Parallel Adaptations to High Temperatures in the Archean Eon

Bastien Boussau, Samuel Blanquart, Anamaria Necsulea, Nicolas Lartillot, Manolo Gouy

LETTERS

Parallel adaptations to high temperatures in the Archean eon

Bastien Boussau[1]*, Samuel Blanquart[2]*, Anamaria Necsulea[1], Nicolas Lartillot[2] & Manolo Gouy[1]

Because fossils from the time when cellular life originated and diversified are scant and difficult to interpret[1], alternative means to investigate the ecology of the Last Universal Common Ancestor (LUCA) and of the ancestors of the three domains of life are of great scientific value. It was recently recognized that footprints of the effect of temperature on ancestral organisms could be uncovered in extant genomes[2–4]. Accordingly, analyses of resurrected proteins predicted that the bacterial ancestor was thermophilic and that Bacteria subsequently adapted to lower temperatures[3,4]. Since the archaeal ancestor is also thought to have been thermophilic[5], LUCA was parsimoniously inferred as thermophilic too. However, an analysis of ribosomal RNAs supported the hypothesis of a non-hyperthermophilic LUCA[2]. Here we show that both rRNA and protein sequences analysed with advanced, realistic models of molecular evolution[6,7] provide independent support for two phases in the history of environmental temperature changes over the tree of life: in the first period, thermotolerance increased from a mesophilic LUCA to thermophilic ancestors of Bacteria and of Archaea-Eukaryota; in the second period, it decreased. Therefore, the two lineages descending from LUCA and leading to the ancestors of Bacteria and Archaea-Eukaryota convergently adapted to high temperatures, maybe in response to a climate change of the early Earth[1,8,9], and/or aided by the transition from an RNA genome in LUCA to organisms with more thermostable DNA genomes[10,11]. This analysis unifies apparently contradictory results[2–4] into a coherent depiction of the evolution of an ecological trait over the entire tree of life.

Was LUCA a hyperthermophilic, i.e., with optimal growth temperature (OGT) $\geq$ 80 °C, thermophilic (80 °C $\geq$ OGT $\geq$ 50 °C), or mesophilic (OGT $\leq$ 50 °C) organism? Several investigations of this question have relied on correlations between the species OGT and the composition of their macromolecular sequences. In extant prokaryotic species, the G+C content of ribosomal RNA (rRNA) stems (*i.e.*, double-stranded parts) has been shown to correlate with OGT[12]. Exploiting this correlation, Galtier *et al.* obtained support for a non-hyperthermophilic LUCA[2]. In contrast, studies based on correlations between the composition of LUCA's proteins and OGT concluded in favour of a hyperthermophilic LUCA[13,14] and of hyperthermophilic ancestors for both Archaea and Bacteria. The discrepancy between these results could come from some unexplained incongruence between rRNA and proteins, or, as we shall see, from differences between employed evolutionary models.

All three investigations[2,13,14] based their conclusions on comparisons of reconstructed ancestral sequence compositions with extant ones. Accurate modelling of the evolution of compositions is therefore crucial for such approaches. On one side, two of these studies[13,14] relied on homogeneous models of evolution which make the simplifying hypothesis that substitutions occur with constant probabilities along time and across all lineages. Had genomes and proteins evolved according to a homogeneous model, they would all share the same base and amino acid compositions. Clearly, rRNA[12] and protein sequences[15] do not. On the other side, Galtier *et al.*[2] relied on a branch-heterogeneous model of RNA sequence evolution. Branch-heterogeneous models are computationally more challenging but more realistic as they allow replacement or substitution probabilities to vary between lineages, and thus explicitly account

for compositional drifts[2,6,7,16,17]. Accordingly, they have been shown to accurately reconstruct ancestral sequence compositions[7].

We recently developed *nhPhyML*[7], an efficient program for the branch-heterogeneous modelling of nucleotide sequence evolution in the maximum likelihood framework, and *nhPhyloBayes*[6], that implements a site- and branch-heterogeneous Bayesian model of protein sequence evolution. The latter combines the Break Point approach[17] to model variations of amino acid replacement rates along branches and the CAT mixture model[18] to model site-wise variations of these rates. These models have been shown to describe the evolution of real sequences more faithfully than homogeneous ones[6,17], although neither homogeneous nor heterogeneous models ensure that inferred ancestral sequences are biologically functional. Using *nhPhyML* and *nhPhyloBayes*, we can, for the first time, reconstruct ancestral sequences of both rRNAs and proteins with branch-heterogeneous models, and estimate sequence compositions of all nodes of the tree of life, including LUCA and its descendants. These compositions can be translated into approximate OGTs using the OGT/composition correlations observed in extant sequences[12,15].

A nucleotide dataset of concatenated small and large subunit rRNAs, restricted to double-stranded regions, from 456 organisms (1,043 sites), and an amino acid dataset of 56 concatenated nearly universal proteins from 30 organisms (3,336 sites) were assembled, each dataset sampling all forms of cellular life. Correspondence analyses of the protein dataset show that eukaryotes and prokaryotes markedly differ in amino acid compositions and that an effect of temperature on proteomes is detectable only among prokaryotic species (Supplementary Figs. 4, 6B). Similarly, the correlation between rRNA G+C content and OGT has only been documented in prokaryotes[12]. The ability to infer ancestral OGTs from rRNA and protein compositions therefore applies only to prokaryotes. However, eukaryotic sequences were kept in the subsequent analyses because they are part of the tree of life and as such provide useful phylogenetic information for ancestral sequence inferences.
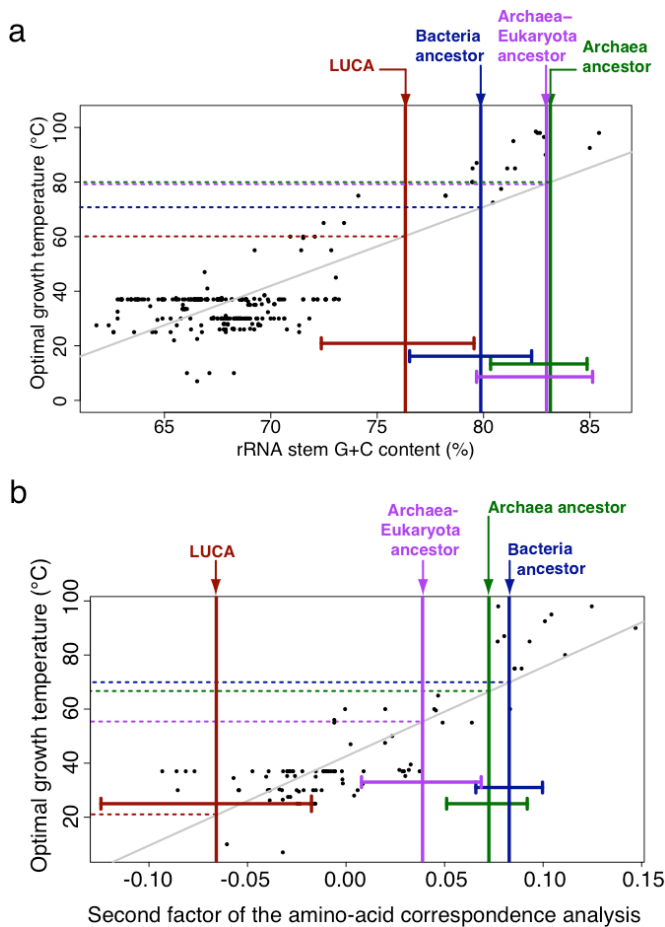
The effect of temperature on prokaryotic proteomes is independent from genomic G+C contents[15], and was summarized in terms of average content in the amino acids IVYWREL. Accordingly, our correspondence analysis identifies two independent factors accounting for most of the variance in amino acid compositions of prokaryotic proteins (Supplementary Fig. 5): the first factor (45.4% of the variance) highly correlates to genome G+C content (r=0.81); the second (13.8% of the variance) is strongly correlated to OGT (r=0.83) and to IVYWREL content (r=0.73, Supplementary Fig. 6). The second factor was therefore used here as a molecular thermometer. The rRNA-based and the protein-based thermometers are thus independent, both because they come from distinct genome parts and because they exploit different effects of temperature on sequence composition. Furthermore, the correlation between rRNA G+C content and OGT is not expected to vary during evolutionary time because it stems from the different thermal stabilities of G-C and A-U RNA base pairs[12]. Thus, assuming the relationship between temperature and amino acid composition of prokaryotes has not varied either since LUCA, the estimations of rRNA G+C content and amino acid compositions through branch-heterogeneous models provide two independent means to analyze the evolution of thermophily.

For each dataset, a phylogenetic tree was inferred and rooted on the branch separating Bacteria from Archaea and Eukaryota (Sup-

[1]Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université de Lyon, Université Lyon I, 43 Boulevard du 11 Novembre, 69622 Villeurbanne, France. [2]LIRMM, CNRS, 161 rue Ada, 34392 Montpellier, France.
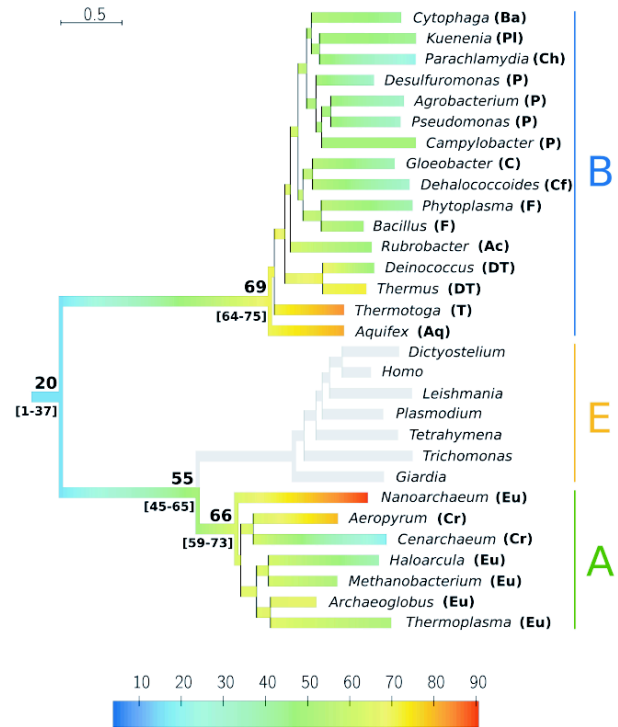*These authors contributed equally to this work

plementary Figs. 7, 8). Because the location of the root in the universal tree remains uncertain[19], the alternative rooting on the eukaryotic branch was also considered. Correlations between G+C content and OGT (Fig. 1a), and between the second axis of the amino acid correspondence analysis and OGT (Fig. 1b), were used to estimate OGTs for LUCA and its progeny (Fig. 2).

Support for the hypothesis of a non-hyperthermophilic LUCA and of subsequent parallel adaptations to high temperatures partly rests on a protein content depleted in IVYWREL for LUCA and subsequently enriched in these amino acids. This is consistent with a recent report that amino acids IVYEW might be under-represented in LUCA's proteins[20]. This finding has been interpreted as evidence that these five amino acids were a late addition to the genetic code, and that LUCA's proteome had not yet reached compositional equilibrium. Although such interpretation in terms of early genetic code evolution is possible, our hypothesis of parallel adaptations to high temperatures has the advantage of explaining the patterns observed with both rRNAs and proteins.

Figure 1 | Correlations between sequence compositions and OGT, and estimates of key ancestral compositions. Black dots: extant prokaryotes positioned according to their sequence composition and OGT. Dashed coloured lines: predicted OGTs for various ancestors. a: Correlation between rRNA G+C content and OGT. Vertical coloured bars: most likely *nhPhyML* estimates of ancestral G+C contents with their 95% confidence intervals. b: Correlation between the second factor of the correspondence analysis on amino acid compositions and OGT. Vertical coloured bars: median ancestral compositions inferred by *nhPhyloBayes* with their 95% confidence intervals. LUCA is significantly less thermophilic than its direct descendants (p-values ≤ 0.005).

Figure 2 | Evolution of thermophily over the tree of life. Protein-derived *nhPhyloBayes* OGT estimates (and their 95% confidence intervals for key ancestors) for prokaryotic organisms are colour-coded from blue to red for low to high temperatures. Colours were interpolated between temperatures estimated at nodes. The eukaryotic domain, in which OGT cannot be estimated, has been shaded. Colour scale in °C; branch length scale in substitutions/site. A: Archaeal, B: Bacterial, E: Eukaryotic domains. Ac: Actinobacteria; Aq: Aquificae; Ba: Bacteroidetes; C: Cyanobacteria; Cf: Chloroflexi; Ch: Chlamydiae; Cr: Crenarchaeota; DT: Deinococcus/Thermus; Eu: Euryarchaeota; F: Firmicutes; P: Proteobacteria; Pl: Planctomycetes; T:Thermotogae.

Proteins and rRNAs support similar patterns of OGT changes for prokaryotes, so the discrepancy between previous rRNA- and protein-based investigations[2,13,14] was not due to incongruence between these molecules. Protein-derived temperature estimates are generally lower than those based on rRNAs (Fig. 1), although some protein and rRNA-based OGT estimates overlap if confidence intervals of ancestral compositions are taken into account (Supplementary Table 3). Both types of data support key conclusions (Fig. 1). Firstly, as Galtier *et al.* reported, LUCA is predicted to be a non-hyperthermophilic organism[2]. Secondly, both archaeal and bacterial ancestors as well as the common ancestor of Archaea and Eukaryota are estimated to have been thermophilic to hyperthermophilic (Fig. 2). This result is in line with those of Gaucher *et al.*[3] and Gribaldo *et al.*[5]. Thirdly, within the bacterial tree, tolerance to heat decreased (Fig. 2). This last result is congruent with recent estimates of the evolution of optimal growth temperatures in the bacterial domain based on ancestral reconstructions and characterizations of Elongation Factor-Tu proteins[4].

Additional experiments suggest that the present analyses of rRNA and protein sequences with branch-heterogeneous models of evolution uncover genuine signals of ancient temperature preferences and are not affected by systematic biases.

First, these results are robust to changes in the topology chosen for inference because analyses with alternative topologies yielded virtually identical OGT estimates (Supplementary Fig. 10). Moreover, trees rooted on the eukaryotic branch also suggest that OGT increased between the universal ancestor and the divergence of Archaea and Bacteria (Supplementary Figs. 13-15).

Second, taxonomic sampling does not strongly affect these results. With rRNA and protein datasets in which eukaryotic sequences were removed, the signal for OGT increases between LUCA and the domain ancestors was essentially unchanged (Supplementary Fig. 36). Moreover, both for rRNAs and proteins, two artificially biased datasets containing sequences from either thermophilic or mesophilic prokaryotes were assembled (see Supplementary Information). The signal for parallel increases in OGT is confirmed in all but one of these four datasets, the mesophilic rRNA

dataset. However, the longest of the two mesophilic alignments, the protein dataset, supports the same pattern of OGT changes as the complete datasets (Supplementary Figs. 16, 17). Importantly, analysis of the protein mesophilic dataset shows that this pattern is independent from the debated position of hyperthermophilic organisms in the tree of life. Furthermore, with all rRNA and protein datasets, even with the sampling limited to thermophilic prokaryotes, LUCA remains predicted as a non-hyperthermophilic organism (Supplementary Figs. 18, 19).

Third, dependence of the results upon models used for ancestral reconstruction was investigated. Additional branch-heterogeneous evolutionary models were applied, two to the rRNA dataset, and one to the protein dataset (see Supplementary Information). All these alternative branch-heterogeneous models confirm our results (Supplementary Figs. 21-23, 29, 30). Compositional analyses were also conducted using branch-homogeneous models of evolution, GTR[21] for rRNA and proteins, and CAT[18] for proteins. Interestingly, all these models tend to predict parallel adaptations to higher temperatures from LUCA to its descendants, suggesting the existence of a genuine signal for such a pattern in the data (Supplementary Figs. 24, 26, 28). However, only when models are realistic enough is LUCA predicted as significantly less thermophilic than its two descendants. For instance, ancestral protein compositions predicted by the GTR model for LUCA and its two descendants strongly overlap, which may explain previously published results[13], while the CAT model better separates these ancestral node distributions, though less clearly than does the CAT-BP branch-heterogeneous model (Supplementary Figs. 26, 28, 29). These experiments show that as the evolutionary process is more accurately modelled, the support for parallel increases in OGT from LUCA to its offspring is strengthened.

Fourth, it is known that the base compositions of fast and slowly evolving sites and, particularly, of single and double-stranded regions of rRNA molecules differ and that this may bias ancestral sequence estimates[16]. To minimize this bias, only double-stranded rRNA regions have been analyzed here. Moreover, if fast-evolving sites are removed, estimates still support parallel adaptations to high temperatures (Supplementary Fig. 33).

Fifth, it has been shown that some ancestral reconstruction methods might improperly estimate the frequencies of rare amino acids[22]. To control for that potential bias, the two rarest amino acids, cystein and tryptophan, were discarded from estimated ancestral sequences: this had essentially no impact on results (Supplementary Fig. 34).

Sixth, the sensitivity, in Bayesian analyses, of the OGT estimates at the tree root to the prior distribution of ancestral amino acid compositions was investigated (Supplementary Fig. 35). This prior distribution induces a flat, uninformative distribution over OGTs, whereas the posterior distributions estimated for LUCA and the bacterial ancestor have small variance, and thus, reflect a genuine signal in the data rather than a bias from the prior. Moreover, even with a strongly informative prior distribution that is biased towards high temperature amino acid distributions, the posterior distribution of LUCA's amino acid composition, although altered, is centred at lower temperatures than that of the bacterial ancestor.

The present use of molecular thermometers requires that evolution of the datasets under analysis can be modelled by a tree structure as far as reconstruction of ancestral compositions is concerned. We emphasize that our protein analyses are based on 56 genes that did not undergo between-domain transfers (see Methods), which precludes that ancestral sequence reconstructions are confounded by such gene exchanges. We do not exclude within-domain lateral transfers of these genes; however, the robustness of the inferred ancestral compositions to alternative domain phylogenies (Refs. 4,7 and Supplementary Figs. 10, 20) suggests that these potential transfers do not fundamentally affect the results for domain ancestors. Finally, because molecular thermometers measure the average environmental temperature of the hosts of ancestral genes, they apply even if ancestral genes of extant prokaryotes originate from diverse organisms[19].

Thus, all our analyses support the hypothesis of a non-hyperthermophilic LUCA and of transitions to higher environmental temperatures for its offspring. Although these organisms have not been anchored in time yet[23], a few geological and biological factors may explain observed changes in temperature preferences. It has already been observed[4] that the general trend of decreasing OGTs from the bacterial ancestor to extant species strikingly parallels recent geological estimates of the progressive cooling down of oceans shifting from about 70 °C 3.5 billion years ago to approximately 10 °C presently[24]. The evolution of thermophily in the bacterial domain might therefore stem from the continuous adjustment of Bacteria to ocean temperatures, although the evidence for hot Archean climate remains debated[25]. A similar conclusion may apply to Archaea as well, but would require confirmation with additional genome sequences from mesophilic archaea. A hot Archean ocean may preclude the existence of a cool 'little pond' where LUCA could have evolved. Therefore, a non-hyperthermophilic LUCA would suggest that moderate temperatures existed earlier in the history of the Earth.

Geological data about paleoclimates that old are very scarce. However, some models of Hadean and early Archean climates (3.5 to 4.2 billion years ago) suggest that the Earth might have been colder than it is today, possibly covered with frozen oceans[1,26]. Moreover, a hypothesis of brutal temperature changes involving meteoritic impacts that boiled the oceans and therefore nearly annihilated all life forms but the most heat-resistant ones has been proposed[1,8,9]. Huge meteorites probably impacted the Earth at least as late as 3.8-4 billion years ago, most notably during the Late Heavy Bombardment[27] and created a series of brief but very hot climates on Earth[1]. As life may have originated more than 3.7 billion years ago[28], it is possible that early organisms, namely LUCA's offspring, experienced such bottlenecks.

Alternatively, under the hypothesis that life originated extra-terrestrially, the transfer of life to the Earth from another planet in ejecta created by meteorite impacts would have also entailed selection of heat resistant cells[1]. Overall, geological knowledge provides several frames that might fit the predictions of our biological thermometers.

Global considerations put aside, a biological hypothesis could provide an internal mechanism to explain the observed pattern. It posits that LUCA had an RNA genome, and that its offspring lineages independently evolved the ability to use DNA for genome encoding[10], possibly by co-opting it from viruses[11]. Although our results do not bring direct evidence in support of this hypothesis, they are compatible with it and could even help explaining such independent acquisitions of DNA in adaptive terms, as DNA is much more thermostable than RNA[29].

Great care is necessary when attempting reconstruction of events that took place more than three billion years ago. However, the strong agreement between results obtained using two types of data (proteins and rRNAs), two independent temperature proxies (protein amino acid composition and rRNA G+C content), and independently developed statistical models, is remarkable. This suggests that a similar approach could successfully be used to gain insight into other ecological features of early life. For example, it has been shown that aerobic and anaerobic bacteria differ in the amino acid composition of their proteome[30]; future ancestral sequence reconstructions could reveal the evolution of aerobiosis along the tree of life in relation with the geological record of oxygen atmospheric concentration.

METHODS SUMMARY

Ribosomal RNA sequences were aligned according to their shared secondary structure. Sites belonging to double-stranded stems were selected to obtain an alignment of 1,043 stem sites for 456 organisms. Protein families with wide species coverage and no or very low redundancy in all species were selected from the HOGENOM database of families of homologous genes. Only sites showing less than 5% gaps were kept, giving an alignment of 3,336 positions for 30 organisms. Phylogenetic trees were inferred using Bayesian or maximum likelihood techniques. Ancestral nucleotide and amino acid compositions were inferred for all tree nodes using the programs *nhPhyML*[7] and *nhPhyloBayes*[6], respectively. The G+C contents of ancestral rRNA sequences were compared to extant rRNA base compositions. The second fac-

tor of the correspondence analysis of amino acid compositions of extant pro-karyotic proteins was used to estimate ancestral environmental temperatures by adding ancestral amino acid compositions as supplementary rows to the correspondence analysis. These two procedures allowed us to estimate ancestral environmental temperatures with the rRNA and the protein data sets, respectively. Confidence intervals for the estimated environmental temperatures were as follows: in the case of rRNAs, they contained 95% of the distribution obtained by a bootstrap procedure (200 replicates); for Bayesian analyses, regular 95% credibility intervals were computed from a sample of 2,000 points drawn from the posterior distribution.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409**, 1083–1091 (2001).
2. Galtier, N., Tourasse, N. & Gouy, M. A nonhyperthermo-philic common ancestor to extant life forms. *Science* **283**, 220–221 (1999).
3. Gaucher, E. A., Thomson, J. M., Burgan, M. F. & Benner, S. A. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**, 285–288 (2003).
4. Gaucher, E. A., Govindara jan, S. & Ganesh, O. K. Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature* **451**, 704–707 (2008).
5. Gribaldo, S. & Brochier-Armanet, C. The origin and evolution of archaea: a state of the art. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1007–1022 (2006).
6. Blanquart, S. & Lartillot, N. A site- and time-heterogeneous model of amino-acid replacement. *Mol. Biol. Evol.* **25**, 842–858 (2008).
7. Boussau, B. & Gouy, M. Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* **55**, 756–768 (2006).
8. Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* **342**, 139–142 (1989).
9. Gogarten-Boekels, M., Hilario, E. & Gogarten, J. P. The effects of heavy meteorite bombardment on the early evolution–the emergence of the three domains of life. *Orig. Life Evol. Biosph.* **25**, 251–264 (1995).
10. Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10268–10273 (1996).
11. Forterre, P. The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* **5**, 525–532 (2002).
12. Galtier, N. & Lobry, J. R. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44**, 632–636 (1997).
13. DiGiulio, M. The universal ancestor and the ancestor of bacteria were hyperthermophiles. *J. Mol. Evol.* **57**, 721–730 (2003).
14. Brooks, D. J., Fresco, J. R. & Singh, M. A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. *Bioinformatics* **20**, 2251–2257 (2004).
15. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3**, 62–72 (2007).
16. Gowri-Shankar, V. & Rattray, M. On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol. Biol. Evol.* **23**, 352–364 (2006).
17. Blanquart, S. & Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23**, 2058–2071 (2006).
18. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
19. Zhaxybayeva, O., Lapierre, P. & Gogarten, J. P. Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **277**, 53–64 (2005).
20. Fournier, G. P. & Gogarten, J. P. Signature of a primitive genetic code in ancient protein lineages. *J. Mol. Evol.* **65**, 425–436 (2007).
21. Lanave, C., Preparata, G., Saccone, C. & Serio, G. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**, 86–93 (1984).
22. Williams, P. D., Pollock, D. D., Blackburne, B. P. & Goldstein, R. A. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput. Biol.* **2**, 598–605 (2006).
23. Graur, D. & Martin, W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* **20**, 242–247 (2004).
24. Robert, F. & Chaussidon, M. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature* **443**, 969–972 (2006).
25. Shields, G. A. & Kasting, J. F. Evidence for hot early oceans? *Nature* **447**, E1 (2007).
26. Kasting, J. F. & Ono, S. Palaeoclimates: the first two billion years. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 917–929 (2006).
27. Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature* **435**, 466–469 (2005).
28. Rosing, M. T. ^{13}C-depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from West Greenland. *Science* **283**, 674–676 (1999).
29. Islas, S., Velasco, A. M., Becerra, A., Delaye, L. & Lazcano, A. Hyperthermophily and the origin and earliest evolution of life. *Int. Microbiol.* **6**, 87–94 (2003).
30. Naya, H., Romero, H., Zavala, A., Alvarez, B. & Musto, H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.* **55**, 260–264 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions B.B. and S.B. contributed equally to this study, designing and conducting experiments. A.N. performed statistical analyses and retrieved optimal growth temperatures. N.L. and M.G. provided guidance throughout the study, and M.G. gave the original idea. All authors participated in manuscript writing.

METHODS

rRNA dataset. Prokaryotic small (SSU) and large (LSU) subunit rRNAs were retrieved in January 2007 from complete genomes available at the National Center for Biotechnology Information (NCBI). SSU and LSU rRNA sequences from ongoing genome projects or from large genomic fragments of important or poorly represented groups (e.g., Archaea or hyperthermophilic bacteria) were added in June 2007. Eukaryotic SSU and LSU rRNA sequences were kindly provided by David Moreira; 65 slowly evolving sequences were selected from this dataset[31]. Sequences were aligned using MUSCLE[32]. Resulting alignments were concatenated and manually improved using the MUST package[33]. Regions of doubtful alignment were removed using the MUST package; 2,239 sites were kept. A distance phylogenetic tree was computed using dnadist (Jukes and Cantor model) and neighbor from the PHYLIP package[34]. The final dataset contained 65 eukaryotic, 60 archaeal and 331 bacterial sequences representative of the molecular diversity in each domain. An additional dataset of 60 sequences sampling the diversity of the full dataset was used in Bayesian analyses. Secondary structure predictions were downloaded from the rRNA database[35]. Sites that were predicted as double-stranded stems in *Saccharomyces cerevisiae*, *Escherichia coli* and *Archaeoglobus fulgidus* were selected to give an alignment of 1,043 sites.

Protein dataset. Nearly universal protein families with one member per genome were used to avoid ill-defined orthology. Protein families from the HOGENOM database of families of homologous genes (release 03, October 2005, Penel S, Duret L, personal communication, http://pbil.univ-lyon1.fr/databases/hogenom3.html) that displayed a wide species coverage with no or very low redundancy in all species were selected. Additional sequences from other genomes whose phylogenetic position was interesting were considered. These were downloaded from the Joint Genome Institute (*Desulfuromonas acetoxidans*), The Institute for Genomic Research (*Giardia lamblia, Tetrahymena thermophila, Trichomonas vaginalis*) or the NCBI (*Kuenenia stuttgartiensis*), and were searched for homologous genes using BLAST[36]; only the best hit was retrieved. The protein families were subsequently aligned using MUSCLE[32] and submitted to phylogenetic analysis using the NJ algorithm[37] with Poisson distances with Phylo_Win[38]. Proteins from mitochondrial or chloroplastic symbioses and families in which horizontal transfers between Bacteria and Archaea may have occurred were discarded, and so were aminoacyl-tRNA synthetases prone to transfers[39]. In the rare families with two sequences from the same species, the sequence showing the longest terminal branch or whose position was most at odds with the biological classification was discarded. This provided 56 protein families (Supplementary Table 2) for 115 species, which were concatenated using ScaFos[40]. From the 9,218 concatenated sites, 3,336 positions with less than 5% gaps were conserved. The whole dataset was used to compute the correspondence analysis and correlations between amino-acid composition and optimal growth temperature. For Bayesian analyses, 30 species among 115 were selected sampling the diversity of cellular life (Supplementary Table 1).

Multivariate data analyses. Correspondence analysis[41] was performed on the amino-acid compositions of the protein dataset, using the ade4 package[42] of the R environment for statistical computing.

Phylogenetic tree construction. An rRNA phylogenetic tree was built from the 456-sequence alignment with both stems and loops with PhyML_aLRT [43,44] with the GTR model, a gamma law with 8 categories and an estimated proportion of invariant sites. The tree for the 60 sequence dataset was obtained in the same manner. The phylogenetic trees for the three protein datasets (Supplementary Table 1) were obtained using MrBayes 3.1.1[45], using the GTR substitution model and a gamma law with 4 categories for rates across sites. Chains were run for 1,000,000 generations and samples were collected each 100 generations, a burn-in of 1,000 samples was discarded. The majority rule consensus was computed from the 9,000 remaining samples.

Identification of fast-evolving rRNA sites. Posterior probabilities for gamma law rate categories were predicted for each site with PhyML_aLRT. Site evolutionary rates were obtained by averaging gamma law rate categories weighted by their posterior probabilities. Sites whose evolutionary rate was above the arbitrarily chosen threshold of 2.0 (Supplementary Fig. 2) were discarded, which left 940 sites.

Estimation of ancestral sequence compositions. For the maximum likelihood approach, *nhPhyML*[7] was applied to the rRNA stem sites alignment and the phylogenetic tree described above, and used to estimate all evolutionary parameter values, except tree topology, which was fixed. Site-specific ancestral nucleotide compositions at tree root and at internal node j descendant of node i were computed by:

$$p_{root}(x) = a(x) \, L_{low}(x \text{ at root}) / L; \quad a(A) = a(T) = (1 - \omega)/2; \quad a(C) = a(G) = \omega/2$$
$$p_j(x) = (\, \Sigma_y L_{upp}(y \text{ at node } i) \, p_{y \to x} \, L_{low}(x \text{ at node } j) \,) / L$$

where x and y are in {A, C, G, T}, L is the total tree likelihood at this site, L_{low} , L_{upp} are site lower and upper conditional likelihoods[7], ω is the maximum likelihood estimate of root G+C content and $p_{y \to x}$ is the probability of the y to x substitution on the i to j branch. For Bayesian analyses, *nhPhy-*

loBayes[6] was applied to trees described above. Ancestral sequence reconstruction started, for each site, by drawing a state x at the root: $x \sim \omega(x) \, L_{low}(x \text{ at root})$, where ω was the Markov Chain Monte Carlo[45] (MCMC) estimate of root amino acid or nucleotide frequencies. Then, states x have been recursively drawn at each node j: $x \sim p_{y \to x} \, L_{low}(x \text{ at } j)$, where y was the parental node state. Given a realization of the model, this permitted to reconstruct ancestral sequences at all nodes. Posterior distributions were sampled by 2, for proteins, or 4, for rRNA, independent MCMC chains, each with 1,000 to 2,000 realizations. Posterior distributions of sequence compositions combined all realizations of all chains. Protein ancestral compositions were projected on the second axis of the correspondence analysis, and rRNA ancestral compositions were summed up as G+C contents.

Statistical tests. In bootstrap analyses, all parameters but topology and branch lengths were estimated under the maximum likelihood criterion for each replicate. In tests of whether LUCA is less thermophilic than one of its descendants, p-values were the fraction of cases where the temperature estimate for LUCA in a bootstrap replicate or in an iteration of an MCMC chain was above the estimate obtained for its descendant.

31. Moreira, D. *et al.* Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. *Mol. Phylogenet. Evol.* **44**, 255–266 (2007).

32. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

33. Philippe, H. MUST, a computer package of management utilities for sequences and trees. *Nucleic Acids Res* **21**, 5264–5272 (1993).

34. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. (2005).

35. Wuyts, J., Perrière, G. & Van De Peer, Y. The european ribosomal RNA database. *Nucleic Acids Res* **32**, D101–D103 (2004).

36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

37. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425 (1987).

38. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12**, 543–548 (1996).

39. Wolf, Y. I., Aravind, L., Grishin, N. V. & Koonin, E. V. Evolution of aminoacyl-trna synthetases–analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* **9**, 689–710 (1999).

40. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol* **7 Suppl 1**, S2 (2007).

41. Hill, M. O. Correspondence analysis: a neglected multivariate method. *Appl. Stat.* **23**, 340-354 (1974).

42. Chessel, D., Dufour, A. B. & Thioulouse, J. The ade4 package -I- one-table methods. *R News* **4**, 5–10 (2004).

43. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).

44. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552 (2006).

45. Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).