these analyses are not reliable, and I critically discuss archaeal ancestor scenarios, as well as fusion scenarios for the origin of eukaryotes. Based on recognized evolutionary trends toward reduction in archaea and toward complexity in eukarya, I suggest that their last common ancestor was more complex than modern archaea but simpler than modern eukaryotes (the bug in-between scenario). I propose that the ancestors of archaea (and bacteria) escaped protoeukaryotic predators by invading high temperature biotopes, triggering their reductive evolution toward the "prokaryotic" phenotype (the thermoreduction hypothesis). Intriguingly, whereas archaea and eukarya share many basic features at the molecular level, the archaeal mobilome resembles more the bacterial than the eukaryotic one. I suggest that selection of different parts of the ancestral virosphere at the onset of the three domains played a critical role in shaping their respective biology. Eukarya probably evolved toward complexity with the help of retroviruses and large DNA viruses, whereas similar selection pressure (thermoreduction) could explain why the archaeal and bacterial mobilomes somehow resemble each other.

1. Introduction

Archaea have been confused with bacteria, under the term prokaryotes, until their originality was finally recognized by 16S rRNA cataloguing [1]. Archaea were previously "hidden before our eyes", strikingly resembling bacteria under the light and electron microscopes. Archaea and bacteria are also quite similar at the genomic level, with small circular genomes, compact gene organization, and functionally related genes organized into operons. At the same time, archaea, unlike bacteria, exhibit a lot of "eukaryotic features" at the molecular level [2–6]. It is often assumed that archaea resemble eukarya when their informational systems (DNA replication, transcription, and translation) are considered but resemble bacteria in terms of their operational systems. This is clearly not the case, since many archaeal operational systems (such as ATP production, protein secretion, cell division and vesicles formation, and protein modification machinery) also use proteins that have only eukaryotic homologues or that are more similar to their eukaryotic rather than to their bacterial homologues [7–14]. The bacterial-like features of some archaeal metabolic pathways could be mostly due to lateral gene transfer (LGT) of bacterial genes into Archaea, driven by their cohabitation in various biotopes [15]. Indeed, beside bacterial-like genes possibly recruited by LGT, metabolic pathways in archaea—such as the coenzyme A or the isoprenoid biosynthetic pathways—also involve a mixture of archaea-specific and eukaryotic-like enzymes [16–18]. Archaea and eukarya share so many features in all aspects of their cellular physiology and molecular fabric that eukaryotes cannot be simply envisioned as a mosaic of archaeal and bacterial features. Archaea and eukarya clearly share a more complex evolutionary relationship that remains to be understood.

Whereas many eukaryotic traits of archaea are ubiquitous or widely distributed in that domain, recent discoveries have identified several new eukaryotic traits that are only present in one phylum, one order, or even in one species of archaea [6, 11, 14]. Phylogenetic analyses suggest that these traits were already present in the last archaeal common

ancestor (LACA) since, in most cases, archaeal and eukaryal sequences form two well separated monophyletic groups [12, 13, 19]. This indicates that these traits have not been sporadically acquired from eukarya by lateral gene transfer but were lost in most members of the archaeal domain after their divergence from LACA [6]. Considering this loss of eukaryotic traits and the gain of bacterial traits by LGT, LACA was probably even more "eukaryotic-like" than modern archaea. However, despite their eukaryotic affinity, archaea lack many eukaryote-specific features (ESFs) at the cellular and/or molecular levels. These, for example, include the spliceosome, mRNA capping, and extensive polyadenylation as well as huge transcriptional machineries with unique components, such as the mediator, endoplasmic reticulum, and derived structures such as lysosomes, the Golgi apparatus and the nuclear membrane, an elaborated cytoskeleton and associated vesicle trafficking system with endosomes and ectosomes, nuclear pores, nucleolus and other nucleus-specific structures, linear chromosomes with centromeres and telomeres, mitosis and associated chromosome segregation system linked to the cytoskeleton, complex and great sex with meiosis derived from mitosis, an incredible machinery for cell division apparatus with synaptonemal complex for meiosis, centrioles and midbodies for cell division, and I probably miss some of them. Archaea not only lack all these ESFs but also lack homologues of most proteins (a few hundreds) that are involved in building and operating them [20]. This remains true even if a few ex-ESPs (e.g., actin, tubulin, and DNA topoisomerase IB) have recently lost this status following the discovery of archaeal homologues [12, 13, 19]. The number, diversity, and complexity of ESFs are impressive and their origin remains a major evolutionary puzzle that should not be underestimated. The puzzle became of even greater magnitude when it was realized during the last decade from phylogenomic analyses that all ESFs (and associated ESPs) were most likely already present in the last common ancestor of all modern eukaryotes, (*the last eukaryotic common ancestor*) (LECA) [20]. In a recent review, Martijn and Ettema called the period that experienced the emergence of ESFs (so before LECA): *"the evolutionary dark ages of eukaryotic cells"* [21]. This denomination well illustrates the complexity of the "complexity problem" in eukaryotic evolution.

Besides lacking (by definition) all ESFs, archaea also fundamentally differ from eukarya in the nature of their membranes (with a unique type of lipids in archaea), and the type of viruses infecting them. The problems raised by the evolution of membranes have been nicely reviewed recently by Lombard et al. and I will refer to their work later on to discuss different models for the origin of archaea [22]. In contrast, the problem raised by the drastic differences between archaeal and eukaryotic viruses has never been really discussed. For instance, Martijn and Ettema never mentioned the word virus in their review on the origin of eukaryotes [21]. Viruses are also completely absent from the papers of Cavalier-Smith or Carl Woese himself. This is probably because, as recently stated by Koonin and Wolf, "*viruses are no part of the traditional narrative of evolutionary biology*" [23].

2. The Bacterial Flavour of Archaeal Viruses and Plasmids: Another Evolutionary Puzzle

Viruses infecting archaea have fascinated for a long time scientists that are aware of their existence by the amazing morphologies of their virions that, in most cases, differ drastically from those produced by bacterioviruses (formerly bacteriophages) or eukaryoviruses [24, 25]. Among the 13–15 families of archeoviruses presently known, most are unique to archaea, and none of them is specifically related to a family of eukaryoviruses. The only archaeal viruses with eukaryovirus relatives are the archeoviruses STIV (*Sulfolobus islandicus* turreted virus) (see below) and Caudovirales, which belong to major lineages of viruses infecting members from the three cellular domains [26]. STIV is the archaeal member of the PRD1/adenovirus lineage that groups bacterial *Tectiviridae* (a group of small membrane-containing viruses resembling STIV) with large DNA viruses infecting eukaryotes, such as adenoviruses and Megavirales (formerly nucleocytoplasmic large DNA viruses, NCLDV) as well as the recently discovered satellite viruses (virophages) of giant Megavirales. Viruses of this lineage are characterized by major capsid proteins containing the so-called double jelly-roll fold. Archaeal and bacterial Caudovirales (head and tailed viruses) belong to the same viral lineage as eukaryoviruses of the family Herpesviridae. Their virions are constructed from the major capsid proteins displaying the so-called Hong-Kong 97 fold (structurally unrelated to the jelly-roll fold). Strikingly, the archaeal viruses in these two lineages are much more similar in virion size and overall structure to their bacterial than to their eukaryotic counterparts. In particular, archaeal and bacterial Caudovirales are identical in terms of virion morphology and genome organization and share several homologous proteins [27]. The three families of Caudovirales (Siphoviridae, Myoviridae, and Podoviridae) first described in bacteria have been now found in archaea [25, 27, 28]. Moreover, Caudovirales were recently found to be more widespread than previously thought among archaea, suggesting that Caudovirales already existed when archaea and bacteria started to diverge from each other [29]. Finally, a recently described family of archaeal pleomorphic viruses, pleolipoviruses, could be related to bacterial pleomorphic viruses of the family Plasmaviridae [30]. In summary, whereas archaea and eukarya share basic molecular biology features for all major ancestral cellular functions, the archaeal virosphere shares much more similarities with the bacterial one than with the eukaryotic one.

Beside common viruses, archaea and bacteria also share similar types of plasmids, insertion sequences (IS), and related transposons [25, 27–35]. In particular, Filée and coworkers were surprised by their observation that *"most of the archaeal ISs fall into family found in bacteria"* and that "*archaeal ISs resemble bacterial ISs rather than those identified in eukaryotes*" [31]. Furthermore, they detected no IS elements in archaeal genomes with significant similarity to the nine known superfamilies of eukaryotic DNA transposons [31]. Plasmids are abundant, diverse in size, and widespread in archaea, as in bacteria, contrasting with the paucity of plasmids in eukaryotes. Moreover, archaeal ISs and plasmids

use bacterial-like proteins for transposition, plasmid resolution, and segregation [31]. Some of these proteins are only present in archaea and bacteria, and when they are universal (for instance initiator proteins for rolling circle replication) the archaeal version is more similar to the bacterial version than to the eukaryotic one [36, 37]. The bacterial affinity of archaeal viruses and plasmids confirms that these mobile elements are evolutionarily related, with plasmids probably being derived from ancient viral lineages [38]. Interestingly, viruses and plasmids encode many proteins that are involved in both bacterial and archaeal chromosome segregation and resolution, such as tyrosine recombinases of the XerCD/XerA family or ATPases of the ParA/SegA family [37, 39]. One can wonder if the similarity between archaeal and bacterial viruses/plasmids could explain the presence in many archaea of bacterial-like proteins involved in chromosome resolution/segregation. This would fit with a provocative scenario in which I suggested that the archaeal and bacterial chromosomes evolved from large DNA plasmids, with divergent replication mechanisms but homologous partition machineries, themselves derived from giant DNA viruses [40]. Finally, it is striking that archaea and bacteria use homologous defence systems against plasmids and viruses (CRISPR, toxin-antitoxin and restriction-modification systems) that are very divergent from the siRNA interference defence systems used by eukaryotes [41, 42]. Homologues of argonaute proteins, the core component of the eukaryotic interference system, have been detected in archaea and bacteria, but it is not yet known if these proteins are involved in an interference pathway [42, 43]. All these observations raise major unresolved questions: why so many archaeal mobile elements (head and tailed viruses and plasmids) are similar to bacterial ones, whereas archaea and eukarya are so similar in terms of molecular biology? Why, on the other hand, so many viruses infecting archaea are unique, having neither bacterial nor eukaryotic counterparts? A good theory for the origin of archaea and their relationships with eukarya should definitely explain these puzzling observations.

3. Different Scenarios for the Origin of Archaea and Eukarya

Several scenarios are in competition to explain the origin of archaea and eukarya [20–22, 44–52]. The most popular presently are the fusion scenarios in which eukarya originated by the intimate association of an archaeon and a bacterium ([48, 49], reviewed in [46]; for a more recent hypothesis see [21]). In these scenarios, the fusion is triggered by the engulfment of one of the two partners (the endosymbiont) by the other (the host). This association is followed by a dramatic reorganization of the structures of the two partners (the fusion), promoting the emergence of a completely new type of cell (eukaryote instead of prokaryote). Several propositions have been made concerning the origin of the two partners (one archaeon and one bacterium) involved. The proposed scenarios also differ by the timing of the mitochondrial endosymbiosis. In some of them, this event takes place after the fusion [48], in others it corresponds

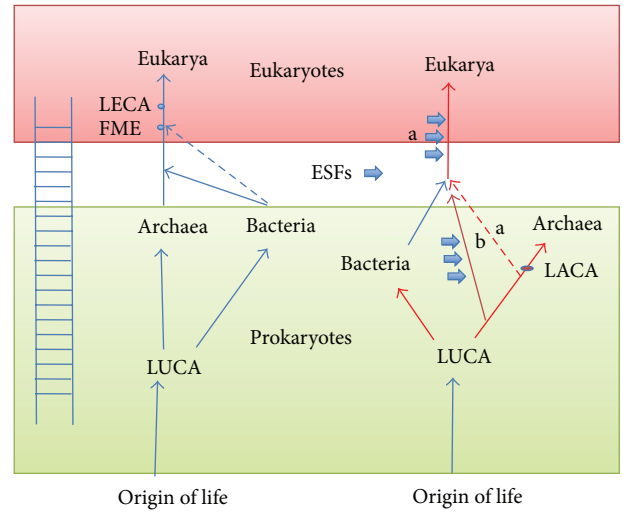


Figure 1: The Aristotle scenarios. (a) The traditional ladder-like evolutionary scenario, in which organisms increased in complexity from the origin of life to prokaryotes and eukaryotes, with archaea being intermediate organisms on the way to become eukaryotes; (b) the classical universal tree of life of Woese et al. [53] in red, combined with the fusion hypothesis (blue line). The LECA is the last eukaryotic common ancestor and FME the first eukaryote harbouring mitochondria; the dotted line refers to the hypothesis in which eukaryotes originated by the association of an archaeon with the mitochondrial bacterial ancestor [49]. Thick arrows indicate the emergence of eukaryotic specific features (ESFs).

to the fusion itself [49]. A common point to most fusion scenarios is that they involve two partners that are very similar to some modern archaea and bacteria [46]. These partners either belong to modern lineages of bacteria and archaea or are derived from an extinct (transient) archaeal lineage that originated from modern-looking archaea (as in the recently proposed *phagocytosing archaeon scenario* of Martijn and Ettema, [21]). This led to the common view that eukaryotes ascend from archaea in an evolutionary ladder (*scala natura*) leading from LUCA to eukarya via archaea (Figure 1(a)).

The fusion scenarios for the origin of eukaryotes are in apparent contrast to the classical tree of life proposed by Woese et al. in which archaea and eukarya are sister groups [53]. However, this tree can be reconciled with fusion scenarios if the archaeal-like partner diverged from the branch leading to archaea before the emergence of LACA. At some point. however, if this archaeal-like partner was very different from modern archaea (a protoarchaeon), with possibly additional eukaryotic traits, and if the bacterial partner was the alpha-proteobacterium at the origin of mitochondria, this "fusion hypothesis" becomes very similar to the "protoeukaryote" hypothesis in which the host of the alpha-proteobacterium was an organism belonging to a third lineage, distinct from lineages leading to archaea and bacteria [19, 50, 51, 57]. Accordingly, there is a continuum of possible scenarios between classical fusion hypotheses involving partners very similar to modern "prokaryotes"

and the idea of a protoeukaryote "fusing" with the alphaproteobacterial ancestor of mitochondria. In the following discussions, I will only refer to those fusion hypotheses that involve partners very similar to modern "prokaryotes."

Beside fusion hypotheses, the classical "three-domain tree" has been challenged by scenarios in which the eukaryotic lineage emerges within archaea [55, 58–61]. To support this view, it has been repeatedly claimed recently that the three-domain tree should be replaced by an updated version of the "*eocyte tree*," proposed by Lake et al. three decades ago [61] (eocyte being the name proposed by Lake et al. for the a group of thermophilic archaea, later on christened Crenarchaeota by Woese and colleagues [53]). The eocyte tree has been rejuvenated because it is apparently validated by phylogenetic analyses of universal proteins in which eukaryotes are nested within archaea [55, 58–60]. These analyses provided apparent support for a clade grouping eukarya with a candidate archaeal superphylum called TACK, that encompasses Thaumarchaeota, Aigiarchaeota, Crenarchaeota, and Korarchaeota [59]. Amazingly, if this is true, eukaryotes are members of the archaea (as we are apes) and should be considered as an archaeal phylum (and the TACK superphylum should be renamed TACKE since it also includes eukarya; otherwise, archaea would be paraphyletic, thus not a valid taxon, since the last common ancestor of archaea was also an ancestor of eukarya.

Here, I will first briefly criticize the fusion scenario and argue in favour of the monophyly of archaea (so that we do not have to worry about TACKE). Then, I will discuss recent results that, in my opinion, strongly suggest that the last common ancestor of archaea and eukarya was more complex than archaea and less complex than eukarya (the "bug in between" scenario). I will call this ancestor thereafter the *last archaeal-eukaryal common ancestor* (LAECA). I will also try to put viruses in the general picture.

4. Criticism of Fusion Scenarios

I have previously proposed my own fusion scenario, as a joke [46], accompanied by criticisms against my new version of the fusion (the association of a thaumarchaeon and a PVC bacterium). Since this paper is now often cited simply as a new fusion hypothesis (!), I have to reiterate here my antifusion arguments (for other critics, see [50–52]). In fusion scenarios, the host cell can be either a bacterium or an archaeon. In both cases, these scenarios first raise major problems for the fate of the ancestral membranes of the host or the endosymbiont, since the putative ancestral archaeal membrane has disappeared in eukaryotes. In all known real cases of symbiosis, the symbiont keeps its membrane, even in the most extreme cases of reduction (e.g., the mitoribosome or the nucleomorph) [62, 63] or when the symbiont uses lipids from their hosts, as in the case of *Nanoarchaeum equitans* [64]. Membrane conservation throughout evolutionary processes seems to be a major feature of cellular organisms [22]. This strongly contradicts with fusion hypotheses in which the host is a bacterium, since the archaeal membrane of the endosymbiont has completely disappeared after the fusion (the nuclear membrane in eukaryotic cells being derived from the bacterial-like membrane of the endoplasmic reticulum). In fact, disappearance of the membrane of an infectious entity is only known in the case of enveloped viruses, when the membrane of the viral particles (virions) disappears in the infected cell (i.e., fuses with the cellular membrane). However, this situation cannot be assimilated to a true endosymbiosis, since the virion is not an organism and should not be confused with the virus itself [65].

Scenarios with an archaeal host raise other problems, since they involve the transformation of the host archaeal membrane into the bacterial-like membrane of eukaryotes. Indeed, there is no obvious selection pressure that could have favoured this transformation. As noticed by Lombard and colleagues, such transformation has never been observed in nature [22]. In particular, acquisition of a bacterial type of membrane never occurred in any archaeal lineage, despite the massive lateral gene transfer (LGT) of bacterial genes into archaea. For instance, Nelson-Sathi and co-workers determined recently that about 1000 bacterial genes have been transferred, probably in a single event, in the ancestral archaeal lineage from which emerged all modern Haloarchaea [66]. Importantly, despite this massive transfer of bacterial genes, Haloarchaea still have "archaeal lipids" and remained *bona fide* archaea in terms of all their major cellular and molecular processes.

Fusion scenarios also involve highly unlikely *ad hoc* hypotheses to explain the emergence of all complex ESFs from the simpler compact and fully integrated molecular machines working in archaea and bacteria (for critics, see [44, 46, 50–52, 57]). There is no general agreement between proponents of fusion scenarios for most of these *ad hoc* hypotheses. For instance, several authors proposed different selection pressures to explain the origin of the eukaryotic nucleus [20, 67–69]. These hypotheses posit that (for various reasons) some kind of barrier was required between the nucleoid provided by one of the two partners and the cytoplasm provided by the other. However, they do not usually explain why all genes from the nucleoid of the host migrated through this barrier into the nucleoid of the endosymbiont (or *vice versa* depending on the scenario) and/or why the circular nucleoid of the endosymbiont/host was transformed into multiple linear chromosomes with telomeres and centromeres. They also do not explain why the simple "prokaryotic" cell division mechanism of the endosymbiont at the origin of the nucleus was replaced by the complex mitotic cell division machinery.

Fundamentally, fusion scenarios posit that modern cells (archaea and bacteria) were transformed by their association into cells of a completely new domain (with an abrupt but transient acceleration of protein evolutionary rates leading to new versions of universal proteins in eukarya). This possibility was strongly rejected by Woese who wrote that: "*modern cells are sufficiently complex, integrated and 'individualized' that further major change in their designs does not appear possible*" [70]. I fully agree with this statement; the observation of nature tells us indeed that such transformation is not possible. In all known cases of endosymbioses or close association between organisms that belong to different domains, both

partners remain members of their respective domains and there is no dramatic acceleration of protein evolutionary rate, especially for universal proteins. For instance, the association between a cyanobacterium that produced chloroplasts did not transform Viridiplantae into a new domain. Viridiplantae remained eukaryotes (with eukaryotic ribosomes), whereas chloroplasts (and mitochondria) can still be recognized as highly derived bacteria, with highly divergent—but still bacterial—ribosomes. As already mentioned, the massive invasion of an Haloarchaeal ancestor by more than one thousand bacterial genes had no effect on the archaeal nature of Haloarchaea.

Finally, another rarely discussed important argument against fusion scenarios is the uniqueness of eukaryotic viruses and related transposons [46]. Indeed, fusion scenarios posit that all modern eukaryotes originated from a unique fusion event between one particular archaeon and one particular bacterium. If the host was an archaeon, all eukaryotic viruses should have originated from those archaeal viruses that were able to specifically recognize the surface of this particular archaeon (or of its immediate descendants). Similarly, if the host was a bacterium, eukaryotic viruses should have originated from bacterial viruses that were able to recognize the surface of this particular bacterium (or its immediate descendants). This seems at odds with the present diversity of eukaryotic viral lineages and transposons, especially with the existence of many lineages of eukaryotic DNA and RNA viruses that have no viral counterparts in bacteria and eukarya, such as Baculoviridae, Megavirales, Retroviridae, and many others. One should posit that all eukaryotic viruses (in particular, most RNA viruses, retroviruses, and pararetroviruses) and transposons originated *de novo* after the fusion event, *in the dark age of eukaryotic evolution*, and/or that the ancestral archaeal or bacterial viruses and transposons evolved so fast after the mitochondrial endosymbiosis event that it is no longer possible today, with few exceptions, to recognize their evolutionary relationships with viruses and transposons infecting bacteria and archaea. These two possibilities seem unlikely. The *de novo* late origin of eukaryotic RNA viruses is at odds with the current assumption that RNA viruses are somehow relics of ancestral viruses from the RNA world. It is in particular appealing to think that ancestral retroviruses and/or retrotransposons played a key role in the transition from RNA to DNA genomes. The rapid and complete transformation of bacterial (or archaeal) viruses into eukaryotic ones (Caudovirales becoming Herpesviridae and *Tectiviridae*/STIV becoming Megavirales) and of archaeal/bacterial transposons into eukaryotic ones after the fusion event while archeoviruses and bacterioviruses, as well as bacterial and archaeal transposons, remained unchanged for billion years seems to me very unlikely.

In summary, fusion scenarios posit a transient but extreme acceleration of protein evolutionary rates and drastic structural changes to take into account the existence of eukaryotic specific versions of universal proteins (e.g., ribosomal proteins) and the rapid emergence of all ESFs in the period between the fusion event and LECA. They should also posit a transient but extreme acceleration of evolution of viral structures and the appearance of many new viral and transposon families in that same period. This does not seem reasonable, even more so if the fusion event is assimilated to the endosymbiosis that produced mitochondria [49], since in that case, all these dramatic evolutionary changes should have occurred between the appearance of the first mitochondrial eukaryote (FME) (i.e., after diversification of all bacterial lineages) and LECA! Such scenarios require no less than several miracles for the emergence of eukaryotes, miracles that occurred only once in 2-3 billion years of coexistence between archaea and bacteria.

5. The Monophyly of Archaea

As previously mentioned, it is commonly assumed that the eocyte tree is now validated by phylogenetic analyses in which eukarya emerge from within archaea [55, 58–60], with the consequence that all eukaryotic ESFs should have originated in a highly divergent archaeal lineage and that archaea are our ancestors. However, these analyses, concerning very deep phylogenies, are prone to many artefacts (for a critical analysis of contradictory results obtained by different authors with more or less the same dataset; see [45]). In particular, phylogenetic analyses of Embley and colleagues [55, 58, 60] include many ribosomal proteins for which there is no significant signal for deep branching because bacterial proteins are too divergent from their archaeal and eukayotic homologues [71]. Elongation factors, amino-acyl tRNA synthetases, or else V-ATPases are also used in these analyses despite the fact that these proteins are heavily saturated with respect to amino acid substitutions [72] and cannot even resolve the phylogeny of eukarya, putting microsporidia (highly derived fungi) at the base of the eukaryotic tree. Several universal proteins used (RNA polymerases, RFC proteins and amino-acyl tRNA synthetases) are also encoded by many viruses (especially Megavirales) and it is unclear if the eukaryotic and archaeal versions are orthologues of if some of them have been independently acquired from viruses ([40, 73, 74]; see discussion below). The phylogeny of RNA polymerases is especially puzzling since the eukaryotic RNA polymerases of type I branch in between bacteria and a clade formed by archaeal and eukaryal RNA pol II and III [75]. In the analysis of Cox et al., the three homologous RNA polymerases are analyzed separately with their archaeal and bacterial homologues with RNA pol I being the only protein whose phylogeny supports the monophyly of archaea [55, Figure S27]. Strikingly, examination one by one of all phylogenies, published by Cox et al., shows that all of them failed to recover correctly the internal branching of the archaeal domain and are plagued with very bad resolution. In a more recent global analysis of a similar set of universal proteins, Lasek-Nesselquist and Gogarten [76] again obtained results favouring the eocyte tree, but they also notice that the method used *"generated trees with known defects, such as the placement of Microsporidia at the root of the eukaryotic tree, a paraphyletic Euryarchaeota, and an attraction of Nanoarchaeota to the base of the TACK + eukarya clade, revealing that this method is still error prone"*. Despite the exhaustive usage of complex alternative models to perform and test

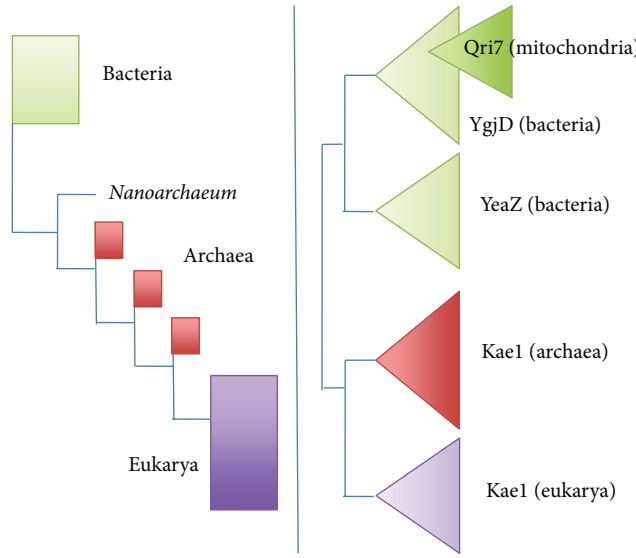


Figure 2: Two contrasting phylogenies of the same universal protein. Schematic representation of two published phylogenies of a universal protein known under different names (YgjD, YeaZ, Kae1/OSGEP, and Qri7/OSGEP L) which is involved in the biosynthesis of the universal tRNA modified base t6A [54]. These simplified phylogenies are adapted from Figure S46 in [55] (left panel) and from Figure S1 in [56] (right panel). Squares indicate unresolved nodes, whereas triangles indicate resolve nodes. The tree on the right is congruent with firmly established biological knowledge such as the monophyly of bacteria and eukarya, the bacterial origin of mitochondria. It favours the classical three-domain tree of Woese and colleagues. The tree on the left, which is not resolved, with aberrant paraphyly of archaea (see for instance the position of *Nanoarchaeum equitans*) was nevertheless used by the authors to support the "eocyte tree". Comparison of these trees clearly reveals that the methodology (data sampling and/or algorithm for tree reconstruction) used by Cox and co-workers [55] for phylogenetic analyses cannot recover correct phylogenies for universal proteins.

them, the phylogenies used in these global analyses cannot provide answer to the question of archaeal monophyly *versus* paraphyly because, in most cases they lack valid phylogenetic signal. Moreover, the quality of these phylogenetic analyses itself is also questionable. To be convinced, compare the phylogenies of the protein YgjD/Kae1 (involved in a universal tRNA modification) published by Cox and co-workers, with those published in Hecker et al. in [56] (Figure S46 in [55], the YgjD/Kae1 protein is named O-sialoglycoprotein endopeptidase, according to an ancient annotation that turned out to be wrong [56]) (Figure 2). In the tree of Cox and co-workers the phylogenies of the bacterial and eukaryotic proteins are not resolved, and the archaea are paraphyletic, with eukarya branching with *Methanopyrus kandleri* and Crenarchaeota! In striking contrast, the phylogeny of all domains is well resolved in the tree published in Hecker et al., and the overall phylogeny exhibits a clear-cut three-domain topology [56]. However, the three can be divided in to five groups because an ancient duplication occurred in the bacterial domain leading to two paralogous proteins, YgjD and YeaZ, and mitochondrial proteins, named Qri7, branch within the YgjD tree, in agreement with their bacterial origin (Figure 3). The surprisingly unresolved YgjD/Kae1 phylogeny published by Embley and colleagues suggests that their analyses favouring the paraphyly of archaea, with the emergence of eukarya within archaea, correspond to the concatenation of poorly resolved phylogenies and are plagued by multiple methodological problems.

Some logical considerations argue in fact against the emergence of eukarya within archaea (thereafter called the "*archaeal* ancestor scenario"). In that scenario, the archaeal ancestor should have contained all eukaryotic features that are presently dispersed in modern archaea. In that case, since LACA was probably more eukaryotic-like than any one of its descendants [6], the archaeal ancestor should have been LACA itself or a descendant of LACA, which, unlike the others, never lost a single eukaryotic feature. At this point one should remind that LACA was not a special (breakthrough) organism but simply the *last* of all common archaeal ancestors that thrived between LUCA and LACA. Why this particular ancestor should have also been the ancestor of eukarya? Another complication for the archaeal ancestor scenario is that LACA was probably a hyperthermophile [6, 80, 81]. The only mesophilic organisms presently known in the putative TACKE phylum are mesophilic Thaumarchaeota [82]. However, all known mesophilic Thaumarchaeota lack some critical eukaryotic features, such as actin or tubulin [6]. Again, one should suppose that eukarya originated from a mesophilic descendant of LACA, which has never lost a single eukaryotic feature and left no other descendants itself besides eukarya! Finally, the archaeal ancestor scenario raises the same problems as those of the fusion/association theory concerning (1) the transformation of the archaeal membrane into the eukaryotic one, (2) the deconstruction of the very efficient and integrated prokaryotic-like molecular biology of archaea, such as the coupling of transcription and translation, into the complex and often odd eukaryotic molecular biology, and (3) the origin of eukaryotic viruses and transposons, which, in that model, should have all

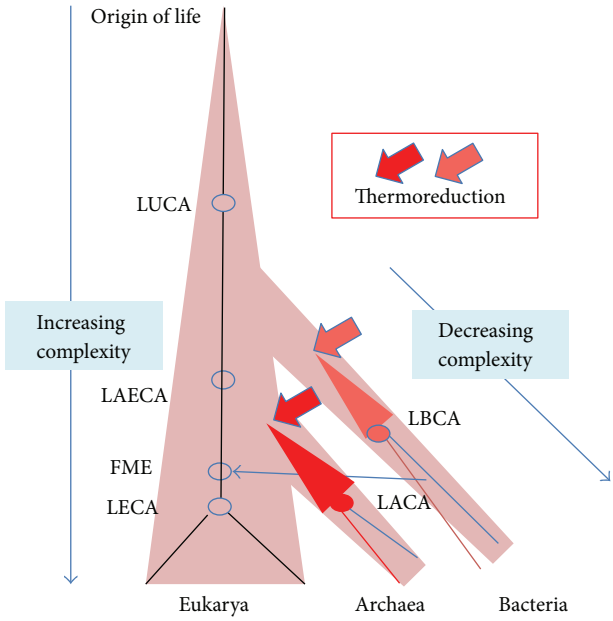


Figure 3: A scenario based on divergent evolutionary trends for "prokaryotes" (archaea and bacteria) and eukaryotes. This scheme is based on the assumption that the universal tree is rooted in the bacterial branch [53]. Complexity increased from the origin of life to LUCA to eukaryotes, via the last archaeal-eukaryal common ancestor (LAECA). Reductive evolution occurred from LAECA to LACA and modern archaea, possibly triggered by thermoreduction [77] indicated by large red triangles and/or as a way to escape protoeukaryotic predators [50]. Bacteria experienced independently a similar evolutionary path. The blue arrow indicates the mitochondrial endosymbiosis.

originated from evolutionary unrelated archaeal viruses and transposons (!) or originated recently, *de novo*, in *the dark age of eukaryotic evolution*.

6. Divergent Evolutionary Trends Shaped the History of Archaea and Eukarya

We have always the tendency to interpret evolution as a general trend from simple to complex because, as *Homo sapiens*, we are still under the spell of the Aristotle's *scala natura*. This is why the idea that human originated from apes looking like chimps was so prevalent in narratives describing our origin. However, it seems now that our common ancestor with chimps was possibly already bipedal, looking more like an ancient *Homo* than a modern chimp [83]. I will bet here that, as our last common ancestor with chimps neither resembled chimps nor *Homo*, LAECA was neither an archaeon nor a eukaryote, but a creature endowed with the property to be at the origin of both (Figure 3). To draw a picture of LAECA, I will use as Ariadne's thread the concept of evolutionary trends, knowing that these trends can be different for different lineages, depending on their ecological setting and "ways of life"; that is, whereas in some lineages organisms become more and more complex by the acquisition of novel traits and the use of more components to perform the same task, in other lineages, they evolve by reduction (streamlining), by losing ancestral traits and simplifying molecular processes [84].

It has been clearly shown recently from comparative genomics that a long suspected trend toward simplification indeed occurs in archaea and bacteria. For instance, Wolf and co-workers have shown that gene losses are estimated to outnumber gene gains at least 4 to 1 in these two domains [85]. Importantly, two independent studies concluded that LACA was an organism of greater complexity than most of the extant archaea [85, 86], in agreement with the observation that the ancestral archaeal ribosome contained more proteins than the ribosomes of modern archaea [87, 88] and that LACA should have combined all eukaryotic traits presently dispersed in various archaea [6]. Phylogenomic analyses focusing on protein structures also detected a trend toward proteome reduction in the archaeal and bacterial lineages, suggesting that both lineages originated from ancestors with more complex proteomes ([89–91] and references therein). All these observations are in agreement with ancient ideas proposed by scientists like Carlile who argued long ago that "prokaryotes" with their highly integrated and efficient molecular biology have evolved by streamlining to increase their reproduction rate and use of resources in rapidly changing environments (r-selection) [92]. There are of course some exceptions to this general trend, especially in bacteria, such as the late evolution of giant bacteria in subgroups of proteobacteria [93], but these are the exceptions that confirm the rule.

In contrast to the overall trend toward simplification observed in archaea and bacteria, an evolutionary trend toward more complex forms, slower growth rates, and larger size is clearly operating in eukaryotes [92]. This trend can be seen of course in animals, with the emergence of the immune and nervous systems and finally brains becoming larger and larger in some lineages, but it can be also observed in plants and fungi, which now rule the macroscopic world and in various protists with primary and secondary endosymbioses producing very sophisticated organisms. The evolutionary trend toward complexity in eukaryotes is very ancient since LECA was already as complex as modern eukaryotes in terms of cellular structure and molecular biology, with, in particular, a genome full of introns [94]. This trend toward complexity is again only an overall trend and reductive evolution in eukaryotes led, for instance, several times independently to the transformation of multicellular fungi into unicellular yeasts [95].

I will thus argue here that LAECA was not an archaeon as is currently assumed, but an organism from which two lineages, whose destiny was shaped by opposite evolutionary forces, have diverged. One led to the emergence of archaea by reduction, the other to eukarya by increasing complexity (Figure 3). Indeed, if evolution by reduction has taken place from LAECA to modern archaea and increasing complexity from LECA to modern eukarya, there is no reason to imagine that these respective trends became effective only at the time of LACA and LECA, respectively. LACA and LECA were not special organisms on the two evolutionary lines that can be drawn from LAECA to modern organisms (Figure 3).

They are only the most recent organisms that all archaea or eukaryotes share as common ancestors, respectively (much like the African Eve and Adam who were not special individuals in the *Homo sapiens* lineage but only those at the origin, respectively, of all women—for Eve—and men—for Adam—living today).

If the two opposite evolutionary trends discussed above are related to ancient differences in the way of life that have very early on fashioned cell structure and function in the two respective lineages, there is therefore no obvious reason why these trends should have changed at the time of LACA and LECA. A weak point of the archaeal ancestor hypothesis is that it precisely involves a dramatic reversal of the reductive trends at work in archaea, as if a particular lineage of yeast started today to evolve back toward extremely complex fungi. As previously discussed, to posit that this trend reversal was triggered by the endosymbiosis of a bacterium is at odds with current observations showing that endosymbioses never modify the basic molecular biology of the host and usually follow a previous evolutionary trend of the host toward complexity in order to capture the symbiont. This is even recognized by recent proponent of fusion models, which now imagine that the host was an archaeon more complex than modern ones, with already elaborated phagocytic capacities, as in the "*phagocytosing archaeon theory*" of Martijn and Ettema [21]. It seems to me more logical to think that eukaryotes did not evolve from a particular archaeal lineage, which was at odds with the archaeal evolutionary trend, but from a LAECA that was more complex than LACA, but less complex than LECA. LAECA was definitely not an archaeon since, by definition, all archaea have originated from LACA, whereas in the scenario proposed here, LACA itself originated from LAECA.

7. The Origin of Archaea: The Thermoreduction Hypothesis

Gouy and his co-workers have shown, using ancestral rRNA and universal protein sequence reconstruction, that LUCA was probably a mesophile, whereas LACA was probably a hyperthermophile (i.e., an organism living at temperature above 80°C, [96]) [80, 81]. Since all known eukaryotes are mesophiles, it is thus more parsimonious to think that LAECA itself lived at moderate temperature, the transition from mesophile to thermophily having taking place between LAECA and LACA (Figure 3). This would provide in particular a highly plausible selection pressure to explain the emergence of the archaeal membrane [97–99]. The unique chemical and structural features of archaeal lipids are indeed perfectly suited to maintain the cytoplasmic membrane functional at high temperature [100]. In particular, the archaeal membrane is much less permeable to protons/ions than the bacterial and eukaryotic one [100]. This last point is critical for the maintenance at high temperature of the transmembrane gradient of ions/protons required for ATP synthesis. Interestingly, phylogenomic analyses strongly suggest that many enzymes involved in the biosynthesis of both bacterial and archaeal building blocks for lipid biosynthesis were already present in LUCA [18, 22]. Lombard and colleagues thus suggest that LUCA had already both types of lipids, having a mixed archaeal/bacterial membrane. However, modern cells also contain much of these enzymes but use a single type of membrane, either an archaeal one for archaea or a "bacterial one" for bacteria and eukarya. The building blocks, such as isoprenoids, used in Archaea for the synthesis of membrane lipids, are used for other tasks in bacteria and eukarya, and *vice versa* [18, 22]. If LAECA was a mesophile, it thus makes sense to imagine that it possessed the "bacterial type" of lipids that are still present today in eukarya and that archaeal lipids emerged specifically in the archaeal lineage under the pressure of adaptation to higher temperatures (Figure 4). In agreement with this view, Lombard et al. have shown that Archaea and eukarya share unique enzymes involved in the synthesis of archaeal lipids that differ from the bacterial ones, but that archaea also use a few archaea-specific enzymes required for the biosynthesis of archaeal membranes [18, 22]. This indicates that important modifications occurred specifically in the archaeal lineage, from LAECA to LACA, and it seems reasonable to suggest that these modifications are those that were involved in the formation of the unique archaeal type of membrane, perfectly suited for life at high temperature (Figure 4).

More generally, the selective pressure behind the trend toward reduction from LAECA to LACA might have been the progressive adaptation of the archaeal ancestors to hotter environments in the framework of the "*thermoreduction hypothesis*" for the origin of "prokaryotes" [77] (Figure 3). The major features of the "prokaryotic" phenotype (coupling of transcription and translation, short half-life of messenger RNAs, small cell, and genome sizes, and high macromolecular turnover) are indeed perfectly suited for life at high temperature [77]. This could explain why the upper temperature limit for eukaryotes, around 60°C, established by Brock in the sixties [101] has never been exceeded from that time, despite the extensive search of hyperthermophilic eukaryotes that ended up with thermophilic protists growing up to 54°C [102]. The adaptation of archaeal ancestors to high temperature might also explain why they got rid of RNA viruses that possibly infected LAECA and still infect eukaryotes, since RNA is very unstable at high temperature [77, 97], making it difficult to imagine RNA genomes remaining intact for long periods in virions bathing in hot ponds. Bolduc and co-workers identified by PCR putative novel positive strand RNA viruses in an archaea-rich hot springs [103], but the archaeal nature of their host cells remains to be demonstrated. In fact, the "thermoreduction hypothesis" that I proposed 20 years ago is now strongly supported by the more recent work of Gouy and colleagues on the reconstructed temperatures of LUCA and LAECA [80, 81] and well explains the evolutionary trend toward reduction now recognized in the archaeal domain [85, 87], suggesting that one can probably extrapolate this trend from LACA back to LAECA. The thermoreduction hypothesis can be also coupled to the "raptor scenario" proposed by Kurland and colleagues, since the possibility to explore hot environments may have been a formidable advantage for prokaryotes in

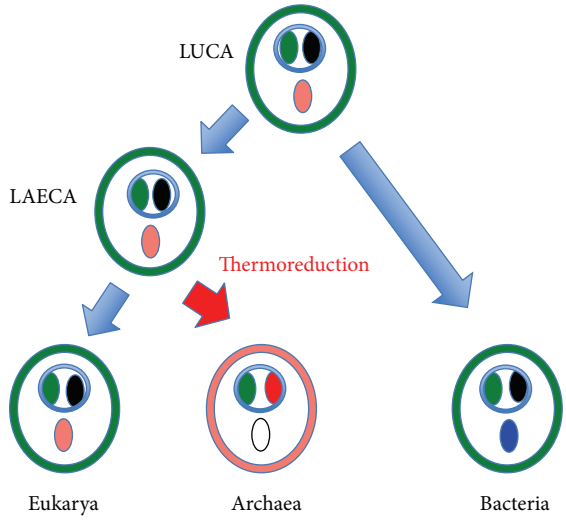


Figure 4: Schematic evolution of cellular membranes. Distribution of proteins involved in phospholipid biosynthesis in the three domains, LUCA, and LACA are depicted according to Lombard et al., [18, 22], but my evolutionary interpretation is different from those proposed by these authors [22]. Bacterial/eukaryal type membranes are in green and archaeal type membrane in red. Green circles correspond to universal enzymes involved in phospholipid biosynthesis (glycerol phosphate dehydrogenases, enzymes linking polar head groups to glycerol). Pink circles correspond to the classical mevalonate pathway for isoprenoid biosynthesis that was probably present in LUCA and lost in bacteria. Red circle represents enzymes of the alternative mevalonate pathway, which are specific of archaea, and involves a mixture of eukaryotic-like and archaeal specific enzymes. Blue circles correspond to the nonhomologous methylerythritol pathway for isoprenoid biosynthesis present in bacteria. Black circles represent the fatty acid biosynthetic pathway, which is no longer used for membrane phospholipid biosynthesis in archaea. Circles corresponding to proteins involved in the biosynthesis of membrane phospholipids are encircled.

hiding themselves from protoeukaryotic raptors unable to follow them to hell [46].

8. The Mechanisms of Eukaryogenesis

We have seen that LECA was already a very complex organism already endowed with all ESFs of modern eukaryotes [20, 104]. Part of this complexity might have been due to the acquisition of mitochondria that allowed a dramatic input in energy available for macromolecular biosynthesis [105]. LECA was thus certainly more complex than the first mitochondrial eukaryote (the FME), that is, the organism harbouring the alpha-proteobacterium at the origin of modern mitochondria (thereafter called the mitochondrial ancestor). It has been shown for instance that the mitochondrial ribosome of LECA has already acquired 19 new proteins of unknown origin (and lost one bacterial protein) compared to the FME [88].

Lane and Martin [105] have argued that the energy produced by mitochondria with their core genome encoding proteins of the respiratory electron transport chain was essential for the emergence of eukaryotic complexity, including the emergence of all ESFs. However, eukaryotes that have lost mitochondrial genomes have not changed their eukaryotic fabric and still harbour most, if not all, ESFs. This suggests that EFS may have appeared before the mitochondrial-driven increase in cellular and genome sizes typical of large complex eukaryotes. The idea that all ESFs originated in the time period from the FME to LECA also implies that the bacterial ancestor of mitochondria was not captured by phagocytosis. However, the presence of an elaborated cytoskeleton and machinery for vesicle formation (exosomes, ectosomes, endosomes, etc.) in the FME seems also logical, considering that actin, tubulin, and G-ATPases were already present in LAECA. LUCA itself might have been compartmentalized, considering the structural similarities detected between eukaryotic coat proteins (involved among other things in nuclear pore formation) and proteins present in compartmentalized bacteria of the superphylum PVC (Planctomycetales, Verrucomicrobiales, and Chlamydiae) [106]. In that case, it is logical to think that the mitochondrial ancestor was obtained by phagocytosis, as it was probably the case later on for the ancestor of chloroplast, and as it is still now for so many intracellular bacteria living in eukaryotic cells. Beside phagocytic capacities, I suggest here that most ESFs were already present in the FME, even if some of them might have been profoundly altered following the assimilation of the mitochondrial ancestor. This is in agreement with the idea that the FME itself was more complex than LAECA because the evolutionary trend toward higher complexity was already operational during the period between LAECA and LECA (via the FME) (Figure 3).

The mechanisms that induced increasing complexity in the eukaryotic lineage before LECA can be inferred from comparative genomics and from the analysis of the mechanisms involved in recent complexity increase in modern eukaryotes. Eukaryotes typically use multiple paralogous proteins to build complexes that are composed of a single (or very few) paralogous protein in archaea. For instance the archaeal MCM helicase is a homohexamer, whereas the eukaryotic

MCM helicase is a heterohexamer made of six paralogous MCM2-7 proteins. Eukaryotes use three paralogous RNA polymerases (I, II, and III) for transcription, whereas archaea use only one. Eukaryotes possess five DNA polymerases of the B family *versus* 1, 2, or 3 in archaea.

The complexity of eukaryotic molecular mechanisms compared to their archaeal homologues has been attributed to extensive gene duplications from LAECA to LECA that roughly double the number of genes in eukaryotic core genomes [20]. Indeed, gene duplications, and even whole genome duplications, have been involved in more recent stages of eukaryotic evolution. Here, however, I would remind that the multiple "paralogous proteins" present in the ancient core genomes of eukaryotes might not be all true paralogues, but in some cases homologues introduced by viral integration into the genomes of protoeukaryotes [46, 74]. This phenomenon has been observed in the case of the evolution of the archaeal DNA replication apparatus. For instance, it was deduced from phylogenetic analyses that 4 out of 6 MCM genes present in the genomes of some Methanococcales were recruited from mobile genetic elements (viruses or plasmids) [107]. Many eukaryotic viruses with large DNA genomes encode transcription or replication proteins homologous to eukaryotic ones and, via integration in the genomes of protoeukaryotes, their ancestors might have been the source of multiple different versions of homologous genes in modern eukaryotic genomes [46]. This would explain for instance the odd phylogenies of eukaryotic RNA or DNA polymerases (see Figure 5 for a schematic tree of RNA polymerases) [75, 108]. In these phylogenies, the multiple versions of the eukaryotic enzymes do not form monophyletic groups, themselves sister groups of their archaeal homologues, as would be expected if they originated by gene duplication in eukaryotes after the divergence of archaea and eukarya. Instead, they are paraphyletic, often nested with enzymes encoded by DNA viruses, and they are intermixed with archaeal enzymes.

Ancestors of Megavirales could have been a major source of new genes in the eukaryotic linage [46]. These viruses, with genome sizes varying from 150 kb to more than 1 Mb, are very ancient and most likely predated LECA [109]. Moreover, integration of Megavirales genomes into eukaryotic genomes has been documented [110]. The abundance of lineage-specific proteins in the various lineages of Megavirales testifies for the genetic creativity of these viruses [111], indicating that they might have been an important source of new genes both before and after LECA. The integration of viral genomes thus provides the hosts with new proteins that can acquire important functions. There are many examples in the evolution of modern eukaryotes that testify for the importance of viral proteins in the evolution of eukaryotes. For instance, exaptation of a retroviral protein, syncytin, has been critical for the formation of placenta in mammals [112]. These phenomena also occur in archaea and bacteria, but considering the extreme abundance of mobile elements in eukaryotic genomes, their importance in eukaryotes is an order of magnitude higher.

The integration of ancient Megavirales can also explain why the eukaryotic core genomes contain many bacterial genes of different origin and without affinity to the mitochondrial alpha-proteobacterium ancestor. These bacterial genes are usually supposed by proponents of fusion hypotheses to derive from the bacterial partner [113–117]. Alternatively (or in addition) these genes are supposed to be derived from genes present in bacteria that were eaten by protoeukaryotes ("*you are what you eat*", [118]) or in bacterial endosymbionts of ancestral eukaryotes [20]. I suggest here that many of these genes have not been directly transferred from bacteria to protoeukaryotes but that these transfers have been mediated by ancient Megavirales. Remarkably, these giant viruses contain today up to 10% of bacterial genes in their genomes [119]. Megavirales with linear genomes have probably recruited these genes from endosymbiotic bacteria living in their hosts via their specific DNA replication mechanism [119]. Bacterial genes are also present in Megavirales with circular genomes, indicating that linear genomes (typical of eukaryotes) could have preceded circular ones (typical of "prokaryotes") in genome evolution [74]. Eukaryotes might have also recruited their messenger RNA capping system from ancient Megavirales [120], since these viruses use for capping the eukaryotic system, whereas other lineages of viruses use different systems [121]. The viral diversity in terms of capping mechanisms suggests that different capping mechanisms originated several times independently in the viral world and that one of them was transferred later on into

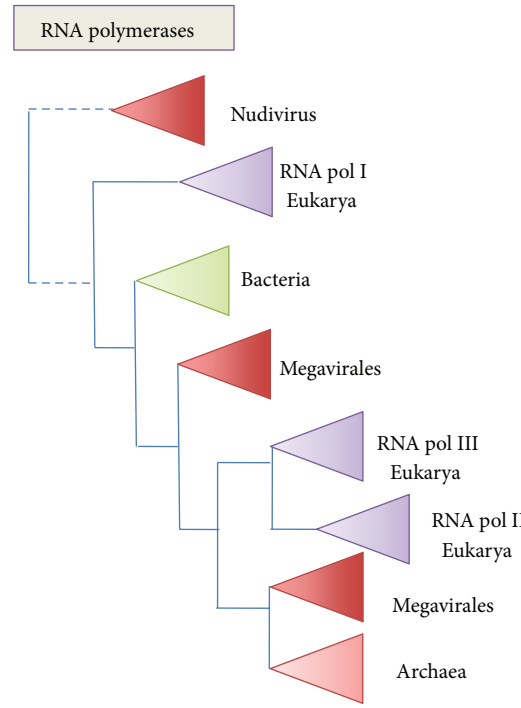


Figure 5: Schematic phylogenies of RNA polymerases.The RNA polymerase tree has been drawn combining and interpreting results from several papers [54, 60, 78, 79]. The position of the Nudivirus RNA polymerase is extrapolated from its high divergence with other homologous RNA polymerases (indicated by dotted lines).An update phylogeny would be welcome but would not change the take-home message, the puzzling mixture of cellular and viral enzymes, suggesting several ancient transfers between viruses and cells.

the protoeukaryotic lineage [120]. The idea to tag chemically its own messenger RNA to discriminate it from the messenger RNA of your host can be viewed as a typical viral trick that was later on stolen by eukarya in the arms race between viruses and cells.

Besides, megavirales, retroviruses, retroelements, and other eukaryotic transposons have been probably a major source of variations at the origin of eukaryote evolution [46, 74]. These genetic elements, which represent a large portion of modern eukaryotic genomes, have been critical factors in recent eukaryotic evolution [122]. The mobility of IS triggers rapid genome rearrangements and modifies genome expression patterns, providing new promoter elements, activating or regulating genes, even creating new genes by interfering with alternative splicing. It is therefore tempting to suggest that retroviruses have been used as toolkits in the formation of some ESFs at the onset of the eukaryotic lineage.

Retroviruses and/or derived retroelements have been probably instrumental in the emergence of eukaryotic chromosomes. Indeed, telomeres are evolutionarily related to retroposons of the Penelope family [123, 124] and telomerases are homologous to reverse transcriptases [125]. Moreover, centromeres are formed by the repetition of numerous retroelements [126]. The cell nucleus itself might have originated as a protective device allowing the cell to hide their chromosomes from viruses [46, 120, 127]. For this purpose, cells might have recruited viral proteins able to manipulate the endoplasmic reticulum membranes, and that were originally used to build viral factories [120]. This scenario is supported by the fact that viruses themselves use viral factories to protect their replication machinery from the defence systems of the host [128]. Protoeukaryotic cells might have learned from viruses how to use this trick against them by protecting their genomes from viral attack, building giant "cellular factories" that became cell nucleus. It is indeed remarkable that the "volcano-like" viral factories of giant Megavirales, such as *Mimivirus*, are in fact as big as the large nucleus of their amoebal host cells [129].

I have previously emphasized the profound difference existing between the eukaryotic defence mechanisms against viruses and those used in common by archaea and bacteria. The arms race between eukaryotes and their specific viruses or else the necessity for eukaryotic cells to somehow control the spread of their specific transposons probably played an immense role in the specific evolution of the eukaryotic cells. It is clear for instance that exaptation of the siRNA antiviral defence system by eukarya to produce various types of microRNA has been decisive in the evolution of eukaryotic cells toward complexity. Modern eukaryotes use DNA methylation and histone modifications to limit the spread of mobile elements [122]. It is possible that these mechanisms that are typical of eukaryotic cells first originated as tools in the arms race between eukaryotes and their viruses and were recruited only later on for gene regulation in eukaryotes.

I think that it is highly significant that Megavirales, retroviruses, and transposons with their associated defence/attack systems, which have been probably essential for the evolution of eukaryotes toward complexity, are missing in archaea and bacteria. Interestingly, these megavirales and retroviruses are also unknown in yeasts. One can wonder if the loss of most viral families in yeasts, especially retroviruses, could explain why these unicellular eukaryotes seem to be trapped in their "prokaryotic lifestyle" and have never evolved back toward complexity. The absence of the siRNA antiviral defence system in archaea and bacteria is also remarkable. For me, the lack of these features could explain why these "prokaryotes" did not evolve toward greater complexity, even when many of them finally escaped the selection pressure toward streamlining linked to thermoreduction. When archaea and bacteria finally invaded mesophilic biotopes, they could not anymore "benefit" from the evolutionary toolkit for increasing complexity provided in eukaryotes by retroviruses, retroelements, and megavirales.

9. A Picture of LAECA

If we agree that LAECA was more complex than archaea and less complex than eukarya, the main question now is how much was it more and less complex than archaea and eukarya, respectively? In one extreme view, close to the *archaeal ancestor* scenario, the period between LAECA and LACA was short and LAECA was only slightly more complex than LACA (the "prokaryotic LACA). In that scenario, thermoreduction played only a minor role in shaping LACA, and all ESFs originated from LAECA to LECA. At the other extreme, most ESFs were already present in LAECA (the "eukaryotic" LAECA) and lost from LAECA to LACA. I suspect that the truth is again "somewhere in between". However, many evolutionists traditionally favour the "prokaryotic LACA" view. To counterbalance, I will try to push somewhat in the other direction. It is now often assumed for instance that the spliceosome machinery of eukarya derived from group II introns present in the mitochondrial ancestor or in ancient endosymbiotic bacteria, because some RNA components of the spliceosomes are evolutionarily related to group II introns [67, 68, 130]. However, this does not explain why intermediates of this evolutionary process have never been observed, despite the fact that eukaryotes have continued to coexist after LECA for 1-2 billion years with intracellular bacteria harbouring group II introns. Moreover, comparative genomics analyses have shown that the LECA had a spliceosome and contained probably a plethora of spliceosomal introns [94] and it seems unlikely that such incredibly complex molecular machine could have emerged in a short time between the FME and LECA. Interestingly, it has been shown recently that the genomes of some nucleomorphs have lost all introns and all genes encoding components of the spliceosomal machinery [131]. The nucleomorphs are remnants of eukaryotic nucleus from eukaryotic endosymbionts that are present in some photosynthetic protists. This observation tells us that the spliceosome (and protein coding genes containing introns) could be in fact an ancestral feature that was completely lost in bacteria and eukarya, while being retained in eukarya. This scenario would make sense since the spliceosome machinery (a giant ribozyme) reminds strikingly the ribosome, suggesting that both are remnants of the RNA world.

Kurland et al. have suggested a few years ago that the "prokaryote" ancestors evolved by streamlining to escape protoeukaryote praying on them by phagocytosis (the phagotrophic unicellular raptor scenario, [50]). This would imply the presence of an already quite elaborated cytoskeleton in LAECA, with possibly already an endoplasmic reticulum.

I have no space here to discuss the timing of appearance of the modern eukaryotic nucleus with typical eukaryotic chromosomes, mitosis and meiosis, and so on. I will argue that we should try to answer the question of their presence/absence in LAECA, by thinking in terms of reversibility or irreversibility. Is it possible or not for a specific ancestral ESF, once established, to have completely disappeared in archaea and bacteria? The nucleomorph story tells us that the answer is yes for the spliceosome. Beside intellectual constructions, we should look for similar examples to try obtaining answers for other ESFs.

A major question whose answer could help us to go further in our scenario is, what kind of viruses infected LAECA? If viruses are very ancient, as now suspected, having emerged well before LUCA [26, 132, 133], the logical conclusion would be that LAECA and closely related organisms living at that time were infected by ancestors of all viruses infecting now archaea and eukaryotes. This would mean that protoretroviruses and protomegavirales were around at that time and have been later on lost in the archaeal lineage. If true, as previously discussed, this might have maintained this lineage irreversibly into the path of reduction. Major questions then remain to be tackled such as, why so many lineages of archaeal viruses, such as Fuselloviridae, Rudiviridae, Lipothrixviridae, Clavaviridae, and so on, have disappeared in eukarya? One possibility may be that invention of the nucleus dramatically reduced the number of viral families capable to survive this invention, because, originally, only a few viruses were able to replicate in the cytoplasm. In that case, this would mean that the nucleus originated after LAECA. One can also wonder why eukarya lost the CRISPR system. This system was possibly more specific to those viruses that were lost in eukaryotes? Similarly, one can wonder why archaea lost the siRNA interference system? A reasonable possibility is that the *raison d'être* of this system, to detect and kill RNA viruses by targeting their genomes, disappeared once "protoarchaea" got rid of RNA viruses. Interestingly, many viral families known in Crenarchaeota, such as Rudiviridae, Lipothrixviridae, and Clavaviridae, are presently unknown in Euryarchaeota. It seems unlikely that these viruses originated *de novo*, in the branch leading to Crenarchaeota. If these viruses were present at the time of LACA, this implies that these vital families were eliminated in the lineage leading to Euryarchaeota. If this is the case (much more work on archaeal viruses will be required to confirm this scenario) this could indicate that loss of viral lineages is indeed a common feature in the emergence of novel cellular lineages.

10. The Path from LUCA to LAECA

I was critical of the traditional rooting of the tree of life in the bacterial branch, because this rooting is not supported by robust phylogenies and is often interpreted as supporting a "prokaryotic" LUCA [134, 135]. These previous observations remain valid but let open the resolution of the rooting problem itself. I used to favour the idea that the more complex molecular features observed in archaea/eukarya compared to archaea (e.g., more RNA polymerase subunits) were possibly ancestral. However, I am changing my mind because scenarios in which all specific archaea/eukaryotic proteins were present in LUCA and replaced later on systematically by the bacterial version seem more and more unlikely with increasing knowledge on the molecular mechanisms involved. For example, nonhomologous archaeal (eukaryal) and bacterial ribosomal proteins or transcription factors often occupy the same site on the ribosome and RNA polymerase, respectively, and it is difficult to imagine how one set of proteins was replaced by the other. It seems more likely that, once established, the bacterial and the archaea/eukaryotic versions of molecular machines could not have changed drastically. In agreement with this view, these machines then remained similar in all lineages that diverged from the ancestors of bacteria and archaea/eukarya, respectively. These considerations argue in favour of a relatively "simple" LUCA, as originally suggested by Woese and Fox under the name progenote [136], possibly still a member of a cellular RNA world [137]. In that scenario, LAECA was certainly more complex than LUCA, since the complex archaeal/eukaryotic versions of molecular system were now all present in that organism (Figure 3). I previously suggested that LAECA had possibly still an RNA genome to explain some major differences between the DNA replication apparatus of archaea and eukarya, such as the presence of very divergent DNA polymerases [46]. However, two more recent findings suggest that LAECA already had a DNA genome: (1) the discovery in Thaumarchaeota of a type IB DNA topoisomerase that was probably present in LAECA [19] and (2) the observation of conserved genomic contexts in archaea suggesting the existence of a regulatory mechanism coupling DNA replication and translation conserved between archaea and eukarya [138]. I thus favour now a scenario in which the RNA to DNA transition (possible mediated by viruses) occurred only twice, once in the bacterial branch and the other in the branch leading from LUCA to LAECA [137]. Importantly, a "relatively simple LUCA" with an RNA genome does not mean necessarily a "very simple" LUCA, since this organism could have harboured endomembrane systems [106], possibly spliceosomes [46], and encoded a rather large amount of proteins [89].

For me, an appealing hypothesis is that the eukaryotic trend toward increasing complexity corresponds in fact to the continuation of the major trend that operated from the origin of life up to modern organisms, via LUCA and LAECA, whereas archaea and bacteria, far from being intermediate "primitive" forms, originated by a reversal of this trend (Figure 3). Importantly, Gouy and colleagues have shown that, similarly to LACA, *the last bacterial common ancestor* (LBCA) was probably also a thermophile [80, 81]. If the universal tree is indeed rooted in the bacterial branch, this implies (LUCA being a mesophile) that adaptation to thermophily has occurred twice independently, once in the branch leading to archaea and once in the branch leading

to bacteria (Figure 3). I wonder if the fact that archaea and bacteria experienced similar selection pressure at their origin (adaptation to high temperature) could explain why they share partly similar types of mobile elements? As for archaea, bacteria might have escaped most RNA viruses (not all in that case) and retroviruses by "thermoelimination".

In the case of bacteria, an important event in the formation of this lineage was the invention of peptidoglycan and thick cell walls. The wide distribution of genes involved in the biosynthesis of peptidoglycan in bacterial genomes [139] suggests that this unique structure was already present in the LBCA. This invention could have dramatically reduced the number of viral lineages effective against bacteria, allowing bacteria to escape those lineages of viruses that are now archaea specific [140–142]. The efficiency of peptidoglycan against some devices produced by archaeal viruses is well illustrated by the failure of archaeal virus-associated pyramids (VAP) expressed in *Escherichia coli* to cross the peptidoglycan [143]. Archaeal VAP accumulate in the periplasm of *E. coli*, whereas those expressed in *Sulfolobus solfataricus* are formed and exposed at cell surface where they open for virions egress [143].

11. An Archaeon Is Born

To conclude this paper, let us have a time vision back to a population of LAECA-like organisms, relatively complex cells with internal membranes and spliced genes, infected by a myriad of diverse DNA and RNA viruses. In that population, a particular bug has two offsprings, each of them gave rise to many lineages by binary fission. In one of these lineages, cells improved their capacity for phagocytosis, increased their size, and became first class predators (the ancestors of eukaryotes). Some of them invent the nucleus and become free from many viruses that infected LAECA, except those that, in a first time, could replicate in the cytoplasm (later on, some viruses will find their way to the nucleus). To escape these big raptors [50], cells from another lineage started to reduce their size and increased their growth rate. Among their descendants, a particular lineage survives all damn big raptors living around by jumping into hot water. In that process, they get rid of many viruses that tortured them before (in particular all RNA viruses), but many viruses succeeded to follow them. Some descendants of these first hot swimmers started to like it very hot; making use of isoprenoids, they built a new type of membrane, and, fusing a helicase and a topoisomerase, they invented an amazing enzyme, reverse gyrase, to stabilize (we still do not know how) their genomes [144]. These superbugs became the only organisms capable of living at temperature near (or above) the boiling point of water. One of them became LACA, the last common ancestor of all modern archaea, organisms that had become so sophisticated in their way of life and physiology that they are now capable of confronting the giant descendants of the big raptors, sometimes even to live inside their guts.

At some point in that story, either before or after the emergence of LACA, archaea and/or protoarchaea have met other microbes in hot springs, bacteria. These fast-growing microbes also had succeeded to escape predators for a while and to get rid of many viruses previously disturbing their ancestors by inventing peptidoglycan (bringing with them some viruses well known by archaea), but they have not invented a new type of membrane. They have just adjusted their classical version to better survive in hot water. They have not invented reverse gyrase, but many of them will capture this amazing enzyme from archaea to thrive happily in hot water [145]. However, bacteria have invented another enzyme, DNA gyrase, which provides them with a dramatic selective advantage by coupling gene expression rapidly to environmental fluctuations via supercoiling-dependent modification of promoter activities [146]. With peptidoglycan as armour and gyrase to adapt rapidly changing environments, bacteria were ready to rule the world; they have now invaded all biotopes in the air, soil, and sea (except when temperature exceed 95°C) and the body of all organisms larger than themselves. However, archaea will survive this bacterial expansion and expand themselves out of their initial hot cradle. Taking benefit of their unique lipids, they will thrive in energy poor biotopes, deep in the ocean, or in soils and lakes with low oxygen content [147]. Later on, catching gyrase from bacteria, some euryarchaea (Haloarchaea, Archaeoglobales, Thermoplasmatales, and Methanogens) will become able to confront and coexist with bacteria with equal efficiency in many different types of environments [148].

12. Conclusion

The Scenario I favoured in this paper for the origin and evolution of archaea is at odds with the traditional view that "prokaryotes" gave rise to "eukaryotes". This traditional paradigm is so entrenched in our minds that it is not surprising that so many scientists endorse now "*fusion scenarios*" or "*archaeal ancestor's scenarios*" despite their many weaknesses. The confusing view that prokaryotes (assimilated to archaea and bacteria) predated eukaryotes (assimilated to modern eukaryotes) is inherent to the nomenclature "prokaryotes", meaning "*before the nucleus*". This is only one of the drawbacks of using the term prokaryote. I agree on this point with Pace who has strongly advocated to completely repel the term prokaryote [149]. However, despite the work of Woese and his followers, the unfortunate term prokaryote is still widely used for its convenience and I use it myself in this paper (although between "brackets"). In some cases, indeed, it is useful to refer to archaea and bacteria as two groups sharing similar traits (the coupling of transcription and translation, for instance) that are characteristic of the "prokaryotic" phenotype. In the future, I will try to replace the term prokaryote by the neutral term "akaryote", meaning without nucleus, that I proposed twenty years ago [150], a term that was reused recently by Harish and colleagues [91]. I proposed in the same 1992 paper to rename in parallel eukaryotes by the neutral term synkaryotes (with a nucleus). Indeed, I think today that it will be very difficult to get rid of the term "prokaryote" as long as we will use the term "eukaryote". However, synkaryote, referring to a phenotypic trait, is not really adequate to name a domain, defined instead

by genotypic traits [53]. At the moment, I am favouring the name *Splicea*, instead of eukarya, since possession of the spliceosome is a unique common trait to all "eukaryotes" derived from LECA. With this name, LECA becomes the LSCA and LAECA the LASCA, why not? Indeed, the origin (and fate) of the spliceosome(s) is, in my opinion, one of the more important questions in the history of life. If you like this novel nomenclature, you can change eukarya by Splicea and eukaryotes by spliceotes in this text, LECA by LSCA, and LAECA by LASCA and read it again with a fresh mind. The proposed hypotheses will possibly then seem less unorthodox to you.

Conflict of Interests

The author declares that there is no conflict of interests.

Acknowledgments

This paper is dedicated to the memory of Carl Woese, who has dramatically changed the life and career of so many biologists all over the world by his visionary work on microbial evolution. The author thanks Mart Krupovic for critical reading and correction of this paper.

References

[1] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.

[2] W. Zillig, "Comparative biochemistry of Archaea and Bacteria," *Current Opinion in Genetics & Development*, vol. 1, no. 4, pp. 544–551, 1991.

[3] G. Olsen and C. R. Woese, "Archaeal genomics: an overview," *Cell*, vol. 89, no. 7, pp. 991–994, 1997.

[4] P. Forterre, "Archaea: what can we learn from their sequences?" *Current Opinion in Genetics and Development*, vol. 7, no. 6, pp. 764–770, 1997.

[5] R. Garrett and H. P. Klenk, *Archaea*, Blackwell, Oxford, UK, 2007.

[6] C. Brochier-Armanet, P. Forterre, and S. Gribaldo, "Phylogeny and evolution of the Archaea: one hundred genomes later," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 274–281, 2011.

[7] G. Grüber and V. Marshansky, "New insights into structure-function relationships between archeal ATP synthase (A1A0) and vacuolar type ATPase (V1V0)," *BioEssays*, vol. 30, no. 11-12, pp. 1096–1109, 2008.

[8] B. Van den Berg, W. M. Clemons Jr., I. Collinson et al., "X-ray structure of a protein-conducting channel," *Nature*, vol. 427, no. 6969, pp. 36–44, 2004.

[9] X. Wang and J. Lutkenhaus, "FtsZ ring: the eubacterial division apparatus conserved in archaebacteria," *Molecular Microbiology*, vol. 21, no. 2, pp. 313–319, 1996.

[10] E. Gérard, B. Labedan, and P. Forterre, "Isolation of a minD-like gene in the hyperthermophilic archaeon pyrococcus AL585, and phylogenetic characterization of related proteins in the three domains of life," *Gene*, vol. 222, no. 1, pp. 99–106, 1998.

[11] K. S. Makarova, N. Yutin, S. D. Bell, and E. V. Koonin, "Evolution of diverse cell division and vesicle formation systems in Archaea," *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 731–741, 2010.

[12] N. Yutin, M. Y. Wolf, Y. I. Wolf, and E. V. Koonin, "The origins of phagocytosis and eukaryogenesis," *Biology Direct*, vol. 4, p. 9, 2009.

[13] N. Yutin and E. V. Koonin, "Archaeal origin of tubulin," *Biology Direct*, vol. 7, p. 10, 2012.

[14] T. Nunoura, Y. Takaki, J. Kakuta et al., "Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group," *Nucleic Acids Research*, vol. 39, no. 8, pp. 3204–3223, 2011.

[15] F. M. Cohan and A. F. Koeppel, "The origins of ecological diversity in prokaryotes," *Current Biology*, vol. 18, no. 21, pp. R1024–R1034, 2008.

[16] J. Armengaud, B. Fernandez, V. Chaumont et al., "Identification, purification, and characterization of an eukaryotic-like phosphopantetheine adenylyltransferase (coenzyme A biosynthetic pathway) in the hyperthermophilic archaeon Pyrococcus abyssi," *Journal of Biological Chemistry*, vol. 278, no. 33, pp. 31078–31087, 2003.

[17] T. Sato and H. Atomi, "Novel metabolic pathways in Archaea," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 307–314, 2011.

[18] J. Lombard, P. López-García, and D. Moreira, "Phylogenomic investigation of phospholipid synthesis in archaea," *Archaea*, vol. 2012, Article ID 630910, 13 pages, 2012.

[19] C. Brochier-Armanet, S. Gribaldo, and P. Forterre, "A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya," *Biology Direct*, vol. 3, p. 54, 2008.

[20] E. V. Koonin, "The origin and early evolution of eukaryotes in the light of phylogenomics," *Genome Biology*, vol. 11, no. 5, p. 209, 2010.

[21] J. Martijn and T. J. Ettema, "From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell," *Biochemical Society Transactions*, vol. 41, no. 1, pp. 451–457, 2013.

[22] J. Lombard, P. López-García, and D. Moreira, "The early evolution of lipid membranes and the three domains of life," *Nature Reviews Microbiology*, vol. 10, no. 7, pp. 507–515, 2012.

[23] E. V. Koonin and Y. I. Wolf, "Evolution of microbes and viruses: a paradigm shift in evolutionary biology?" *Frontiers in Cellular and Infection Microbiology*, vol. 2, p. 119, 2012.

[24] M. Pina, A. Bize, P. Forterre, and D. Prangishvili, "The archeoviruses," *FEMS Microbiology Reviews*, vol. 35, no. 6, pp. 1035–1054, 2011.

[25] M. Krupovic, D. Prangishvili, R. W. Hendrix, and D. H. Bamford, "Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere," *Microbiology and Molecular Biology Reviews*, vol. 75, no. 4, pp. 610–635, 2011.

[26] N. G. Abrescia, D. H. Bamford, J. M. Grimes, and D. I. Stuart, "Structure unifies the viral universe," *Annual Review of Biochemistry*, vol. 81, pp. 795–822, 2012.

[27] M. Krupovič, P. Forterre, and D. H. Bamford, "Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria," *Journal of Molecular Biology*, vol. 397, no. 1, pp. 144–160, 2010.

[28] N. S. Atanasova, E. Roine, A. Oren, D. H. Bamford, and H. M. Oksanen, "Global network of specific virus-host interactions in hypersaline environments," *Environmental Microbiology*, vol. 14, no. 2, pp. 426–440, 2012.

[29] M. Krupovic, A. Spang, S. Gribaldo, P. Forterre, and C. Schleper, "A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea," *Biochemical Society Transactions*, vol. 39, no. 1, pp. 82–88, 2011.

[30] M. K. Pietilä, N. S. Atanasova, V. Manole et al., "Virion architecture unifies globally distributed pleolipoviruses infecting halophilic archaea," *Journal of Virology*, vol. 86, no. 9, pp. 5067–5079, 2012.

[31] J. Filée, P. Siguier, and M. Chandler, "Insertion sequence diversity in archaea," *Microbiology and Molecular Biology Reviews*, vol. 71, no. 1, pp. 121–157, 2007.

[32] N. Soler, M. Gaudin, E. Marguet, and P. Forterre, "Plasmids, viruses and virus-like membrane vesicles from Thermococcales," *Biochemical Society Transactions*, vol. 39, no. 1, pp. 36–44, 2011.

[33] M. Krupovic, M. Gonnet, W. B. Hania, P. Forterre, and G. Erauso, "Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids," *PLoS One*, vol. 8, no. 1, Article ID e49044, 2013.

[34] B. O. Greve, S. Jensen, K. Brügger, W. Zillig, and R. A. Garrett, "Genomic comparison of archaeal conjugative plasmids from Sulfolobus," *Archaea*, vol. 1, no. 4, pp. 231–239, 2004.

[35] B. Greve, S. Jensen, H. Phan et al., "Novel RepA-MCM proteins encoded in plasmids pTAU4, pORA1 and pTIK4 from Sulfolobus neozealandicus," *Archaea*, vol. 1, no. 5, pp. 319–325, 2005.

[36] D. Cortez, S. Quevillon-Cheruel, S. Gribaldo et al., "Evidence for a Xer/dif system for chromosome resolution in archaea," *PLoS Genetics*, vol. 6, no. 10, Article ID e1001166, 2010.

[37] N. Soler, A. Justome, S. Quevillon-Cheruel et al., "The rolling-circle plasmid pTN1 from the hyperthermophilic archaeon Thermococcus nautilus," *Molecular Microbiology*, vol. 66, no. 2, pp. 357–370, 2007.

[38] P. Forterre, "Evolution, viral," in *Encyclopedia of Microbiology*, M. Schaechter, Ed., pp. 370–389, Elsevier, New York, NY, USA, 3rd edition, 2009.

[39] A. K. Kalliomaa-Sanford, F. A. Rodriguez-Castañeda, B. N. McLeod et al., "Chromosome segregation in Archaea mediated by a hybrid DNA partition machine," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 10, pp. 3754–3579, 2012.

[40] P. Forterre, "Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 10, pp. 3669–3674, 2006.

[41] R. Sorek, C. M. Lawrence, and B. Wiedenheft, "CRISPR-mediated adaptive immune systems in Bacteria and Archaea," *Annual Review of Biochemistry*, vol. 82, pp. 237–266, 2013.

[42] K. S. Makarova, Y. I. Wolf, and E. V. Koonin, "Comparative genomics of defense systems in archaea and bacteria," *Nucleic Acids Research*, vol. 41, no. 8, pp. 4360–4377, 2013.

[43] K. S. Makarova, Y. I. Wolf, J. van der Oost, and E. V. Koonin, "Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements," *Biology Direct*, vol. 4, p. 29, 2009.

[44] A. M. Poole and D. Penny, "Evaluating hypotheses for the origin of eukaryotes," *BioEssays*, vol. 29, no. 1, pp. 74–84, 2007.

[45] S. Gribaldo, A. M. Poole, V. Daubin, P. Forterre, and C. Brochier-Armanet, "The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse?" *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 743–752, 2010.

[46] P. Forterre, "A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong," *Research in Microbiology*, vol. 162, no. 1, pp. 77–91, 2011.

[47] N. Yutin, K. S. Makarova, S. L. Mekhedov, Y. I. Wolf, and E. V. Koonin, "The deep archaeal roots of eukaryotes," *Molecular Biology and Evolution*, vol. 25, no. 8, pp. 1619–1630, 2008.

[48] P. López-Garćia and D. Moreira, "Metabolic symbiosis at the origin of eukaryotes," *Trends in Biochemical Sciences*, vol. 24, no. 3, pp. 88–93, 1999.

[49] W. Martin and M. Müller, "The hydrogen hypothesis for the first eukaryote," *Nature*, vol. 392, no. 6671, pp. 37–41, 1998.

[50] C. G. Kurland, L. J. Collins, and D. Penny, "Genomics and the irreducible nature of eukaryote cells," *Science*, vol. 312, no. 5776, pp. 1011–1014, 2006.

[51] C. De Duve, "The origin of eukaryotes: a reappraisal," *Nature Reviews Genetics*, vol. 8, no. 5, pp. 395–403, 2007.

[52] T. Cavalier-Smith, "Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution," *Biology Direct*, vol. 5, p. 7, 2010.

[53] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[54] B. El Yacoubi, I. Hatin, C. Deutsch et al., "Role for the universal Kae1/Qri7/YgjD (COG0533) family in tRNA modification," *EMBO Journal*, vol. 30, no. 5, pp. 882–893, 2011.

[55] C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, and T. M. Embley, "The archaebacterial origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 51, pp. 20356–20361, 2008.

[56] A. Hecker, N. Leulliot, D. Gadelle et al., "An archaeal orthologue of the universal protein Kae1 is an iron metalloprotein which exhibits atypical DNA-binding properties and apurinic-endonuclease activity in vitro," *Nucleic Acids Research*, vol. 35, no. 18, pp. 6042–6051, 2007.

[57] A. M. Poole and N. Neumann, "Reconciling an archaeal origin of eukaryotes with engulfment: a biologically plausible update of the Eocyte hypothesis," *Research in Microbiology*, vol. 162, no. 1, pp. 71–76, 2011.

[58] P. G. Foster, C. J. Cox, and T. Martin Embley, "The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods," *Philosophical Transactions of the Royal Society B*, vol. 364, no. 1527, pp. 2197–2207, 2009.

[59] L. Guy and T. J. G. Ettema, "The archaeal 'TACK' superphylum and the origin of eukaryotes," *Trends in Microbiology*, vol. 19, no. 12, pp. 580–587, 2011.

[60] T. A. Williams, P. G. Foster, T. M. Nye, C. J. Cox, and T. M. Embley, "A congruent phylogenomic signal places eukaryotes within the Archaea," *Proceedings of the Royal Society B*, vol. 279, no. 1749, pp. 4870–4879, 2012.

[61] J. A. Lake, E. Henderson, M. Oakes, and M. W. Clark, "Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, no. 12, pp. 3786–3790, 1984.

[62] C. P. Vivarès, M. Gouy, T. Thomarat, and G. Méténier, "Functional and evolutionary analysis of a eukaryotic parasitic genome," *Current Opinion in Microbiology*, vol. 5, no. 5, pp. 499–505, 2002.

[63] B. A. Curtis, G. Tanifuji, F. Burki et al., "Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs," *Nature*, vol. 492, no. 7427, pp. 59–65, 2012.

[64] U. Jahn, R. Summons, H. Sturt, E. Grosjean, and H. Huber, "Composition of the lipids of Nanoarchaeum equitans and their origin from its host Ignicoccus sp. strain KIN4/I," *Archives of Microbiology*, vol. 182, no. 5, pp. 404–413, 2004.

[65] P. Forterre, "The virocell concept and environmental microbiology," *The ISME Journal*, vol. 7, no. 2, pp. 233–236, 2013.

[66] S. Nelson-Sathi, T. Dagan, G. Landan et al., "Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 50, pp. 20537–20542, 2012.

[67] P. López-García and D. Moreira, "Selective forces for the origin of the eukaryotic nucleus," *BioEssays*, vol. 28, no. 5, pp. 525–533, 2006.

[68] W. Martin and E. V. Koonin, "Introns and the origin of nucleus-cytosol compartmentalization," *Nature*, vol. 440, no. 7080, pp. 41–45, 2006.

[69] G. Jékely, "Origin of the nucleus and Ran-dependent transport to safeguard ribosome biogenesis in a chimeric cell," *Biology Direct*, vol. 3, p. 31, 2008.

[70] C. R. Woese, "Interpreting the universal phylogenetic tree," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 15, pp. 8392–8396, 2000.

[71] O. Matte-Tailliez, C. Brochier, P. Forterre, and H. Philippe, "Archaeal phylogeny based on ribosomal proteins," *Molecular Biology and Evolution*, vol. 19, no. 5, pp. 631–639, 2002.

[72] H. Philippe and P. Forterre, "The rooting of the universal tree of life is not reliable," *Journal of Molecular Evolution* , vol. 49, no. 4, pp. 509–523, 1999.

[73] L. P. Villarreal and V. R. DeFilippis, "A hypothesis for DNA viruses as the origin of eukaryotic replication proteins," *Journal of Virology*, vol. 74, no. 15, pp. 7079–7084, 2000.

[74] P. Forterre, "Giant viruses: conflicts in revisiting the virus concept," *Intervirology*, vol. 53, no. 5, pp. 362–378, 2010.

[75] L. G. Pühler, H. Leffers, F. Gropp et al., "Archaebacterial DNA-dependent RNA polymerase testify to the evolution of the eukaryotic nuclear genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 12, pp. 4569–4573, 1989.

[76] E. Lasek-Nesselquist and J. P. Gogarten, "The effects of model choice and mitigating bias on the ribosomal tree of life," *Molecular Phylogenetics and Evolution*, vol. 69, no. 1, pp. 17–38, 2013.

[77] P. Forterre, "Thermoreduction, a hypothesis for the origin of prokaryotes," *Comptes Rendus de l'Academie des Sciences*, vol. 318, no. 4, pp. 415–422, 1995.

[78] D. Raoult, S. Audic, C. Robert et al., "The 1.2-megabase genome sequence of Mimivirus," *Science*, vol. 306, no. 5700, pp. 1344–1350, 2004.

[79] T. A. Williams, T. M. Embley, and E. Heinz, "Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses," *PLoS One*, vol. 6, no. 6, Article ID e21080, 2011.

[80] B. Boussau, S. Blanquart, A. Necsulea, N. Lartillot, and M. Gouy, "Parallel adaptations to high temperatures in the Archaean eon," *Nature*, vol. 456, no. 7224, pp. 942–945, 2008.

[81] M. Groussin and M. Gouy, "Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea," *Molecular Biology and Evolution*, vol. 28, no. 9, pp. 2661–2674, 2011.

[82] C. Brochier-Armanet, B. Boussau, S. Gribaldo, and P. Forterre, "Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota," *Nature Reviews Microbiology*, vol. 6, no. 3, pp. 245–252, 2008.

[83] C. O. Lovejoy, "Reexamining human origins in light of Ardipithecus ramidus," *Science*, vol. 326, no. 5949, pp. 74e2–74e8, 2009.

[84] Y. I. Wolf and E. V. Koonin, "Genome reduction as the dominant mode of evolution," *Bioessays*, vol. 35, no. 9, pp. 829–837, 2013.

[85] Y. I. Wolf, K. S. Makarova, N. Yutin, and E. V. Koonin, "Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer," *Biology Direct*, vol. 7, p. 46, 2012.

[86] M. Csurös and I. Miklós, "Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model," *Molecular Biology and Evolution*, vol. 26, no. 9, pp. 2087–2095, 2009.

[87] O. Lecompte, R. Ripp, J. Thierry, D. Moras, and O. Poch, "Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale," *Nucleic Acids Research*, vol. 30, no. 24, pp. 5382–5390, 2002.

[88] E. Desmond, C. Brochier-Armanet, P. Forterre, and S. Gribaldo, "On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes," *Research in Microbiology*, vol. 162, no. 1, pp. 53–70, 2011.

[89] M. Wang, C. G. Kurland, and Caetano-Anollés, "Reductive evolution of proteomes and protein structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 29, pp. 11954–11958, 2011.

[90] A. Nasir, K. M. Kim, and G. Caetano-Anolles, "Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya," *BMC Evolutionary Biology*, vol. 12, p. 156, 2012.

[91] A. Harish, A. Tunlid, and G. C. Kurland, "Rooted phylogeny of the three superkingdoms," *Biochimie*, vol. 95, no. 8, pp. 1593–1604, 2013.

[92] M. Carlile, "Prokaryotes and eukaryotes: strategies and successes," *Trends in Biochemical Sciences*, vol. 7, no. 4, pp. 128–130, 1982.

[93] E. R. Angert, "DNA replication and genomic architecture of very large bacteria," *Annual Review of Microbiology*, vol. 66, pp. 197–212, 2012.

[94] I. B. Rogozin, L. Carmel, M. Csuros, and E. V. Koonin, "Origin and evolution of spliceosomal introns," *Biology Direct*, vol. 7, p. 11, 2012.

[95] B. Dujon, "Yeast evolutionary genomics," *Nature Reviews Genetics*, vol. 1, no. 7, pp. 512–524, 2010.

[96] K. O. Stetter, "A brief history of the discovery of hyperthermophilic life," *Biochemical Society Transactions*, vol. 41, no. 1, pp. 416–420, 2013.

[97] P. Forterre, "A hot topic: the origin of hyperthermophiles," *Cell*, vol. 85, no. 6, pp. 789–792, 1996.

[98] N. Glansdorff, Y. Xu, and B. Labedan, "The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner," *Biology Direct*, vol. 3, p. 29, 2008.

[99] N. Glansdorff, Y. Xu, and B. Labedan, "The origin of life and the last universal common ancestor: do we need a change of perspective?" *Research in Microbiology*, vol. 160, no. 7, pp. 522–528, 2009.

[100] Y. Koga, "Thermal adaptation of the archaeal and bacterial lipid membranes," *Archaea*, vol. 2012, Article ID 789652, 6 pages, 2012.

[101] T. D. Brock, "Life at high températures," *Science*, vol. 158, no. 3804, pp. 1012–1019, 1967.

[102] J. F. De Jonckheere, M. Baumgartner, S. Eberhardt, F. R. Opperdoes, and K. O. Stetter, "Oramoeba fumarolia gen. nov., sp. nov., a new marine heterolobosean amoeboflagellate growing at 54°C," *European Journal of Protistology*, vol. 47, no. 1, pp. 16–23, 2011.

[103] B. Bolduc, D. P. Shaughnessy, Y. I. Wolf, E. V. Koonin, F. F. Roberto, and M. Young, "Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot Springs," *Journal of Virology*, vol. 86, no. 10, pp. 5562–5573, 2012.

[104] E. V. Koonin, "The incredible expanding ancestor of eukaryotes," *Cell*, vol. 140, no. 5, pp. 606–608, 2010.

[105] N. Lane and W. Martin, "The energetics of genome complexity," *Nature*, vol. 467, no. 7318, pp. 929–934, 2010.

[106] P. Forterre and S. Gribaldo, "Bacteria with a eukaryotic touch: a glimpse of ancient evolution?" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 29, pp. 12739–12740, 2010.

[107] M. Krupovič, S. Gribaldo, D. H. Bamford, and P. Forterre, "The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with Hitchhiking of mobile genetic elements," *Molecular Biology and Evolution*, vol. 27, no. 12, pp. 2716–2732, 2010.

[108] J. Filée, P. Forterre, T. Sen-Lin, and J. Laurent, "Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins," *Journal of Molecular Evolution*, vol. 54, no. 6, pp. 763–773, 2002.

[109] E. V. Koonin and N. Yutin, "Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses," *Intervirology*, vol. 53, no. 5, pp. 284–292, 2010.

[110] J. M. Cock, L. Sterck, P. Rouzé et al., "The Ectocarpus genome and the independent evolution of multicellularity in brown algae," *Nature*, vol. 465, no. 7298, pp. 617–621, 2010.

[111] H. Ogata and J. Claverie, "Unique genes in giant viruses: regular substitution pattern and anomalously short size," *Genome Research*, vol. 17, no. 9, pp. 1353–1361, 2007.

[112] A. Dupressoir, C. Lavialle, and T. Heidmann, "From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation," *Placenta*, vol. 33, no. 9, pp. 663–671, 2012.

[113] A. C. Esser, N. Ahmadinejad, C. Wiegand et al., "A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes," *Molecular Biology and Evolution*, vol. 21, no. 9, pp. 1643–1660, 2004.

[114] M. C. Rivera and J. A. Lake, "The ring of life provides evidence for a genome fusion origin of eukaryotes," *Nature*, vol. 431, no. 7005, pp. 152–155, 2004.

[115] D. Pisani, J. A. Cotton, and J. O. McInerney, "Supertrees disentangle the chimerical origin of eukaryotic genomes," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1752–1760, 2007.

[116] T. Thiergart, G. Landan, M. Schenk, T. Dagan, and W. F. Martin, "An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin," *Genome Biology and Evolution*, vol. 4, no. 4, pp. 466–485, 2012.

[117] D. Alvarez-Ponce, P. Lopez, E. Bapteste, and J. O. McInerney, "Gene similarity networks provide tools for understanding eukaryote origins and évolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 7, pp. 1594–1603, 2013.

[118] W. F. Doolittle, "You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes," *Trends in Genetics*, vol. 14, no. 8, pp. 307–311, 1998.

[119] J. Filée and M. Chandler, "Convergent mechanisms of genome evolution of large and giant DNA viruses," *Research in Microbiology*, vol. 159, no. 5, pp. 325–331, 2008.

[120] P. Forterre and D. Prangishvili, "The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties," *Annals of the New York Academy of Sciences*, vol. 1178, pp. 65–77, 2009.

[121] E. V. Koonin and B. Moss, "Viruses know more than one way to don a cap," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 8, pp. 3283–3284, 2010.

[122] I. R. Arkhipova, M. A. Batzer, J. Brosius et al., "Genomic impact of eukaryotic transposable elements," *Mobile DNA*, vol. 3, no. 1, p. 19, 2012.

[123] E. A. Gladyshev and I. R. Arkhipova, "Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 22, pp. 9352–9357, 2007.

[124] I. R. Arkhipova, "Distribution and phylogeny of penelope-like elements in eukaryotes," *Systematic Biology*, vol. 55, no. 6, pp. 875–885, 2006.

[125] M. F. Singer, "Unusual reverse transcriptases," *Journal of Biological Chemistry*, vol. 270, no. 42, pp. 24623–24626, 1995.

[126] A. C. Chueh, E. L. Northrop, K. H. Brettingham-Moore, K. H. A. Choo, and L. H. Wong, "LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin," *PLoS Genetics*, vol. 5, no. 1, Article ID e1000354, 2009.

[127] C. I. Bandea, "A unifyed scenario on the origin and évolution of cellular and viral domains," *Nature Preceedings*, 2009, http://precedings.nature.com/.

[128] S. Miller and J. Krijnse-Locker, "Modification of intracellular membrane structures for virus replication," *Nature Reviews Microbiology*, vol. 6, no. 5, pp. 363–374, 2008.

[129] M. Suzan-Monti, B. La Scola, L. Barrassi, L. Espinosa, and D. Raoult, "Ultrastructural characterization of the giant volcano-like virus factory of Acanthamoeba polyphaga Mimivirus," *PLoS One*, vol. 2, no. 3, Article ID e328, 2007.

[130] T. Cavalier-Smith, "Intron phylogeny: a new hypothesis," *Trends in Genetics*, vol. 7, no. 5, pp. 145–148, 1991.

[131] C. E. Lane, K. Van Den Heuvel, C. Kozera et al., "Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 50, pp. 19908–19913, 2007.

[132] D. H. Bamford, "Do viruses form lineages across different domains of life?" *Research in Microbiology*, vol. 154, no. 4, pp. 231–236, 2003.

[133] P. Forterre and M. Krupovic, "The origin of virions and virocells: the escape hypothesis revisited," in *Viruses: Essential Agents of Life*, G. Witzany, Ed., pp. 43–60, Springer Science+Business Media, Dordrecht, The Netherlands, 2012.

[134] P. Forterre and H. Philippe, "Where is the root of the tree of life," *Bioessays*, vol. 21, no. 10, pp. 871–879, 1999.

[135] P. Forterre, "The universal tree of life and the Last Universal Cellular Ancestor (LUCA): revolution and counter-revolutions," in *Evolutionary Genomics and Systems Biology*, Caetano-Anollés, Ed., pp. 43–62, 2010.

[136] C. R. Woese and G. E. Fox, "The concept of cellular évolution," *Journal of Molecular Evolution*, vol. 10, no. 1, pp. 1–6, 1977.

[137] P. Forterre, "The origin of DNA genomes and DNA replication proteins," *Current Opinion in Microbiology*, vol. 5, no. 5, pp. 525–532, 2002.

[138] J. Berthon, R. Fujikane, and P. Forterre, "When DNA replication and protein synthesis come together," *Trends in Biochemical Sciences*, vol. 34, no. 9, pp. 429–434, 2009.

[139] C. Cayrou, B. Henrissat, P. Gouret, P. Pontarotti, and M. Drancourt, "Peptidoglycan: a post-genomic analysis," *BMC Microbiology*, vol. 12, p. 294, 2012.

[140] M. Jalasvuori and J. K. H. Bamford, "Structural co-evolution of viruses and cells in the primordial world," *Origins of Life and Evolution of Biospheres*, vol. 38, no. 2, pp. 165–181, 2008.

[141] D. Prangishvili, "The wonderful world of archaeal viruses," *Annual Review of Microbiology*, vol. 67, pp. 565–585, 2013.

[142] P. Forterre and D. Prangishvili, "The major role of viruses in cellular evolution: facts and hypotheses," *Current Opinion in Virology*, 2013.

[143] T. E. F. Quax, S. Lucas, J. Reimann et al., "Simple and elegant design of a virion egress structure in Archaea," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 8, pp. 3354–3359, 2011.

[144] P. Forterre, "A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein," *Trends in Genetics*, vol. 18, no. 5, pp. 236–238, 2002.

[145] C. Brochier-Armanet and P. Forterre, "Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers," *Archaea*, vol. 2, no. 2, pp. 83–93, 2007.

[146] P. Forterre and D. Gadelle, "Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms," *Nucleic Acids Research*, vol. 37, no. 3, pp. 679–692, 2009.

[147] D. L. Valentine, "Adaptations to energy stress dictate the ecology and evolution of the Archaea," *Nature Reviews Microbiology*, vol. 5, no. 4, pp. 316–323, 2007.

[148] P. Forterre, S. Gribaldo, D. Gadelle, and M. Serre, "Origin and evolution of DNA topoisomerases," *Biochimie*, vol. 89, no. 4, pp. 427–446, 2007.

[149] N. R. Pace, "Time for a change," *Nature*, vol. 441, no. 7091, p. 289, 2006.

[150] P. Forterre, "Neutral terms," *Nature*, vol. 335, p. 305, 1992.

Hindawi Publishing Corporation
Archaea
Volume 2013, Article ID 680436, 9 pages
http://dx.doi.org/10.1155/2013/680436

Research Article

Comparative Analysis of Barophily-Related Amino Acid Content in Protein Domains of *Pyrococcus abyssi* and *Pyrococcus furiosus*

Liudmila S. Yafremava,[1,2] Massimo Di Giulio,[3] and Gustavo Caetano-Anollés[1]

[1] *Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[2] *National Center for Supercomputing Applications, Urbana, IL 61801, USA*
[3] *Laboratory of Molecular Evolution, Institute of Genetics and Biophysics "Adriano Buzzati-Traverso", CNR, 80131 Napoli, Italy*

Correspondence should be addressed to Gustavo Caetano-Anollés; gca@illinois.edu

Received 4 July 2013; Revised 21 August 2013; Accepted 23 August 2013

Academic Editor: Kyung Mo Kim

Copyright © 2013 Liudmila S. Yafremava et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Amino acid substitution patterns between the nonbarophilic *Pyrococcus furiosus* and its barophilic relative *P. abyssi* confirm that hydrostatic pressure asymmetry indices reflect the extent to which amino acids are preferred by barophilic archaeal organisms. Substitution patterns in entire protein sequences, shared protein domains defined at fold superfamily level, domains in homologous sequence pairs, and domains of very ancient and very recent origin now provide further clues about the environment that led to the genetic code and diversified life. The pyrococcal proteomes are very similar and share a very early ancestor. Relative amino acid abundance analyses showed that biases in the use of amino acids are due to their shared fold superfamilies. Within these repertoires, only two of the five amino acids that are preferentially barophilic, aspartic acid and arginine, displayed this preference significantly and consistently across structure and in domains appearing in the ancestor. The more primordial asparagine, lysine and threonine displayed a consistent preference for nonbarophily across structure and in the ancestor. Since barophilic preferences are already evident in ancient domains that are at least ~3 billion year old, we conclude that barophily is a very ancient trait that unfolded concurrently with genetic idiosyncrasies in convergence towards a universal code.

1. Introduction

The biophysical properties of amino acids determine their use in proteins. Amino acid polarity and molecular volume are especially important for protein stability and function in hyper thermophilic and barophilic conditions. These two properties have been associated to the origins of the genetic code [1]. Thus, tracing amino acids with common physicochemical properties may help derive the conditions in which the genetic code originated.

A method was developed previously to assign temperature and pressure asymmetry indices to amino acids [2]. These indices are based on patterns of amino acid substitution within homologous sequences of phylogenetically related organisms living in two different environmental conditions, including barophilic versus nonbarophilic and thermophilic versus nonthermophilic conditions. The temperature asymmetry index (TAI) reflects the extent to which an amino acid is preferred by hyper thermophiles and was studied in *Deinococcus radiodurans*, *Thermus thermophilus* [3], *Methanococci*, and *Bacilli* [2]. The hydrostatic pressure asymmetry index (PAI) reflects the extent to which an amino acid is preferred by barophiles; it was studied in *Pyrococcus furiosus* and *P. abyssi* [4] and recently extended to the *P. furiosus—P. yayanosi* and *Thermococcus kodakarensis—T. barophilus* pairs [5]. The strength of statistical significance of this preference allows ranking amino acids for their propensities to be used in hyper thermophilic [6] and barophilic organisms [7]. Under the hypothesis that life may have arisen in a thermobarophilic environment, such as hydrothermal vents where hot volcanic exhalations clashed with circulating hydrothermal water flows and primed early

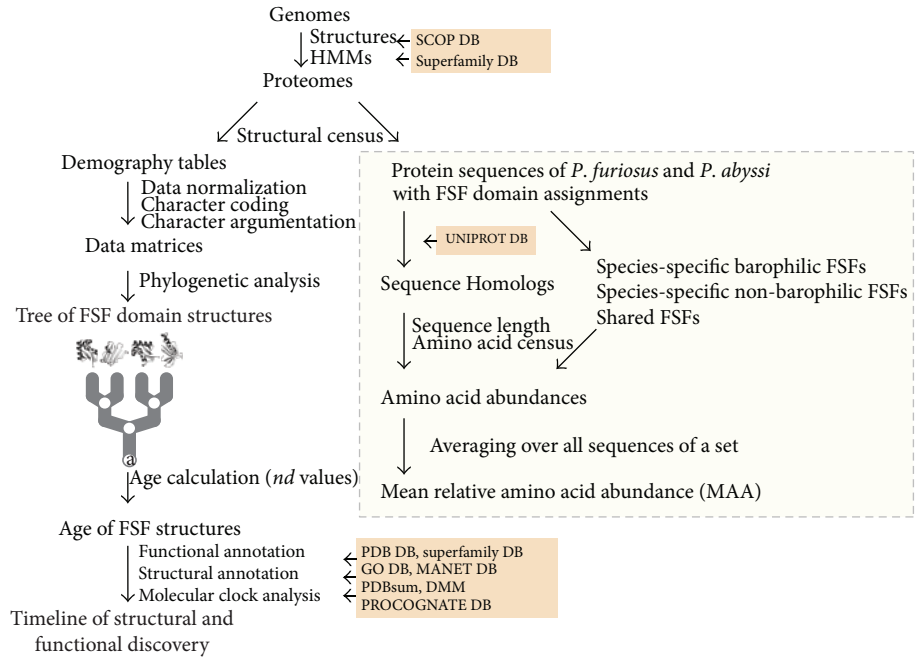


Figure 1: Flow chart describing the experimental strategy of the study.

metabolic chemistries [8], we set out to use these measures of thermobarophily as tools for further exploration.

If ancestral organisms originated in extremophilic conditions and genetics has recorded ancestral history, how did their descendants diverge over the course of evolutionary time? This divergence must be reflected in the function of proteins they use, which is nested in their three-dimensional structure [9]. The structural classification of proteins (SCOP) groups protein domains into a structural hierarchy that includes several levels of complexity: class, fold superfamily and family, from top to bottom [10]. Studies of domains, especially of fold superfamilies (FSFs), demonstrate their usefulness in investigating the divergence of organisms at the levels of molecular organization and physiological function (e.g., [11–18]). Thus, it is likely that the bias in the amino acid composition of protein sequences is reflected in protein structure, which could allow the study of the evolutionary divergence of organisms from a common thermo-barophilic ancestor. The hyper thermophilic barophile *Pyrococcus abyssi* and hyper thermophilic nonbarophile *P. furiosus* are two closely related organisms that are very suitable for such analysis due to the extreme similarity of their physiology [19] and the very early evolutionary origins of its lineages [16].

In this study we analyze the amino acid distributions within and outside the domain regions of protein sequences, with domains defined at FSF level of the SCOP hierarchy. We also explore the relationship between barophilic amino acid distributions in domain regions and the evolutionary age of the respective FSFs. Amino acid barophily and thermophily ranks tend to be specific to the pair of organisms for which they are documented. As mentioned above, barophily ranks have been previously described for *Pyrococcus*, but thermophily ranks have not. Thus, we focus our analysis on barophily. Results are interpreted in terms of ecological and physiological differences between the organisms to understand the process of divergence of the two species and in the context of protein history.

2. Materials and Methods

FSF assignments and their respective sequences were obtained from a structural genomic census in 749 organisms [20] that used advanced linear HMMs of structural recognition in superfamily [21], probability cutoffs e of 10^{-4}, and domain definitions from SCOP version 1.73 [10] (Figure 1). FSFs were segregated into 3 classes: (1) those present only in the barophile (species-specific barophilic FSFs), (2) those present only in the nonbarophile (species-specific nonbarophilic FSFs), and (3) those present in both species (shared FSFs). Statistical analysis of amino acid content was performed on each group, as well as on those parts of each sequence that were found outside domain assignments. A separate analysis was performed on identical FSF assignments of homologous sequences. Sequence homology was determined by mining the Uniprot database [22]. The evolutionary age of FSFs was derived directly from the phylogenomic tree of FSF domains reconstructed from the global census of protein structures (for methods see [20]). Because trees of domains are rooted and are highly unbalanced, we unfolded the relative age of protein domains directly for the phylogeny as a distance in nodes (*nd*, node distance) from the hypothetical ancestral structure at the base of the tree. *nd* was calculated by counting the number of internal nodes along a lineage from the root to a terminal node (a leaf) of the tree on a relative 0-1 scale with the following equation: nd_a = (number of internal nodes between nodes r and a)/(number of internal nodes between nodes r and m), where a is the target leaf node, r is

Table 1: Protein sequences and FSF domain statistics.

Statistics	*P. furiosus* DSM 363	*P. abyssi* GE5
Number of sequences with FSF assignment	1,339 (63%)	1,229 (65%)
Average sequence length in base pairs ± SD (minimum–maximum length)	276 ± 181 (21–1,740)	289 ± 190 (18–2,122)
Fraction of amino acids found in FSFs (not in intervening sequences)	61%	61%
Total FSFs	495	472
Species-specific FSFs	43	20
Average number of amino acids in FSFs (minimum–maximum length)	267 (5–1,107)	272 (5–1,107)

the hypothetical root node, and m is a leaf node that has the largest possible number of internal nodes from r. Consequently, the nd value of the most ancestral taxon is 0, whereas that of the most recent one is 1. The molecular clock for protein domains at FSF level ($t = -3.831\,nd + 3.628$) [20] was used to calculate the geological ages of selected FSFs in billions of years (Gy). Calibrations showed a significant linear correlation ($P < 0.0001$) between FSF age (nd) and geological time. The P-loop hydrolase fold structure, the most ancient FSF of our timeline, is used as lower boundary and linked to the earliest evidence of biological activity derived from ion microprobe analysis and isotopic composition of carbonaceous inclusions in 3.8 Gy-old banded iron rock formations. Other FSFs were linked to the biosynthesis of porphyrins (spectroscopic identification of vanadyl-porphyrin complexes in carbonaceous matter embedded in 3.49 Gy-old polycrystalline rocks), enzymes of nitrogen assimilation (with ages inferred mostly from biogeochemical evidence), lipid biomarkers such as hopanoids and biphytanes recovered from kerogen, bitumens and hydrocarbons, markers of bacterial and eukaryotic diversification episodes with times established from microfossil evidence (e.g., unicellular cyanobacterial coccoids in 1.9 Gy-old tidal sedimentary rock and acritarchs in 1.5 Gy-old rocks from Northern Australia) integrated with molecular, physiological, paleontological and geochemical data, folds linked to biological processes and lineages (e.g., biosynthesis of flavonoids and red algae, hemocyanins and mollusks), and finally present day boundary FSFs [20].

The relative abundance of each amino acid in a sequence was calculated as the number of amino acid instances divided by the total length of that sequence. The relative abundances of amino acids in domain sequences or entire protein sequences were normalized by the length of the domains or the length of entire protein sequences, respectively. These numbers were then averaged over all sequences in a group under consideration, obtaining mean relative amino acid abundance (MAA) measures specific for each amino acid. Analysis for homologous sequences was slightly different. Only the homologous sequences with identical FSF domain assignments in the two organisms were used. First, the relative amino acid abundance was calculated for each FSF assignment in each sequence. If a sequence had multiple repeats of the same FSF, the MAA values were averaged within that sequence. MAA differences were calculated for the homologous pair of sequences from the two organisms. If many sequence pairs contained a particular FSF, these MAA differences were subsequently averaged over the dataset. Statistical analyses were performed in *R* and Instat using the Welch 2-sample t-test.

3. Results and Discussion

3.1. Exploring General Tendencies in Amino Acid Use. Previous work established that *P. abyssi* tends to substitute arginine (Arg), serine (Ser), valine (Val), aspartate (Asp), and glycine (Gly) for amino acids in sequences homologous to *P. furiosus* [4]. Henceforth we will refer to them as barophilic amino acids. Conversely, *P. furiosus* tends to substitute asparagine (Asn), lysine (Lys), proline (Pro), isoleucine (Ile), threonine (Thr), glutamine (Gln), and tyrosine (Tyr) for amino acids in sequences homologous to *P. abyssi*. Henceforth we will refer to them as nonbarophilic amino acids. The substitution of leucine (Leu), histidine (His), phenylalanine (Phe), methionine (Met), glutamate (Glu), alanine (Ala), cysteine (Cys), and tryptophan (Trp) appears to be unbiased. These preferences were determined based on the statistical significance of the underlying patterns of substitution in homologous sequences. They allow grouping amino acids into three broad categories: barophilic, nonbarophilic, and indifferent. Our task is to establish patterns of amino acid use depending on their barophilic group and location in the protein sequence. We started from the known point of reference, documenting the amino acid use in the entire protein sequences of the two organisms. From there we progressively focused on the ancestral set of functional protein sequences by studying amino acid use in domain sequences and regions that intervene between domains, then comparing amino acid counts in species-specific and shared FSF domain structures, and finally performing pairwise comparisons of amino acid use in matching domain regions of homologous sequence pairs. Finally, we compared the amino acid preferences of FSF domains that are considered most ancient and most recent from an evolutionary point of view, using ages of domain structures inferred from a structural phylogenomic census that is very well indexed [20].

3.2. P. furiosus and P. abyssi Are Very Similar at the Level of Protein Domain Structure and Share a Very Ancient Proteomic Ancestor. The two species of *Pyrococci* share the majority (452) of their FSF domain structures (Table 1). *P. abyssi* has fewer FSFs (472) than *P. furiosus* (495), probably because it inhabits an extremophilic niche that combines extreme pressure and temperature, both of which have been shown to put limits on viable protein structures (e.g., [23, 24]). The similarities in FSF content of proteomes are explained by

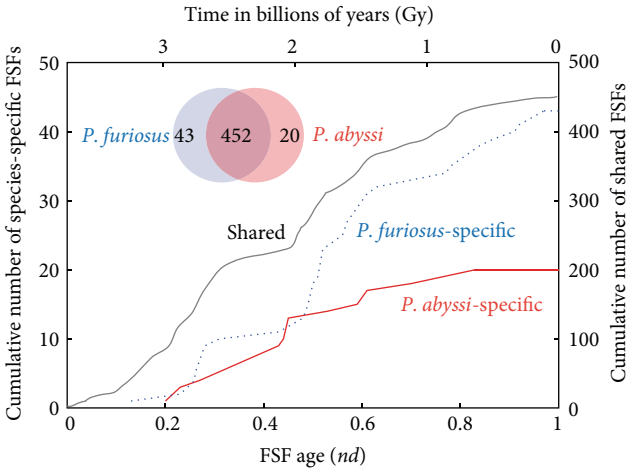


FIGURE 2: Evolutionary accumulation of protein domain structures defined at FSF level of the structural hierarchy in SCOP. The cumulative plot divides FSFs into those that are shared between *Pyrococcus* and those that are species-specific. Domain age is provided in a relative scale (*nd*, node distance) or in timescales in billions of years (Gy) according to a molecular clock of FSFs. The Venn diagram describes the actual FSF counts in the proteomes of *P. furiosus* and *P. abyssi*.

the common biological heritage and similar physiology of the two organisms, which differ mainly in the utilization of metabolic substrates [19]. These differences must be reflected in the function of FSFs that are specific to each of the two organisms. To evaluate this functional link, we annotated molecular functions using FSF assignment definitions of the superfamily database [15]. Out of 20 FSFs specific to *P. abyssi*, 8 participate in "metabolism", 6 in "regulation", 2 in "intracellular processes", and 4 belong to the "general" and "other" categories. Out of 43 FSFs specific to *P. furiosus*, 21 participate in "metabolism", 6 in "regulation", 5 in "intracellular processes" and 1 in "extracellular processes", and 6 belong to the "general", "other," and "unknown" categories. As expected, the majority (30–40%) of the species-specific FSFs participate in metabolic functions.

The history of divergence of protein structure and function has been previously studied with cumulative plots of protein domain structures that display the rate at which organisms accumulate domains over the course of evolutionary history [16, 25, 26]. This method is useful for the study of the diversification of organisms (Figure 2). The cumulative plot of species-specific FSFs for the two *Pyrococci* indicates early divergence both in structure and in function. This result is congruent with other studies that suggest early divergence based on sequence phylogeny and gene loss [19] and a very ancestral trend of loss in protein domain repertoires of Archaea responsible for their very early origin [16]. Since the age of FSF domain structures follows a tight molecular clock [20], it was possible to establish a timeframe for the appearance of lineage-specific FSFs in evolution. *P. furiosus*-specific FSFs appeared for the first time ~3.2 Gy ago and were closely followed by *P. abyssi*-specific FSFs, which appeared ~3 Gy ago. The timeline therefore suggests an early hyper thermophilic origin of the barophilic organism (by domain loss [16]), regardless of the accuracy of geological time assignments. However, lineage-specific FSFs are indicative of the origin of the organism and not the origin of barophilic traits. We also note that the acquisition of lineage-specific FSFs by *P. abyssi* is slower than *P. furiosus*, probably due to greater constraints imposed on *P. abyssi* by its barophilic environment. A fully enzymatic biosynthetic pathway for purine biosynthesis and a functional ribosome were already in place ~3 Gy ago during the rise of the *P. abyssi* lineage, fulfilling the expanding matter-energy and processing needs of genomic information [26]. During that time, aminoacyl-tRNA synthetases accreted anticodon-binding domains [27], unfolding the specificity of the genetic code and biasing amino acid composition of flexible regions in protein structure [28]. We note that Archaea suffered very early and impactful evolutionary episodes of genomic reduction [16, 17], setting archaeal organisms apart from those in other superkingdoms. These episodes and their associated "loser trends" most likely manifested differently but consistently in the archaeal lineages that were compared, without artificially pushing the age of lineage-specific domain repertoires significantly back in time [16].

3.3. The Two Pyrococci Display Expected Bias in Their Use of Amino Acids. The mean relative amino acid abundances (MAA) were plotted against barophily rank (BR) for the sequence of entire polypeptide chains. There does not seem to be a particular preference within *P. abyssi* for using more barophilic amino acids than nonbarophilic ones or within *P. furiosus* for using more nonbarophilic amino acids than barophilic ones (Figure 3(a)). However, BR was determined as a measure of comparison between the two species, not within them. Thus, we plotted mean MAA differences between the two *Pyrococci* by subtracting MAA of the nonbarophile from those of the barophile. A positive value thus indicates a bias for using an amino acid in *P. abyssi* relative to *P. furiosus*; a negative value indicates the bias for the reverse. Figure 3(b) demonstrates a clear and statistically significant bias toward barophilic amino acids in *P. abyssi* and a significant bias toward nonbarophilic amino acids in *P. furiosus*. No such bias exists for the "indifferent" amino acids. This is an expected result congruent with the definition of barophily rank. Further exploration focuses on functionally important portions of the protein sequences, progressively moving the analysis closer to the evolutionary ancestor of the two species.

3.4. Bias in the Use of Amino Acids between the Two Pyrococci Is due to Their Shared FSF Repertoires. The apparent functional similarities pose a question: does the above bias of amino acid abundance within complete protein sequences arise from the parts of those sequences that correspond to protein domains or the intervening "connecting" sequences between domains? It is logical to predict that the latter have smaller, if any, bias, compared to the FSF domain regions. Indeed, the surmised purpose of the amino acid substitutions in barophiles is to stabilize domain structure against penetration by water, which tends to be forced into the protein core under the high pressure of the ocean abyss

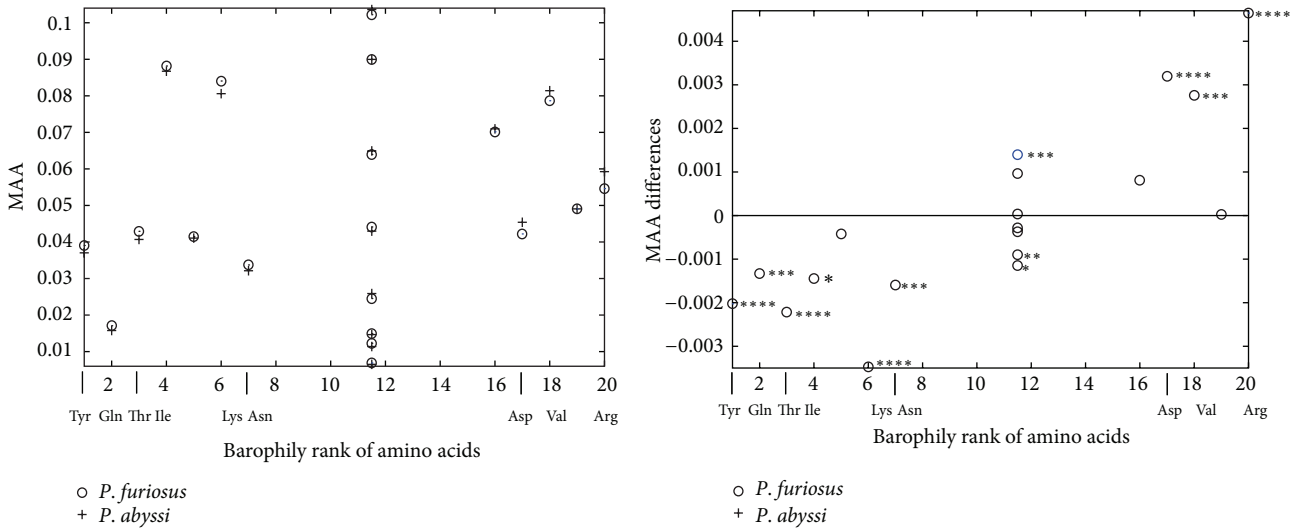


Figure 3: Mean relative amino acid abundance (MAA) analysis of the set of entire polypeptide sequences. (a) On average, both *Pyrococci* seem to use barophilic and nonbarophilic amino acids to the same extent. (b) The barophilic Asp, Val, and Arg are used significantly more in sequences of the barophilic *P. abyssi* compared to the nonbarophilic *P. furiosus* (positive difference), whereas the nonbarophilic Tyr, Gln, Thr, Ile, Lys, and Asn are used significantly more in *P. furiosus* compared to *P. abyssi* (negative difference). Statistical significance is marked with stars according to Welch's two-sided test: $^{*}P < 0.1$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$. A total of 1,896 *P. abyssi* and 2,125 *P. furiosus* sequences were analyzed.

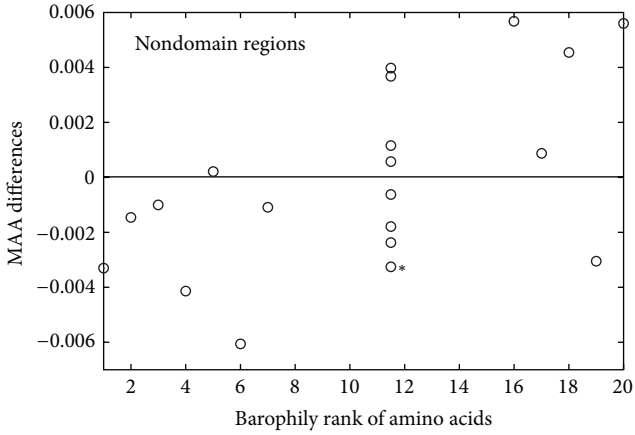


Figure 4: MAA analysis of nondomain regions shows there is no significant difference in barophilic and nonbarophilic amino acid use between the two organisms. Statistical significance is marked with stars as described in Figure 2. A total of 1,229 *P. abyssi* and 1,339 *P. furiosus* sequences were analyzed.

[23]. The effect on the intervening and more flexible regions should therefore be negligible.

As expected, the statistics on MAA differences computed within the intervening sequences showed no significant difference in amino acid use between the two organisms (Figure 4). In contrast, regions of FSFs shared by the two organisms show significant bias toward using barophilic amino acids in *P. abyssi* and nonbarophilic amino acids in *P. furiosus* (Figure 5(a)). This stands for all amino acids displaying statistical significance in whole sequences, except Gln and Tyr. It is interesting to note that this bias is broken within the species-specific FSF regions (Figure 5(b)), where nonbarophilic Ile and Gln change alliances, being present in higher abundance in the barophilic *Pyrococcus*. These patterns may be explained by two non-competing hypotheses. (1) The specific FSF domains likely arose in these organisms after the divergence from the common ancestor, and consequently, their evolution proceeded de novo according to the idiosyncratic ecological conditions of each species. Thus, their use of amino acids is not subject to the rules of homologous amino acid substitution, and the BR measure is inapplicable to their case. (2) The shared FSFs are more likely to belong to homologous sequences that both organisms inherited from their common ancestor. Their evolution proceeded from a certain starting point, which may have not been optimal for functioning in their eventual ecological niche. Amino acid substitutions were therefore used to stabilize the function of respective proteins, resulting in the observed patterns. These hypotheses naturally lead us to the exploration of the homologous sequences and their FSF domains.

3.5. Analysis of Homologous Sequence Pairs Sharing FSF Domains. We found a total of 359 pairs of homologous sequences in the two *Pyrococci*. These sequences fall into 5 different categories:

(1) 29 pairs in which neither sequence has FSF assignments,

(2) 9 pairs in which FSF assignments were made in one sequence but not in the other,

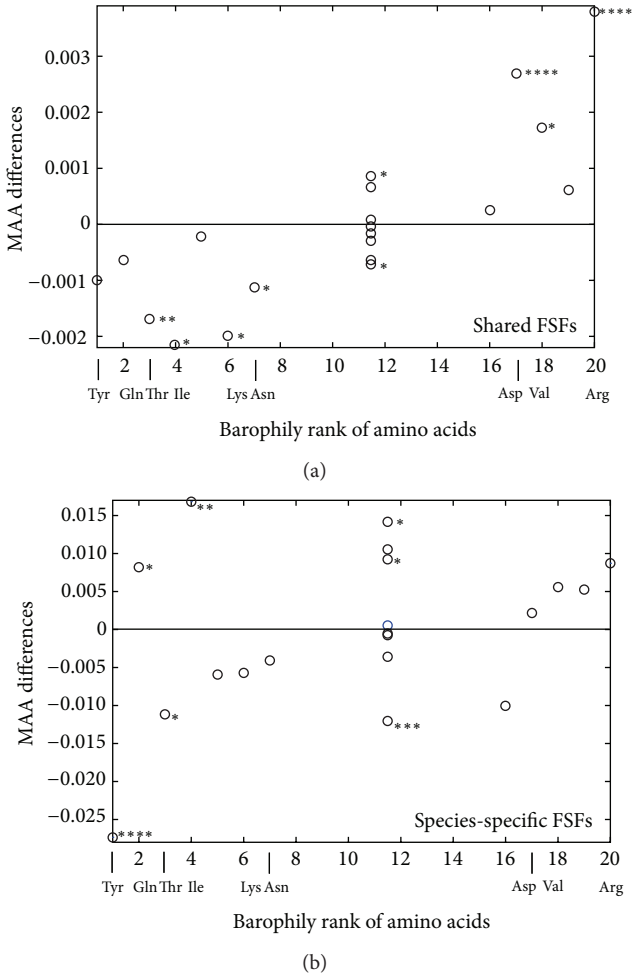


Figure 5: MAA analysis of shared FSF domains. (a) Although Tyr and Gln were used significantly more by *P. furiosus* in entire sequences, their use within the shared FSF domains seems to be statistically the same between the two *Pyrococci*. (b) Amino acids do not follow their patterns of homologous substitution within the species-specific FSFs. Statistical significance is marked with stars as described in Figure 2. A total of 1,627 *P. abyssi* and 1,739 *P. furiosus* sequences were analyzed.

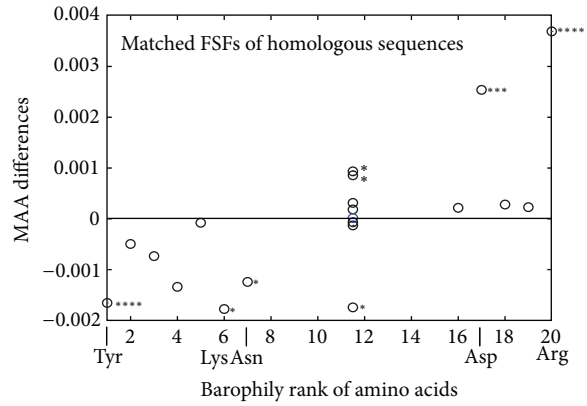


Figure 6: MAA analysis of matching FSF domains within pairs of homologous sequences confirms significant preferences of the barophilic organism for Arg and Asp and a preference of the nonbarophilic organism for Tyr, Lys, and Asn. Other barophilic and nonbarophilic amino acids do not display a significant difference in use. Statistical significance is marked with stars as described in Figure 2. A total of 241 matching FSFs were analyzed.

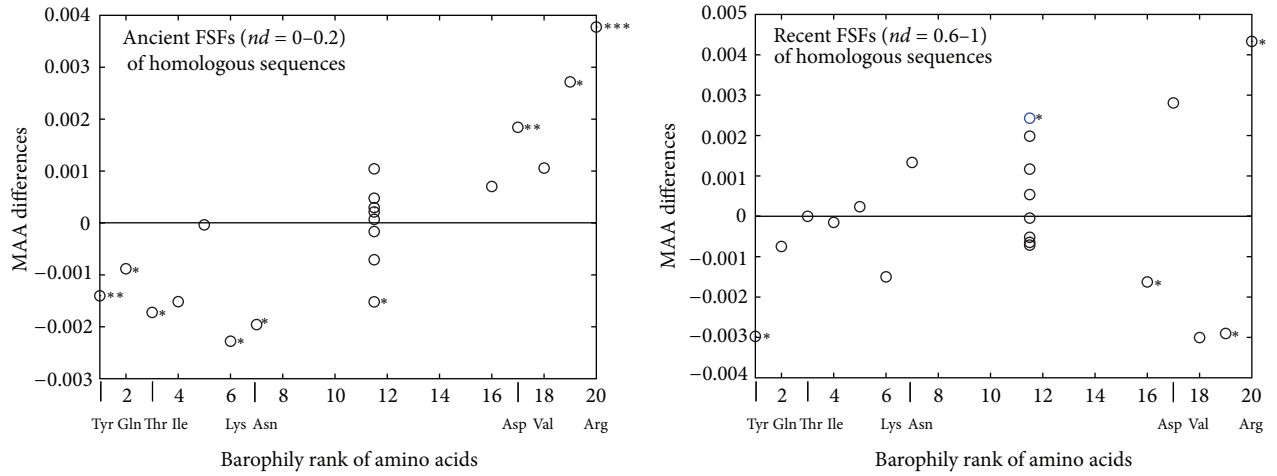
(3) 10 pairs in which extra FSF assignments were made in one of the sequences of a pair,

(4) 3 pairs with completely different FSF assignments,

(5) 309 pairs with identical assignments.

Categories (1) and (2) are not useful for comparing amino acid content of FSF domains. None of the species-specific FSFs were found in categories (3) and (4), supporting our first hypothesis that specific FSF were developed de novo. Thus, we proceeded to investigate MAA differences within the matching FSFs of the homologous sequences from category (5). Results demonstrate similar biases in use of amino acids as were found in the preceding comparisons (Figure 6). However, Val, Ile, Thr, and Gln lost the statistical significance they had in the total set of FSFs. This may have happened for at least two reasons: either these amino acids are the most representative of the ancestral sequences and the most difficult to change or they are important for novel functions introduced at later time during the process of divergence.

To tease these two possibilities apart, we identified a set of the most ancient FSFs and a set of the most recent ones. The evolutionary age of FSF domains has been established previously in the works of Wang et al. [16, 20, 25] by building a phylogenetic tree of FSFs based on their global abundance in organisms. The evolutionary age of each FSF was calculated by tracing the number of internal nodes (*nd*) for each FSF lineage, starting from the root of the phylogenomic tree and expressed on a relative scale from 0 (the origin of protein domains) to 1 (the present). In Wang et al. [16] the most ancient FSFs have been defined as those that emerged before the beginning of massive FSF loss in the three superkingdoms due to reductive evolution (*nd* = 0–0.2). The choice of ancestry values for ancient FSFs in *Pyrococcus* is also supported by the history of their divergence: at *nd* = 0.2 both species have acquired unique FSFs (Figure 2). For the group of more recent FSFs we chose those that emerged after the "big bang" of domain combination [25], which is mostly driven by Eukarya (*nd* = 0.6–1). Comparison of MAA differences between these two groups was instructive (Figure 7). It demonstrated that some amino acids have different patterns of use in FSFs of different age. While nonbarophilic Gln, Thr, Lys, and Asn were used more by *P. furiosus* in the ancient FSFs, their use became more balanced in the new FSFs. Similarly, the barophilic Asp, and Arg were used more by *P. abyssi* in the old FSFs, but they became more balanced in the recent FSFs. Interestingly, bias reversed completely for the barophilic Ser, which only showed significant difference in use in ancient FSFs of homologous sequence pairs. Thus, the loss of statistical significance of barophilic Val and nonbarophilic Thr, Gln, and Ile of Figure 6 can be explained by their likely relevance to unfold newly introduced functions late in evolution, even when portraying



(a) (b)

Figure 7: MAA analysis of matching FSFs of different evolutionary age. (a) Organisms display the same preferences for amino acids within ancient FSFs (*nd* = 0–0.2) of homologous sequences as they do in entire sequences. However, here Ser is also preferred more by the barophile. (b) Many amino acid preferences are erased in evolutionarily recent FSFs (*nd* = 0.6–1.0) of homologous sequences. Some preferences were even reversed; the normally barophilic Gly and Ser are now used significantly more by the nonbarophilic organism. Statistical significance is marked with stars as described in Figure 2. A total of 55 recent and ancient FSFs were analyzed.

an ancestral composition that is refractory to change (e.g., Thr and Gln).

We note that the impact of horizontal gene transfer (HGT) or other convergent evolutionary processes should be considered minimal at the FSF structural level and should not affect the conclusions of this study. The structures of protein domains are retained over long evolutionary time frames and their gain or loss have been shown to be rare in lineages at FSF and other levels of the structural abstraction (e.g., [29, 30]). Structural cores evolve linearly with amino acid substitutions per site and at 3–10 times slower rates than sequences [31]. This high conservation highlights the slow dynamics of structural change in proteins and the clock-like behavior of FSF evolution [20]. The effects that lateral transfer could have on the highly conserved sequences of shared repertoires must therefore be considered negligible. Furthermore, the divergence time between the archaeal lineages examined in this study is too short compared to divergences occurring across the entire evolutionary spectrum. This fact diminishes any possible effects that HGT could have on relative amino acid abundances of lineage-specific sequences.

4. Conclusions

The patterns in amino acid use presented here suggest further depth to our understanding of the barophilic rank of amino acids. Previous studies designated five amino acids to be preferentially barophilic, based on the significance of their substitution patterns: Arg, Ser, Val, Asp and Gly. In our analysis only two out of these amino acids display this preference consistently and significantly in entire sequences, shared FSFs, matched FSFs of homologous sequence pairs, and ancient FSFs: Asp and Arg (Table 2). Since barophilic preferences are already evident in the ancient set of FSFs that are at least ~3 Gy-old, but not in species-specific and young domains, we conclude that barophily is a very ancient trait that goes back to the start of organismal diversification. These two amino acids appear late in the evolution of the amino acid charging function of the genetic code as judged by the age of isoacceptor tRNA [32] and coevolving synthetase domains with amino acid editing functions [28]. However, a consensus chronology of amino acid evolution placed Asp and Arg fourth and tenth in the timeline [33], suggesting that the coding of Asp could have unfolded early. In fact, Asp belongs to an early group of amino acid codified by codons of the GNN type; an observation that supports the coevolutionary theory of the origin of the genetic code [34]. While the apparent mismatch of data can be explained by separate histories of amino acid charging and encoding [32], it is clear that the primordial barophilic trait has impacted the early evolution of genetics. Since tRNA, the genetic code and Archaea appear to have polyphyletic origins (e.g., [30, 35]); results suggest the colonization of barophilic environments by the ancestors of the emerging archaeal lineages as these were unfolding genetic idiosyncrasies in convergence towards a universal code.

Seven amino acids, Asn, Lys, Pro, Ile, Thr, Gln, and Tyr, have been previously designated as nonbarophilic. Asn, Lys, and Thr displayed a preference for non-barophily most consistently, in entire sequences, shared FSFs, matched FSFs of homologous sequence pairs, and ancient FSFs. It is possible that these "faithful" amino acids are either easy to swap within the homologous sequences as they diverge or contribute the most to the stability and function of proteins at their respective environmental conditions. Their presence in the ancient FSFs confirms the ancestrality of nonbarophilic traits, which does not invalidate the primordial nature of barophily.

An archaic nonbarophilic trait challenges the views of an abiotic start of genetics in deep vent environments occurring prior to organismal diversification [36], supporting instead

Table 2: Barophily rank (BR) and preference of the barophilic (B) and non-barophilic amino acids (N). Bold font identifies amino acids that have most consistent and significant preference that is congruent with their barophily or non-barophily group.

Amino acid	BR	Complete sequence	Shared FSFs	Specific FSFs	Homologous sequences		
					All FSFs	Ancient FSFs	Recent FSFs
Arg (B)	**20**	**B**	**B**		**B**	**B**	
Ser (B)	19					B	N
Val (B)	18	B	B				
Asp (B)	**17**	**B**	**B**		**B**	**B**	
Gly (B)	16						N
Asn (N)	**7**	**N**	**N**		**N**	**N**	
Lys (N)	**6**	**N**	**N**		**N**	**N**	
Pro (N)	5						
Ile (N)	4	N	N	B			
Thr (N)	**3**	**N**	**N**	**N**		**N**	
Gln (N)	2	N		B		N	
Tyr (N)	**1**	**N**		**N**	**N**	**N**	**N**

the view that ocean abysses played an important role in tailoring the diversification of genetics and life [4, 5]. Consequently, the origin of primordial life (prior to the genetic code) is more likely in nonbarophilic environments such as terrestrial anoxic geothermal fields [37] or alkaline aquifers generated by serpentinizing rocks [38]. We stress that this does not nullify the fact of a crucial involvement of barophily during the rise of cellular organisms, modern biochemistry, and genetics.

The other amino acids displayed more balanced use in the sequences we inspected. This may indicate that they are most representative of the composition of the ancestral organism and are most difficult to change for structural or functional reasons. The "changeling" amino acids Ser, Ile, and Gln were preferentially used according to their BR in shared/ancient FSFs but had opposite patterns of use in specific/recent FSFs. This may indicate that their BR values have more to do with the legacy left over from the common pyrococcal ancestor, yet they are actually more useful under conditions opposite than those suggested by BR. Finally, the nonbarophilic Pro demonstrated no significant difference in any of our tests. Pro is enriched in structured regions that involve turns, which have been suggested important for primordial coding [39] and linked to loops and protein flexibility [28]. Their role in protein structure and genetics may be archaic and hardwired into the make up of proteins.

Acknowledgments

The authors thank Minglei Wang for providing phylogenomic data. Research was supported by awards from the National Science Foundation (MCB-0749836) and CREES-USDA to GCA. Travel of MDG was supported in part by the Center for Advanced Study of the University of Illinois. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

References

[1] M. Di Giulio, "Some aspects of the organization and evolution of the genetic code," *Journal of Molecular Evolution*, vol. 29, no. 3, pp. 191–201, 1989.

[2] J. H. McDonald, A. M. Grasso, and L. K. Rejto, "Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*," *Molecular Biology and Evolution*, vol. 16, no. 12, pp. 1785–1790, 1999.

[3] J. H. McDonald, "Patterns of temperature adaptation in proteins from the bacteria *Deinococcus radiodurans* and *Thermus thermophilus*," *Molecular Biology and Evolution*, vol. 18, no. 5, pp. 741–749, 2001.

[4] M. Di Giulio, "A comparison of proteins from *Pyrococcus furiosus* and *Pyrococcus abyssi*: barophily in the physicochemical properties of amino acids and in the genetic code," *Gene*, vol. 346, pp. 1–6, 2005.

[5] M. Di Giulio, "The origin of the genetic code in the ocean abysses: new comparisons confirm old observations," *Journal of Theoretical Biology*, vol. 333, pp. 109–116, 2013.

[6] M. Di Giulio, "The late stage of genetic code structuring took place at a high temperature," *Gene*, vol. 261, no. 1, pp. 189–195, 2000.

[7] M. Di Giulio, "The ocean abysses witnessed the origin of the genetic code," *Gene*, vol. 346, pp. 7–12, 2005.

[8] C. Huber, F. Kraus, M. Hanzlik, W. Eisenreich, and G. Wächtershaüser, "Elements of metabolic evolution," *Chemistry European Journal*, vol. 18, no. 7, pp. 2063–2080, 2013.

[9] G. Caetano-Anollés, M. Wang, D. Caetano-Anollés, and J. E. Mittenthal, "The origin, evolution and structure of the protein world," *Biochemical Journal*, vol. 417, no. 3, pp. 621–637, 2009.

[10] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[11] G. Caetano-Anollés and D. Caetano-Anollés, "An evolutionarily structural universe of protein architecture," *Genome Research*, vol. 13, no. 7, pp. 1563–1571, 2003.

[12] G. Caetano-Anollés and D. Caetano-Anollés, "Universal sharing patterns in proteomes and evolution of protein fold architecture and life," *Journal of Molecular Evolution*, vol. 60, no. 4, pp. 484–498, 2005.

[13] C. Vogel, C. Berzuini, M. Bashton, J. Gough, and S. A. Teichmann, "Supra-domains: evolutionary units larger than single protein domains," *Journal of Molecular Biology*, vol. 336, no. 3, pp. 809–823, 2004.

[14] C. Vogel, S. A. Teichmann, and J. Pereira-Leal, "The relationship between domain duplication and recombination," *Journal of Molecular Biology*, vol. 346, no. 1, pp. 355–365, 2005.

[15] C. Vogel and C. Chothia, "Protein family expansions and biological complexity," *PLoS Computational Biology*, vol. 2, no. 5, article e48, 2006.

[16] M. Wang, L. S. Yafremava, D. Caetano-Anollés, J. E. Mittenthal, and G. Caetano-Anollés, "Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world," *Genome Research*, vol. 17, no. 11, pp. 1572–1585, 2007.

[17] K. M. Kim and G. Caetano-Anollés, "Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data," *Molecular Biology and Evolution*, vol. 27, no. 7, pp. 1710–1733, 2010.

[18] H. Fang, M. E. Oates, R. B. Pethica et al., "A daily-updated tree of (sequenced) life as a reference for genome research," *Scientific Reports*, vol. 3, articel 2015, 2013.

[19] K. V. Gunbin, D. A. Afonnikov, and N. A. Kolchanov, "Molecular evolution of the hyperthermophilic archaea of the Pyrococcus genus: analysis of adaptation to different environmental conditions," *BMC Genomics*, vol. 10, article 639, 2009.

[20] M. Wang, Y.-Y. Jiang, K. M. Kim et al., "A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation," *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 567–582, 2011.

[21] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure," *Journal of Molecular Biology*, vol. 313, no. 4, pp. 903–919, 2001.

[22] E. Jain, A. Bairoch, S. Duvaud et al., "Infrastructure for the life sciences: design and implementation of the UniProt website," *BMC Bioinformatics*, vol. 10, article 136, 2009.

[23] T. Ghosh, A. E. García, and S. Garde, "Molecular dynamics simulations of pressure effects on hydrophobic interactions," *Journal of the American Chemical Society*, vol. 123, no. 44, pp. 10997–11003, 2001.

[24] I. N. Berezovsky and E. I. Shakhnovich, "Physics and evolution of thermophilic adaptation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12742–12747, 2005.

[25] M. Wang and G. Caetano-Anollés, "The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world," *Structure*, vol. 17, no. 1, pp. 66–78, 2009.

[26] K. Caetano-Anollés and G. Caetano-Anollés, "Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism," *PLoS One*, vol. 8, no. 3, Article ID e59300, 2013.

[27] G. Caetano-Anollés, K. M. Kim, and D. Caetano-Anollés, "The phylogenomic roots of modern biochemistry: origins of proteins, cofactors and protein biosynthesis," *Journal of Molecular Evolution*, vol. 74, pp. 1–34, 2012.

[28] G. Caetano-Anollés, M. Wang, and D. Caetano-Anollés, "Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility," *PLoS One*, vol. 8, no. 8, Article ID e72225, 2013.

[29] J. Gough, "Convergent evolution of domain architectures (is rare)," *Bioinformatics*, vol. 21, no. 8, pp. 1464–1471, 2005.

[30] K. M. Kim and G. Caetano-Anollés, "The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms," *BMC Evolutionary Biology*, vol. 12, no. 1, article 13, 2012.

[31] K. Illergård, D. H. Ardell, and A. Elofsson, "Structure is three to ten times more conserved than sequence—a study of structural response in protein cores," *Proteins: Structure, Function and Bioinformatics*, vol. 77, no. 3, pp. 499–508, 2009.

[32] F.-J. Sun and G. Caetano-Anollés, "Evolutionary patterns in the sequence and structure of transfer RNA: a window into early translation and the genetic code," *PLoS One*, vol. 3, no. 7, Article ID e2799, 2008.

[33] E. N. Trifonov, "The triplet code from first principles," *Journal of Biomolecular Structure and Dynamics*, vol. 22, no. 1, pp. 1–11, 2004.

[34] M. Di Giulio, "An extension of the coevolution theory of the origin of the genetic code," *Biology Direct*, vol. 3, article 37, 2008.

[35] M. Di Giulio, "A polyphyletic model for the origin of tRNA has more support than a monophyletic model," *Journal of Theoretical Biology*, vol. 318, pp. 124–128, 2013.

[36] W. Martin and M. J. Russell, "On the origin of biochemistry at an alkaline hydrothermal vent," *Philosophical Transactions of the Royal Society B*, vol. 362, no. 1486, pp. 1887–1925, 2007.

[37] A. Y. Mulkidjanian, A. Y. Bychkov, D. V. Dibrova, M. Y. Galperin, and E. V. Koonin, "Origin of first cells at terrestrial, anoxic geothermal fields," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 14, pp. E821–E830, 2012.

[38] S. A. Benner, H.-J. Kim, and M. A. Carrigan, "Asphalt, water, and the prebiotic synthesis of ribose, ribonucleosides, and RNA," *Accounts in Chemical Research*, vol. 45, pp. 2025–2034, 2012.

[39] J. E. Zull and S. K. Smith, "Is genetic code redundancy related to retention of structural information in both DNA strands?" *Trends in Biochemical Sciences*, vol. 15, no. 7, pp. 257–261, 1990.

Hindawi Publishing Corporation
Archaea
Volume 2013, Article ID 373275, 14 pages
http://dx.doi.org/10.1155/2013/373275

Review Article

Protein Adaptations in Archaeal Extremophiles

Christopher J. Reed,[1] Hunter Lewis,[1] Eric Trejo,[1,2] Vern Winston,[2] and Caryn Evilia[1,2]

[1] *Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA*
[2] *Department of Biological Sciences, Idaho State University, Pocatello, ID 83209, USA*

Correspondence should be addressed to Caryn Evilia; evilcary@isu.edu

Received 24 June 2013; Revised 26 July 2013; Accepted 14 August 2013

Academic Editor: Kyung Mo Kim

Copyright © 2013 Christopher J. Reed et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Extremophiles, especially those in Archaea, have a myriad of adaptations that keep their cellular proteins stable and active under the extreme conditions in which they live. Rather than having one basic set of adaptations that works for all environments, Archaea have evolved separate protein features that are customized for each environment. We categorized the Archaea into three general groups to describe what is known about their protein adaptations: thermophilic, psychrophilic, and halophilic. Thermophilic proteins tend to have a prominent hydrophobic core and increased electrostatic interactions to maintain activity at high temperatures. Psychrophilic proteins have a reduced hydrophobic core and a less charged protein surface to maintain flexibility and activity under cold temperatures. Halophilic proteins are characterized by increased negative surface charge due to increased acidic amino acid content and peptide insertions, which compensates for the extreme ionic conditions. While acidophiles, alkaliphiles, and piezophiles are their own class of Archaea, their protein adaptations toward pH and pressure are less discernible. By understanding the protein adaptations used by archaeal extremophiles, we hope to be able to engineer and utilize proteins for industrial, environmental, and biotechnological applications where function in extreme conditions is required for activity.

1. Introduction

Archaea thrive in many different extremes: heat, cold, acid, base, salinity, pressure, and radiation. These different environmental conditions over time have allowed Archaea to evolve with their extreme environments so that they are adapted to them and, in fact, have a hard time acclimating to less extreme conditions. This is reflected in current taxonomy in Archaea [1, 2]. Archaea are presently partitioned into four branches: the halophiles, the psychrophiles, the thermophiles, and the acidophiles. While we typically think about the methanogens as a distinct group, they are, in fact, spread among all the other branches in Archaea. For the purposes of this review, we have included them in their principle branch (e.g., the thermophiles) where appropriate.

The branches of Archaea intersect in interesting ways. For example, alkaliphiles (which are not one of the branches mentioned above) are grouped with the halophiles because the two archaeal groups not only are found together in saline environments but also share genome similarities. Thermophiles and acidophiles branches are also clustered together, not only because most acid environments are hot but because these groups also share genome similarities. Many archaeal piezophiles (pressure-loving organisms) are found at deep sea thermal vents, leading them to have many similarities to hyperthermophiles. Psychrophiles also share branches with the halophiles; again for similar reasons, psychrophilic environments can be hypersaline.

In researching different Archaea, we have divided Archaeal protein adaptations into three general categories: thermophilic, psychrophilic, and halophilic. Organisms that could be classified into more than one branch could show one or more of these three major adaptations with minor adjustments to accommodate environmental conditions. For example, haloalkaliphiles, like *Natronomonas pharaonis*, have their primary protein adaptation as halophilic with no clear adaptation for the extreme basic environment (pH > 11) in which they live [3]. For *N. pharaonis*, this fits with their cell conditions, which are only mildly basic but extremely saline. *N. pharaonis* allows salt across the cell membrane, for a high internal salt concentration, but the intracellular pH is mildly basic (pH ~ 8).

Acidophilic proteins, some of which show increased negative surface charge, also show thermophilic adaptations [4, 5]. Acidophiles pump protons out of the cell to maintain a mildly acidic cytoplasm. pH varies from 5.0 to 6.5, which allows their primary adaptation to be for their thermophilic environment. Mesophilic acidophiles prove this point as their proteins have small changes that could account for their activity in the acidic cytoplasm [4, 5].

Thermophiles, psychrophiles, and halophiles, on the other hand, have evolved to live within their environmental conditions, rather than to adapt ways to circumvent it. Obviously, thermophiles and psychrophiles cannot shut out heat or cold, so, besides cellular adaptations like secondary metabolites which maintain overall cell stability, this required novel protein adaptations to survive. Halophiles had to evolve a system to deal with extreme osmotic stress. To facilitate this, they possess a membrane system to pump potassium in while pumping sodium out [6]. The intracellular concentration of potassium, depending on which species of Halobacteria, can vary from 1.2 to 4.5 M [6–9]. This functions to maintain the osmotic balance in the cell. However, their proteins require certain features that allowed them to work under such extreme ionic conditions.

While the three categories (thermophiles, psychrophiles and halophiles) of Archaea show the most obvious protein adaptations to their environments, those adaptations are not necessarily uniform throughout all of their proteins. This makes studying protein adaptations in all extremophiles, especially Archaea, difficult because one is not simply looking for a single trend or feature. In fact, variability in adaptations has been noted multiple times throughout studies of archaeal and extremophile proteins [4, 12–14]. There have been many reasons proposed for these differences. One of the more convincing ideas is that, by only having a few protein modifications, the enzyme might have activity over a range of conditions [4, 14–17]. This gives the organism some flexibility to grow in a range of different conditions. Another idea is that having various protein adaptations could be an alternative to regulatory pathways. Along these lines, a particular protein would become optimally active only under certain conditions, which would save the organism from having to regulate that protein through cell signaling pathways [18]. This supports the notion that Archaea take advantage of simple adaptations to reap the benefits of their extreme environments.

Not all adaptations are hard coded into the protein sequence though. This follows because, in order for proteins to function under extreme conditions, multiple structural considerations must be accommodated to balance activity, flexibility, and stability [19–21]. Some protein structure/function issues under environmental extremes can be accommodated by flexible folding [22]. Protein folding states are dynamic; they have to change in order for the protein fold to accommodate different conditions and remain active. However, there is a limit to how many folding events can be accomplished in order to meet environmental challenges [23, 24]. For example, the cysteinyl-tRNA synthetase from *Halobacterium salinarum* sp. NRC-1 shows little change in activity and global structure when the salt concentration varies from 3.5 M to 2 M. This change would be a large decrease in salinity for the organism, which would cause it to lose the integrity of the cell membrane and S-layer, but this enzyme is tolerant of the change. The enzyme probably remains active and structurally sound due to local folding events that accommodate the change in conditions. When the salinity is further decreased from 2 M to 0.5 M, the enzyme loses activity and its structure, indicating that folding states have their limit and other forces need to be at work to get the enzyme to remain stable [16, 18]. This reflects that sequence changes over time have led to protein features that protect or preserve a function under greater extremes.

In this review, we will summarize the current known protein adaptations for thermophilic, psychrophilic, and halophilic Archaea. Along the way, we will discuss other extreme conditions, such as acid, base, and pressure, for which their adaptations are considered secondary to that of the main adaptation. For example, thermoacidophiles, thermopiezophiles, and haloalkaliphiles will be discussed with thermophiles and halophiles. This was done as an attempt to sort out "minor" adaptations into their defining category while not ignoring them.

2. Thermophilic Proteins

While thermal vents and hot springs are considered to be some of the most extreme environments on Earth, several organisms are able to thrive in these hostile locations where most life would perish. Among these are thermophiles and hyperthermophiles. While the two share similar adaptations to survive in these extremes, they differ in their temperature growth optimum. Hyperthermophiles can grow optimally up to 105°C, whereas thermophiles are classified as growing between 50°C and 70°C. At such extreme temperatures, proteins lacking the necessary adaptations undergo irreversible unfolding, exposing the hydrophobic cores, which causes aggregation [25]. Thermophilic proteins have several adaptations that give the protein the ability to retain structure and function in extremes of temperature. Some of the most prominent are increased number of large hydrophobic residues, disulfide bonds, and ionic interactions.

2.1. Oligomerization and Large Hydrophobic Core. Observed within many thermostable proteins are deviations from standard quaternary organization seen in their mesophilic counterparts. This strategy is thought to increase the rigidity of the individual subunits, promote tighter packing of the hydrophobic core, and reduce exposure of hydrophobic residues to solvent [14]. Three acetyl-CoA synthetases and one amylase from thermophilic Archaea highlight the argument that aberrancies in quaternary structure are the causative agents in these enzymes' thermostability and others as well.

Recent characterizations of two acetyl-CoA synthetases (ACS) from *Ignicoccus hospitalis* [26] and *Pyrobaculum aerophilum* [27] have uncovered novel structural adaptations, namely, a difference in oligomeric state from that of the mesophilic variants. Compared to their mesophilic counterparts, these hyperthermophilic enzymes form octomers

whereas the aforementioned mesophiles follow the general trend of being monomers or homodimers. However, the ACS from *Archaeoglobus fulgidus*, a hyperthermophile with lower temperature growth optimum, is a trimer [28].

The hydrophobic effect becomes increasingly more important in protein stability and folding as temperature increases. This has been observed in a phosphotriesterase from *Sulfolobus solfataricus* where tighter packing is observed due to favorable hydrophobic interactions at the dimer interface [29]. This strategy could have also been adopted by hyperthermophilic ACS proteins in order to decrease the overall ratio of surface area to volume in regard to individual subunits and solvent exposed hydrophobic regions. As a consequence this would result in tighter packing of the hydrophobic core, a general feature of all thermostable proteins.

Contrary to the increase in oligomeric state in ACS, the study of a thermostable amylase from *Pyrococcus furiosus* (Pf) showed a lack of oligomerization compared to the mesophilic homologues [30]. This is the first example of a functional monomeric version of a cyclodextrin hydrolyzing enzyme. Bacterial homologues require dimerization before activity is seen [31]. A novel domain on the N-terminus (N′) allows Pf amylase to be active as a monomer, although as a consequence it lacks the transferring activity seen in CD-hydrolyzing enzymes. The bacterial N-domain possesses a loop that extends over the active site that acts as structural "lid" and functions to stabilize certain substrates such as maltose, whereas Pf amylase does not.

While the absence of quaternary structure and the N-terminal loop in Pf amylase alters substrate specificity, it would appear that this modification is important in regard to overall stability. It is widely accepted that structural flexibility in protein structure is unfavorable in thermostable enzymes, even though there is no generally accepted mechanism by which rigidity and by proxy stability is achieved [32]. In the case of thermostable ACS, a higher oligomeric state is favorable, whereas Pf amylase implements the opposite strategy: organizing all the necessary components into a single subunit, creating structural rigidity, and promoting tighter packing of the hydrophobic core. Both cases lend credence to the hypothesis that changes in quaternary structure can be advantageous; however, there is no discernible trend within this strategy.

2.2. Increased Number of Disulfide Bonds. Disulfide bridging between cysteine residues is an important tertiary structural element that is paramount in determining the overall structure of a protein. Organisms within all domains of life have adapted their own systems of keeping proper bridging in check, as some are favorable and some completely inactivate enzymes. Within thermostable enzymes these structural elements are significant, since they have been shown to increase stability within thermophilic proteins and play a role in preventing alteration of quaternary structure. Studies by Cacciapuoti et al. [12, 33] and separately by Boutz et al. [34] provide evidence for these claims.

One example of the use of disulfide bridging in thermostability is 5′-deoxy-5′methylthioadenosine phosphorylase II which was used to study the CXC motif and intrasubunit disulfide bonds within thermophilic proteins [12, 33]. Using circular dichroism spectroscopy, under reducing conditions, the hexameric protein was seen to disassociate into its monomeric state in a reversible fashion. Chemical and thermal denaturation resulted in irreversible degradation of structure. Single and double mutants within key cysteine residues demonstrated an appreciable change in thermostability. The native protein was observed to be almost completely denatured at 108°C, whereas the single mutant (C262S) was shifted to 102°C and 99°C for the CXC double mutant (C259S/C261S). These results elucidated that mutations within these residues decreased thermal stability inferring that disulfide bridging is a structural adaptation [12]. It is noteworthy that the CXC motif forms a strained 11-member disulfide ring, which has been implicated as a useful redox reagent [35]. This novel adaptation parallels that of disulfide isomerases, whose function is to maintain proper disulfide bridging within proteins [36].

Disulfide bonds have also been shown to be important in oligomerization. The citrate synthetase from *Pyrobaculum aerophilum* illustrated the use of disulfide bonds to create cyclized protein chains that topologically interlinks two monomeric subunits of the homodimer [34]. This novel structural feature confers stability within the dimer by disallowing separation of the individual subunits. These two examples demonstrate the role of disulfide bridging in thermostability, either from increased rigidity or the interlocking of adjacent chains between monomeric subunits.

2.3. Increased Salt-Bridging. Salt-bridging is a prevalent feature of most thermophilic enzymes compared to their mesophilic variants [37]. This is in contrast to findings that salt-bridging may destabilize mesophilic proteins and are unfavorable compared to hydrophobic interactions [38]. The desolvation penalty and entropic cost associated with ion pairing found in salt bridges is more easily overcome at higher temperatures [39]. When these thermodynamic considerations are negated, salt bridges become a structurally stabilizing element, increasing the thermal capacity of proteins using favorable charge-charge interactions.

Experimental circular dichroism studies within a thermophilic ribosomal protein, L30e, from *Thermococcus celer* produced a noticeable change in thermal capacity without causing major structural changes [39]. Mutations from charged residues involved in salt-bridging to hydrophobic residues increased the heat capacity change of unfolding, ΔC_p. Lowering of ΔC_p could be a strategy used to increase thermostability in proteins, favoring the natively folded state over that of the unfolded. This demonstrates that favorable interactions of charged residues (salt bridges) improve the thermal stability of proteins [39].

2.4. Increased Surface Charges. Ubiquitous within thermostable proteins is the increase of charged residues on the surface of proteins [40]. Replacement of polar uncharged surface residues with polar charged residues can result in an overall increased stability due to several factors. At higher temperatures, polar residues such as asparagine and

glutamine can undergo deamination which would reduce stability [40]. Replacement of these and other thermolabile residues increase both short- and long-range charge interactions which, generally, help to protect against thermal denaturation [41].

To further explain the roles of charge-charge interactions outside of those involved in salt bridges, multiple single-point mutations of surface charged residues to alanine were made on the ribosomal protein, L30e, from *Thermococcus celer* [41]. It was found that the thermal capacity of this thermophilic protein could be further increased by favorable mutations to charged residues and could be decreased when surface charges were replaced with alanine. Long-range charge-charge interactions were found to be a large determining factor in the stability of L30e, as removal of these electrostatic interacts caused greater susceptibility to thermal and chemical denaturation.

The effects of too much surface charge were observed in a putative DNA binding protein from *Methanothermobacter thermautotrophicus*, MTH10b [42]. While the protein has an unknown activity, the thermal capacity was greatly reduced in the absence of salt. In the crystal structure a highly charged surface was observed, which provided insight to the salt-dependent stability. Uneven surface charge distribution, the vast majority of which is positively charged, is attributed with causing the protein to lack the intrinsic thermostability possessed by others within this protein family. MTH10b demonstrates that the presence of salt acts to offset the repulsive forces that act to destabilize the protein bestowing thermostability [42].

While higher content of surface charged residues can serve to stabilize proteins by preventing aggregation at higher temperatures, it can also serve to destabilize the structure [43]. The necessity of inorganic salts for protein stability and functionality has also been observed for other enzymes in *M. kandleri* [44]. This suggests that an extremely charged protein surface may require some form of compensation in order to reap the positive benefits of this structural adaptation.

2.5. Industrial Applications. Thermophilic enzymes show a high potential for biotechnological and industrial application because they are optimally active at high temperatures, where the kinetics and thermodynamics of the catalyzed reaction are more favorable [45]. This allows for a more efficient reaction and higher product yield. Secondary benefits that accompany thermostability include a lower chance of bacterial contamination (important in food and drug applications) and the reduction of operating costs from constant enzyme replacement due to thermal denaturation [45, 46]. The first application of thermophilic enzymes was Taq DNA polymerase from the bacterium *Thermus aquaticus*. The clever use of this enzyme reduced the cost and allowed the automation of PCR, which greatly advanced research in biochemistry and molecular biology laboratories [45, 46]. Today, a number of thermophilic DNA polymerases from archaeal species are used instead of *T. aquaticus* including *pfu*Turbo, DeepVent$_R$, Therminator, among others (Stratagene Inc., and New England BioLabs Inc.). A potential application of archaeal thermophilic enzymes was discovered in mutational studies done on a thermostable amylase from *Pyrococcus furiosus*. A mutation in Pf amylase caused an increase in the production of maltoheptaose from β-cyclodextrin. Maltoheptaose and other linear maltooligosaccharides are of high value in the food, cosmetic, and pharmaceutical industries where they can be used as carriers [30]. Other potential uses of thermophilic enzymes exploit not their activity at high temperatures but their lack of catalytic activity at ambient temperatures. Thermophilic enzymes can act as optical nanosensors which could bind a substrate but not turn over a product [46]. The substrate-enzyme complex could then be detected by measuring a variation of enzyme fluorescence, which in turn could allow for quantification of the amount of substrate in a sample. Such innovations have the potential to become important tools in biotechnology, medical testing, and drug discovery [46].

3. Piezophilic Proteins

Piezophiles are organisms that live under extremely high hydrostatic pressure often in other extremes, like high or low temperature. Their typical habitat is deep in the ocean, under extreme pressure, and in the extreme heat of hydrothermal vents or in the cold of the ocean. Most archaeal piezophiles, such as *Pyrococcus abyssi* or *Sulfolobus solfataricus*, are thermophilic, while psychrophilic piezophiles are usually, but not strictly, bacterial [47]. Adaptations in their proteins to the extreme pressure appear to be secondary to their adaptations to temperature [48]. The general adaptations for archaeal and bacterial piezophiles, outside of their temperature adaptation, are a compact, dense hydrophobic core, the prevalence of smaller hydrogen-bonding amino acids and increased multimerization [49–51].

One example of this is seen in *Pyrococcus abyssi*, a hyperthermophilic piezophile. There is an increase in small amino acids across its proteome when compared to that of the related archaeon but non-piezophile, *Pyrococcus furiosus* [50]. Overall reduced amino acid size leads to a reduction in the number of large hydrophobic residues, such as tryptophan and tyrosine, in the core of its proteins. This is contrary to the composition of the hydrophobic core seen in most thermophilic proteins, which contain a higher percentage of large amino acids. Nevertheless, such an adaptation is advantageous because it allows for tighter packing, creating a more pressure stable protein [50]. Another example of piezophilic adaptation of a compact hydrophobic core was a study done with the Sso7d, a DNA binding protein from *Sulfolobus solfataricus* (Ss) [52, 53]. Using mutagenesis and structural studies with NMR, it was demonstrated that any change, that either created a cavity in the protein or disrupted the hydrophobic nature of the protein's core, decreased the pressure stability as well as the thermostability of the protein [52, 53]. Similar results were seen in glutamate dehydrogenase from *Thermococcus litoralis* [54].

Another way proteins can cope with pressure is to form multimeric proteins. The piezophlic protein, TET3 peptidase (TET3) from *Pyrococcus horikoshii*, forms a discreet dodecamer, rather than a barrel-shaped multimer, and demonstrates increased stability at high pressure [55]. The fact that it

formed a dodecamer was important for this protein, since its formation makes the individual monomers more compact in shape. By making its monomers more compact, there is less chance for water to penetrate the core of the protein when high pressures are applied. The trapped water would then disrupt the structure of the protein [49, 55].

Multimerization also protects the hydrogen bonding between the protein subunits, which, in general, are not as susceptible to pressure as ionic interactions [49, 55]. Ionic interactions, especially electrostatic interactions, are more susceptible to solvation which disrupts these intraprotein interactions at higher pressure [49, 51]. The strength of the hydrogen bonds between the subunits is enough to mitigate the salt bridge instability [55]. Some thermophilic adaptations, such as an increase in basic amino acids, especially arginine, were found to be beneficial for a protein in both extreme environments. This has also been observed in proteins from *P. abyssi* [50].

While archaeal psychrophilic piezophiles do exist, they are relatively unstudied in terms of their protein adaptations. However, the bacteria that do occupy this niche have important and sometimes similar adaptations to thermopiezophiles. In particular, psychrophilic piezophiles do not rely on salt bridges for protein stability, like the thermopiezophiles, which helps them adapt to low temperature and high pressure [48].

3.1. Possible Industrial Applications. Little research has been done on piezophilic enzymes; however, there is great potential industrial applications. There are many industrial processes that use high pressure coupled with high or low temperature, especially within the food industry. High pressure is not only sterilizing but also preserves the color and flavor of foods. Enzymes isolated from psychrophilic or thermophilic piezophiles could function under these conditions [56].

Another potential to exploit with piezophilic proteins is the bias in chemistry that has been seen with some piezophilic enzymes. Abe and Horikoshi discussed a porcine α-amylase that demonstrated a higher production of maltotriose instead of maltose and maltotetraose, when maltohexaose was used as a substrate in high pressure [57]. Conversely, in low pressure, all three products were produced at the same rate [57]. Other piezophilic enzymes could show similar properties, and this could be exploited to select for certain products that are applicable to industry.

4. Acidophilic Proteins

Acidophiles are defined as organisms that grow in the lower extremes of pH. Acidophilic enzymes have optimal structure and stability in acidic environments and have been shown to be catalytically active at pHs as low as 1. Most known acidophiles are also thermophiles, and hence their proteins reflect thermophilic features. Interestingly, the adaptation of acidophilic proteins to pH is unclear and inconsistent.

Acidophilic proteins must adapt to the low pH because acid interferes with the charge on many residues. At low pH many polar charged residues become protonated and, therefore, their charges change. This has the possibility of disrupting stabilizing structural interactions, unfolding the protein.

While the specific adaptation has not been explored in great detail, the activity of these proteins at low pH seems to be attributed to the prevalence of acidic (negatively charged at a neutral pH) amino acids on the surface of these enzymes and proteins.

4.1. Negative Surface Charge. Research has shown that a number of acidophilic enzymes have optimal activity at a pH significantly lower than the intracellular pH where that enzyme is located. One explanation for pH stability is offered from research conducted on acidophilic and thermostable endo-β-glucanase from *Sulfolobus solfataricus*. This enzyme has an optimum pH of 1.8 [58]. A prominent feature is the excess of glutamic and aspartic acid surface residues on the enzyme when modeled. This results in a highly negative surface at a pH of 7. Huang et al. also noted that many acidic surface residues have been attributed to instability at high pH because of the repulsion of these excess negative charges. However, at a lower pH of 2, endo-β-glucanase would not have the excess negative charge seen at higher pH, which could help stabilize it in the acidic conditions. These extra acidic residues would also correspond to a lower isoelectric point (pI) for the endo-β-glucanase. However, it should be noted that nonacidophilic β-glucanases have theoretical pI values very similar to that from *S. solfataricus*, while having optimal activity at neutral to only slightly acidic pH. This suggests the abundance of acidic surface residues cannot be the only factor in determining acid stability of endo-β-glucanase [58].

4.2. Possible Explanations for Discrepancies in pH Optima. Another example of low pH stability was demonstrated by the α-glucosidase from *Ferroplasma acidiphilum*. α-Glucosidase demonstrated a preference for pH of 3 instead of 5.6, which is the internal average cytoplasmic pH of *F. acidiphilum* [59]. Similarly carboxylesterase in *F. acidiphilum* was also shown to have a pH optimum of approximately 2. Several other cytoplasmic enzymes also showed similar pH optima. All of these enzymes showed significantly lower activity after the pH was higher than 5, except *F. acidiphilum* α-glucosidase which was still ~60% active [59]. Such low pH optima would be expected in excreted enzymes, due to the acidic environment they are subjected to, but not enzymes that are in the cytoplasm or membrane. Golyshina and Timmis proposed two possible explanations for these discrepancies. It is possible that these enzymes are localized to highly acidic "compartments" within the cytoplasm, even though there is little evidence to support this claim. Another suggestion is that these enzymes form multienzyme complexes which raise the pH optima closer to that of the cytoplasm (5.8). However, no multienzyme complexes like those have been observed in *F. acidiphilum* or other acidophiles [59].

Not all proteins from acidophiles have a preference for low pH as seen in the previous examples. This would be expected since the intracellular pH is not as acidic as external environment. An example of this is seen in the ATP-dependent DNA ligase in *Ferroplasma acidarmanus*.

While the glucosidase and other enzymes from *F. acidiphilum* have a pH optima ranging from 2 to 3, *F. acidarmanus* DNA ligase prefers a more neutral environment. It has optimal nick joining activity at pH 6-7, which is similar to DNA ligases from nonacidophiles [60, 61]. This begs the question why some intracellular acidophilic enzymes have such a low pH optima while others, like the DNA ligase, do not. The answer could be related to the substrate of the enzyme; DNA has decreased stability at acidic pH [60, 61]. Therefore, it would be disadvantageous for the *F. acidarmanus* DNA ligase to be optimally active at a low pH.

4.3. Possible Industrial Applications. Many of these acidophilic enzymes also fall into the thermophilic category and have potential for biotechnological and industrial applications. One such example is in biofuels production where currently high sugar compounds (e.g., corn) are used for ethanol production. Polymeric and oligomeric sources provide a large but unfortunately inaccessible carbon source. For example, if cellulases and xylanolytic enzymes could be used in a hot acidic environment, then the high temperature and acidity could help hydrolyze lignocellulosic materials, making them more accessible [5]. This could help improve ethanol yields from these carbon sources. Another potential application could be in the food industry where glucoamylases could be used to break down complex polysaccharides into basic dextrose and fructose sugars [5]. If these enzymes were heat and acid stable, this could improve the efficiency of monosaccharide production. Further applications of thermal/acid stable enzymes could be in mining industries. The release of acid and metal contaminants from mining sites could damage the environment [4]. The technique known as bioleaching utilizes microorganisms and their enzymes to harvest metals such as copper, nickel, cobalt, zinc, and uranium [62]. This could reduce the environmental damage done by these mining operations.

5. Psychrophilic Proteins

Psychrophiles are a class of extremophiles that grow at temperatures below 20°C [63]. A majority of research on protein adaptations in psychrophiles has been done with bacterial and eukaryotic proteins [64]. Nevertheless, there have been a number of studies that have been done on archaeal organisms living in extremely cold environments; most of the research on archaeal psychrophiles has been done on methanogens growing in Alaska and the Antarctic [65].

A typical protein has extremely low activity at temperatures below 20°C, which is unsuitable for a growing cell [66]. Enzyme activity decreases at lower temperatures due to a lower mean kinetic energy; the conformational movements of a protein become slower and therefore enzymatically less efficient [67]. Also, at low temperatures, the energy barrier of activation for catalysis becomes too great for a protein, further reducing the enzyme's activity [66]. Adaptations in psychrophilic proteins allow them to have enough activity in low temperatures for psychrophilic organisms to thrive in the cold, even though the optimal activity for these proteins is at a temperature above their physiological temperature [66]. Psychrophilic proteins have high activity at low temperatures because they are better able to move and change conformation due to a structure that is more flexible [64].

5.1. Weak Protein Interactions. Psychrophilic proteins have greater flexibility due to a lower energy barrier between the various conformations of the protein [66]. This is because of a difference in the amino acid composition from mesophilic proteins. In general, the stabilizing interactions typically found within a protein are weakened or removed in cold-active proteins. In an excellent review of cold- and heat-active enzymes by Feller, the following adaptations in psychrophilic proteins were summarized: (i) glycine residues are increased, which provide greater conformational mobility in psychrophilic proteins, (ii) proline residues, which provide conformational rigidity, are reduced in loop regions, (iii) salt bridge and hydrogen bond forming arginine residues are reduced, (iv) the size of nonpolar residues in the protein core is reduced to create weaker hydrophobic interactions [66]. As an example of these features, proteins from the archaeal cold-adapted halophile *Halorubum lacusprofundi* display a decrease in large hydrophobic amino acids, such as tryptophan, and in hydrogen bond forming residues, like glutamic acid. In the *H. lacusprofundi* β-galactosidase, there was an increasing hydrophobicity observed on the protein surface, which replaced anionic electrostatic interactions which are usually abundant on halophilic proteins [68, 69]. These types of amino acid trends have also been reported in the elongation factor 2 proteins of psychrophilic methanogens [70].

Genomes from a number of archaeal methanogens across a wide range of optimal growth temperatures were examined by Saunders and coworkers in 2003. Using the draft genome sequences from two psychrophilic methanogens, *Methanogenium frigidum* and *Methanococcoides burtonii*, three-dimensional models of proteins were constructed and compared to other modeled proteins from mesophilic and thermophilic methanogens [71]. As expected, a decrease in the number of charged residues on the amino acid surface was observed on the cold-adapted proteins. Furthermore, an increase in glutamine and threonine residues was seen in these proteins. This is thought to reduce the charge on the protein surface without causing aggregation by creating a surface that is too hydrophobic [71]. This research was examining psychrophilic adaptations in a large number of molecular models (141), and it supported adaptations that have been seen in studies from single proteins [71].

5.2. Lower Thermal Stability. Weaker interactions between amino acid residues in a psychrophilic protein prevent it from being "frozen" in a particular conformation and make the molecular motions needed for catalysis possible. A consequence of these weaker interactions is a less stable protein; thus, cold-adapted proteins unfold at lower temperatures than mesophilic proteins [66, 72, 73]. The thermal unfolding of psychrophilic proteins has been reported to occur through a single transition. This is because the weaker interactions in cold-adapted proteins have greater influence on the overall stability, and local unfolding greatly destabilizes the protein

due to fewer stabilizing interactions [66]. These characteristics have been observed in an archaeal cold-shock protein from *Methanogenium frigidum*, which was shown to be less stable at its optimal temperature than its mesophilic homologue from *E. coli* [74].

5.3. Increased Specific Activity. The catalytic activity of a psychrophilic enzyme, due to the more flexible structure, is much greater at low temperatures than the same enzyme from a mesophile. In fact, despite decreased reaction rates at low temperatures, the specific activity (k_{cat}) of a psychrophilic enzyme is typically 10 times greater than a mesophilic enzyme [66, 73]. A typical observation that is made to explain the greater k_{cat} is the increase in binding site size in psychrophilic proteins [64]. In psychrophilic enzymes, the substrate binding area is enlarged by a number of mechanisms while the catalytic residues are unchanged [66]. Some of the mechanisms by which this area is enlarged include deletion of loops near the binding site [75], strategic glycine residues near the functional sites [66], and pulling the protein backbone out to increase substrate accessibility [76]. As a result, substrates are not able to bind as well to a psychrophilic enzyme, and, therefore, the Michaelis-Menten constant (K_m) of psychrophilic enzymes is high [66, 77]. Poor substrate affinity improves enzyme activity at low temperatures because it reduces the energy of activation for the enzyme [66].

5.4. Industrial Applications. Psychrophilic enzymes have found useful applications in the biotechnical industry. Due to their higher activity at low temperatures, cold-adapted lipases from bacterial psychrophiles are used in commercial detergents [63]. Likewise, cellulases find use due to their reduced thermal stability, making it easier to inactivate the enzyme after a certain amount of time. This is important for stone washing in the textile industry, where if the cellulases are active too long, the mechanical resistance of the cotton is lost [63, 78]. Archaeal cold-adapted enzymes are not as widely used as enzymes from bacteria. Nevertheless, they still have many possible applications in industry due to their adaptations.

6. Halophilic Proteins

Salt has significant effects on the solubility, stability, and conformation of a protein, which ultimately affects its ability to function. Organisms that thrive in extremely salty environments like the Great Salt Lake or the Dead Sea have two major ways through which they adapt to the extreme salt. Some halophiles, mostly halophilic bacteria and eukaryotes, prevent the entry of the inorganic salts (such as NaCl) into the cell and synthesize small organic molecules (like ectoine), known as osmolytes, to balance the osmotic pressure [8]. Halophilic Archaea, though, survive by taking in high concentrations of inorganic salts, requiring their proteins to carry adaptations that allow them to remain stable and functional. At high salt concentrations (higher than 0.1 M), water is less available to protein because most water is surrounding salt in an ionic lattice [8]. The lower availability of water can cause hydrophobic amino acids in a protein to lose hydration and aggregate. Therefore, high salt concentrations strengthen hydrophobic interactions in a protein. Salt also interferes with the electrostatic interactions between charged amino acids [79]. Nonhalophilic proteins cannot function in high salt concentrations because the hydrophobic and electrostatic interactions they normally rely on for proper folding and for maintaining stability are greatly altered. This can even lead to destabilization of the protein, potentially causing global unfolding and aggregation, ultimately leading to precipitation. Archaeal halophilic proteins have a number of adaptations that allow them to utilize the high concentrations of inorganic salt to stabilize their native fold.

6.1. Increased Acidic Residues. One of the most notable differences between halophilic and nonhalophilic proteins is the large increase in acidic residues, like glutamic and aspartic acid, on the protein's surface. This is almost ubiquitous with halophilic proteins and can distinguish between halophilic and nonhalophilic protein sequences [80]. There are a number of possible roles for these acidic residues. It is thought that the increased negative charge on the protein's surface allows the protein to compete with ions for water molecules and, therefore, keep the protein in solution [79, 81–83]. This is supported by the crystal structures of halophilic proteins that show water binding with these acidic surface residues [8, 83, 84]. Bioinformatics analysis of halophilic proteins has shown that their sequences also consistently contain less serine. Serine is good at interacting with water but not at competing with charged ions, so it is thought that serine is less useful for proteins at high salt concentrations [85]. An alternative to increased water binding would be that the acidic residues on halophilic proteins bind hydrated cations which would maintain a shell of hydration around the protein [8, 79, 83, 86–88]. Crystal structures showing specific cation-protein binding are known [83, 84, 89]. The prevalence of protein-cation binding is not well understood, mainly because crystal structures of halophilic proteins are not able to distinguish between salt and water. To distinguish between sodium ions and water (which both have 10 electrons), data on its coordination geometry is required, which requires a structure of high resolution (below 2.4 A) [8].

Recently, Qvist et al. have suggested that, despite crystal structures, halophilic proteins do not have increased waters of hydration due to their greater negative charge [90]. They studied a mutant (Kx6E) of a domain in protein L (immunoglobulin G binding B1 domain) from *Streptococcus magnus*, which contained a number of salt-dependent features seen with normal halophiles (large negative charge and salt-dependent folding and stability). Using an ^{17}O magnetic spin relaxation technique to monitor water associating with the protein or returning to more mobile bulk solvent, they determined that there was no difference in the amount of water bound to the halophilic over the mesophilic versions of protein L [90]. Furthermore, homology-modeled structures of halophilic dihydrofolate reductases show a similar number of hydrogen bonding networks as their nonhalophilic counterparts [86]. This raises questions on how acidic residues, then, are able to keep halophilic proteins soluble. In explaining the hydrating shell of waters seen in crystal structures, Madern et al. note

that crystalizing conditions for proteins involve salting-out conditions, which cause the exclusion of salt and improve water binding [84]. The role of the acidic residues in a halophilic protein may be to increase the proteins flexibility by having a large number of nearby negative charges that repel each other [8]. The repelling charges would make it easier for a halophilic protein to change its conformation despite having a more rigid hydrophobic core (discussed below).

6.2. Decreased Hydrophobic Residues. Other than the larger number of acidic residues in halophilic proteins, bioinformatics studies of halophilic protein sequences have shown that they also contain different hydrophobic residues than mesophilic protein sequences. Using the known crystal structures of 15 pairs of halophilic and nonhalophilic proteins, Siglioccolo et al. determined that the hydrophobic contact in the core of halophilic proteins, exposed to molar concentrations of inorganic salt, is consistently smaller than that in mesophilic proteins (but, interestingly, not for halophilic proteins that are exposed to the organic salts) [91]. They propose that the lower hydrophobic contact in the core may counterbalance the increased strength of hydrophobic interactions in high salt concentrations [91]. Most halophilic proteins contain less of the large, aromatic hydrophobic amino acids [85]. In the homology-modeled structure of halophilic dihydrofolate reductase, there was a decrease in the number of large hydrophobic amino acids, and a reduction of the enzyme core was observed [86]. Weaker hydrophobic interactions due to smaller hydrophobic residues can increase the flexibility of protein in high salt, since it prevents the hydrophobic core from becoming too rigid [8].

6.3. Salt-Dependent Folding. An important advance in understanding halophilic protein adaptation has been the evidence that these proteins rely on salt to fold [92]. This research demonstrates that salt adaptation by halophiles is not only to have proteins that survive the high salt environment but that actually utilize it to function [8]. Our study of the cysteinyl-tRNA synthetase in *H. salinarum* NRC-1 shows how the enzyme not only folds from increasing salt concentrations, but it also becomes more stable and resists thermal denaturation (paper in preparation).

Salt-dependent folding may have been important for very early proteins. The typical amino acid adaptations seen in halophiles (greater acidic residues and smaller hydrophobic amino acids) have also been observed recently in constructed prebiotic proteins [93]. There are, currently, ten known amino acids that could have been created without biosynthetic pathways: alanine, aspartic acid, glutamic acid, glycine, isoleucine, leucine, proline, serine, threonine, and valine. Research by Longo et al. shows that a foldable set of these amino acids leads to a protein with halophilic features and could use high salt concentrations to fold. This suggests that a halophilic environment may have been important for biogenesis [93].

6.4. Halophilic Peptide Insertions. Protein adaptations to high salt are not always found throughout the entire protein sequence. In some cases, halophilicity has been significantly increased by a peptide insertion in the protein [16, 18, 94, 95]. These insertions typically contain a large number of acidic amino acids, and, as seen with cysteinyl-tRNA synthetase from *H. salinarum* NRC-1, the insertion greatly increased the catalytic turnover of the enzyme [18]. Serinyl-tRNA synthetase in *Haloarcula marismortui* also has an insertion sequence, speculated to improve enzyme flexibility [94, 96]. Ferredoxin from the same organism was shown to have an N-terminal extension that contained 15 negatively charged amino acids. This insertion is thought to improve the enzyme's solvent-accessible surface area [83, 84, 97]. These insertion sequences are proposed to have a number of possible functions and could be a way to quickly impart halophilic adaptations to a protein, evolutionarily [97].

6.5. Possible Industrial Applications. Halophilic proteins, so far, have found little use in industry, but there is much interest in finding an application for salt-functioning enzymes. One of these possible applications for halophilic enzymes is in treating highly saline wastewater, such as the waste created by the pickling industry, which has a saline content up to 10%. A number of other possible industrial applications for halophiles have been recently reviewed [98].

Some current work has gone into changing the halophilic features of some enzymes. Ishibashi et al. were able to raise and lower the salt-dependent refolding of *H. salinarum* nucleoside diphosphate kinase with only one amino acid substitution [99]. Mutating asparagine-111 to leucine (N111L) eliminated a hydrogen bond between basic dimeric units of the protein, supposedly making the formation of the functional enzyme more dependent on hydrophobic interactions. This modified the enzyme's optimum activity from 0.45 M NaCl to 1.35 M NaCl, since a higher salt concentration improves the hydrophobic interactions in the nucleoside diphosphate kinase mutant. They were also able to create the reverse effect by substituting glycine 114 to arginine (G114R). This created a new hydrogen bond between basic dimeric units and required less salt to form a functional protein [99]. Tokunaga et al. were able to impart halophilic properties to the same enzyme from the nonhalophilic *Pseudomonas aeruginosa* by only changing two adjacent residues from alanine to glutamic acid [100]. If improving an enzyme's activity in salt is as simple as changing one or two residues, or adding an insertion peptide, this means it could soon be easy to modify almost any protein to function in extreme concentrations of salt for industrial purposes.

7. Haloalkaliphilic Proteins

Because halophilic environments vary in pH, subsets of these environments are highly alkaline. A number of haloalkaliphilic species have been discovered in soda lakes in Egypt, Kenya, China, India, and the western United States [101]. All archaeal alkaliphiles are halophiles [102, 103]. Protein adaptations to alkaline pH in haloalkaliphiles are subtle, due to the fact that these organisms have cellular mechanisms to maintain a more neutral pH in their cytoplasm, usually within a range from 7 to 8.5 [104]. A complex cellular envelope, with a large number of glycosylated proteins, helps maintain a neutral intracellular pH [3, 105]. Also, it appears

that protein adaptations to pH in haloalkaliphiles are secondary to their halophile adaptations. It was observed that the proteins from haloalkaliphiles contained a high proportion of acidic residues that is typically seen with halophilic proteins [3, 106]. Currently, there is no commercial use of archaeal haloalkaliphilic enzymes, though a number of enzymes from bacterial alkaliphiles have found use in industry, including proteases, cellulases, lipases, xylanase, pectinases, and chitinases [104].

8. Summary of Archaeal Adaptations

To illustrate these various protein adaptations, we surveyed differences, with homology modeling, among extremophiles using the enzyme cysteinyl-tRNA synthetase (CysRS). This enzyme catalyzes a highly conserved reaction, the coupling of the amino acid cysteine to its cognate tRNA, which is then used by the ribosome for protein synthesis. Because of its importance in translation, the structure of CysRS is highly conserved, and the regions of the protein sequence that are involved in tRNA binding, anticodon recognition, and catalysis are identical among all organisms. Differences in the models of CysRS between extremophiles highlight the types of adaptations that are seen in these organisms.

Homology models were made using MODELLER [107] with the *E. coli* CysRS-tRNA crystal structure (PDB: 1U0B, [10]) as a template and the amino acid sequence of CysRS from representative halophilic, thermophilic, and psychrophilic organisms.

The sequences used for the alignments and the models were *E. coli*, AP_001173.1, *H. salinarum* sp. NRC-1, NP_280014.1, *P. furiosus*, NP_578753.1, and *M. psychrophilus*, WP_015053952.1. MODELLER generated the following data regarding the models: *H. salinarum* was 39% identical to the *E. coli* CysRS and generated a GA341 score of 1, *P. furiosus* was 48% identical to the *E. coli* CysRS and generated a GA341 score of 1, and *M. psychrophilus* was 44% identical to the *E. coli* CysRS and generated a GA341 score of 1. A GA341 score of 1 is the highest score generated by MODELLER and indicates an acceptable model. The models were then aligned in VMD [108] to further refine the models. No energy minimization was done. Rendering was done using Chimera [11]. All models are drawn with a Coulombic surface map (Figures 1(a) and 1(c)) and a customized homology map (Figures 1(b) and 1(d)). The Coulombic surface map colors the amino acid electrostatic potential (according to Coulomb's law) on surface residues.

As can be seen in the Coulombic surface model of *E. coli* (Ec) CysRS, there is relatively even distribution of positive and negative charges, which is typical of a mesophilic, nonextremophilic protein structure. Highlighted in green on the models of the protein are the conserved regions of CysRS required for proper enzyme function.

The most dramatic change from the Ec CysRS Coulombic surface model is in the halophilic model (Hs), which displays a substantial negative potential from many acidic acid residues (aspartic acid and glutamic acid) and residues with an overall negative surface potential. This is the most common feature of halophilic proteins and enzymes. In the homology model and supplementary figure S1 of the halophilic CysRS, a peptide insertion, which is an additional 20 residues, is near the enzyme's active site. By having the insertion at this location, it is thought that it imparts additional flexibility to the enzyme around the active site [18]. In the back of the molecule, extra acidic residues dot the surface, which might function to pull positively charged ions away from the active site and tRNA binding site.

The thermophilic CysRS model (Pf) displays a more basic and positively charged surface compared to Ec and also possesses a larger hydrophobic core seen near the active site. These features are generally associated with thermophilic proteins. The homology model and supplementary figure S1 highlights additional cysteine, proline, hydrophobic, and charged residues (in red). These residues, which are unique to the thermophilic enzyme compared to the other organisms, are seen on both sides of the enzyme, possibly indicating that these features would provide greater overall stability to the molecule.

The psychrophilic CysRS (Mp) surface potential model shows a small reduction in surface charge, despite an unexpected acidic patch on the back of the molecule. The reduced charge is consistent with the common psychrophilic adaptation of increased surface hydrophobic residues. Other unique features observed in the homology model and supplementary figure S1 were additional glycines and hydrophobic patches (blue). A majority of these adaptations are proximal to the active site of the protein, which could impart greater flexibility in this region, improving catalytic activity at lower temperatures.

9. Concluding Statements

In this review we have discussed the major protein adaptations observed in archaeal organisms that thrive in vastly different extreme environments. While not all adaptations are known, it appears that, for some proteins, subtle changes in the amino acid composition are all that is needed to remain functional in an extreme environment. These differences are reflected as changes in charge, hydrophobicity, and subtle changes in structure. It is also clear that the organisms have evolved ways to manipulate these changes to optimize the protein or enzyme activity. These adaptations allow the organism and their proteins to take advantage of their environment. This has led to much interest in understanding these extreme adaptations and in manipulating these changes to find applications for these biological molecules.

Acknowledgments

C. Reed was supported through the S-STEM program at Idaho State University, which is funded through a Grant from the National Science Foundation (NSF no. 0965939) and through the BS/MS program through the Department of Chemistry at ISU. H. Lewis and E. Trejo were supported though the Career Path Internship program at Idaho State University. E. Trejo was also supported

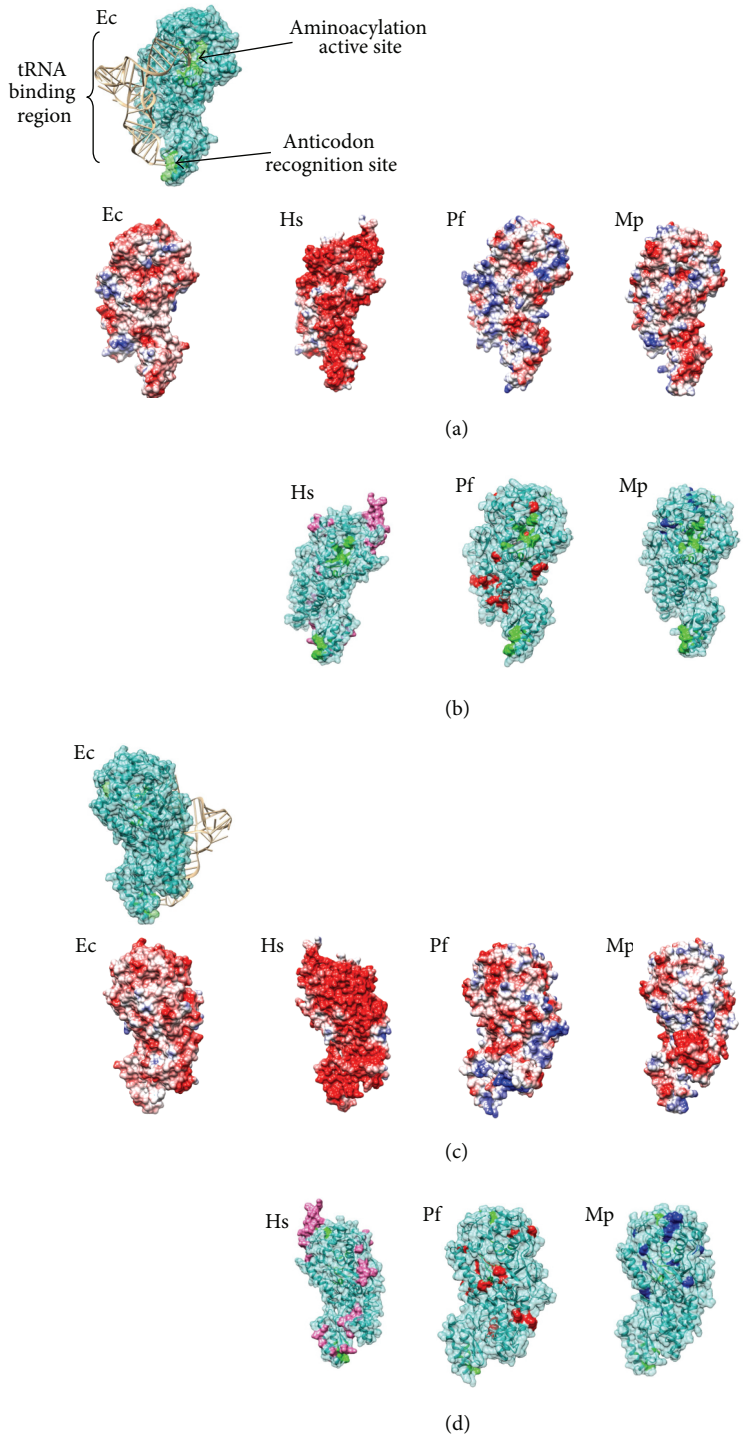


Figure 1: Graphical view of cysteinyl-tRNA synthetase with extremophilic protein adaptations. The homology models of *Halobacterium salinarum* (Hs), *Pyrococcus furiosus* (Pf), and *Methanolobus psychrophilus* (Mp) CysRS were generated based on the structure of *Escherichia coli* CysRS (see text for details). In the upper corner, the crystal structure of the Ec CysRS (PDB 1U0B, [10]) is provided for orientation and description of the enzyme's features. (a) and (c) Coulombic surface map of the models on the tRNA side and back of the molecule, respectively. The Coulombic surface maps the amino acid electrostatic potential (according to Coulomb's law) on surface residues: red is a negative potential, blue is a positive potential, and white indicates a relatively nonpolar potential. (b) and (d) The conserved features (in green) and unique adaptations highlighted on the surface of the models on the tRNA side and back of the molecule, respectively. The corresponding adaptations have been noted in the sequence alignment in Figure S1 (See Figure S1 in the Supplementary Material available online at http://dx.doi.org/10.1155/2013/373275). Unique features are highlighted in different colors for the different extremes: halophilic adaptations are in pink, the thermophilic adaptations are in red, and the psychrophile adaptations are in blue. The molecular graphics were created with the USCF Chimera package [11].

through the summer INBRE Program, NIH, Grant nos. P20 RR016454 (National Center for Research Resources) and P20 GM103408 (National Institute of General Medical Sciences). The authors' models of the Hs, Pf, and Mp CysRS were made with MODELLER and VMD. MODELLER was made and developed by Andrej Sali, the Program for Comparative Protein Structure Modeling by Satisfaction of Spatial Restraints at the University of California, San Francisco. VMD was developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign. Our models were visualized and analyzed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). The authors would like to acknowledge the use of the GALAXY [109] computational cluster at the Molecular and Bioinformatics core facility, which ran the HMMER3 algorithm [110]. GALAXY is maintained by Dr. Michael Thomas and Dr. Luobin Yang, Department of Biological Sciences, ISU.

References

[1] A. Kletzin, *General Characteristics and Important Model Organisms*, ASM Press, Washington, DC, USA, 2007, Archaea: Molecular and Cellular Biology, Edited by: R. Cavicchioli.

[2] C. Schleper, G. Jurgens, and M. Jonuscheit, "Genomic studies of uncultivated archaea," *Nature Reviews Microbiology*, vol. 3, no. 6, pp. 479–488, 2005.

[3] M. Falb, F. Pfeiffer, P. Palm et al., "Living with two extremes: conclusions from the genome sequence of *Natronomonas pharaonis*," *Genome Research*, vol. 15, no. 10, pp. 1336–1343, 2005.

[4] C. Baker-Austin and M. Dopson, "Life in acid: pH homeostasis in acidophiles," *Trends in Microbiology*, vol. 15, no. 4, pp. 165–171, 2007.

[5] A. Sharma, Y. Kawarabayasi, and T. Satyanarayana, "Acidophilic bacteria and archaea: acid stable biocatalysts and their potential applications," *Extremophiles*, vol. 16, no. 1, pp. 1–19, 2012.

[6] W. D. Grant, "Life at low water activity," *Philosophical Transactions of the Royal Society B*, vol. 359, no. 1448, pp. 1249–1266, 2004.

[7] C. Ebel, L. Costenaro, M. Pascu et al., "Solvent interactions of halophilic malate dehydrogenase," *Biochemistry*, vol. 41, no. 44, pp. 13234–13244, 2002.

[8] M. Mevarech, F. Frolow, and L. M. Gloss, "Halophilic enzymes: proteins with a grain of salt," *Biophysical Chemistry*, vol. 86, no. 2-3, pp. 155–164, 2000.

[9] D. B. Wright, D. D. Banks, J. R. Lohman, J. L. Hilsenbeck, and L. M. Gloss, "The effect of salts on the activity and stability of *Escherichia coli* and *Haloferax volcanii* dihydrofolate reductases," *Journal of Molecular Biology*, vol. 323, no. 2, pp. 327–344, 2002.

[10] S. Hauenstein, C.-M. Zhang, Y.-M. Hou, and J. J. Perona, "Shape-selective RNA recognition by cysteinyl-tRNA synthetase," *Nature Structural and Molecular Biology*, vol. 11, no. 11, pp. 1134–1141, 2004.

[11] E. F. Pettersen, T. D. Goddard, C. C. Huang et al., "UCSF chimera—a visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[12] G. Cacciapuoti, F. Fuccio, L. Petraccone et al., "Role of disulfide bonds in conformational stability and folding of 5'-deoxy-5'-methylthioadenosine phosphorylase II from the hyperthermophilic archaeon *Sulfolobus solfataricus*," *Biochimica et Biophysica Acta*, vol. 1824, no. 10, pp. 1136–1143, 2012.

[13] A. Szilágyi and P. Závodszky, "Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey," *Structure*, vol. 8, no. 5, pp. 493–504, 2000.

[14] C. Vieille and G. J. Zeikus, "Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability," *Microbiology and Molecular Biology Reviews*, vol. 65, no. 1, pp. 1–43, 2001.

[15] H. Connaris, J. B. Chaudhuri, M. J. Danson et al., "Expression, reactivation, and purification of enzymes from *Haloferax volcanii* in *Escherichia coli*," *Biotechnology and Bioengineering*, vol. 64, no. 1, pp. 38–45, 1999.

[16] C. Evilia, X. Ming, S. Dassarma, and Y.-M. Hou, "Aminoacylation of an unusual tRNACys from an extreme halophile," *RNA*, vol. 9, no. 7, pp. 794–801, 2003.

[17] Y. Yonezawa, H. Tokunaga, M. Ishibashi, S. Taura, and M. Tokunaga, "Cloning, expression, and efficient purification in *Escherichia coli* of a halophilic nucleoside diphosphate kinase from the moderate halophile *Halomonas* sp. #593," *Protein Expression and Purification*, vol. 27, no. 1, pp. 128–133, 2003.

[18] C. Evilia and Y.-M. Hou, "Acquisition of an insertion peptide for efficient aminoacylation by a halophile tRNA synthetase," *Biochemistry*, vol. 45, no. 22, pp. 6835–6845, 2006.

[19] E. Bae and G. N. Phillips Jr., "Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases," *The Journal of Biological Chemistry*, vol. 279, no. 27, pp. 28202–28208, 2004.

[20] S. Kumar and R. Nussinov, "Different roles of electrostatics in heat and in cold: adaptation by citrate synthase," *ChemBioChem*, vol. 5, no. 3, pp. 280–290, 2004.

[21] M. Tehei and G. Zaccai, "Adaptation to high temperatures through macromolecular dynamics by neutron scattering," *FEBS Journal*, vol. 274, no. 16, pp. 4034–4043, 2007.

[22] E. T. Powers and W. E. Balch, "Diversity in the origins of proteostasis networks—a driver for protein function in evolution," *Nature Reviews Molecular Cell Biology*, vol. 14, no. 4, pp. 237–248, 2013.

[23] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annual Review of Biophysics*, vol. 37, pp. 289–316, 2008.

[24] N. C. Fitzkee, P. J. Fleming, H. Gong, N. Panasik Jr., T. O. Street, and G. D. Rose, "Are proteins made from a limited parts list?" *Trends in Biochemical Sciences*, vol. 30, no. 2, pp. 73–80, 2005.

[25] S. J. Tomazic and A. M. Klibanov, "Mechanisms of irreversible thermal inactivation of *Bacillus* α-amylases," *The Journal of Biological Chemistry*, vol. 263, no. 7, pp. 3086–3091, 1988.

[26] F. Mayer, U. Küper, C. Meyer et al., "AMP-forming acetyl coenzyme a synthetase in the outermost membrane of the hyperthermophilic crenarchaeon *Ignicoccus hospitalis*," *Journal of Bacteriology*, vol. 194, no. 6, pp. 1572–1581, 2012.

[27] C. Bräsen, C. Urbanke, and P. Schönheit, "A novel octameric AMP-forming acetyl-CoA synthetase from the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*," *FEBS Letters*, vol. 579, no. 2, pp. 477–482, 2005.

[28] C. Ingram-Smith and K. S. Smith, "AMP-forming acetyl-CoA synthetases in Archaea show unexpected diversity in substrate utilization," *Archaea*, vol. 2, no. 2, pp. 95–107, 2007.

[29] P. Del Vecchio, M. Elias, L. Merone et al., "Structural determinants of the high thermal stability of SsoPox from the hyperthermophilic archaeon *Sulfolobus solfataricus*," *Extremophiles*, vol. 13, no. 3, pp. 461–470, 2009.

[30] J. T. Park, H. N. Song, T. Y. Jung et al., "A novel domain arrangement in a monomeric cyclodextrin-hydrolyzing enzyme from the hyperthermophile *Pyrococcus furiosus*," *Biochimica Et Biophysica Acta*, vol. 1834, no. 1, pp. 380–386, 2013.

[31] K.-H. Park, T.-J. Kim, T.-K. Cheong, J.-W. Kim, B.-H. Oh, and B. Svensson, "Structure, specificity and function of cyclomaltodextrinase, a multispecific enzyme of the α-amylase family," *Biochimica et Biophysica Acta*, vol. 1478, no. 2, pp. 165–185, 2000.

[32] M. Vihinen, "Relationship of protein flexibility to thermostability," *Protein Engineering*, vol. 1, no. 6, pp. 477–480, 1987.

[33] G. Cacciapuoti, M. Porcelli, C. Bertoldo, M. De Rosa, and V. Zappia, "Purification and characterization of extremely thermophilic and thermostable 5'-methylthioadenosine phosphorylase from the archaeon *Sulfolobus solfataricus*. Purine nucleoside phosphorylase activity and evidence for intersubunit disulfide bonds," *The Journal of Biological Chemistry*, vol. 269, no. 40, pp. 24762–24769, 1994.

[34] D. R. Boutz, D. Cascio, J. Whitelegge, L. J. Perry, and T. O. Yeates, "Discovery of a thermophilic protein complex stabilized by topologically interlinked chains," *Journal of Molecular Biology*, vol. 368, no. 5, pp. 1332–1344, 2007.

[35] K. J. Woycechowsky and R. T. Raines, "The CXC motif: a functional mimic of protein disulfide isomerase," *Biochemistry*, vol. 42, no. 18, pp. 5387–5394, 2003.

[36] B. Wilkinson and H. F. Gilbert, "Protein disulfide isomerase," *Biochimica et Biophysica Acta*, vol. 1699, no. 1-2, pp. 35–44, 2004.

[37] A. Karshikoff and R. Ladenstein, "Ion pairs and the thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads," *Trends in Biochemical Sciences*, vol. 26, no. 9, pp. 550–556, 2001.

[38] Z. S. Hendsch and B. Tidor, "Do salt bridges stabilize proteins? A continuum electrostatic analysis," *Protein Science*, vol. 3, no. 2, pp. 211–226, 1994.

[39] C.-H. Chan, T.-H. Yu, and K.-B. Wong, "Stabilizing salt-bridge enhances protein thermostability by reducing the heat capacity change of unfolding," *PLoS ONE*, vol. 6, no. 6, Article ID e21624, 2011.

[40] S. Fukuchi and K. Nishikawa, "Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria," *Journal of Molecular Biology*, vol. 309, no. 4, pp. 835–843, 2001.

[41] C.-F. Lee, G. I. Makhatadze, and K.-B. Wong, "Effects of charge-to-alanine substitutions on the stability of ribosomal protein L30e from *Thermococcus celer*," *Biochemistry*, vol. 44, no. 51, pp. 16817–16825, 2005.

[42] Y. F. Liu, N. Zhang, X. Liu et al., "Molecular mechanism underlying the interaction of typical Sac10b family proteins with DNA," *PLoS ONE*, vol. 7, no. 4, Article ID e34986.

[43] B. Mamat, A. Roth, C. Grimm et al., "Crystal structures and enzymatic properties of three formyltransferases from archaea: environmental adaptation and evolutionary relationship," *Protein Science*, vol. 11, no. 9, pp. 2168–2178, 2002.

[44] J. Breitung, G. Borner, S. Scholz, D. Linder, K. O. Stetter, and R. K. Thauer, "Salt dependence, kinetic properties and catalytic mechanism of *N*-formylmethanofuran: tetrahydromethanopterin formyltransferase from the extreme thermophile *Methanopyrus kandleri*," *European Journal of Biochemistry*, vol. 210, no. 3, pp. 971–981, 1992.

[45] L. D. Unsworth, J. Van Der Oost, and S. Koutsopoulos, "Hyperthermophilic enzymes—stability, activity and implementation strategies for high temperature applications," *FEBS Journal*, vol. 274, no. 16, pp. 4044–4056, 2007.

[46] M. De Champdoré, M. Staiano, M. Rossi, and S. D'Auria, "Proteins from extremophiles as stable tools for advanced biotechnological applications of high social interest," *Journal of the Royal Society Interface*, vol. 4, no. 13, pp. 183–191, 2007.

[47] J. Fang, L. Zhang, and D. A. Bazylinski, "Deep-sea piezosphere and piezophiles: geomicrobiology and biogeochemistry," *Trends in Microbiology*, vol. 18, no. 9, pp. 413–422, 2010.

[48] S. Hay, R. M. Evans, C. Levy et al., "Are the catalytic properties of enzymes from piezophilic organisms pressure adapted?" *ChemBioChem*, vol. 10, no. 14, pp. 2348–2353, 2009.

[49] B. B. Boonyaratanakornkit, C. B. Park, and D. S. Clark, "Pressure effects on intra- and intermolecular interactions within proteins," *Biochimica et Biophysica Acta*, vol. 1595, no. 1-2, pp. 235–249, 2002.

[50] M. Di Giulio, "A comparison of proteins from *Pyrococcus furiosus* and *Pyrococcus abyssi*: barophily in the physicochemical properties of amino acids and in the genetic code," *Gene*, vol. 346, pp. 1–6, 2005.

[51] E. Mombelli, E. Shehi, P. Fusi, and P. Tortora, "Exploring hyperthermophilic proteins under pressure: theoretical aspects and experimental findings," *Biochimica et Biophysica Acta*, vol. 1595, no. 1-2, pp. 392–396, 2002.

[52] R. Consonni, L. Santomo, P. Fusi, P. Tortora, and L. Zetta, "A single-point mutation in the extreme heat- and pressure-resistant Sso7d protein from *Sulfolobus solfataricus* leads to a major rearrangement of the hydrophobic core," *Biochemistry*, vol. 38, no. 39, pp. 12709–12717, 1999.

[53] P. Fusi, K. Goossens, R. Consonni et al., "Extreme heat- and pressure-resistant 7-kDa protein P2 from the archaeon *Sulfolobus solfataricus* is dramatically destabilized by a single-point amino acid substitution," *Proteins*, vol. 29, no. 3, pp. 381–390, 1997.

[54] M. M. C. Sun, R. Caillot, G. Mak, F. T. Robb, and D. S. Clark, "Mechanism of pressure-induced thermostabilization of proteins: studies of glutamate dehydrogenases from the hyperthermophile *Thermococcus litoralis*," *Protein Science*, vol. 10, no. 9, pp. 1750–1757, 2001.

[55] E. Rosenbaum, F. Gabel, M. A. Durá et al., "Effects of hydrostatic pressure on the quaternary structure and enzymatic activity of a large peptidase complex from *Pyrococcus horikoshii*," *Archives of Biochemistry and Biophysics*, vol. 517, no. 2, pp. 104–110, 2012.

[56] F. Simonato, S. Campanaro, F. M. Lauro et al., "Piezophilic adaptation: a genomic point of view," *Journal of Biotechnology*, vol. 126, no. 1, pp. 11–25, 2006.

[57] F. Abe and K. Horikoshi, "The biotechnological potential of piezophiles," *Trends in Biotechnology*, vol. 19, no. 3, pp. 102–108, 2001.

[58] Y. Huang, G. Krauss, S. Cottaz, H. Driguez, and G. Lipps, "A highly acid-stable and thermostable endo-β-glucanase from the thermoacidophilic archaeon *Sulfolobus solfataricus*," *Biochemical Journal*, vol. 385, no. 2, pp. 581–588, 2005.

[59] O. V. Golyshina and K. N. Timmis, "Ferroplasma and relatives, recently discovered cell wall-lacking archaea making a living in extremely acid, heavy metal-rich environments," *Environmental Microbiology*, vol. 7, no. 9, pp. 1277–1288, 2005.

[60] B. R. Jackson, C. Noble, M. Lavesa-Curto, P. L. Bond, and R. P. Bowater, "Characterization of an ATP-dependent DNA ligase from the acidophilic archaeon "*Ferroplasma acidarmanus*" Fer1," *Extremophiles*, vol. 11, no. 2, pp. 315–327, 2007.

[61] S. Magnet and J. S. Blanchard, "Mechanistic and kinetic study of the ATP-dependent DNA ligase of *Neisseria meningitidis*," *Biochemistry*, vol. 43, no. 3, pp. 710–717, 2004.

[62] T. Rohwerder, T. Gehrke, K. Kinzler, and W. Sand, "Bioleaching review part A: progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation," *Applied Microbiology and Biotechnology*, vol. 63, no. 3, pp. 239–248, 2003.

[63] M. Luisa Tutino, G. Di Prisco, G. Marino, and D. De Pascale, "Cold-adapted esterases and lipases: from fundamentals to application," *Protein and Peptide Letters*, vol. 16, no. 10, pp. 1172–1180, 2009.

[64] A. O. Smalas, H. K. Leiros, V. Os et al., "Cold adapted enzymes," *Biotechnology Annual Review*, vol. 6, pp. 1–57, 2000.

[65] X. Dong and Z. Chen, "Psychrotolerant methanogenic archaea: diversity and cold adaptation mechanisms," *Science China Life Sciences*, vol. 55, no. 5, pp. 415–421, 2012.

[66] G. Feller, "Protein stability and enzyme activity at extreme biological temperatures," *Journal of Physics*, vol. 22, no. 32, Article ID 323101, 2010.

[67] R. Cavicchioli, T. Thomas, and P. M. G. Curmi, "Cold stress response in Archaea," *Extremophiles*, vol. 4, no. 6, pp. 321–331, 2000.

[68] S. Dassarma, M. D. Capes, R. Karan et al., "Amino acid substitutions in cold-adapted proteins from *Halorubrum lacusprofundi*, an extremely halophilic microbe from antarctica," *PLoS ONE*, vol. 8, no. 3, Article ID e58587, 2013.

[69] R. Karan, M. D. Capes, P. DasSarma et al., "Cloning, overexpression, purification, and characterization of a polyextremophilic β-galactosidase from the Antarctic haloarchaeon *Halorubrum lacusprofundi*," *BMC Biotechnology*, vol. 13, no. 3, 2013.

[70] T. Thomas and R. Cavicchioli, "Archaeal cold-adapted proteins: structural and evolutionary analysis of the elongation factor 2 proteins from psychrophilic, mesophilic and thermophilic methanogens," *FEBS Letters*, vol. 439, no. 3, pp. 281–286, 1998.

[71] N. F. W. Saunders, T. Thomas, P. M. G. Curmi et al., "Mechanisms of thermal adaptatation revealed from genomes of the anatarctic *Archaea Methanogenium frigidum* and *Methanococcoides burtonii*," *Genome Research*, vol. 13, no. 7, pp. 1580–1588, 2003.

[72] S. D'Amico, C. Gerday, and G. Feller, "Structural determinants of cold adaptation and stability in a large protein," *The Journal of Biological Chemistry*, vol. 276, no. 28, pp. 25791–25796, 2001.

[73] D. Georlette, B. Damien, V. Blaise et al., "Structural and functional adaptations to extreme temperatures in psychrophilic, mesophilic, and thermophilic DNA ligases," *The Journal of Biological Chemistry*, vol. 278, no. 39, pp. 37015–37023, 2003.

[74] L. Giaquinto, P. M. G. Curmi, K. S. Siddiqui et al., "Structure and function of cold shock proteins in archaea," *Journal of Bacteriology*, vol. 189, no. 15, pp. 5738–5748, 2007.

[75] R. J. M. Russell, U. Gerike, M. J. Danson, D. W. Hough, and G. L. Taylor, "Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium," *Structure*, vol. 6, no. 3, pp. 351–362, 1998.

[76] N. Aghajari, F. Van Petegem, V. Villeret et al., "Crystal structures of a psychrophilic metalloprotease reveal new insights into catalysis by cold-adapted proteases," *Proteins*, vol. 50, no. 4, pp. 636–647, 2003.

[77] S. D'Amico, J. S. Sohier, and G. Feller, "Kinetics and energetics of ligand binding determined by microcalorimetry: insights into active site mobility in a psychrophilic α-amylase," *Journal of Molecular Biology*, vol. 358, no. 5, pp. 1296–1304, 2006.

[78] F. Hasan, A. A. Shah, and A. Hameed, "Industrial applications of microbial lipases," *Enzyme and Microbial Technology*, vol. 39, no. 2, pp. 235–251, 2006.

[79] R. Karan, M. D. Capes, and S. Dassarma, "Function and biotechnology of extremophilic enzymes in low water activity," *Aquatic Biosystems*, vol. 8, no. 1, p. 4, 2012.

[80] G. Zhang, G. Huihua, and L. Yi, "Stability of halophilic proteins: from dipeptide attributes to discrimination classifier," *International Journal of Biological Macromolecules*, vol. 53, pp. 1–6, 2013.

[81] K. L. Britton, P. J. Baker, M. Fisher et al., "Analysis of protein solvent interactions in glucose dehydrogenase from the extreme halophile *Haloferax mediterranei*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 13, pp. 4846–4851, 2006.

[82] O. Dym, M. Mevarech, and J. L. Sussman, "Structural features that stabilize halophilic malate dehydrogenase from an archaebacterium," *Science*, vol. 267, no. 5202, pp. 1344–1346, 1995.

[83] F. Frolow, M. Harel, J. L. Sussman, M. Mevarech, and M. Shoham, "Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin," *Nature Structural Biology*, vol. 3, no. 5, pp. 452–458, 1996.

[84] D. Madern, C. Ebel, and G. Zaccai, "Halophilic adaptation of enzymes," *Extremophiles*, vol. 4, no. 2, pp. 91–98, 2000.

[85] G. Zhang and H. Ge, "Protein hypersaline adaptation: insight from amino acids with machine learning algorithms," *The Protein Journal*, vol. 32, no. 4, pp. 239–245, 2013.

[86] P. L. Kastritis, N. C. Papandreou, and S. J. Hamodrakas, "Haloadaptation: insights from comparative modeling studies of halophilic archaeal DHFRs," *International Journal of Biological Macromolecules*, vol. 41, no. 4, pp. 447–453, 2007.

[87] J. Soppa, "From genomes to function: haloarchaea as model organisms," *Microbiology*, vol. 152, no. 3, pp. 585–590, 2006.

[88] X. Tadeo, B. López-Méndez, T. Trigueros, A. Laín, D. Castaño, and O. Millet, "Structural basis for the aminoacid composition of proteins from halophilic archea," *PLoS Biology*, vol. 7, no. 12, Article ID e1000257, 2009.

[89] S. B. Richard, D. Madern, E. Garcin, and G. Zaccai, "Halophilic adaptation: novel solvent protein interactions observed in the 2.9 and 2.6 Å resolution structures of the wild type and a mutant of malate dehydrogenase from *Haloarcula marismortui*," *Biochemistry*, vol. 39, no. 5, pp. 992–1000, 2000.

[90] J. Qvist, G. Ortega, X. Tadeo, O. Millet, and B. Halle, "Hydration dynamics of a halophilic protein in folded and unfolded states," *Journal of Physical Chemistry B*, vol. 116, no. 10, pp. 3436–3444, 2012.

[91] A. Siglioccolo, A. Paiardini, M. Piscitelli, and S. Pascarella, "Structural adaptation of extreme halophilic proteins through decrease of conserved hydrophobic contact surface," *BMC Structural Biology*, vol. 11, article 50, 2011.

[92] M. Müller-Santos, E. M. de Souza, F. D. O. Pedrosa et al., "First evidence for the salt-dependent folding and activity of an esterase from the halophilic archaea *Haloarcula marismortui*," *Biochimica et Biophysica Acta*, vol. 1791, no. 8, pp. 719–729, 2009.

[93] L. M. Longo, J. Lee, and M. Blaber, "Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 6, pp. 2135–2139, 2013.

[94] C. M.-J. Taupin, M. Härtlein, and R. Leberman, "Seryl-tRNA synthetase from the extreme halophile *Haloarcula marismortui*. Isolation, characterization and sequencing of the gene and its expression in *Escherichia coli*," *European Journal of Biochemistry*, vol. 243, no. 1-2, pp. 141–150, 1997.

[95] C. M.-J. Taupin and R. Leberman, "Archaeabacterial seryl-tRNA synthetases: adaptation to extreme environments and evolutionary analysis," *Journal of Molecular Evolution*, vol. 48, no. 4, pp. 408–420, 1999.

[96] G. Zaccai, F. Cendrin, Y. Haik, N. Borochov, and H. Eisenberg, "Stabilization of halophilic malate dehydrogenase," *Journal of Molecular Biology*, vol. 208, no. 3, pp. 491–500, 1989.

[97] B.-L. Marg, K. Schweimer, H. Sticht, and D. Oesterhelt, "A two-α-helix extra domain mediates the halophilic character of a plant-type ferredoxin from Halophilic Archaea," *Biochemistry*, vol. 44, no. 1, pp. 29–39, 2005.

[98] A. Oren, "Industrial and environmental applications of halophilic microorganisms," *Environmental Technology*, vol. 31, no. 8-9, pp. 825–834, 2010.

[99] M. Ishibashi, T. Hayashi, C. Yoshida et al., "Increase of salt dependence of halophilic nucleoside diphosphate kinase caused by a single amino acid substitution," *Extremophiles*, vol. 17, no. 4, pp. 585–591, 2013.

[100] H. Tokunaga, T. Arakawa, and M. Tokunaga, "Engineering of halophilic enzymes: two acidic amino acid residues at the carboxy-terminal region confer halophilic characteristics to *Halomonas* and *Pseudomonas* nucleoside diphosphate kinases," *Protein Science*, vol. 17, no. 9, pp. 1603–1610, 2008.

[101] M. Enache, T. Itoh, T. Fukushima, R. Usami, L. Dumitru, and M. Kamekura, "Phylogenetic relationships within the family *Halobacteriaceae* inferred from rpoB′ gene and protein sequences," *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. 10, pp. 2289–2295, 2007.

[102] B. Ollivier, P. Caumette, J.-L. Garcia, and R. A. Mah, "Anaerobic bacteria from hypersaline environments," *Microbiological Reviews*, vol. 58, no. 1, pp. 27–38, 1994.

[103] W. D. Grant, *Half A Lifetime in soda lakes*, Springer, New York, NY, USA, 2004, Halophilic Microorganisms, Edited by: A. Ventosa.

[104] K. Horikoshi, "Alkaliphiles: some applications of their products for biotechnology," *Microbiology and Molecular Biology Reviews*, vol. 63, no. 4, pp. 735–750, 1999.

[105] S. Siddaramappa, J. F. Challacombe, R. E. De Castro et al., "A comparative genomics perspective on the geneticcontent of the alkaliphilic haloarchaeon Natrialbamagadii ATCC 43099T," *BMC Genomics*, p. 165, 2012.

[106] M. I. Giménez, C. A. Studdert, J. J. Sánchez, and R. E. De Castro, "Extracellular protease of *Natrialba magadii*: purification and biochemical characterization," *Extremophiles*, vol. 4, no. 3, pp. 181–188, 2000.

[107] N. Eswar, B. Webb, M. A. Marti-Renom et al., "Comparative protein structure modeling using Modeller," *Current Protocols in Bioinformatics*, vol. 5, 2006.

[108] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[109] J. Goecks, A. Nekrutenko, J. Taylor, and T. Galaxy Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, p. R86, 2010.

[110] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.