Towards a systematic understanding of differences between archaenal and bacterial diversity

Anja Spang1,2* and Pierre Offre1**

1NIOZ, Royal Netherlands Institute for Sea Research, Department of Marine Microbiology and Biogeochemistry, and Utrecht University, NL-1790 AB, Den Burg, The Netherlands.
2Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123, Uppsala, Sweden.

Summary

In this crystal ball, we discuss emerging methodologies that can help reaching a synthesis on the biodiversity of Archaea and Bacteria and thereby inform a central enigma in microbiology, i.e. the fundamental split between these primary domains of life and the apparent lower diversity of the Archaea.

Humans have long marvelled at the variety of life forms that populate the Earth. This sense of an aesthetic of life together with the impending need to manage and preserve biological resources and ecosystem services is one of the driving forces that keeps pushing forward the exploration of life’s biodiversity. Yet, in spite of continuing efforts to describe and categorize life forms, a complete inventory of all extant taxa and a comprehensive and systematic synthesis integrating the multiple aspects of biodiversity still lay far ahead. The invisible realm of microorganisms constitutes the largest and least explored reservoir of biodiversity and it is widely recognized that Bacteria represent an extraordinarily diverse group of organisms. In contrast, Archaea, which form a separate domain of life distinct from Bacteria and Eukarya, have only been described four decades ago and have been assumed to comprise much fewer taxa, many of which were initially thought to have only minor roles in global biogeochemical cycles. This point of view emerged from early discoveries, which described archaenal organisms thriving in environments characterized by extreme temperature, salt or pH regimes. The use of cultivation-independent approaches, most notably meta- and single cell genomics, has recently opened a window into the hidden world of microorganisms and has considerably expanded the number of known archaenal and bacterial taxa, revealing novel lineages of high taxonomic rank that are broadly distributed throughout the Biosphere (Adam et al., 2017; Spang et al., 2017; Castelle and Banfield, 2018). Yet, in spite of tremendous progress in sampling bacterial and archaenal genomes, archaenal taxa still appear to be outnumbered by bacterial counterparts, though estimates vary (Hug et al., 2016; Parks et al., 2018). This raises a number of fundamental questions, which need to be addressed carefully as the diversity of archaenal and bacteria has deep connections to their ecology and mode of evolution and sampling microbial diversity is essential to uncover the deepest transitions in the evolution of life on Earth.

How accurate are the estimates of archaenal and bacterial taxonomic richness and how much diversity still defies discovery? Is the discrepancy of observed biodiversity between archaenal and bacteria merely a result of evolutionary contingency or is there a deterministic explanation? For example; do archaenal have a late origin from within bacteria, suggesting that bacteria had more time to evolve and diversify? Or are archaenal characterized by different evolvability, e.g. do they experience different average rates of mutations, substitutions, recombination and horizontal gene transfer? Alternatively, did fundamental traits such as a different degree of adaptation to chronic energy stress (Valentine, 2007) of the primordial archaenal and bacteria result in niche differentiation towards two distinct categories of environments? Finally, one may wonder, whether the biosphere realizes fewer niches in which archaenal organisms have a competitive advantage.

We want to take this opportunity to look into the crystal ball for discussing methodologies that can help reaching a synthesis on the biodiversity of archaenal and bacteria and thereby inform these questions. Furthermore, we highlight some central insights that could emerge from such a synthesis.

The total number of archaenal and bacterial species (defined as groups of 16S amplicons sharing 97% sequence identity) has recently been predicted to be in the order of 1 trillion, less than 107 of which have been
catalogued thus far (Locy and Lennon, 2016). While this estimate provides a powerful message in terms of commonness and rarity of 16S phylotypes, it is unclear how well it reflects the taxonomic richness of archaea and bacteria and to which extent it represents their genetic diversity. Certainly, the use of PCR-based 16S rRNA gene surveys often results in a biased picture of the microbial diversity in a sample, as for example revealed through primer-independent approaches. Thus, even though the use of 16S amplicon-based surveys is still commonly applied to assess microbial diversity, successes of PCR-independent approaches advocate for their widespread adoption. However, several challenges still stand in the way of a complete shift in methodology, particularly if metagenome sequencing takes over 16S amplicon surveys. For example, integrating the wealth of available genomic data will depend on the establishment of an improved computational infrastructure, that enables a centralized data storage and coherent data analysis (Koonin, 2018). Further, it remains to be proven that recovering the genome sequence of rare microorganisms can be routinely achieved. Yet, we are confident that the coming years will see further developments and standardization of metagenomics and single cell genomics helped by the introduction of novel library preparation protocols, sequencing technologies and bioinformatics tools and resources. We expect continuing discoveries of novel archaeal and bacterial taxa but also wonder whether broader taxonomic groups such as the CPR (Brown et al., 2015), DPANN (Rinke et al., 2013; Castelle et al., 2015) and Asgard archaea (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017) still await discovery.

Beyond the sampling of bacterial and archaeal genome sequences, evaluating and comparing the taxonomic richness of the Archaea and Bacteria depends ultimately on the use of normalized taxonomic ranks and a resolved phylogeny that adequately reconstructs the evolutionary history of Bacteria and Archaea. For example, most taxonomies in common use are largely unstandardized: e.g. different genera may represent non-equivalent categories of organisms, which should be assigned to different ranks. Recently, however, a novel standardized taxonomy was proposed for Bacteria, which took into account the varying degrees of evolutionary divergence of distinct taxa for the assignment of ranks (Parks et al., 2018). This taxonomy established 111 phyla including 9911 species based on 94,759 genomes, 58% of which had a change to their existing taxonomy (Parks et al., 2018). For comparison, a similar approach established 11 phyla and 509 species for the archaea based on 2075 genomes (http://gtdb.ecogenomic.org/). Although this approach certainly represents a significant step towards an adequate taxonomy, it relies on the assumption that the topology, branch length and rooting of the phylogeny used for deriving the taxonomy are correct. Furthermore, it remains to be established whether the taxonomic ranks, which were assigned based on separately inferred phylogenies for bacteria and archaea, are comparable. Inferring the phylogeny of all archaea and bacteria with high accuracy remains a major challenge. For example, while models of protein evolution taking into account the compositional heterogeneities of amino acid sequences across branches and sites have already been developed (Blanchart and Artiçiot, 2008; Heaps et al., 2014), current implementations of these models are computationally intensive and can only be applied to relatively small sets of archaeal and bacterial taxa. We foresee that the need to organize the current and forthcoming wealth of genome sequences and the perspective of a coherent and powerfully predictive taxonomy together with the wish of elucidating major evolutionary transitions will stimulate novel developments in phylogenetics and phylogenomics. This may include the computationally efficient implementation of models of evolution that better capture the different constraints on gene and protein sequences (Koonin and Wolf, 2010) and the development of novel methodologies that integrate the horizontal and vertical components of genome evolution and therefore provide a comprehensive representation of the relatedness of bacterial and archaeal genomes simultaneously (Dagan, 2011; Chan et al., 2013). We also hope that this will facilitate agreements upon a conceptual framework for the standardized ranking of taxa using genome sequence data (Konstantinidis et al., 2017).

These approaches will help to represent and organize the diversity of archaeal and bacterial taxa based on the genetic variation within a small subset of genes shared by these organisms (Fig. 1). However, defining biologically and ecologically cohesive units will ultimately require the complementary use of bottom-up approaches to the definition of minimum units of diversity (Achtman and Wagner, 2008). For instance, various studies suggest that a considerable amount of biological and ecological variation almost invariably exists within any taxon (Cordero and Polz, 2014; Larkin and Martiny, 2017). Population genomics has emerged more than a decade ago as a powerful approach to identify genetically cohesive units within natural populations but also to investigate diversification and speciation events as well as study the genetic basis of adaptation (Luikart et al., 2003; Shapiro and Polz, 2014). The general approach of population genomics to the ordering of microbial genetic diversity consists in comparing the genomes of closely related organisms (i.e. having identical or near identical 16S rRNA genes) derived from one or several populations followed by the clustering of the genomes based on single nucleotide polymorphisms within shared genes (Shapiro and Polz, 2014). This forms the basis for further
principles of microbial genome evolution and speciation. And, while a detailed description of these methods is beyond the scope of this crystal ball, we believe that a holistic understanding of microbial diversity will also require the integration of approaches aiming at a systematic and high-throughput phenotyping of both archaea and bacteria as envisioned within the phenomics discipline (Houle et al., 2010). Finally, we believe that substantial progress will rely on the collective effort of a large number of networked research teams and the development of interoperable standards for collecting, analyzing and reporting data as for instance applied for the Earth microbiome project (Thompson et al., 2017).

The subsequent integration of knowledge from these complementary approaches will eventually bring us a step closer to shed light on a central enigma in microbiology, i.e. the fundamental split between the primary domains of life – Bacteria and Archaea – and the apparently lower diversity of the latter. We feel privileged to live in this ‘era of big data’ and methodological breakthroughs and are hopeful, that the coming years will witness substantial progress towards a better understanding of the diversity, nature and evolutionary trajectories of archaea and bacteria.

**Originality – significance statement**

In this crystal ball article, we point out a number of questions that connect the diversity of archaea and bacteria to their ecology, evolution and emphasize processes that may be the cause of their varying degree of diversity. We discuss emerging methodological approaches that can help establish a systematic understanding of archaeal and bacterial diversity.

**Acknowledgments**

This work was supported by a grant of the Swedish Research Council (VR starting grant 2016-03559 to A.S.), the NWO-I foundation of the Netherlands Organization for Scientific Research (WISE fellowship to A.S.).

**References**


