Cell, Vol. 89, 991–994, June 27, 1997, Copyright ©1997 by Cell Press

Archaeal Genomics: An Overview

Minireview

Gary J. Olsen and Carl R. Woese
Department of Microbiology
University of Illinois
Urbana, Illinois 61801

A few years ago, before the first complete genome sequence of an organism was known, a friend and colleague of the authors made the outrageous prediction that by the turn of the century, 100 microbial genomes will have been sequenced. At the time, microbial genome sequencing (except for "the representative" prokaryote, Escherichia coli) was considered mere distraction from sequencing the important genomes, those of eukaryotes. However, in 1995, J. C. Venter and his associates stunned Biology by rapidly and efficiently sequencing the genome of Haemophilus influenzae (Fleischmann et al., 1995), using a simple shotgun sequencing strategy. Two more genomes followed in rapid succession (again from Venter and company), including the first archaeal genome sequence, Methanococcus jannaschii, which provided our first good look at the least understood and most enigmatic of the three primary lines of descent (Bult et al., 1996). Suddenly "prokaryotic" and eukaryotic biologists alike began to care whether their favorite genes were present in this group of exotic microorganisms; certain complex eukaryotic functions can be effectively studied in simpler archaeal systems, molecular structures can be inferred from thermostable archaeal proteins, and the functional essence of an enzyme or system can be revealed by a broader comparative analysis. Today, "one hundred genomes by the turn of the century" seems remarkably insightful.

The importance of microbial genomics is implicit in a universal phylogenetic tree (Figure 1). All of the planet's early evolutionary history and well over 90% of life's phylogenetic diversity lie in the microbial world, as does the bulk of its metabolic, molecular, and ecological diversity. Archaeal genomics is particularly productive because there is so much to be learned—about the Archaea themselves, thermophily, their relationship to the eukaryotic cell, the origin of the three primary lines of descent, and the nature of the most recent universal ancestor.

But the world is not so simple as Figure 1 might suggest. The cell we know is a highly integrated self-sustaining metabolic machine (a phenotype) controlled by an underlying genetic description (a genotype). The cell's precise and powerful information processing systems provide the defining link between the two. As modern cells emerged some billions of years ago, the informational systems that maintain and express the genome were themselves nascent; neither the emerging genotype nor its linkage to the corresponding phenotype could have been precise (Woese and Fox, 1977). Genomic studies reveal a profound evolutionary distinction between the informational and the metabolic facets of the cell: most informational aspects of Archaea are closely allied with those seen in Eucarya, whereas most metabolic aspects of Archaea more strongly resemble those seen in Bacteria. Understanding this difference is one of Biology's great challenges, but one for which there are currently more questions (and speculations) than answers.

In the following sections, we focus on the informational systems associated with the genome and its expression. In the final section, we return briefly to the broader picture.

The Genome

It is self-evident that the genome (its structure and organization) and the mechanism for replicating that genome coevolved. And it is becoming evident that the mechanism of genome replication seen in the Archaea and Eucarya is quite different from that seen in the Bacteria: the only recognizable DNA polymerase in the M. jannaschii genome is homologous to a eukaryal nuclear genome replication polymerase (delta), but these enzymes are unrelated to their functional counterpart in E. coli and other Bacteria, DNA polymerase III (Pol III). (E. coli DNA polymerase II [Pol II], which is a homolog of the archaeal/eukaryotic replication polymerase, is not used in bacterial genome replication, and is not present in many, probably most, Bacteria). Yet, the current picture of archaeal DNA replication is fragmentary. Many ostensibly necessary components have not been revealed by either bacterial or eukaryotic gene homologies. Also, the in vitro DNA synthesis rate of the commercial archaeal polymerases is remarkably slow. Does this

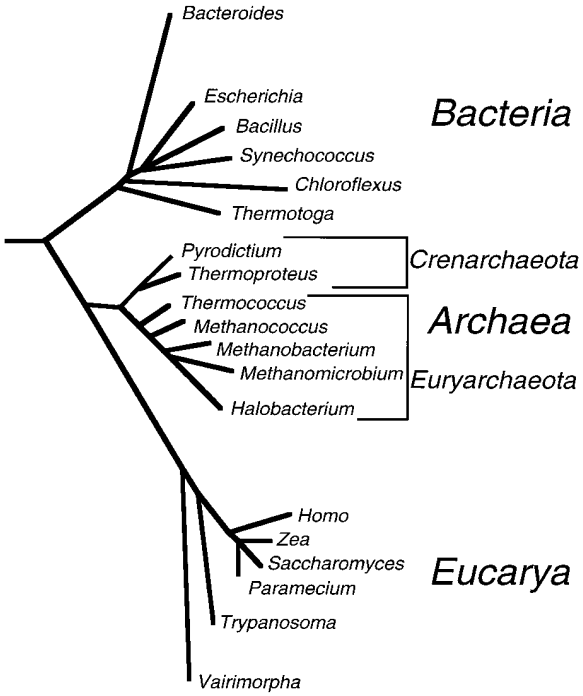

Figure 1. A Universal Phylogenetic Tree

A phylogenetic tree based on small subunit ribosomal RNA sequences. The tree has been rooted by analysis of duplications in protein sequences (Iwabe et al., 1989). (Adapted from Woese et al., 1990.)

reflect missing critical components in the in vitro system, or could it be that archaeal genome replication (like eukaryotic genome replication) involves multiple origins? Or, might it be that the archaeal replication polymerase has not yet been discovered?

Bacterial and archaeal genomes show operonal organization, something not characteristic of eukaryotes. Does this reflect a specific relationship between Archaea and Bacteria, or is it merely the ancestral condition, lost in eukaryotes? On a deeper level, what, if anything, does it tell us about the evolution of genomic organization? Despite impressive detailed parallels between certain operons in Bacteria and Archaea, operonal organization is evolutionarily fluid; reorganization, loss or gain of genes, and loss or creation of operons themselves are well documented. And, although some of the archaeal ribosomal protein genes (and others) are organized into "bacteria-like" operons, the corresponding amino acid sequences are most similar to those of their eukaryotic, not bacterial, counterparts. Ancestral or not, it is evident that operonal organization can be convergent.

Questions about genome structure and organization inevitably lead back to the most recent common ancestor (Universal Ancestor). Of the three main cellular information processing systems, genome replication, transcription, and translation, the first is the only one whose central component(s) are not universally conserved. This strongly implies that the mechanism for genome replication was the least developed of the three at the time of the Universal Ancestor. If central components of the machinery for handling the genome of the Universal Ancestor were nascent, this must also be true of the genome itself, although the evidence here is less direct. Perhaps the Universal Ancestor did not have a consolidated genome, but rather a disperse collection of "operons" (a situation analogous to that seen in the macronucleus of some modern ciliates). To accommodate a complete ribosomal RNA gene, or the largest subunit of RNA polymerase, the size of such "minichromosomes" would need to be several thousand base pairs; thus one could also envision operons of comparable size, for example groupings of ribosomal protein genes.

At the extreme, one could question whether the Universal Ancestor had a DNA-based genome at all. The case has been made that DNA is the most dispensable of the Central Dogma's molecular trinity. The present-day universality of DNA topoisomerases and DNA-dependent RNA polymerases (in self-replicating systems) strongly suggests the presence of DNA in the Universal Ancestor (for more details, see minireview by Edgell and Doolittle, 1997 [this issue of *Cell*]). RNA is the most popular alternative to an ancestral DNA-based genome, but RNA suffers from chemical lability. If the genome were highly segmented and redundant, this might present less of a problem. Alternatively, a genome composed of 2′-O-methyl RNA (a modified nucleotide of universal distribution) would be more resistant to both chemical hydrolysis and depurination, and would also have increased resistance to denaturation—perhaps too much so. Even if the Universal Ancestor had a DNA-based genome, these conjectures would still apply to some earlier stage in the evolution of life.

Transcription

While the core components of the transcription apparatus (the three largest subunits of the RNA polymerase) are universal in distribution, the archaeal and eukaryotic versions are decidedly more similar to one another than either is to their bacterial counterpart. The archaeal holoenzyme contains a number of additional subunits that have counterparts only among eukaryotes (Langer et al., 1995; Bult et al., 1996). The mechanism for transcription initiation in Archaea looks like a simple form of that seen in eukaryotes; neither of which resemble the bacterial mechanism (see minireview by Reeve et al., 1997 [this issue of *Cell*]).

The two general archaeal and eukaryotic transcription initiation factors, TATA-binding protein (TBP) and the family that includes transcription factors IIB and IIIB (generically TFB), are of particular interest. The main section of each molecule comprises a tandem repeat. Although each molecule has diverged in sequence significantly over the phylogenetic course, it is not possible to ascertain with certainty whether the first (or second) repeat in archaeal TBP is the ortholog of the first or second repeat in the eukaryotic case. This lack of differentiation of the halves of the TBP repeat suggests that the evolutionary split between the archaeal and eukaryal lineages occurred relatively soon after the gene duplication event that gave rise to TBP's repeat structure. Interestingly, the presumed progenitor of TBP, a homodimer of half-size molecules, would be symmetrical in structure and hence incapable of defining the direction of transcription. Was it defined instead by the ancestral TFB?

TFB, unlike TBP, shows some evidence for the orthology of the archaeal repeats with their eukaryotic counterparts. Moreover, the archaeal and eukaryotic TFBs share a homologous N-terminal domain, indicating the protein was essentially in its present form at the time of the archaeal/eukaryal split.

The above situation has several evolutionary implications. First, the universal RNA polymerase subunits are most consistent with a DNA genome at the time of the Universal Ancestor. Second, the lack of a universal transcription initiation mechanism suggests a lack of initiation factors as we know them—seductively compatible with a world in which minichromosomes were somehow transcribed from one end to the other. Transcriptional regulation of gene expression would, of course, be problematic in such a world. The specific commonality between archaeal and eukaryotic versions of transcription predicts the two shared an important evolutionary history subsequent to the Universal Ancestral stage.

RNA Splicing

The history of RNA splicing and its role in early evolution is frequently debated, but the archaeal data are unlikely to resolve the issue. Several distinct types of splicing must be considered (see minireview by Belfort and Weiner, 1997 [this issue of *Cell*]). Archaeal genomes have not provided direct evidence of early spliceosomal introns; none of the spliceosome components are evident. Nor have group I or group II self-splicing introns been found—a surprise given their presence in both Bacteria and Eucarya. The more interesting story lies in the splicing of archaeal tRNAs and rRNAs by mechanisms related to those used for eukaryotic tRNA intron

removal. In spite of differences in the details of the reactions, the archaeal splicing protein is a distant relative of two of the proteins in the corresponding yeast reaction. As with many other components of the information systems, there is no known bacterial homolog, hence no evidence that it was present in the Universal Ancestor.

Translation

Although not finalized, translation was a highly developed process at the stage of the Universal Ancestor. The rRNAs and most of the ribosomal proteins, the tRNAs and their charging enzymes, and the major elongation factors are universally distributed. An important unanswered question is whether translation at this early stage was sufficiently developed to have accuracy comparable to that seen today—a necessity for evolving the large proteins characteristic of modern cells (Woese and Fox, 1977). Much of the answer turns on the functional significance of those translation components that are not universally distributed. For example, almost all the translation initiation factors and a fair number of ribosomal proteins qualitatively distinguish the Archaea and Eucarya from the Bacteria. However, if we are correct in our supposition that the largest subunits of RNA polymerase were present (and intact) in the Universal Ancestor, then there must have been sufficient accuracy to make large proteins at that stage. The issue of intactness is important for, in the Archaea, one or both of the subunits exist in two parts.

The fact that asparagine and glutamine do not have aminoacyl-tRNA synthetases in the Archaea (see minireview by Dennis, 1997), and that, more surprisingly, two others, those for lysine and cysteine, could not be identified by normal database searching methods (Bult et al., 1996), suggests more complexity to the evolution of these enzymes than conventional wisdom now allows. It also raises the question as to the state of tRNA charging at the Universal Ancestor stage. How could major variation in these essential enzymes evolve when the genetic code was already solidified in detail?

Protein Splicing

Although introns have not been found in any protein coding genes of Archaea, inteins (self-splicing protein sequences) have. These little understood entities pose intriguing problems in their distribution, evolution, and in vivo excision. Why, for example, if they occur in all three primary lineages and contain homing endonucleases, which provide the potential for rapid horizontal transfer (Perler et al., 1997), are they not uniformly distributed, even within a genus? The answers to, and even the framing of, these key questions will have to await a much larger collection of examples. Yet the fact that the M. jannaschii genome contains 18 inteins (Bult et al., 1996), with as many as three in one gene, replication factor C, is an indicator of the interest that these enigmatic entities will generate.

Final Thoughts

A striking feature of the M. jannaschii genome was the high fraction of genes, about 50%, that match nothing in the sequence databases. Although some of these are likely to remain unique to M. jannaschii, others should turn out to be characteristic of the Archaea as a whole or various major subgroups thereof. Currently, a number of examples are known of genes that span the phylogenetic breadth of the Archaea, but have no counterparts in the Bacteria or eukaryotes. When a full crenarchaeal genome sequence becomes available, many more such cases are expected. The large number of genes confined to Archaea attests to the group's uniqueness and to the challenge that archaeal metabolism presents to the biochemist.

In the broader perspective, group-specific genes point the way to taxonomy's future. The day is ending when relationships among major taxa can be based on the phylogeny inferred for one or a few molecules (a method that all too often focuses on arcane debates over details, while failing to reinforce the remarkably broad areas of consensus). A systematics based on molecular phylogenies must cope with the facts that not all genes in an organism's genome share the same evolutionary history and that not all genes can be effectively used for phylogenetic analysis—some trees differ because the gene histories are different, others because they are merely uncertain. With full genomes, many organismal groups will be clearly defined in terms of large sets of shared unique genetic traits; organisms belonging to a taxon so defined would each possess a substantial fraction of the genes in the set, while those not within the group would possess few if any of them. In this respect, we think it improbable that the numerous genes shared by Archaea and Eucarya, but absent in Bacteria, were all present in the Universal Ancestor and subsequently lost in the ancestral bacterial stem (sometimes to be replaced by a separately invented functional analog).

In the present context we have made no attempt to summarize or explain the similarities of the metabolic genes that seem to specifically unite Archaea and Bacteria. In some instances there is no recognizable homolog in the Eucarya, while in others there is simply a greater similarity of the archaeal-bacterial pair. Numerically, these Bacteria-related genes constitute a greater fraction of the M. jannaschii genome than do the genes (including most of those for replication, transcription and translation) that appear to unite the Archaea and the Eucarya. We have suggested that the former could result from the degeneration of metabolic capacity in the eukaryotic lineage (Olsen and Woese, 1996). It is important to ask whether this interpretation can be reconciled with our argument that the bacterial information processing system is not a simplified version of a more complex ancestral form (as seen in the Archaea and Eucarya). Most of our reasons for accepting a eukaryotic metabolic simplification, but not a bacterial informational streamlining, are that the former presents a relatively uniform, consistent picture, while the latter comprises numerous, often conflicting, vignettes. A broader, more representative sampling of genomes will be needed to resolve this issue.

The evolutionary picture emerging from genomics, as we retrace our ancestry further and further back into the past, is an unsettling, yet entrancing one. Looking back toward the Universal Ancestor, the simple world of distinct organismal lineages and a robust, well-defined connection between genotype and phenotype—foundation stones of biology—loses its substance, dissolving in the turbulent evolutionary dynamic that shaped genotype, phenotype, and their connection. From the

information processing systems seems to emerge a consistent, informative, and satisfying picture of cellular evolution and that of the primary lines of descent. Although far from chaotic, metabolism seems an evolutionary morass—a sure sign of conceptual revolution and enlightenment to come. In this history-laden world of microbial genomics, the evolutionary perspective necessarily changes from one of teasing ourselves with speculation to that of teasing out the grandest history of all.

Selected Reading

Belfort, M., and Weiner, A. (1997). Cell, this issue, *89*, 1003–1006.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., et al. (1996). Science *273*, 1058–1073.

Dennis, P.P. (1997). Cell, this issue, *89*, 1007–1010.

Edgell, D.R., and Doolittle, W.F. (1997). Cell, this issue, *89*, 995–998.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., et al. (1995). Science *269*, 496–512.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. (1989). Proc. Natl. Acad. Sci. USA *86*, 9355–9359.

Langer, D., Hain, J., Thuriaux, P., and Zillig, W. (1995). Proc. Natl. Acad. Sci. USA *92*, 5768–5772.

Olsen, G.J., and Woese, C.R. (1996). Trends Genet. *12*, 377–379.

Perler, F.B., Olsen, G.J., and Adam, E. (1997). Nucleic Acids Res. *25*, 1087–1093.

Reeve, J.N., Sandman, K., and Daniels, C.J. (1997). Cell, this issue, *89*, 999–1002.

Woese, C.R., and Fox, G.E. (1977). J. Mol. Evol. *10*, 1–6.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Proc. Natl. Acad. Sci. USA *87*, 4576–4579.