Opinion

Comparative genomics of archaea: how much have we learned in six years, and what's next?

Kira S Makarova and Eugene V Koonin

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Correspondence: Eugene V Koonin. E-mail: koonin@ncbi.nlm.nih.gov

Published: 16 July 2003

Genome ***Biology*** 2003, **4:**115

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/8/115

Abstract

Archaea comprise one of the three distinct domains of life (with bacteria and eukaryotes). With 16 complete archaeal genomes sequenced to date, comparative genomics has revealed a conserved core of 313 genes that are represented in all sequenced archaeal genomes, plus a variable 'shell' that is prone to lineage-specific gene loss and horizontal gene exchange. The majority of archaeal genes have not been experimentally characterized, but novel functional pathways have been predicted.

"A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (i) the eubacteria, comprising all typical bacteria; (ii) the archaebacteria, containing methanogenic bacteria; and (iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells."

CR Woese and GE Fox, 1977 [1]

Archaea before and after genomes

The quotation above neatly summarizes what is arguably one of the most important scientific discoveries of the twentieth century (rather remarkably, this quote is the entire abstract of Woese and Fox's groundbreaking article [1]). So profound are its implications that the debate rages to this day: did Carl Woese and George Fox really discover a new domain of life, which is equal in status to bacteria and eukaryotes [2,3], or is it 'merely' an unusual branch of bacteria [4-7]? This debate is reflected even in the different names that, 25 years after their description as a distinct, third line of the evolution of life, are still applied to this group of organisms: on the one hand, archaea, in adherence with the three-domain interpretation, and on the other archaeabacteria, emphasizing the purported affinity with bacteria. Of course, Woese and Fox did not actually discover these unusual organisms; some of the would-be archaea have been known for decades and their unusual properties, such as extreme halophilic and extreme thermophilic phenotypes, have been described in considerable detail (see, for example, [8-10]). The revolutionary aspect of Woese and Fox's work was subtler and more profound: by comparing certain parts of the genomic sequences of various organisms, they came up with a three-domain classification of life, in which a group of prokaryotes they designated archaebacteria has been accorded the status of a distinct domain (subsequently renamed archaea, to emphasize the fundamental separation from other domains), on an equal footing with bacteria and eukaryotes. Numerous microbiologists had seen archaea before, but without Woese and Fox's foray into genome analysis no-one recognized these organisms for what they really were. Their way of comparing genome sequences was, by today's standards, extremely crude, as they analyzed not even sequences but oligonucleotide catalogues of rRNA genes. It is all the more astounding that the principal conclusion achieved with this 'primitive' approach stands to this day, 25 years and 16 complete (and several more nearly complete) archaeal genome sequences later (Table 1).

In the years following Woese and Fox's breakthrough [1], many unique features of archaea have become apparent. To

Table 1

Completely sequenced archaeal genomes

Species	Abbreviation	Optimal growth temperature (°C)	Lifestyle and other features	Number of proteins*	Number (%) proteins in COGs	Date of genome release	Reference
Euryarchaeota							
Archaeoglobus fulgidus DSM	**Afu**	83	Anaerobic, sulfate-reducing chemolito- or chemorgano-autotroph, motile	2,420	1,953 (81%)	1997	[124]
Halobacterium sp. NRC-1	**Hsp**	37	Aerobic chemorganotroph, obligate halophile, with a cell envelope; motile; two extrachromosomal elements	2,622	1,809 (69%)	2000	[125]
Methanocaldococcus jannaschii	**Mja**	85	Chemolitoautotroph, strict anaerobe, methanogen, motile; two extrachromosomal elements	1,758	1,448 (82%)	1996	[27]
Methanopyrus kandleri AV19	**Mka**	110	Chemolitoautotroph, strict anaerobe, methanogen, with high cellular salt concentration	1,691	1,253 (74%)	2002	[45]
Methanosarcina acetivorans C2A	**Mac**	37	Chemolitoautotroph, anaerobe possibly capable of aerobic growth; nitrogen-fixing, versatile methanogen; motile, and able to form multicellular structures	4,540	3,142 (69%)	2002	[55]
Methanosarcina mazei Goe1	**Mma**	37	As for **Mac**	3,371	N/A	2002	[54]
Methanothermobacter thermoautotrophicus delta H	**Mth**	65	Chemolitoautotroph, strict anaerobe, nitrogen-fixing, methanogen	1,873	1,500 (80%)	1997	[126]
Pyrococcus horikoshii	**Pho**	96	Anaerobic heterotroph, sulfur enhances growth; motile	1,801	1,425 (79%)	1998	[127]
Pyrococcus abyssi	**Pab**	96	As for **Pho**	1,769	1,506 (85%)	2001	[128]
Pyrococcus furiosus DSM 3638	**Pfu**	96	As for **Pho**	2,065	N/A	2001	[129]
Thermoplasma acidophilum	**Tac**	59	Facultative anaerobe, chemorganotroph, thermoacidophilic, anaerobically able to metabolize sulfur; motile, with a plasma membrane	1,482	1,261 (85%)	2000	[96]
Thermoplasma volcanium	**Tvo**	60	As for **Tac**	1,499	1,277 (85%)	2000	[130]
Crenarchaeota							
Pyrobaculum aerophilum	**Pae**	100	Facultative nitrate-reducing anaerobe	1,840	1,236 (67%)	2002	[131]
Aeropyrum pernix	**Ape**	90	Aerobic chemorganotroph; sulfur enhances growth	2,605	1,529 (59%)	1999	[132]
Sulfolobus solfataricus	**Sso**	80	Aerobe metabolizing sulfur; thermo-acidophilic chemorganotroph; motile	2,977	2,207 (74%)	2001	[97]
Sulfolobus tokodaii	**Sto**	80	As for **Sso**	2,826	N/A	2001	[133]

*According to the original genome annotation.

begin with, many of these organisms thrive under conditions that, by the usual standards of biology, seem unimaginable, such as in the water in the vicinity of the hydrothermal vents called 'black smokers' heated to over-boiling temperatures and saturated with hydrogen sulfide, or in extreme salinity [11-13]. In the most extreme hyperthermophilic habitats, archaea are, in fact, the only detectable life forms. In more moderate environments, archaea coexist with bacteria and eukaryotes, and their ecological importance is being increasingly recognized [14]. The first molecular biological studies showed that archaea are highly unusual and clearly distinct from bacteria at the molecular level. In particular, the structure of the membrane glycerolipids in archaea is different from that of bacterial and eukaryal cells, and archaea do not contain murein, the predominant component of bacterial cell walls [15,16].

But the most striking differences between archaea and bacteria are seen in the organization of their information-processing systems. The structures of ribosomes and chromatin, the presence of histones, and sequence similarity between proteins involved in translation, transcription, replication and DNA repair all point to a closer relationship between archaea and eukaryotes than between either of these and bacteria [17-21]. Moreover, the key components of the DNA replication machinery - such as the polymerases involved in elongation and initiation and the replicative helicases - are not homologous, or at least not orthologous, in archaea and eukaryotes on the one hand, and bacteria on the other [17,22]. This observation led to the hypothesis that replication of double-stranded DNA as the principal form of replication of the genetic material was 'invented' twice, independently: once in bacteria and once in the ancestor of archaea and eukaryotes [22,23]. In contrast many - although not all - of the metabolic pathways of archaea more closely resemble their bacterial rather than eukaryotic counterparts [24-26]. These studies support the status of archaea as a distinct domain of life with specific connections to eukaryotes, and emphasize the unusual and unique nature of archaeal genomes.

The new age of archaea began in 1996 with the whole-genome shotgun sequencing of the first archaeal genome, that of *Methanococcus* (now *Methanocaldococcus*) *jannaschii* [27]. The *Methanococcus* 'genomescape' at first looked largely mysterious, with clear functional assignments produced for only 38% of the genes [27]. A more detailed computational analysis that pushed the methodology available at the time to its limits yielded general functional predictions for up to 70% of the genes, showing that a solid connection between the genomes of archaea and those of other, better known forms of life did exist [24]. Nevertheless, the fact remained that, more than anything, the first sequenced archaeal genome revealed the depth of our ignorance of the biology of this remarkable group of organisms. Subsequent genome sequencing, while certainly less extensive than the devoted 'archaeologists' would wish, produced a rich sampling of genomes of taxonomically diverse archaea (Table 1). This set of completely sequenced genomes includes multiple representatives of the two major divisions of the archaea established by phylogenetic analysis of rRNA, namely the Euryarchaeota and the Crenarchaeota [3], as well as the principal ecological types of archaea, such as hyperthermophiles, moderate thermophiles, and mesophiles, as well as halophiles and methanogens; autotrophic and heterotrophic forms, and anaerobes and aerobes are also represented by multiple species (Table 1).

Some potentially important branches of archaea are still missing from sequence databases, however, such as the mysterious Korachaeota, which might have branched off the trunk of the phylogenetic tree prior to the divergence of the remainder of the archaea [28], and the equally intriguing Nanoarchaea that so far seem to have the smallest genomes of all known cellular life forms [29,30]. These lacunae notwithstanding, the available sampling of archaeal genomes is substantial and is complemented by an even greater diversity of bacterial and eukaryotic genomes that are available for comparative analysis. This article critically assesses the contribution of comparative genomics to our understanding of the functional systems of archaeal cells and their evolution. We pose the following question: what have we learned from comparisons of archaeal genomes that could not easily have been learned by other, more traditional approaches? We suggest some tentative answers, as we see them. What follows is a viewpoint from behind a computer terminal; we realize that, from the experimenter's bench, the perspective might be somewhat different.

Evolutionary archaeogenomics

From the beginning of comparative genomics, it has been obvious that genome comparisons will yield valuable functional and evolutionary information only within a framework of the rational classification of genes and proteins. In our view, perhaps the most natural form of such a classification is a system of orthologous gene sets, which allows a researcher to analyze the evolutionary fate of each individual gene [31]. Orthologs are homologous genes that evolved from a single ancestral gene in the last common ancestor of the compared genomes, whereas paralogs are genes related via duplication within a genome [32-34]. When duplication(s) succeeds speciation, a family of paralogs in one species should be considered orthologous to the corresponding family in the other species [34]. Insomuch as orthologous relationships are correctly defined, phyletic (or phylogenetic) patterns of orthologous gene sets help in the prediction of gene functions and provide clues to the prevailing trends in genome evolution (a phyletic pattern is defined, simply, as the pattern of representation of genomes in each orthologous set) [26,31,35,36]. These phyletic patterns are captured in the database of Clusters of Orthologous Groups of proteins (COGs) [37], and here we use COGs for a

Table 2

The top 15 phyletic patterns in proteins from archaea

Pattern* AHMMMMTTPPPSA fbatjkavhaasp uschaacoobeoe	Number of COGs (and of the complementary pattern, CP)	Comments and examples
+++++++++++++	**313** (0)	Archaeal core, including 200 COGs present in both **B**† and **E**, 34 present in at least one **B**, 63 present in at least one **E**, 16 unique for **A** **CP:** Only COG0564, pseudouridylate synthase, 23S RNA-specific pseudouridylate synthase present in all **E** (in which it has an apparently mitochondrial origin) and **B**, but not in **A**. In all **A** another specific pseudouridylate synthase is present (COG1258)
−−+−−−−−−−−−−	**163** (3)	This pattern reflects a large number of genes acquired via HGT† in **Mac** (see [55]), including F_0F_1-type ATP synthase and NADH:ubiquinone oxidoreductase, and a specific signal transduction system based on several apoptosis-related domains **CP:** The small number of such COGs indicates that the archaeal core is almost fully conserved in **Mac**
−+−−−−−−−−−−−	**79** (14)	This pattern reflects a substantial amount of HGT in **Hsp**; see [125]
+−++++−−−−−−−	**47** (7)	This pattern consists of COGs including four methanogens and **Afu**; these organisms specifically share several metabolic pathways (see [45]). The set includes subunits of coenzyme F420-reducing hydrogenase, formylmethanofuran dehydrogenase, CO dehydrogenase/acetyl-CoA synthase and other enzymes of energy metabolism. These might have originally evolved in methanogens and subsequently transferred to **Afu** **CP:** Sugar ABC transporter and some fatty acid biosynthesis enzymes are missing from methanogens and **Afu**
−−++++−−−−−−−	**40** (2)	This pattern is specific for four methanogens, including unique pathways for coenzyme M biosynthesis and reduction and 14 uncharacterized proteins, many of which are likely to be unique enzymes involved in biosynthesis of other specific coenzymes and their utilization **CP:** COG2096, cob(I)alamin adenosyltransferase and COG1058, predicted nucleotide-utilizing enzyme related to molybdopterin-biosynthesis enzyme MoeA, for which functional substitutes remain to be identified
−−−+++−−−−−−−	**33** (16)	A pattern specific for thermophilic methanogens (**Mth**, **Mja** and **Mka**), comprising mostly uncharacterized COGs, it includes a specific membrane complex EhaA-EhaP (approximately 18 components) involved in hydrogen production and possibly electron transfer [45,134] **CP:** Specific gene loss: peptide ABC-type transporter, NADH:ubiquinone oxidoreductase, malic enzyme (COG0281), and cysteinyl-tRNA synthetase (COG0215; see text)
−−−−−−−−−−−+−	**28** (6)	This pattern reflects a substantial amount of HGT in **Sso**, including several enzymes of carbohydrate metabolism (beta-glucosidase, alpha-L-fucosidase, and malto-oligosyl trehalose synthase) [97]
+−−−−−−−−−−−−	**27** (1)	This reflects a substantial amount of HGT in **Afu** **CP:** COG0449, glucosamine 6-phosphate synthetase, which catalyzes the first step in hexosamine metabolism. A functional substitute remains to be identified
−++−−−−−−−−−−	**25** (4)	A pattern specific for two mesophilic archaea, probably resulting from independent HGT
−−−−+−−−−−−−−	**23** (7)	This pattern includes genes that might have been acquired via HGT in **Mja**, in particular three enzymes of biotin biosynthesis: pimeloyl-CoA synthetase (COG1424), dethiobiotin synthetase (COG0132), and adenosylmethionine-8-amino-7-oxononanoate aminotransferase (COG0161)
−−−−−−−−−−+++	**21** (13)	A crenarchaea-specific pattern, including 11 COGs that do not have orthologs outside this lineage. Among genes shared with bacteria but not euryarchaeota are three subunits of aerobic-type CO dehydrogenase and CO dehydrogenase maturation factor. Genes specifically shared with eukaryotes are three ribosomal proteins (S30, S25 and L13E) **CP:** Euryarchaea-specific pattern, including two subunits of archaeal DNA polymerase II and ERCC4-like helicase, division GTPase FtsZ (COG0206) and ATP-dependent protease LonB (COG1067) plus six COGs that do not have orthologs outside this lineage
+−+−−−−−−−−−−	**20** (0)	Apparent independent HGT to **Mac** and **Afu**
++++++−−+++++	**19** (16)	Apparent specific gene loss in the *Thermoplasma* lineage: two subunits of topoisomerase VI (COG1389, 1697), adenylate cyclase of class 2 (COG1437), and predicted exosome subunits (COG1325, COG1931). **CP:** genes apparently acquired via HGT in *Thermoplasma*, including bacterial nucleoid DNA-binding protein HU (COG0776). See also [96]
++++++++++++−	**18** (6)	Apparent gene loss in **Ape**, including 9 enzymes of purine biosynthesis [135]
−−−−−−−−++−−−	**17** (11)	Apparent HGT in *Pyrococci*. Includes two subunits of allophanate hydrolase (COG1984, 2049), two enzymes of carbohydrate metabolism, β-galactosidase (COG1874) and endoglucanase (COG2730) **CP:** Specific gene loss in the *Pyrococcus* lineage includes five enzymes of heme biosynthesis

*The pattern of appearance within the 13 sequenced archaeal species currently available in the COG database. Species abbreviations are as given in Table 1 and are written vertically. †Abbreviations: **A**, archaea; **B**, bacteria; **E**, eukaryotes; **CP**, complementary pattern; HGT, horizontal gene transfer.

systematic survey of archaeal genomes (most of the phyletic pattern analyses can be done directly on the COG website by using the phyletic pattern search tool [38]).

The most common phyletic patterns found in archaea are shown in Table 2. Not unpredictably, the top pattern consists of the 313 COGs that are represented in all archaeal genomes sequenced so far. What is more remarkable is that this apparent conserved core of archaeal genomes has undergone only limited shrinkage since the time it was first defined by comparative analysis of four archaeal genomes [39] (Figure 1). Extrapolating from the effect (or rather the near lack thereof) of the latest additions to the collection of archaeal genomes on the size of the conserved core of archaeal genes, we are compelled to conclude that around 300 genes are shared by all archaea, encode essential functions and have not been subject to non-orthologous gene displacement during archaeal evolution (non-orthologous gene displacement is a widespread phenomenon whereby a gene responsible for an essential function is displaced by an unrelated or distantly related gene responsible for the same function [40]).

Of the COGs represented in all archaea, 16 so far have no members from other domains of life and comprise a unique archaeal genomic signature, whereas 61 are exclusively archaeo-eukaryotic. The majority of the pan-archaeal genes are known to be involved in, or are implicated in, information processing, particularly translation and RNA modification (Figure 2). Strikingly, among the 61 COGs that are uniquely shared by archaea and eukaryotes, only two do not, technically, belong to the information-processing machinery (COG1936, a nucleotide kinase, and COG3642, a protein kinase typically fused to a metalloprotease domain); the 10 uncharacterized COGs in this category consist of proteins whose predicted biochemical activity (GTPase, methyltransferase or RNA-binding protein) suggests a role in translation or RNA modification.

Thus, phyletic pattern analysis strongly supports the identity of archaea as a distinct group of organisms with a stable, conserved core of genes that primarily encodes proteins involved in the replication and expression of the genome. Furthermore, there is clearly a subset of genes, again primarily associated with information processing, that is shared by archaea and eukaryotes, to the exclusion of bacteria; this is compatible with the archaeo-eukaryotic affinity suggested by phylogenetic analyses of rRNA and proteins involved in translation, transcription and replication. The fact that this archaeo-eukaryotic component is quantitatively small, however, shows that the process of evolution has been more complex than simple vertical inheritance and has involved extensive horizontal gene transfer (HGT) between archaea and bacteria, at least outside the core gene set [24,25,41]. An intensely mixing pool of genes coding for metabolic enzymes, structural components of the cell and other proteins outside the central information-processing machinery might have existed after the divergence of bacteria and archaea but prior to the separation of the major archaeal and bacterial lineages.

More recent HGT, which has emerged as a major aspect of prokaryotic evolution in general [26,42-44], was apparently prominent in all archaea, although gene exchange with bacteria seems to have been much less extensive in hyperthermophiles than in mesophiles such as *Methanosarcina* or even *Halobacterium* [44,45]. Apparent preferential HGT has been noticed between archaea and hyperthermophilic

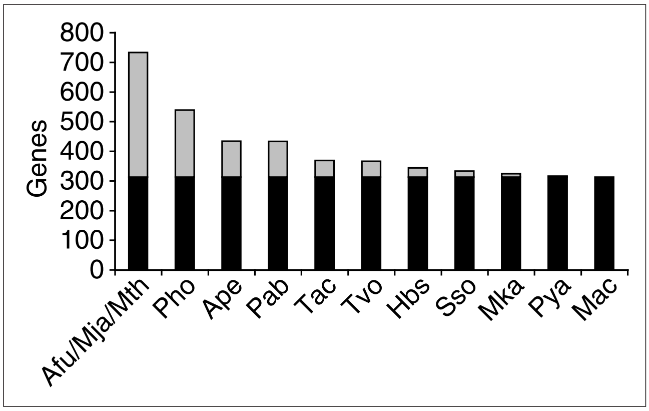

Figure 1
The archaeal gene core: changes resulting from the appearance of new genome sequences. Black bars indicate the current set of pan-archaeal genes (313 COGs); gray indicates COGs that are not part of the current pan-archaeal core but are seen to be conserved after the addition of the given genome sequence. The genomes are listed from left to right in chronological order of release of the complete sequence; species name abbreviations are as in Table 1.

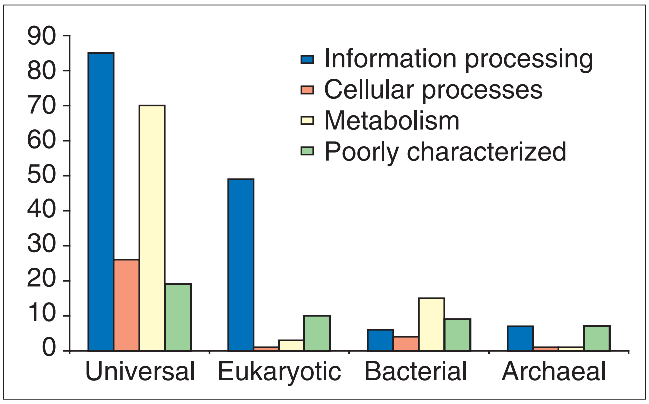

Figure 2
Functional breakdown of genes within the conserved archaeal core. 'Universal' indicates genes with orthologs in both bacteria and eukaryotes; 'eukaryotic', genes with orthologs only in eukaryotes; 'bacterial', genes with orthologs only in bacteria; 'archaeal', genes without non-archaeal orthologs. The data on orthology and functional classification are derived from the COGs.

bacteria, such as *Aquifex* and *Thermotoga*; when compared to bacterial mesophiles these bacteria have many more proteins with greater similarity to archaeal than to bacterial homologs [46,47]. With HGT, or more precisely the pivotal role of HGT in evolution, remaining a controversial subject [48], this conclusion has been disputed on the grounds that *Aquifex* and *Thermotoga* might be early-branching bacteria retaining ancestral features in many protein sequences [49]. But this argument seems untenable simply because of the obvious split of the gene complements of these bacteria into 'garden variety' bacterial genes and 'archaeal' genes [50]. The reality of horizontal gene flow from archaea to thermophilic bacteria becomes even more tangible upon examination of the proteins encoded in the genome of *Thermoanaerobacter tengcongensis* [51,52], which contains many more 'archaeal' genes than appear in other bacteria of the *Bacillus-Clostridium* group and to which the early-branching argument would not apply.

Although archaeal hyperthemophiles do not appear to have many genes acquired via HGT from bacteria, at least after the divergence of the archaeal lineages, horizontal gene exchange between archaea themselves might have been extensive. Strikingly, even within the conserved core of archaeal genes, major diversity of phylogenetic tree topologies has been observed ([53] and Y.I. Wolf and E.V.K., unpublished observations). As noted by Nesbo and coworkers [53], "the notion that there is a core of non-transferable genes...has not been proven and may be unprovable". These findings do not invalidate the notion of a core of indispensable genes that are conserved across archaea but suggest a wide spread of xenologous gene displacement, whereby an essential gene is displaced by an ortholog from a distant lineage, typically via an intermediate stage of redundancy [44].

Other phyletic patterns that are common among archaea seem primarily to reflect HGT or gene loss prevalent in individual archaeal lineages (Table 2). Thus, *Methanosarcina*, a mesophile with by far the largest genome among the sequenced archaeal genomes, is represented in numerous COGs that have no other archaeal members but are present in various groups of bacteria. This organism, which coexists with a diverse bacterial biota, appears to be a veritable sink for horizontally acquired bacterial genes [54,55]. Similar, if less dramatic, evidence of apparent horizontal gene transfer was seen in *Halobacterium*, *Sulfolobus*, and *A. fulgidus* (Table 2; [44]). Of further note are the patterns of genes that are ubiquitous in one of the major branches of archaea, namely Euryarchaeota or Crenarchaeota, but are missing from the other branch. While quantitatively small, the set of euryarchaea-specific genes includes those for several crucial cellular functions, such as the two subunits of DNA polymerase II and the FtsZ GTPase that is required for cell division in Euryarchaeota and bacteria but missing from Crenarchaeota and eukaryotes.

Phyletic patterns can be used for interesting and potentially useful forays into functional genomics - more specifically for the identification of the genomic cognates of particular phenotypes. The most dramatic phenotypic characteristic of archaea is hyperthermophily, and attempts have been made to use the phyletic pattern approach to identify a gene set typical of hyperthermophiles. Strikingly, there is only one COG that is represented in all hyperthermophiles (both bacteria and archaea) but not in any other sequenced genomes, the reverse gyrase ([56]; COG1110). Reverse gyrase consists of a topoisomerase and a helicase domain and functions to introduce negative supercoiling into DNA; this activity is apparently required for DNA replication and gene expression at extreme high temperatures [57]. But 'clean' phyletic patterns that have an unequivocal association with a given phenotype are an exception rather than the rule, so flexible pattern selection approaches have been employed. Our recent analysis of phyletic patterns enriched in archaeal and bacterial hyperthermophiles yielded around 60 COGs potentially related to this phenotype [58]. About one quarter of these COGs encode parts of a predicted DNA repair system that is largely characteristic of thermophiles ([59] and see below). The remaining COGs in this set suggest the existence of a transcriptional regulator that might be involved in adaptation to hyperthermal environments, and a distinct class of enzymes, the S-adenosyl methionine (SAM)-radical enzymes, whose chemistry is likely to be particularly efficient under these conditions [58]. Finally, a substantial number of COGs are specific for methanogens or shared by the methanogens and *A. fulgidus* (Table 2 and [45]). Many of these include known or predicted enzymes involved in methanogenesis and associated metabolic pathways [45,60]; others remain to be characterized and are likely to encode additional components of these pathways.

Further functional and evolutionary information can be extracted from complementary phyletic patterns, which are the signature of non-orthologous gene displacement [26,61]. Although the complementarity is, most often, only partially due to redundancy in some species, several cases of near-perfect complementarity among archaea are notable, such as the two classes of unrelated lysyl-tRNA synthetases [62,63], and two forms of thymidylate synthase that are also unrelated to each other [61,64]. Below, when discussing functional genomics of the archaea, we return to the use of conserved and complementary phyletic patterns for functional prediction.

Genome-wide phylogeny of archaea and reconstruction of archaeal ancestors

Comparative genomics nowadays includes a new variety of phylogenetic analysis, which for short has been dubbed genome-tree construction. Under this approach, phylogenetic trees are built not from the sequences of a single gene (such as an rRNA) but from concatenated sequences of multiple genes (proteins), from other, integral measures of the evolutionary distance between genomes (for example, the median of the distribution of evolutionary rates between orthologs), or from non-sequence-based measures such as the similarity of gene repertoire and gene orders [65]. Generally, it appears that trees produced from concatenated alignments of gene products that are not particularly prone to HGT yield the best resolution [66-68]. All genome-tree analyses unequivocally supported the monophyly of archaea and the monophyly of Crenarchaeota. Beyond that, however, the genome-tree topology is not necessarily compatible with that of rRNA-based trees. Thus, genome-tree analysis cast doubt on the bifurcation of Euryarchaeota and Crenarchaeota being the first split in archaeal evolution; in some of these analyses, *Halobacterium* and *Thermoplasma* branch off first, suggesting that Crenarchaeota are a highly derived lineage that evolved from within Euryarchaeota [66]. The same versions of genome-trees strongly suggest monophyly of methanogens, which is compatible with their distinct gene repertoire and life style [45]; but alternative trees constructed from concatenated multiple alignments of a different assortment of translation machinery components support the original divergence of Crenarchaeota and Euryarchaeota but reject the monophyly of methanogens [21,69]. It appears that a robust phylogeny of archaea will require many additional genome sequences and perhaps also further refinement of phylogenetic methods dealing with long branches and with large amounts of data. The reconstruction of the best approximation of archaeal phylogeny is of interest not so much in and of itself, but more in terms of clarifying the tempo and mode of evolution of this remarkable group of organisms. A definitive tree topology will help answer fundamental questions, such as whether methanogenesis evolved only once or several times, whether the role of histones in chromatin formation is ancestral or derived in the archaeo-eukaryotic lineage, and even the exact evolutionary relationship between archaea and eukaryotes.

Phylogenetic trees can also be employed for reconstruction of the gene sets of ancestral life forms. Given a species tree topology and phyletic patterns of the maximum possible number of orthologous gene sets (or COGs), the most parsimonious evolutionary scenario, which includes the minimum possible number of elementary events, can be reconstructed using various parsimony algorithms [70,71]. The elementary events included in this type of analysis are gene gain and gene loss. Gene gain in a given lineage may occur either as emergence of new genes (COGs), primarily via duplication with subsequent radical divergence, or as HGT from other lineages. The relative likelihood of gene loss and gene gain (the gain penalty) substantially affects the reconstructed evolutionary scenario and the gene composition of the reconstructed ancestral genomes - but this parameter is a major unknown. Nevertheless, examination of the gene sets for the last universal common ancestor (LUCA) derived with different gain penalties showed, perhaps rather unexpectedly, that the assumption of equal probabilities of gains and losses (a gain penalty of 1) yields a reasonable reconstruction of the main functional systems of the cell [71].

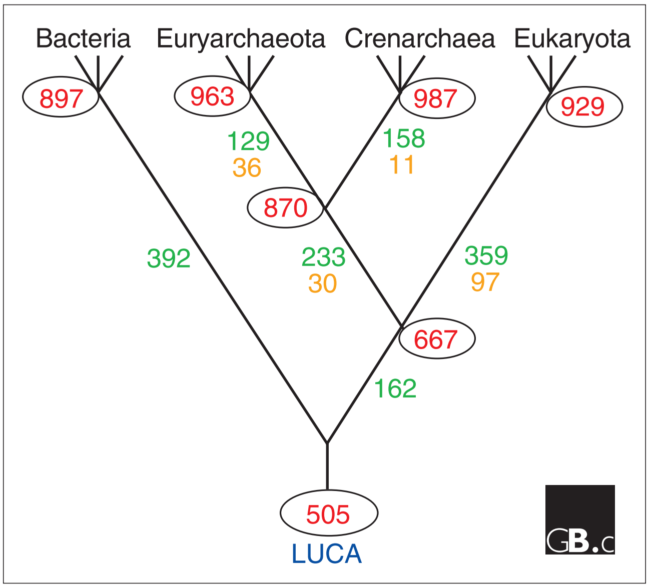

Figure 3
The most parsimonious scenario for the evolution of the main lineages of life. The red numbers in ovals near the internal nodes show the size of the reconstructed gene sets of the respective ancestral forms. Green numbers show gene gains and brown numbers gene losses assigned to each of the branches in the tree. LUCA, last universal common ancestor.

We therefore applied our version of the weighted parsimony algorithm [70], with that assumption, to the updated set of bacterial, archaeal and eukaryotic genomes (also assuming the dichotomy of Euryarchaeota and Crenarchaeota suggested by rRNA trees and some of the genome-trees) and the results are schematically shown in Figure 3 (see also additional data file). This reconstruction suggests that the common ancestor of archaea could have had around 900 genes, with substantial gene gain but only minimal gene loss compared to the more ancient common ancestor of the archaeo-eukaryotic lineage. Obviously, the conserved core of the pan-archaeal genes is a subset of the reconstructed ancestral gene set, but it seems striking that approximately two thirds of the ancestral genes have been lost from at least one of the sequenced archaeal genomes (Figure 3).

From genome comparisons to functional and structural genomics of the archaea

In the era of comparative genomics, experimental studies on a genomic scale lag woefully behind computational studies. The great majority of the genes in most species will never be studied experimentally, and our understanding of the biochemistry and physiology of the respective organisms therefore

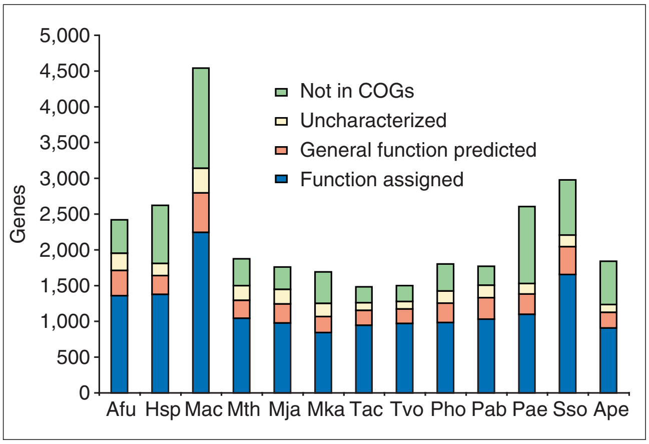

Figure 4
Functional breakdown of genes in each of the sequenced archaeal genomes. The data are from COGs; species name abbreviations are as in Table 1.

depends on the transfer of information from functionally characterized orthologs [26,72]. For both bacteria and eukaryotes, such transfer is facilitated by the availability of a vast body of experimental data on model organisms, such as *Escherichia coli*, *Bacillus subtilis*, the yeast *Saccharomyces cerevisiae* or the fruit fly *Drosophila melanogaster*. The situation is quite different for archaea because, some genetic studies of mesophilic archaeal species notwithstanding [73], there is, so far, no satisfactory model system; this results primarily from the fact that most of these organisms grow slowly and are hard to cultivate. The functions of most of the archaeal genes have therefore been predicted by sequence analysis. Moreover, on many occasions the similarity between an archaeal protein and its functionally characterized homolog is so low that computational methods for sequence analysis have to be extended to the limit of their power.

A substantial fraction of the functional predictions for archaeal proteins appear 'trivial' in the sense that the respective proteins are highly conserved orthologs of well-characterized proteins from model organisms and, for all practical purposes, the validity of the prediction is beyond reasonable doubt (which is not to say that there are no important details of the functions of these proteins that can be uncovered only by experiment). For many other proteins, however, the prediction remains only a pointer to the probable biochemical function while the biology remains a mystery. A rough breakdown of the state of functional characterization of several archaea with sequenced genomes is given in Figure 4. The substantial fraction of genes for which only general, typically biochemical, prediction is available, is testimony to the current limited understanding of archaeal biology. Moreover, even some of the more definitive predictions only serve to emphasize the biological differences between archaea and the bacterial or eukaryotic models from which the predictions are inferred (Table 3). A good example is the archaeal ortholog of the bacterial DNA primase (DnaG), which is a highly conserved protein present in all archaea [24]. The discovery of a predicted bacterial-type primase in archaea was unexpected, given that the archaeal replication system is orthologous to that of eukaryotes and, in particular, archaea encode the two subunits of the eukaryotic-type primase (COG1467 and COG2219; it should be noted parenthetically that detection of the large primase subunit itself required extremely careful sequence analysis due to the low similarity to the eukaryotic ortholog [22]). Given that the niche of the replicative primase seems to be occupied by the eukaryotic-type enzyme [74,75], the DnaG ortholog is likely to have a critical role in repair, but beyond this general idea its function has yet to be determined by direct experimentation; such experiments have the potential to reveal completely new repair systems and pathways. Other proteins implicated in repair as a result of exhaustive sequence analysis, such as the putative nucleases encoded by COG1833 and COG1628 (Table 3), illustrate the same point: the biochemical activities are predicted but the biology remains to be investigated experimentally.

Some of the other functional predictions inferred from sequence analysis directly help filling glaring gaps in otherwise well-characterized pathways of archaeal metabolism. A good example of such focused prediction is the identification of an archaeal fructose-1,6-bisphophate aldolase, an indispensable glycolytic enzyme, which was first predicted computationally to be a member of the DhnA family of aldolases by our group [76] and subsequently identified experimentally [77]. In the same vein, during work for this article, we predicted the missing archaeal aconitase, an essential enzyme of the tricarboxylic acid cycle (Table 3; K.S.M. and E.V.K., unpublished observations).

The identities of a considerable number of proteins responsible for essential functions in archaea remain a mystery. Perhaps the most notable case is the missing cysteinyl-tRNA synthetase of thermophilic methanogens. Cysteine is incorporated into the proteins of these organisms as readily as in any others, but they lack an ortholog of cysteinyl-tRNA synthetase. Two different solutions for this paradox have been proposed, one involving an uncharacterized protein that has been proposed to be a 'third class' of aminoacyl-tRNA synthetases [78], and the other based on the apparent ability of the archaeal prolyl-tRNA synthetase to couple tRNA^{Cys} with cysteine [79]. The first hypothesis has been refuted by our group upon more detailed sequence analysis [80], however, and the second did not seem to be compatible with subsequent structural studies [81]. The real cysteinyl-RNA synthetase of methanogens seems still to be hiding among uncharacterized proteins. Gaping holes also remain in archaeal pathways of isoleucine biosynthesis [82], heme biosynthesis [83], biotin biosynthesis [26], and several others.

Table 3

Examples of computational and experimental discovery of unexpected functions in archaea

COG numbers [37,38]	Function and comments	References
Computational predictions		
0012, 1325, 1603, 1369, 0638, 1500, 1097, 689, 2123, 1996, 2136, 2892, 0618, 1782, 1096, 3286, 1761 and more	Archaeal exosome. Orthologs of eukaryotic exosome subunits form the largest conserved superoperon in archaea, after the ribosomal superoperon, suggesting the existence of a physical complex	[88]
1769, 1336, 3337, 1583, 1367, 1604, 1517, 1857, 1688, 1203, 1468, 1518, 2254, 1343, 1353, 1421, 1337, 1567, 1332, 4343	DNA repair system represented primarily in thermophiles	[59]
0358	Bacterial-type DNA primase (DnaG orthologs)	[24]
1311	Small subunit of euryarchaeal DNA polymerase II, predicted PHP family phosphohydrolase (probably phosphatase); eukaryotic homologs appear to be inactivated	[123]
1833	Uri superfamily endonuclease	[136]
1628	Endonuclease V homologs	K.S.M. and E.V.K., unpublished observations
1679,1786	Aconitase catalytic core and an interacting 'swiveling domain'	K.S.M. and E.V.K., unpublished observations
1711	Possible subunit of the DNA replication machinery	K.S.M. and E.V.K., unpublished observations
1310	Zn^{2+}-dependent hydrolase homologous to the eukaryotic ubiquitin isopeptidase contained in the proteasome and COP9 signalosome	[137,138]
Computational predictions validated by experiments		
1708	'Minimal' nucleotidyltransferases	[100,139]
1830	Fructose-1,6-bisphosphate aldolases (DhnA family)	[76,77]
1351	Thymidylate synthase	[61,64]
1685	Shikimate kinase (predicted on the basis of operon organization)	[140]
3635	Phosphoglycerate mutase	[24,141]
Experimental discovery of unexpected protein functions in archaea		
1384	Class I lysyl-tRNA synthetase	[62]
1933	DNA polymerase II	[104]
1980	Fructose 1,6-bisphosphatase	[142]
1630	NurA, a novel 5′-3′ nuclease encoded next to Rad50 and Mre11 orthologs; present in all sequenced archaeal genomes and some bacteria	[143] and K.S.M. and E.V.K., unpublished observations
1812	*S*-adenosylmethionine synthetase, was identified by mass tags	[144]
1591	Holliday junction resolvase	[101]
1581	Alba, a major DNA-binding chromatin protein in Crenarchaeota	[106]
1945	Pyruvoyl-dependent arginine decarboxylase (PvlArgDC), involved in polyamine biosynthesis	[145]

Beyond straightforward (even if highly sensitive) sequence analysis, a powerful approach to the prediction of functions involves analysis of various forms of genomic context, or establishing 'guilt by association' [26,84-87]. The associations employed to infer gene functions may be manifest at different levels, including the phyletic patterns discussed above, juxtaposition of domains in multidomain proteins, clustering of genes in (predicted) operons, co-expression, and protein-protein interaction. The last two of these types of data, obtained through transcriptomic and proteomic efforts, are becoming increasingly important in the functional genomics of eukaryotes and, to a somewhat lesser extent, bacteria, but are so far unavailable for archaea. The main type of context information in archaea has therefore been obtained by analyzing conserved elements of gene order and multidomain proteins. Only a relatively small fraction (10-15%) of each archaeal genome is covered by evolutionarily conserved gene strings that can be predicted to form operons [87]. Nevertheless, by comparing gene orders in multiple genomes, partially conserved gene neighborhoods can be reconstructed and examination of some of these leads to predictions of functional systems whose existence has not previously been suspected (Table 3).

The most notable illustrations of this approach (both from our own group) are the prediction of the archaeal exosome [88] and a potential new repair system typical of archaeal and bacterial thermophiles [59]. The eukaryotic exosome is a multisubunit complex that consists of RNAses, helicases and RNA-binding proteins and is involved in the exonucleolytic degradation of various classes of RNA [89-91]. During comparative analysis of gene order in prokaryotic genomes, it was observed that a distinct set of genes, some of which encode orthologs of eukaryotic exosome components, form a partially conserved predicted superoperon, which includes in total over 15 genes (although none of the archaeal genomes contains every one of these within the predicted superoperon). In addition to RNAses and RNA-binding proteins (with an RNA helicase apparently encoded in a separate operon), the exosomal superoperon also encodes a proteasome subunit and a subunit of prefoldin, a co-translational molecular chaperone ([88] and Figure 5a). Thus, these observations point to the existence of a multifunctional macromolecular complex that could couple post-translational protein folding with regulated, ATP-dependent degradation of RNA and proteins. This complex remains to be discovered experimentally, and the potential implications for new functional and physical interactions in eukaryotes are also open to experimental study.

A more sophisticated comparison of gene orders, which required special algorithms for delineation of partially conserved genomic neighborhoods [92], led us to predict a distinct DNA repair system that is most prevalent in thermophiles and includes genes for a predicted novel DNA polymerase, a helicase, two nucleases and several uncharacterized genes, at least one of which could encode a novel nuclease ([59] and Figure 5b). Furthermore, this neighborhood contains multiple, diverged versions of a gene coding for a protein with a probable structural role dubbed RAMP (repair-associated mysterious protein). The proliferation of RAMP genes (Figure 5b) is an example of a potentially adaptive lineage-specific expansion of a gene family; such expansions are discussed below in greater detail.

Additional, simpler cases of functional prediction via 'guilt by association' are illustrated in Figure 5c-e. The gene for the uncharacterized protein represented by COG1711 (Figure 5c) forms an evolutionarily highly conserved gene pair with the gene for the clamp subunit of DNA polymerase (ortholog of the eukaryotic PCNA). The orthologs of COG1711 proteins are conserved in all eukaryotes, and this protein might be an essential but still uncharacterized component of the archaeo-eukaryotic DNA replication machinery (K.S.M. and E.V.K., unpublished observations). The gene represented by uncharacterized COG1909 is squeezed between genes for RNA polymerase subunits and that for a ribosomal protein (Figure 5d). Examination of the multiple alignments that lead to this COG shows conservation of polar residues compatible with an enzymatic function (K.S.M. and E.V.K., unpublished observations). There are no readily detectable eukaryotic orthologs for this protein, which is therefore likely to be an archaea-specific enzyme with a house-keeping function.

Finally, uncharacterized COG1545 consists of genes encoding putative zinc-ribbon-containing proteins that form a stable gene pair with the gene for acetyl-CoA acetyltransferase, a central enzyme of fatty acid biosynthesis (Figure 5e). Both these genes show remarkable paralogous expansion in several archaea, probably as a result of a series of duplications of the gene doublet. It appears likely that proteins from COG1545 form a complex with acetyl-CoA acetyltransferase, with the zinc-ribbon protein regulating and/or stabilizing the enzyme. The predictions depicted in Figure 5c-e and other similar ones ([87]; and K.S.M. and E.V.K., unpublished observations) are not particularly precise, even in terms of the biochemical activity of the respective proteins. Nevertheless, guilt by association implicates each of these proteins in specific biological functions, and the evolutionary conservation of both the proteins themselves and the gene order all but proves that their functions are essential. Thus, these proteins appear to be excellent targets for experimental studies, which have the potential to reveal new facets of central cellular processes in archaea.

Comparative-genomic analysis of prokaryotes and eukaryotes points to lineage-specific expansion (proliferation) of paralogous gene families as a major means by which organisms adapt to their specific environment and lifestyle [93-95]. A number of such expansions are seen in archaea but in most cases we have, at best, only a vague understanding of the associated biology; several examples are given in Figure 6. The expansion of two groups of permeases in *Thermoplasma* and *Sulfolobus* (Figure 6a) clearly reflects the heterotrophic metabolism of the former [96] and the chemo-organotrophic lifestyle of the latter [97]. The specific proliferation of ferredoxin in methanogens (Figure 6b) is also easily explained by the role of these proteins in the oxido-reduction reactions of methanogenesis [98]. The remaining two cases in Figure 6(c,d) are much more enigmatic. The congruent proliferation of the transcription-initiation factors TFIIB and TFIID in *Halobacterium* (Figure 6c) might point to unusual aspects of transcription regulation in this archaeon but the details remain obscure. The proliferation of

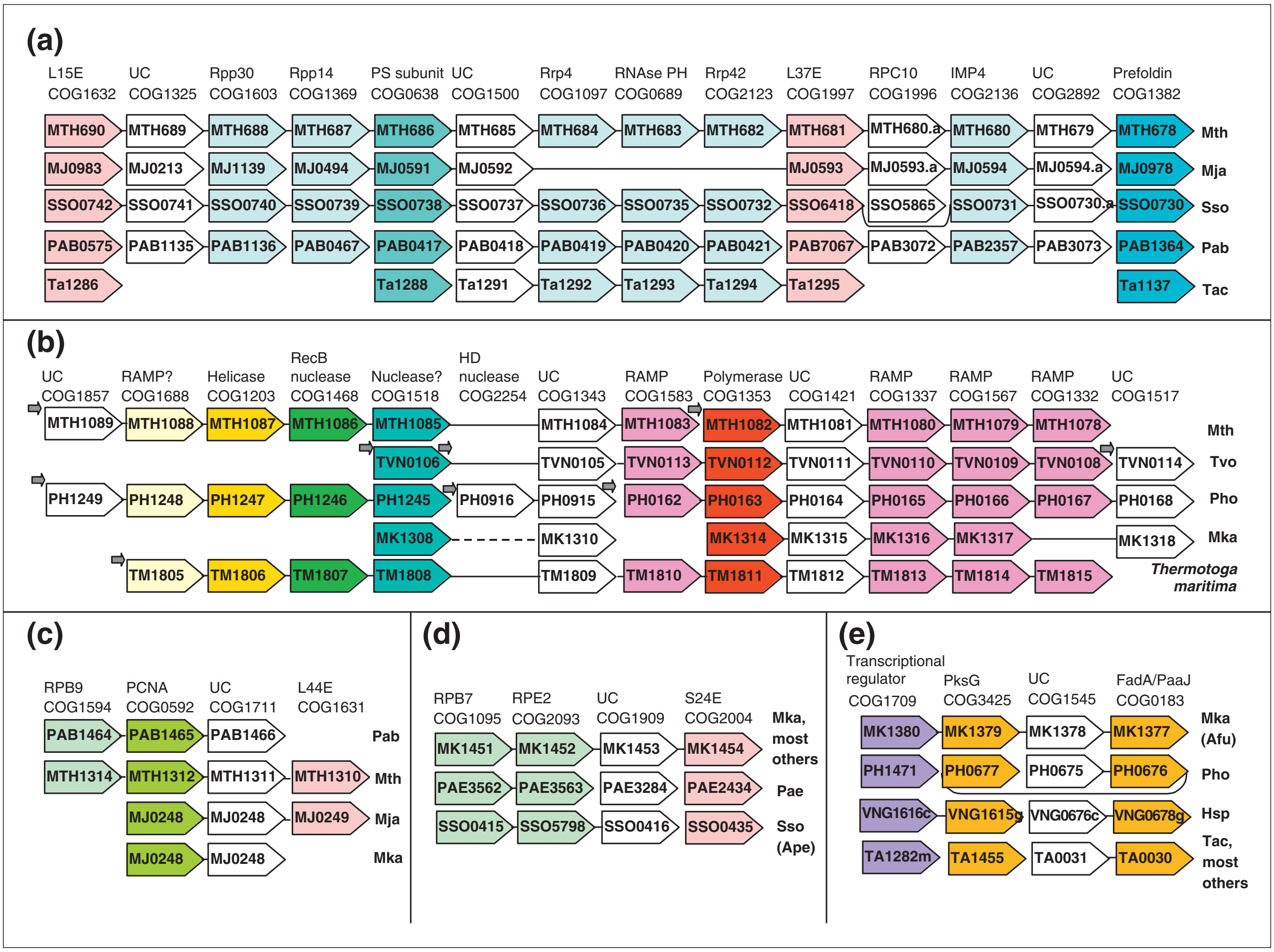
(a)
L15E COG1632
UC COG1325
Rpp30 COG1603
Rpp14 COG1369
PS subunit COG0638
UC COG1500
Rrp4 COG1097
RNAse PH COG0689
Rrp42 COG2123
L37E COG1997
RPC10 COG1996
IMP4 COG2136
UC COG2892
Prefoldin COG1382
MTH690 MTH689 MTH688 MTH687 MTH686 MTH685 MTH684 MTH683 MTH682 MTH681 MTH680.a MTH680 MTH679 MTH678 Mth
MJ0983 MJ0213 MJ1139 MJ0494 MJ0591 MJ0592 MJ0593 MJ0593.a MJ0594 MJ0594.a MJ0978 Mja
SSO0742 SSO0741 SSO0740 SSO0739 SSO0738 SSO0737 SSO0736 SSO0735 SSO0732 SSO6418 SSO5865 SSO0731 SSO0730.a SSO0730 Sso
PAB0575 PAB1135 PAB1136 PAB0467 PAB0417 PAB0418 PAB0419 PAB0420 PAB0421 PAB7067 PAB3072 PAB2357 PAB3073 PAB1364 Pab
Ta1286 Ta1288 Ta1291 Ta1292 Ta1293 Ta1294 Ta1295 Ta1137 Tac
(b)
UC COG1857
RAMP? COG1688
Helicase COG1203
RecB nuclease COG1468
Nuclease? COG1518
HD nuclease COG2254
UC COG1343
RAMP COG1583
Polymerase COG1353
UC COG1421
RAMP COG1337
RAMP COG1567
RAMP COG1332
UC COG1517
MTH1089 MTH1088 MTH1087 MTH1086 MTH1085 MTH1084 MTH1083 MTH1082 MTH1081 MTH1080 MTH1079 MTH1078 Mth
TVN0106 TVN0105 TVN0113 TVN0112 TVN0111 TVN0110 TVN0109 TVN0108 TVN0114 Tvo
PH1249 PH1248 PH1247 PH1246 PH1245 PH0916 PH0915 PH0162 PH0163 PH0164 PH0165 PH0166 PH0167 PH0168 Pho
MK1308 MK1310 MK1314 MK1315 MK1316 MK1317 MK1318 Mka
TM1805 TM1806 TM1807 TM1808 TM1809 TM1810 TM1811 TM1812 TM1813 TM1814 TM1815 *Thermotoga maritima*
(c)
RPB9 COG1594
PCNA COG0592
UC COG1711
L44E COG1631
PAB1464 PAB1465 PAB1466 Pab
MTH1314 MTH1312 MTH1311 MTH1310 Mth
MJ0248 MJ0248 MJ0249 Mja
MJ0248 MJ0248 Mka
(d)
RPB7 COG1095
RPE2 COG2093
UC COG1909
S24E COG2004
MK1451 MK1452 MK1453 MK1454 Mka, most others
PAE3562 PAE3563 PAE3284 PAE2434 Pae
SSO0415 SSO5798 SSO0416 SSO0435 Sso (Ape)
(e)
Transcriptional regulator COG1709
PksG COG3425
UC COG1545
FadA/PaaJ COG0183
MK1380 MK1379 MK1378 MK1377 Mka (Afu)
PH1471 PH0677 PH0675 PH0676 Pho
VNG1616c VNG1615g VNG0676c VNG0678g Hsp
TA1282m TA1455 TA0031 TA0030 Tac, most others

Figure 5

Prediction of gene functions in archaea by genomic context analysis. **(a)** The superoperon coding for the predicted archaeal exosome (see [88]). **(b)** The partially conserved gene neighborhood coding for the predicted repair system found in archaeal and bacterial thermophiles (see [59] for details). **(c-e)** Predicted operons containing uncharacterized genes in the neighborhood of genes from the following COGs: COG1594, DNA-directed RNA polymerase, subunit M, and transcription elongation factor TFIIS (RPB9); COG0592, encoding a DNA polymerase sliding clamp subunit (PCNA ortholog); COG1631, ribosomal protein L44E; COG1095, DNA-directed RNA polymerase, subunit E′ (RPB7); COG2093, DNA-directed RNA polymerase, subunit E′′ (RPE2); COG2004, ribosomal protein S24E; COG1709, transcriptional regulator; COG3425, 3-hydroxy-3-methylglutaryl CoA synthase (PksG); COG0183, acetyl-CoA acetyltransferase (Fad A/PaaJ orthologs). UC, uncharacterized, shown by white arrows. Species abbreviations are as in Table 1. Genes are shown not to scale and are denoted by their respective genes names (some are discussed further in the text); arrows indicate the direction of transcription. A solid line connects genes in a predicted operon. Species that have the same operon organization as the listed species are indicated in parentheses. Orthologous genes are aligned. Genes with similar general functions are shown by the same shading. Broken lines show that genes are in the same predicted operon but are not adjacent. Small arrows indicate the presence of additional functionally related genes in the same predicted operon; these genes are not shown for lack of space.

two subunits of a predicted nucleotidyltransferase in several archaea [99,100] (Figure 6d) is of special interest and might have something to do with thermal adaptation, but the actual functions and even the substrates of these enzymes remain a mystery. Other lineage-specific expansions, such as that of distinct families of predicted ATPases in *Methanocaldococcus* and *Pyrococcus*, or a specific family of RadA(RecA)-like ATPases and the UspA-family of NTP-binding proteins in several archaeal species [39], suggest the existence of unusual pathways, perhaps involved in stress response and signal transduction, but the actual biology associated with these expansions can only be uncovered experimentally.

Archaeal comparative genomics is a young field and so far, as we have seen, largely predictive. But a few experimental studies have already been instigated as a result of comparative-genomic predictions. The discovery of the archaeal fructose-1,6-bisphosphate aldolase mentioned above [76,77] is a case in point, and several other examples of experimental validation of predictions are given in Table 3. It does not seem to be chance that these examples all involve metabolic enzymes for which the specific reaction could be predicted precisely. Validation is likely to be much more difficult for proteins of other functional groups, such as putative repair enzymes, for which the actual substrates are harder to predict.

For some conserved archaeal proteins, functions cannot be predicted computationally despite considerable effort. Several important discoveries have been made by experimental characterization of such mysterious proteins. The most notable cases include the archaeal Holliday-junction resolvase, which is not related to its functional analog in bacteria [101-103], and DNA polymerase II, a highly conserved euryarchaeal protein that is not found outside this lineage and shows no detectable sequence similarity to any other proteins [104,105]. Additional examples of direct experimental determination of the functions of archaeal proteins that could not be predicted by computational techniques (at least not before the experiment had been reported) are given in Table 3.

Especially notable is the story of the Alba protein, a DNA-binding component of chromatin in Crenarchaeota [106,107]. As noted above, crenarchaea lack histones and in these organisms Alba appears to be the main chromatin protein, in a striking case of non-orthologous gene displacement. But orthologs of Alba are also present in thermophilic Euryarchaeota and in some eukaryotic lineages, where its functions remain to be elucidated. The most remarkable discovery regarding Alba is the regulation of its interaction with DNA and with the chromatin-associated protein deacetylase Sir2 via lysine acetylation and deacetylation [106,108]. In eukaryotes, regulation of chromatin dynamics via acetylation and deacetylation occurs through histone tails [109]. Thus, a special case of non-orthologous gene displacement seems to have taken place whereby the regulation mechanism is conserved but the actual substrates are different in archaea and eukaryotes. To add an extra twist to the story, *Thermoplasma* lacks both histones and Alba but has the bacterial DNA-binding protein HU, pointing to three distinct solutions to the problem of chromatin organization in archaea [107].

The last subject we have to briefly touch upon is structural genomics of the archaea. The ultimate goal of the structural genomics enterprise is determining the three-dimensional structure for all proteins, or at least for all sufficiently different proteins encoded in the genomes of diverse life forms [110]. This goal is far from being reached, and targets for

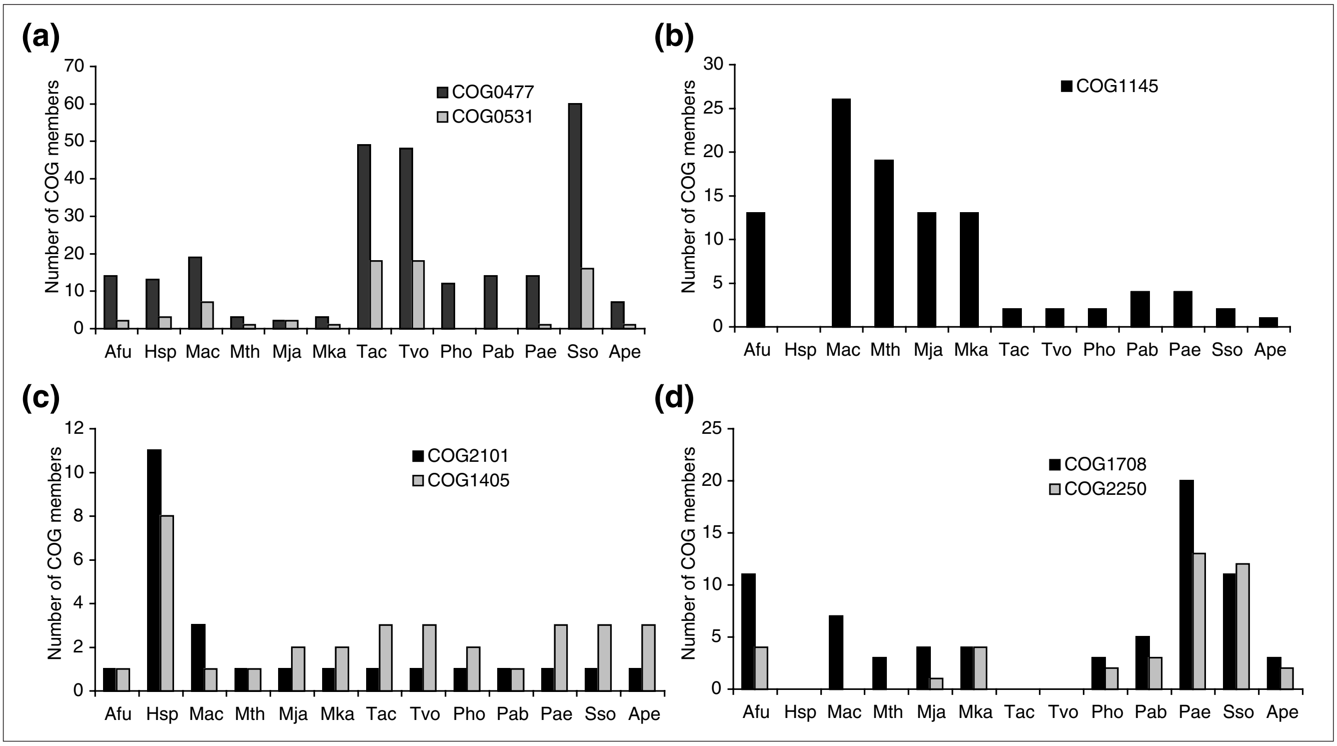

Figure 6

Lineage-specific expansions of paralogous gene families in archaea. The vertical axis shows the number of members of the indicated COGs. **(a)** COG0477, permeases of the major facilitator superfamily; COG0531, amino-acid transporters. **(b)** COG1145, ferredoxin. **(c)** COG2101, TATA-box binding protein (TBP), a component of transcription initiation factors TFIID and TFIIIB; COG1405, Brf1 subunit of transcription-initiation factor TFIIIB and transcription-initiation factor TFIIB. **(d)** COG1708, 'minimal' nucleotidyltransferase catalytic subunit; COG2250, 'minimal' nucleotidyltransferase accessory subunit. Species abbreviations are as in Table 1.

structural determination have been prioritized by different researchers on the basis of different principles, from nearly random choice to relatively elaborate strategies, including the use of the COG database [111-115]. The development of structural genomics so far has been a mixture of success, when informative and interesting structures have been solved, and mild disappointment in cases when the structure determination did not seem to shed any light on a protein's function. Structural genomics could be particularly important in the case of archaea, for which a miniscule number of structures had been solved prior to the launch of structural genomic initiatives, and in which proteins often show low similarity to bacterial or eukaryotic homologs, making homology modeling difficult.

Notable developments that illustrate both the benefits and the pitfalls of structural genomics, are the concerted effort on 'structural proteomics' of *Methanothermobacter thermoautotrophicus* [116] and a similar project on *M. jannaschii* [117]. The elucidation of the structure of the *M. jannaschii* protein MJ0577 [117] is an excellent case for the power of structural genomics. Analysis of this structure and accompanying biochemical experiments revealed a distinct nucleotide-binding domain that is distantly related to the catalytic domains of class I aminoacyl-tRNA synthetases and belongs to the so-called HUP fold of nucleotide-binding domains [118]. Together with comprehensive sequence analysis, the determination of this structure provided the structural, functional and evolutionary context for the UspA protein family, which is specifically expanded in archaea [39]. The exact function(s) of these proteins remains unknown but, in this case, structural genomics ensured a substantial functional insight. On several other occasions, however, determination of the structures of archaeal proteins has failed to provide clear functional clues; these remain structures in search of a function.

What's around the corner?

The first sequenced archaeal genome was a veritable *terra incognita*. Six years after that sequence appeared, the archaeal genomescape looks quite different. The principal landmarks have been mapped and now, when a new archaeal genome is released, we largely know what to expect from it. Computational approaches to comparative genomics, combining in-depth sequence and structure comparison with genome context analysis, have led to the reconstruction of the central functional systems of archaeal cells. But these approaches have also produced numerous isolated predictions of biochemical activities of archaeal proteins that remain to be fitted into a general picture, and this can be done only through 'wet' experiments, although new genome sequences will substantially help by enriching the genomic context. A shrinking but still notable set of archaeal genes includes those that encode highly conserved proteins without any clue to function; solving these mysteries has the potential to bring out truly new biology. Furthermore, in this article we have not even touched upon important aspects of archaeal genomics, such as the in-depth studies of the translation system, which have revealed several highly unusual, remarkable mechanisms and enzymatic systems [63,119] or the identification of regulatory sites in DNA and patterns of transcription regulation [120,121]. The latter avenue of research is still in its infancy but will certainly grow in scale once more archaeal genomes, and in particular closely related ones, are sequenced.

Because of the lack of established model systems for archaeal experimental biology and the resulting difficulty with large-scale experimentation, clues from genome comparison are even more crucial for archaeal functional genomics than they are in the case of bacteria or eukaryotes. So far, the input of comparative genomics into actual experiments has been less prominent than we would hope. Simply put, it is not often that experimenters rush to test predictions produced by *in silico* genome comparison and, furthermore, it is even rarer that targets for functional characterization are carefully prioritized on the basis of how unusual and fundamental the predictions are. As discussed above, however, the few cases when such tests have been performed are encouraging. It is our hope that the future belongs to a much tighter integration of comparative, structural and functional genomics.

Beyond functional studies, archaeal genomics is fundamental to our understanding of two critical transitions in the evolution of life. The first is the primary split between the bacterial and archaeo-eukaryotic lineages, which might have involved the origin of the DNA-replication machinery and of the large, double-stranded DNA genomes themselves [22,23], and the second is the origin of eukaryotes [122]. With regard to the latter problem, archaea are a particularly valuable source of information because, on many occasions, they seem to have retained primitive traits while eukaryotes have undergone major changes. A characteristic example is the small DNA polymerase subunit, which has all the hallmarks of an active phosphatase in archaea, but not in eukaryotes, in which the phosphatase activity is predicted to be inactivated [123]. Indubitably, archaea resemble the common ancestor of the archaeo-eukaryotic line of descent more closely than eukaryotes do, so archaeal genomics is our best chance to reconstruct this critical intermediate in the evolution of life. We are confident that comparative archaeogenomics has a bright future, with major progress in both the functional and the evolutionary avenues of research expected within the next few years.

Additional data file

The list of genes in the reconstructed gene set of the last common ancestor of archaea is available with the complete version of this article, online.

Acknowledgements

We thank Boris Mirkin for producing the data used for Figure 3 and Stephen Bell, Michael Galperin, Dieter Söll and Yuri Wolf for useful discussions.

References

1. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74:**5088-5090.
2. Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, *et al.*: **The phylogeny of prokaryotes.** *Science* 1980, **209:**457-463.
3. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proc Natl Acad Sci USA* 1990, **87:**4576-4579.
4. Woese CR, Gupta R: **Are archaebacteria merely derived 'prokaryotes'?** *Nature* 1981, **289:**95-96.
5. Mayr E: **Two empires or three?** *Proc Natl Acad Sci USA* 1998, **95:**9720-9723.
6. Woese CR: **Default taxonomy: Ernst Mayr's view of the microbial world.** *Proc Natl Acad Sci USA* 1998, **95:**11043-11046.
7. Gupta RS: **Life's third domain (Archaea): an established fact or an endangered paradigm?** *Theor Popul Biol* 1998, **54:**91-104.
8. Kushner DJ: **Lysis and dissolution of cells and envelopes of an extremely halophilic bacterium.** *J Bacteriol* 1964, **87:**1147-1156.
9. Langworthy TA, Smith PF, Mayberry WR: **Lipids of *Thermoplasma acidophilum*.** *J Bacteriol* 1972, **112:**1193-1200.
10. Brock TD, Brock KM, Belly RT, Weiss RL: ***Sulfolobus*: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature.** *Arch Mikrobiol* 1972, **84:**54-68.
11. Stetter KO: **Extremophiles and their adaptation to hot environments.** *FEBS Lett* 1999, **452:**22-25.
12. Segerer AH, Burggraf S, Fiala G, Huber G, Huber R, Pley U, Stetter KO: **Life in hot springs and hydrothermal vents.** *Orig Life Evol Biosph* 1993, **23:**77-90.
13. DeLong EF: **A phylogenetic perspective on hyperthermophilic microorganisms.** *Methods Enzymol* 2001, **330:**3-11.
14. DeLong EF, Pace NR: **Environmental diversity of bacteria and archaea.** *Syst Biol* 2001, **50:**470-478.
15. Hanford MJ, Peeples TL: **Archaeal tetraether lipids: unique structures and applications.** *Appl Biochem Biotechnol* 2002, **97:**45-62.
16. Engelhardt H, Peters J: **Structural research on surface layers: a focus on stability, surface layer homology domains, and surface layer-cell wall interactions.** *J Struct Biol* 1998, **124:**276-302.
17. Edgell DR, Doolittle WF: **Archaea and the origin(s) of DNA replication proteins.** *Cell* 1997, **89:**995-998.
18. Sandman K, Pereira SL, Reeve JN: **Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome.** *Cell Mol Life Sci* 1998, **54:**1350-1364.
19. Sandman K, Bailey KA, Pereira SL, Soares D, Li WT, Reeve JN: **Archaeal histones and nucleosomes.** *Methods Enzymol* 2001, **334:**116-129.
20. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O: **Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale.** *Nucleic Acids Res* 2002, **30:**5382-5390.
21. Forterre P, Brochier C, Philippe H: **Evolution of the Archaea.** *Theor Popul Biol* 2002, **61:**409-422.
22. Leipe DD, Aravind L, Koonin EV: **Did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27:**3389-3401.
23. Forterre P: **The origin of DNA genomes and DNA replication proteins.** *Curr Opin Microbiol* 2002, **5:**525-532.
24. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25:**619-637.
25. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96:**3801-3806.
26. Koonin EV, Galperin MY: *Sequence - Evolution - Function. Computational Approaches in Comparative Genomics*. New York: Kluwer Academic; 2002.
27. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, *et al.*: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*.** *Science* 1996, **273:**1058-1073.
28. Pace NR: **A molecular view of microbial diversity and the biosphere.** *Science* 1997, **276:**734-740.
29. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO: **A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont.** *Nature* 2002, **417:**63-67.
30. Huber H, Hohn MJ, Stetter KO, Rachel R: **The phylum Nanoarchaeota: present knowledge and future perspectives of a unique form of life.** *Res Microbiol* 2003, **154:**165-171.
31. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
32. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19:**99-113.
33. Fitch WM: **Homology: a personal view on some of the problems.** *Trends Genet* 2000, **16:**227-231.
34. Sonnhammer EL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genet* 2002, **18:**619-620.
35. Gaasterland T, Ragan MA: **Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes.** *Microb Comp Genomics* 1998, **3:**199-217.
36. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.
37. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.
38. **Prokaryotic COGs project phyletic pattern search** [http://www.ncbi.nlm.nih.gov/COG/new/release/phylox.cgi]
39. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV: **Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9:**608-628.
40. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12:**334-336.
41. Doolittle WF, Logsdon JM, Jr.: **Archaeal genomics: do archaea have a mixed heritage?** *Curr Biol* 1998, **8:**R209-R211.
42. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284:**2124-2129.
43. Doolittle WF: **Lateral genomics.** *Trends Cell Biol* 1999, **9:**M5-M8.
44. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes - quantification and classification.** *Annu Rev Microbiol* 2001, **55:**709-42.
45. Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, *et al.*: **The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens.** *Proc Natl Acad Sci USA* 2002, **99:**4644-4649.
46. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: **Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles.** *Trends Genet* 1998, **14:**442-444.
47. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, *et al.*: **Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*.** *Nature* 1999, **399:**323-329.
48. Brown JR: **Ancient horizontal gene transfer.** *Nat Rev Genet* 2003, **4:**121-132.
49. Kyrpides NC, Olsen GJ: **Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry?** *Trends Genet* 1999, **15:**298-299.
50. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: **Reply. Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry?** *Trends Genet* 1999, **15:**299-300.
51. Bao Q, Tian Y, Li W, Xu Z, Xuan Z, Hu S, Dong W, Yang J, Chen Y, Xue Y, *et al.*: **A complete sequence of the *T. tengcongensis* genome.** *Genome Res* 2002, **12:**689-700.
52. ***Thermoanaerobacter tengcongensis* proteins** [http://www.ncbi.nlm.nih.gov/sutils/taxik.cgi?gi=237]
53. Nesbo CL, Boucher Y, Doolittle WF: **Defining the core of nontransferable prokaryotic genes: the euryarchaeal core.** *J Mol Evol* 2001, **53:**340-350.

54. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Baumer S, Jacobi C, *et al.*: **The genome of *Methanosarcina mazei:* evidence for lateral gene transfer between bacteria and archaea.** *J Mol Microbiol Biotechnol* 2002, **4:**453-461.
55. Galagan JE, Nusbaum C, Roy A, Endrizzi MG, Macdonald P, FitzHugh W, Calvo S, Engels R, Smirnov S, Atnoor D, *et al.*: **The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity.** *Genome Res* 2002, **12:**532-542.
56. Forterre P: **A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein.** *Trends Genet* 2002, **18:**236-237.
57. Forterre P, Bergerat A, Lopez-Garcia P: **The unique DNA topology and DNA topoisomerases of hyperthermophilic archaea.** *FEMS Microbiol Rev* 1996, **18:**237-248.
58. Makarova KS, Wolf YI, Koonin EV: **Potential genomic determinants of hyperthermophily.** *Trends Genet* 2003, **19:**172-176.
59. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30:**482-496.
60. White RH: **Biosynthesis of the methanogenic cofactors.** *Vitam Horm* 2001, **61:**299-337.
61. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol* 2000, **18:**609-613.
62. Ibba M, Morgan S, Curnow AW, Pridmore DR, Vothknecht UC, Gardner W, Lin W, Woese CR, Soll D: **A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases.** *Science* 1997, **278:**1119-1122.
63. Praetorius-Ibba M, Ibba M: **Aminoacyl-tRNA synthesis in archaea: different but not unique.** *Mol Microbiol* 2003, **48:**631-637.
64. Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U: **An alternative flavin-dependent mechanism for thymidylate synthesis.** *Science* 2002, **297:**105-107.
65. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life.** *Trends Genet* 2002, **18:**472-479.
66. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evol Biol* 2001, **1:**8.
67. Clarke GD, Beiko RG, Ragan MA, Charlebois RL: **Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores.** *J Bacteriol* 2002, **184:**2072-2080.
68. Korbel JO, Snel B, Huynen MA, Bork P: **SHOT: a web server for the construction of genome phylogenies.** *Trends Genet* 2002, **18:**158-162.
69. Matte-Tailliez O, Brochier C, Forterre P, Philippe H: **Archaeal phylogeny based on ribosomal proteins.** *Mol Biol Evol* 2002, **19:**631-639.
70. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Res* 2002, **12:**17-25.
71. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: **Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes.** *BMC Evol Biol* 2003, **3:**2.
72. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *J Mol Biol* 2000, **297:**233-249.
73. Luo Y, Wasserfallen A: **Gene transfer systems and their applications in Archaea.** *Syst Appl Microbiol* 2001, **24:**15-25.
74. Liu L, Komori K, Ishino S, Bocquier AA, Cann IK, Kohda D, Ishino Y: **The archaeal DNA primase: biochemical characterization of the p41-p46 complex from *Pyrococcus furiosus*.** *J Biol Chem* 2001, **276:**45484-45490.
75. Bocquier AA, Liu L, Cann IK, Komori K, Kohda D, Ishino Y: **Archaeal primase: bridging the gap between RNA and DNA polymerases.** *Curr Biol* 2001, **11:**452-456.
76. Galperin MY, Aravind L, Koonin EV: **Aldolases of the DhnA family: a possible solution to the problem of pentose and hexose biosynthesis in archaea.** *FEMS Microbiol Lett* 2000, **183:**259-264.
77. Siebers B, Brinkmann H, Dorr C, Tjaden B, Lilie H, van der Oost J, Verhees CH: **Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase.** *J Biol Chem* 2001, **276:**28710-28718.
78. Fabrega C, Farrow MA, Mukhopadhyay B, de Crecy-Lagard V, Ortiz AR, Schimmel P: **An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes.** *Nature* 2001, **411:**110-114.
79. Stathopoulos C, Li T, Longman R, Vothknecht UC, Becker HD, Ibba M, Soll D: **One polypeptide with two aminoacyl-tRNA synthetase activities.** *Science* 2000, **287:**479-482.
80. Iyer LM, Aravind L, Bork P, Hofmann K, Mushegian AR, Zhulin IB, Koonin EV: ***Quod erat demonstrandum?* The mystery of experimental validation of apparently erroneous computational analyses of protein sequences.** *Genome Biol* 2001, **2:**research0051.1-0051.11
81. Kamtekar S, Kennedy WD, Wang J, Stathopoulos C, Soll D, Steitz TA: **The structural basis of cysteine aminoacylation of tRNAPro by prolyl-tRNA synthetases.** *Proc Natl Acad Sci USA* 2003, **100:**1673-1678.
82. Xie G, Forst C, Bonner C, Jensen RA: **Significance of two distinct types of tryptophan synthase beta chain in Bacteria, Archaea and higher plants.** *Genome Biol* 2002, **3:**research0004.1-0004.13
83. Panek H, O'Brian MR: **A whole genome view of prokaryotic haem biosynthesis.** *Microbiology* 2002, **148:**2273-2282.
84. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10:**366-370.
85. Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10:**1204-1210.
86. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome Res* 2000, **10:**1074-1077.
87. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11:**356-372.
88. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.** *Genome Res* 2001, **11:**240-252.
89. Decker CJ: **The exosome: a versatile RNA processing machine.** *Curr Biol* 1998, **8:**R238-R240.
90. van Hoof A, Parker R: **The exosome: a proteasome for RNA?** *Cell* 1999, **99:**347-350.
91. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D: **The exosome: a conserved eukaryotic RNA processing complex containing multiple 3′—>5′ exoribonucleases.** *Cell* 1997, **91:**457-466.
92. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV: **Connected gene neighborhoods in prokaryotic genomes.** *Nucleic Acids Res* 2002, **30:**2212-2223.
93. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes.** *Genome Res* 2002, **12:**1048-1059.
94. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV: **Lineage-specific gene expansions in bacterial and archaeal genomes.** *Genome Res* 2001, **11:**555-565.
95. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314:**1041-1052.
96. Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W: **The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*.** *Nature* 2000, **407:**508-513.
97. She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ, Chan-Weiher CC, Clausen IG, Curtis BA, De Moors A, *et al.*: **The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2.** *Proc Natl Acad Sci USA* 2001, **98:**7835-7840.
98. Deppenmeier U: **The unique biochemistry of methanogenesis.** *Prog Nucleic Acid Res Mol Biol* 2002, **71:**223-283.
99. Aravind L, Koonin EV: **DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history.** *Nucleic Acids Res* 1999, **27:**1609-1618.
100. Lehmann C, Lim K, Chalamasetty VR, Krajewski W, Melamud E, Galkin A, Howard A, Kelman Z, Reddy PT, Murzin AG, Herzberg O: **The HI0073/HI0074 protein pair from *Haemophilus influenzae* is a member of a new nucleotidyltransferase family: structure, sequence analyses, and solution studies.** *Proteins* 2003, **50:**249-260.

101. Komori K, Sakae S, Shinagawa H, Morikawa K, Ishino Y: **A Holliday junction resolvase from *Pyrococcus furiosus*: functional similarity to *Escherichia coli* RuvC provides evidence for conserved mechanism of homologous recombination in Bacteria, Eukarya, and Archaea.** *Proc Natl Acad Sci USA* 1999, **96:**8873-8878.
102. Aravind L, Makarova KS, Koonin EV: **Survey and summary: Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.** *Nucleic Acids Res* 2000, **28:**3417-3432.
103. Daiyasu H, Komori K, Sakae S, Ishino Y, Toh H: **Hjc resolvase is a distantly related member of the type II restriction endonuclease family.** *Nucleic Acids Res* 2000, **28:**4540-4543.
104. Ishino Y, Komori K, Cann IK, Koga Y: **A novel DNA polymerase family found in Archaea.** *J Bacteriol* 1998, **180:**2232-2236.
105. Ishino Y, Ishino S: **DNA polymerases from euryarchaeota.** *Methods Enzymol* 2001, **334:**249-260.
106. Bell SD, Botting CH, Wardleworth BN, Jackson SP, White MF: **The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation.** *Science* 2002, **296:**148-151.
107. White MF, Bell SD: **Holding it together: chromatin in the Archaea.** *Trends Genet* 2002, **18:**621-626.
108. Wardleworth BN, Russell RJ, Bell SD, Taylor GL, White MF: **Structure of Alba: an archaeal chromatin protein modulated by acetylation.** *EMBO J* 2002, **21:**4654-4662.
109. Kurdistani SK, Grunstein M: **Histone acetylation and deacetylation in yeast.** *Nat Rev Mol Cell Biol* 2003, **4:**276-284.
110. Vitkup D, Melamud E, Moult J, Sander C: **Completeness in structural genomics.** *Nat Struct Biol* 2001, **8:**559-566.
111. Elofsson A, Sonnhammer EL: **A comparison of sequence and structure protein domain families as a basis for structural genomics.** *Bioinformatics* 1999, **15:**480-500.
112. Gaasterland T: **Structural genomics: bioinformatics in the driver's seat.** *Nat Biotechnol* 1998, **16:**625-627.
113. Brenner SE: **Target selection for structural genomics.** *Nat Struct Biol* 2000, **7 Suppl:**967-969.
114. Gerstein M: **Integrative database analysis in structural genomics.** *Nat Struct Biol* 2000, **7 Suppl:**960-963.
115. Koonin EV, Wolf YI, Aravind L: **Protein fold recognition using sequence profiles and its application in structural genomics.** *Adv Protein Chem* 2000, **54:**245-275.
116. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, *et al.*: **Structural proteomics of an archaeon.** *Nat Struct Biol* 2000, **7:**903-909.
117. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH: **Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics.** *Proc Natl Acad Sci USA* 1998, **95:**15189-15193.
118. Aravind L, Anantharaman V, Koonin EV: **Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA.** *Proteins* 2002, **48:**1-14.
119. Woese CR: **Translation: in retrospect and prospect.** *RNA* 2001, **7:**1055-1067.
120. Gelfand MS, Koonin EV, Mironov AA: **Prediction of transcription regulatory sites in Archaea by a comparative genomic approach.** *Nucleic Acids Res* 2000, **28:**695-705.
121. Rodionov DA, Mironov AA, Gelfand MS: **Conservation of the biotin regulon and the BirA regulatory signal in eubacteria and archaea.** *Genome Res* 2002, **12:**1507-1516.
122. Dacks JB, Doolittle WF: **Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help.** *Cell* 2001, **107:**419-425.
123. Aravind L, Koonin EV: **Phosphoesterase domains associated with DNA polymerases of diverse origins.** *Nucleic Acids Res* 1998, **26:**3746-3752.
124. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, *et al.*: **The complete genome sequence of the hyperthermophilic, sulphate- reducing archaeon *Archaeoglobus fulgidus*.** *Nature* 1997, **390:**364-370.
125. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, *et al.*: **Genome sequence of *Halobacterium* species NRC-1.** *Proc Natl Acad Sci USA* 2000, **97:**12176-12181.
126. Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, *et al.*: **Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics.** *J Bacteriol* 1997, **179:**7135-7155.
127. Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, *et al.*: **Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3.** *DNA Res* 1998, **5:**55-76.
128. Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Querellou J, Ripp R, *et al.*: **An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*.** *Mol Microbiol* 2003, **47:**1495-1512.
129. Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB, Dunn DM: **Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology.** *Methods Enzymol* 2001, **330:**134-157.
130. Kawashima T, Amano N, Koike H, Makino S, Higuchi S, Kawashima-Ohya Y, Watanabe K, Yamazaki M, Kanehori K, Kawamoto T, *et al.*: **Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*.** *Proc Natl Acad Sci USA* 2000, **97:**14257-14262.
131. Fitz-Gibbon ST, Ladner H, Kim UJ, Stetter KO, Simon MI, Miller JH: **Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*.** *Proc Natl Acad Sci USA* 2002, **99:**984-989.
132. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, *et al.*: **Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1.** *DNA Res* 1999, **6:**83-101.
133. Kawarabayasi Y, Hino Y, Horikawa H, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, Kosugi H, Hosoyama A, *et al.*: **Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7.** *DNA Res* 2001, **8:**123-140.
134. Tersteegen A, Hedderich R: ***Methanobacterium thermoautotrophicum* encodes two multisubunit membrane-bound [NiFe] hydrogenases. Transcription of the operons and sequence analysis of the deduced proteins.** *Eur J Biochem* 1999, **264:**930-943.
135. Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV: **Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs).** *Genome Biol* 2000, **1:**research0009.1-0009.19.
136. Aravind L, Walker DR, Koonin EV: **Conserved domains in DNA repair proteins and evolution of repair systems.** *Nucleic Acids Res* 1999, **27:**1223-1242.
137. Cope GA, Suh GS, Aravind L, Schwarz SE, Zipursky SL, Koonin EV, Deshaies RJ: **Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1.** *Science* 2002, **298:**608-611.
138. Verma R, Aravind L, Oania R, McDonald WH, Yates JR, 3rd, Koonin EV, Deshaies RJ: **Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome.** *Science* 2002, **298:**611-615.
139. Aravind L, Koonin EV: **DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history.** *Nucleic Acids Res* 1999, **27:**1609-1618.
140. Daugherty M, Vonstein V, Overbeek R, Osterman A: **Archaeal shikimate kinase, a new member of the GHMP-kinase family.** *J Bacteriol* 2001, **183:**292-300.
141. van der Oost J, Huynen MA, Verhees CH: **Molecular characterization of phosphoglycerate mutase in archaea.** *FEMS Microbiol Lett* 2002, **212:**111-120.
142. Rashid N, Imanaka H, Kanai T, Fukui T, Atomi H, Imanaka T: **A novel candidate for the true fructose-1,6-bisphosphatase in archaea.** *J Biol Chem* 2002, **277:**30649-30655.
143. Constantinesco F, Forterre P, Elie C: **NurA, a novel 5'-3' nuclease gene linked to rad50 and mre11 homologs of thermophilic Archaea.** *EMBO Rep* 2002, **3:**537-542.
144. Graham DE, Bock CL, Schalk-Hihi C, Lu ZJ, Markham GD: **Identification of a highly diverged class of S-adenosylmethionine synthetases in the archaea.** *J Biol Chem* 2000, **275:**4055-4059.
145. Graham DE, Xu H, White RH: ***Methanococcus jannaschii* uses a pyruvoyl-dependent arginine decarboxylase in polyamine biosynthesis.** *J Biol Chem* 2002, **277:**23500-23507.