Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea

Mathieu Groussin*[,1] and Manolo Gouy[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université de Lyon, Université Lyon I, Villeurbanne, France

***Corresponding author:** E-mail: mathieu.groussin@etu.univ-lyon1.fr.

Associate editor: Hervé Philippe

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

Research article

Abstract

Methods to infer the ancestral conditions of life are commonly based on geological and paleontological analyses. Recently, several studies used genome sequences to gain information about past ecological conditions taking advantage of the property that the G+C and amino acid contents of bacterial and archaeal ribosomal DNA genes and proteins, respectively, are strongly influenced by the environmental temperature. The adaptation to optimal growth temperature (OGT) since the Last Universal Common Ancestor (LUCA) over the universal tree of life was examined, and it was concluded that LUCA was likely to have been a mesophilic organism and that a parallel adaptation to high temperature occurred independently along the two lineages leading to the ancestors of Bacteria on one side and of Archaea and Eukarya on the other side. Here, we focus on Archaea to gain a precise view of the adaptation to OGT over time in this domain. It has been often proposed on the basis of indirect evidence that the last archaeal common ancestor was a hyperthermophilic organism. Moreover, many results showed the influence of environmental temperature on the evolutionary dynamics of archaeal genomes: Thermophilic organisms generally display lower evolutionary rates than mesophiles. However, to our knowledge, no study tried to explain the differences of evolutionary rates for the entire archaeal domain and to investigate the evolution of substitution rates over time. A comprehensive archaeal phylogeny and a non homogeneous model of the molecular evolutionary process allowed us to estimate ancestral base and amino acid compositions and OGTs at each internal node of the archaeal phylogenetic tree. The last archaeal common ancestor is predicted to have been hyperthermophilic and adaptations to cooler environments can be observed for extant mesophilic species. Furthermore, mesophilic species present both long branches and high variation of nucleotide and amino acid compositions since the last archaeal common ancestor. The increase of substitution rates observed in mesophilic lineages along all their branches can be interpreted as an ongoing adaptation to colder temperatures and to new metabolisms. We conclude that environmental temperature is a major factor that governs evolutionary rates in Archaea.

Key words: Archaea, evolutionary rates, optimal growth temperature, ancestral sequence reconstruction, nonhomogeneous models.

Introduction

Bacteria and Archaea show adaptations to many kinds of environments and especially to a wide range of temperatures. Several recent studies have attempted to reconstruct ancestral environmental temperatures using molecular sequence data (Galtier et al. 1999; Boussau and Gouy 2006; Boussau et al. 2008; Gaucher et al. 2008). These analyses exploited the signal left by environmental temperature on both extant and ancestral sequences in terms of base and amino acid compositions. Boussau et al. (2008) studied the evolution of thermophily along the universal tree of life using two molecular thermometers based on the compositions of ribosomal RNA (rRNA) and protein sequences. They concluded firstly that LUCA (the Last Universal Common Ancestor) lived at low temperatures, secondly that parallel adaptations to high temperatures occurred from LUCA to the last common ancestor of Bacteria and to that of Archaea, and thirdly that optimal growth temperatures (OGTs) decreased with time in the bacterial domain. These results were obtained with 30 organisms, among which only seven archaeal species. This limitation prevented a precise study of the evolution of OGT in the archaeal domain. Presently available fully sequenced archaeal genomes give the opportunity to investigate in greater detail the evolutionary history of OGT in this domain.

It has long been observed that thermophilic lineages tend to have shorter branches in archaeal phylogenetic trees than do mesophilic lineages (Stetter 2006). Several factors have been proposed to explain why molecular evolutionary rates vary between organisms (Bromham 2009). In vertebrates, the generation time is critical in the determination of evolutionary rates (Bromham et al. 1996). In mammals, it has been shown that population size, body size, and metabolic rates are probably involved in shaping molecular evolutionary rates (Bromham 2009). Concerning Archaea and Bacteria, few factors are known to explain the differences of evolutionary rates between species. Nevertheless, it seems clear that for all species, and particularly in Bacteria where it has been shown, the efficiency of DNA replication and DNA repair machineries is under selection and determines substitution rates (Denamur and Matic 2006). Archaeal and bacterial species are considerably sensitive to the variations of their environment and to

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

the variations of mutagen concentrations (e.g., UV, temperature) (Foster 2007). Thus, Valentine (2007) proposed that chronic energy stress, from a metabolic and thermodynamic point of view, is the major selective pressure that governs evolutionary rates in Archaea. A physical factor that could cause such energy stress is environmental temperature. Thus, previous studies showed that thermophilic species are characterized by a stronger purifying selection than mesophiles. Indeed, Friedman et al. (2004) showed that thermophiles display a lower ratio of nonsynonymous to synonymous substitutions than mesophiles. This is consistent with the idea that proteins of species living in hot environments are more functionally constrained (Vetriani et al. 1998). Drake suggested that thermophiles exhibit very low genomic mutation rates and that this phenomenon could be explained by an adaptation to avoid deleterious mutations at high temperatures (Drake 2009). However, most studies that focused on mutational and evolutionary rates in Archaea or in thermophiles were restricted to few species (Grogan et al. 2001; Friedman et al. 2004; Mackwan et al. 2007, 2008; Drake 2009) making it impossible to have a vision at the scale of the entire domain. In this work, we attempt to reconstruct the evolutionary history of environmental temperatures at the level of the entire archaeal domain and investigate whether there is evidence that evolutionary rates are constrained by environmental temperatures.

The reconstruction of ancestral environmental temperatures and evolutionary rates using extant molecular data requires statistical models of the molecular evolutionary process and a phylogenetic tree of the organisms under study. Phylogenetic relationships between all archaeal species remain debated. Two major phyla have long been recognized (Gribaldo and Brochier-Armanet 2006): Euryarchaea, which is composed of thermoacidophiles, methanogens, extreme halophiles, and a few hyperthermophiles, and Crenarchaea, which were believed to be restricted to hyperthermophiles until mesophilic crenarchaeal species were discovered (DeLong 1992). These mesophilic species were grouped with Crenarchaea on the basis of 16S rRNA phylogenies. Brochier-Armanet et al. (2008) recently questioned the dichotomy between Euryarchaea and Crenarchaea in an analysis of *Cenarchaeum symbiosum*, the first mesophilic crenarchaeon entirely sequenced. They proposed that this group of mesophilic organisms should not be considered as Crenarchaea but rather as a third phylum, named Thaumarchaea, which diverged first in the archaeal tree. However, this conclusion remains uncertain. The evolutionary origins of other archaeal species are also unresolved, for example, that of the recently sequenced *Candidatus Korarchaeum cryptofilum* (Elkins et al. 2008).

We used here nonhomogeneous evolutionary models, which have been shown to be more realistic than homogeneous models (Dutheil and Boussau 2008). These models were used to infer base and amino acid compositions at each ancestral node of the archaeal tree. Through the use of appropriate molecular thermometers, these compositions permitted us to deduce OGT along the tree. We inferred that the ancestors of Archaea, Crenarchaea, and Euryarchaea were hyperthermophiles, and therefore, that ancestral archaeal species were adapted to hot environments. Furthermore, a strong relationship between environmental temperature and molecular evolutionary rates in Archaea has been identified. This implies that environmental temperature has been a major determinant of evolutionary rates in the archaeal domain.

Materials and Methods

Data Retrieval

Thirty-five completely sequenced archaeal genomes were selected for the phylogenetic studies to represent all known archaeal biodiversity. However, all 56 genomes available in GenBank as of February 2009 were used to construct the rRNA and protein data sets (see rRNA and Protein Data Sets). Protein sequences were downloaded from the Hogenom database (Penel et al. 2009) when possible. The genomes of *Ignicoccus hospitalis*, *Desulfurococcus kamchatkensis*, *Metallosphaera sedula*, *Caldivirga maquilingensis*, *Pyrobaculum arsenaticum*, *Thermoproteus neutrophilus*, *Halobacterium salinarum*, *Methanococcus voltae*, *Methanobrevibacter smithii*, *C. Korarchaeum cryptofilum* and *Nitrosopumilus maritimus* were retrieved from GenBank database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Two thaumarchaeal fosmids sequences, also extracted from GenBank, were added to the study to increase the diversity within this group: *uncultured crenarchaeote 74A4* and *uncultured crenarchaeote KM3-34-D9*. For these two mesophilic species, both small and large rRNA subunits were used in the rRNA data set, whereas only the elongation factor G protein for *KM3-34-D9* and the 30S ribosomal protein S10 for *74A4* were available and used in the protein data set.

All bacterial and eukaryal genomes were retrieved from the Hogenom database, with the exception of the *Giardia lamblia* genome, which was extracted from the GiardiaDB database (http://giardiadb.org/giardiadb/). The complete list of genomes with their origin is in supplementary table 1 (Supplementary Material online).

rRNA and Protein Data Sets

The rRNA and protein alignments were constructed as follows: the 56 species were used in order to improve the quality of the alignment with as much information as possible. Then, species not exploited in the following steps were removed from the final alignments, which contain 35 species. The small and the large subunits (SSUs and LSUs) of archaeal, bacterial, and eukaryal rRNAs were extracted from GenBank. Archaeal rRNA SSUs and LSUs were aligned separately with the Silva aligner (http://www.arb-silva.de/aligner/), which takes secondary structures into account. Bacterial rRNAs were aligned following the same procedure. Then, archaeal rRNAs were aligned with bacterial rRNAs, using the "profile alignment" function of the ClustalW program (Thompson et al. 2002). For the data sets containing the three domains

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

of life, archaeal and bacterial rRNAs were first aligned together and then aligned by profile against eukaryal rRNAs with ClustalW. Eukaryal 5.8S rRNAs were added upstream from large eukaryal subunits because they are homologous to the 5′ end of the large prokaryotic subunits (Nazar 1980). Then, SSUs and LSUs rRNAs were concatenated with ScaFos (Roure et al. 2007). Fast-evolving sites were subsequently removed from the alignment by the Gblocks program, with standard options, allowing gap positions (Castresana 2000). The final archaeal + bacterial rRNA data set contains 3,719 sites. With the three domains, Gblocks retained 3,629 sites. Protein gene families extracted from the Hogenom database (Penel et al. 2009) were selected with different selection criteria. First, universal and single-copy gene families for all 56 archaeal genomes were retrieved. Second, gene families that are universal and single-copy only for Euryarchaea or Crenarchaea were also selected. Gene families affected by "distant" horizontal gene transfers (HGTs) were removed from the selection, distant HGTs being defined by topologies of single-gene phylogenies that do not respect the monophyly of Crenarchaea and Euryarchaea, as presented in the consensus archaeal phylogeny of Brochier-Armanet et al. (2008). *Nanoarchaeum equitans*, *C. Korarchaeum*, and the two thaumarchaeal species were not taken into account in this approach as their position is highly controversial (Brochier-Armanet et al. 2008). As a result, 72 gene families were conserved for the archaeal + bacterial phylogenies and 68 gene families for the universal tree (supplementary table 2, Supplementary Material online). We stress here that the aim of this study is to investigate the evolution of the adaptation to OGT in Archaea at the compositional level and not to completely solve the archaeal phylogeny. Indeed, it is likely that many gene families retained for this analysis are affected by HGT, but we hypothesize that these HGT did not shape the long-term evolution of proteins at the compositional level. Each family was aligned by Muscle (Edgar 2004) and treated by Gblocks (Castresana 2000) with standard parameters and all gaps allowed. Overall, 9,799 and 8,598 sites were retained for the two-domain and three-domain alignments, respectively.

Phylogenetic Reconstructions

In Archaea, three taxa were defined a priori: Crenarchaea, Euryarchaea, and Thaumarchaea. The monophyly of these three phyla was strongly supported by the analysis of Brochier-Armanet et al. (2008). The TreeFinder program was used to resolve multifurcations within each of these taxa. As it is extremely difficult to place *N. equitans* (its genome is highly degenerated because of its parasitic way of life (Hubert et al. 2002; Forterre et al. 2009)), it was deliberately placed within Euryarchaea, based on previous results (Brochier et al. 2006). For rRNAs, the GTR+Γ_8+I model was used. Concerning proteins, the LG substitution model was employed (Le and Gascuel 2008) with a gamma law (four categories). No proportion of invariant was considered (this proportion was at first estimated and was revealed to be negligible). Bootstrap analysis was computed with PhyML (Guindon and Gascuel 2003) (100 replicates). To reduce risks of long-branch attraction, PhyML-CAT (Le et al. 2008) was used to confirm the protein results obtained with PhyML. We chose 20 profiles (model C20) and applied a gamma law (four categories).

Nonhomogeneous Models of Evolution

All nonhomogeneous experiments were carried out with BppML, belonging to the BppSuite of Programs (Dutheil and Boussau 2008). The following options were used: all sites were taken into account with no restriction on the percentage of gaps (maximum amount of allowed gaps of 100%) and all root frequencies were initially set to one per size of the alphabet (4 for RNA, 20 for proteins). A gamma law was added to all models that were tested, with eight and four categories for rRNAs and proteins, respectively. A proportion of invariants was also considered for rRNAs. We chose a simple likelihood recursion with a recursive site compression. All other options were set to default values. BppML allowed to estimate evolutionary parameters such as substitution and rate distribution parameters, ancestral frequencies, and branch lengths from the reference topology, which remains fixed. Different models of substitutions have been tested: T92, HKY85, and GTR models for the rRNA data set and JTT92, WAG, and LG models for the protein data set. The aim of this process was to fit as well as possible the compositional heterogeneity of the data set and to improve the estimation of evolutionary parameters (e.g., branch lengths). For the nonhomogeneous approach, we defined for each model several submodels in which parameters are either shared by the whole tree or assigned to one branch or to a specific group of branches. Three approaches have been used. The first assigned one substitution model per branch. The second approach assigned one substitution model to each phylum (Crenarchaea, Euryarchaea, Thaumarchaea, and Korarchaea), plus one to Bacteria (and one to Eukarya when present). The third approach considered again each phylum separately. However, inside each phylum, a further distinction between thermophiles and mesophiles has been added to the model. Concerning the universal tree, one specific model has been assigned to the GC-rich *G. lamblia* species.

The inference of ancestral rRNA and protein sequences at each node of the tree was performed by bppAncestor with previously computed parameters (Dutheil and Boussau 2008). Concerning rRNAs, BppML was first run with the whole alignment (3,719 sites or 3,629 for the two-domain or the three-domain alignment, respectively); the reconstruction of ancestral sequences was performed with an rRNA data set restricted to double-stranded regions. This second rRNA data set (1,801 sites or 1,142 sites for the two-domain or the three-domain alignment, respectively) was obtained by eliminating single-stranded sites manually with SeaView (Gouy et al. 2010). One hundred ancestral sequences for each node of the tree were inferred, and their average G+C content or amino acid composition were computed. Confidence

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

intervals (95%) of ancestral OGTs were computed following Boussau et al. (2008) with 100 bootstrap replicates.

Statistics

Statistical computations were performed using R (http://www.R-project.org). Multivariate analyses were realized using the ade4 package (Thioulouse et al. 1997). All correlation coefficients presented in this study are statistically different from zero. The phylogenetic independent contrasts (PICs) analysis was performed using the ape package (Paradis et al. 2004). Bowker's tests were computed with the R scripts made available by Ababneh et al. (2006) at http://www.maths.usyd.edu.au/u/johnr/testsym/.

Optimal Growth Temperatures

OGTs of Bacteria and Archaea were extracted from the German National Resource Centre for Biological Material (DSMZ, http://www.dsmz.de/). We referred to the literature for two bacteria, *Pseudomonas entomophila* and *Anaeromyxobacter dehalogenans*, because the DSMZ database does not provide such data (He and Sanford 2003; Hegan et al. 2007). Following Boussau et al. (2008), we defined three temperature classes: mesophiles when OGT $\leq$ 50 °C, thermophiles when OGT is between 50 °C and 80 °C, and hyperthermophiles when OGT $\geq$ 80 °C. The complete list of OGTs for all species is available in supplementary table 1 (Supplementary Material online).

Results

Domain-Scale Archaeal Phylogeny

Before inferring ancestral archaeal sequences and compositions to study the evolution of thermophily within Archaea, a phylogenetic reconstruction was carried out to obtain a reliable topology. Previous results (Brochier-Armanet et al. 2008; Elkins et al. 2008) revealed some uncertainty concerning the positions in the archaeal tree of key species in respect to the adaptation to OGT, like Thaumarchaea (originally named mesophilic crenarchaea) and the recently sequenced *C. Korarchaeum cryptophilum*. Cox et al. (2008) and Foster et al. (2009) recently questioned the monophyly of Archaea with phylogenetic results supporting the grouping of Eukarya and Crenarchaea to the exclusion of Euryarchaea. Lake et al. (1984) first proposed this hypothesis, known as the "eocyte" hypothesis. Therefore, Bacteria were chosen as the outgroup of Archaea to build rRNA and protein trees. Assuming the monophyly of Crenarchaea and Euryarchaea, we defined four major archaeal groups (Euryarchaea, Crenarchaea, Thaumarchaea, and Korarchaea)—because major topology ambiguities concern the positions of these phyla—and explored the 15 topologies defined by all possible arrangements between these groups (supplementary fig. 1, Supplementary Material online). We identified the best topology supported by RNA and protein data, based on the sum of rRNA and protein log-likelihoods and results of the expected-likelihood weights (ELW) statistical test (Strimmer and Rambaut 2002). Likelihoods for each possible topology were estimated with TreeFinder (Jobb et al. 2004), which optimized the tree within each predefined group.

Supplementary figure 1 (Supplementary Material online) summarizes the results and reveals that topology no. 14, where Thaumarchaea are a sister group of Korarchaea, both of them being a sister group of Crenarchaea, possesses the best maximum likelihood for both rRNAs and proteins. The ELW test cannot statistically rule out topologies no. 13 and 15. In these two topologies, the affinity between Thaumarchaea and Korarchaea disappears but Thaumarchaea never branch deeply in the tree. This maximum likelihood topology was also found using PhyML-CAT (Le et al. 2008) (which implements a rough approximation of the site-heterogeneous mixture model CAT (Lartillot and Philippe 2004)). The CAT model has been proven to be less sensitive to long-branch attraction (Lartillot et al. 2007) and confirmed the affinity between Thaumarchaea and Korarchaea. Several positions of Thaumarchaea have been proposed so far, for instance, a deep branching in the archaeal tree (Brochier-Armanet et al. 2008) or a branching within Crenarchaea (Elkins et al. 2008). Our results do not support these hypotheses but do not allow ruling them out because our protein data set is likely to be affected by HGT (see Material and Methods and Discussion). However, rRNA and protein phylogenies converge toward the same tree of the four predefined groups, and topologies within Crenarchaea and Euryarchaea are very similar for the rRNA and protein data sets. In order to check whether our results are robust to uncertainties in the phylogenetic tree of the archaeal domain, several alternative topologies were also used to infer ancestral OGTs (see The Influence of the Input Topology). To control whether the presence of eukaryotic sequences changes the inferences of ancestral OGTs in the archaeal domain (see below), the same approach was used to determine the best universal tree of life. Supplementary figure 1 (Supplementary Material online) shows that the eocyte topology (topology A), which clusters Eukarya and the association between Crenarchaea, Thaumarchaea, and Korarchaea, obtained the best maximum likelihood scores. This result is in agreement with recent propositions (Cox et al. 2008; Foster et al. 2009).

Nonhomogeneous Modeling of the Molecular Evolution of Archaea

The chosen topology (no. 14 in supplementary fig. 1, Supplementary Material online) was used as input tree to run nonhomogeneous models of evolution implemented in the BppML program (Dutheil and Boussau 2008). See Material and Methods for a full description of the homogeneous and nonhomogeneous models used. Table 1 sums up all the results obtained with the HKY85 and LG models (The results obtained with the T92, GTR, JTT92, and WAG models are shown in supplementary table 3, Supplementary Material online.). Usually, models that are more parameter rich will have a higher likelihood than a more restricted model (Felsenstein 2004). Here, as we have different models,

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

Table 1. Estimation of the Best Nonhomogeneous Model of Evolution for rRNAs and Proteins.

Substitution Model	Model Attribution	Parameters Shared in the Whole Tree	Number of Parameters	ln L	BIC values
HKY85 (rRNAs)	Homogeneous	All	7	−69578	139188.9
	Per Branch	None	355	−67866	138650.5
		θ^a/κ^b	105	−69361	139585.2
		κ	268	−68089	138381.3
		θ	268	−68955.7	140114.7
		$\theta_1{}^c/\theta_2{}^d$	181	−67981.7	137451.4
		$\kappa/\theta_1/\theta_2$	**94**	**−68196.9**	**137166.6**
		κ/θ_1	181	−68147.1	137782.2
		κ/θ_2	181	−68138.1	137764.2
	Per Phylum	None	27	−69077.1	138376.2
		κ	22	−69131	138442.9
		$\kappa/\theta_1/\theta_2$	12	−69140.6	138379.9
		κ/θ_1	17	−69133.6	138407
		κ/θ_2	17	−69138	138415.8
	Per group of extant species sharing similar OGT[e]	None	59	−68484.6	137454.3
		κ	46	−68559.5	137497.2
		$\kappa/\theta_1/\theta_2$	20	−68596.9	136358.2
		κ/θ_1	33	−68578.9	137429.1
		κ/θ_2	33	−68575.1	137421.5
LG + F (Proteins)	Homogeneous	—	20	−447106	894395.8
	Per Phylum	—	134	−446059	893349.5
	Per group of extant species sharing similar OGT[e]	—	324	−444820	892617.7

[a] Equilibrium G+C content ($\pi G+\pi C$).
[b] Transition/transversion ratio.
[c] $\theta_1 = \pi A/(\pi A+\pi T)$.
[d] $\theta_2 = \pi G/(\pi G+\pi C)$.
[e] Within a phylum, a further distinction is made between mesophilic and thermophilic species for the attribution of the sets of equilibrium frequencies. The best model for rRNAs and proteins is in bold characters.

selecting the model that has the highest likelihood could lead to the choice of an unreasonably complex model. Thus, bayesian information criterion (BIC) tests have been carried out for each model to balance the effect of the number of parameters on the final likelihoods. BIC has been preferred to akaike information criterion because it penalizes more parameter-rich models (Ripplinger and Sullivan 2008), as occurs in this study. For rRNAs, the best BIC score is attained by the HKY85 model with κ (transition/transversion ratio), θ_1 ($= \pi\tilde{A}\pi\underset{\cdot}{A}\pi\underset{\cdot}{T}$), and θ_2 ($= \pi G/(\pi G + \pi C)$) shared in the whole tree and one θ ($= \pi G + \pi C$) per branch. In the rest of the study, this model will be referred to as HKY85θpb (HKY85 with one θ per branch). It suggests that the HKY85θpb model represents a good compromise between the number of parameters and the ability to fit the data. Thus, homogeneous models do not fit properly the data because of their simplicity and, conversely, the assumption of one GTR model per branch is too complex and overparameterized (supplementary table 3, Supplementary Material online). As already mentioned by Dutheil and Boussau (2008), we did not observe a significant improvement of the results by allowing different κ on each branch or group. Concerning the protein data set, the same approach has been performed, and we observed that the model with one set of LG-based equilibrium frequencies per group of mesophilic or thermophilic organisms obtained the best statistical scores among all tested models. Finally, we ran nonhomogeneous experiments for the 14 other topologies used as input trees with the two selected protein and rRNA models described above. Topology no. 14 remains the maximum likelihood topology with this nonhomogeneous approach (data not shown).

To assess if the HKY85θpb model properly fits the heterogeneity of the rRNA data set, we used a parametric bootstrapping method based on Bowker's test designed by Dutheil and Boussau (2008). This is a pairwise test that allows to detect whether two sequences evolved under two different processes (Ababneh et al. 2006). Bowker's tests have been performed for all rRNA sequence pairs. The number of significant Bowker's tests defined the heterogeneity of the alignment. We used the parameters of the HKY85θpb model that had been previously estimated by BppML to simulate 10,000 data sets with BppSeqGen (Dutheil and Boussau 2008). For each simulated alignment, the heterogeneity was calculated and the distribution of heterogeneity values for the whole simulated sequences was obtained. Finally, the heterogeneity value of the initial data set was compared with this distribution. Clearly, the

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

HKY85 homogeneous model does not fit the data because the simulated distribution significantly underestimates the heterogeneity of sequences (supplementary fig. 2, Supplementary Material online). However, the HKY85θpb model produced simulated sequences with heterogeneity values that are representative of the intrinsic heterogeneity of the original rRNA data set (P value = 0.188) (supplementary fig. 2, Supplementary Material online).

Inference of Ancestral OGTs

Previous studies proved that rRNA G+C content and OGT were strongly correlated in Bacteria and Archaea (Galtier and Lobry 1997) and used this correlation as a molecular thermometer to infer ancestral OGTs. To establish this relationship with our rRNA data set, we retained only double-stranded regions because it has been shown that equilibrium frequency estimations were biased toward frequencies at slowly evolving sites (single-stranded regions in rRNAs) when heterogeneous models of evolution are used (Gowri-Shankar and Rattray 2006). We obtained a double-stranded regions data set of 1,801 sites. The G+C content of these regions is highly correlated to OGT (fig. 1A, r = 0.95, P value < 0.001). Concerning our protein data set, a correspondence analysis has been performed on the amino acid compositions of our alignment. This procedure, introduced by Boussau et al. (2008), produced another molecular thermometer. The results (fig. 1B) show that OGT and amino acid compositions are strongly linked together. Two major independent factors explain most of the variance in amino acid compositions in archaeal proteins. The first factor (41.5% of the total variance) highly correlates with genomic G+C content ($r = 0.81$, P value < 0.001) and the second factor (26.7% of the total variance) with OGT ($r = 0.84$, P value < 0.001).

However, the regressions of figure 1A and B are made with data points that are not statistically independent. Indeed, each data point being one species, the nonindependence arises from the fact that all species share a common ancestry and are not independently drawn from the same distribution. Thus, if a strong phylogenetic inertia exists in the traits under study, closely related species will tend to have similar values for the two traits and, consequently, will tend to cluster together in a regression diagram, increasing the correlation coefficient (Felsenstein 1985; Harvey and Pagel 1991). This problem has been noticed before, and several methods have been proposed to take the nonindependence of taxa into account (Lanfear et al. 2010). One of these methods, the PICs, proposed early on by Felsenstein (1985), has been employed here. The PIC approach uses the original values of each trait for all species and transforms them to produce new values, called contrasts, that are statistically independent and identically distributed and that can be compared by a correlation test. New correlation coefficients were calculated from the contrasts in OGT and in G+C content on one side ($r = 0.85$, P value < 0.001) and in second factor values ($r = 0.7$, P value < 0.001) on the other,

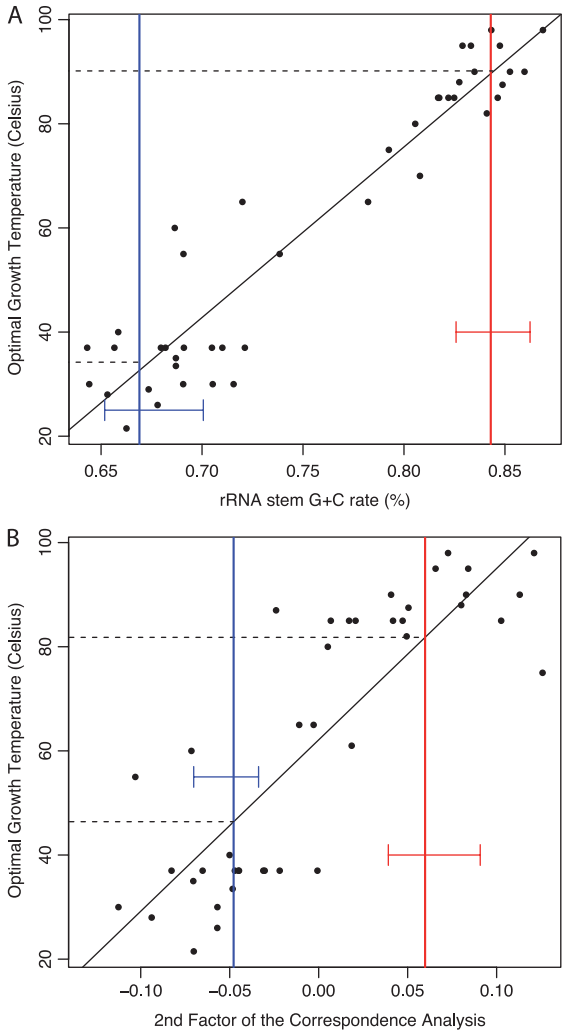

Fig. 1. Correlations between nucleotide or amino acid compositions and OGT. (*A*) rRNA thermometer. (*B*) Protein thermometer. In each plot, black dots indicate the positions of extant archaea and bacteria. For rRNAs, the linear correlation coefficient between OGT and rRNA stem G+C content is 0.95 (P value < 0.001). For proteins, the second factor values of the correspondence analysis are strongly correlated with OGT ($r = 0.84$, P value < 0.001). Vertical lines represent the inferred compositions for the ancestor of Thaumarchaea (blue) and for the HACA (red) with their 95% confidence interval. Dashed lines represent the projection of ancestral compositions on the OGT axis. The HACA is predicted to be hyperthermophile, by both rRNAs and proteins.

confirming that the strong relationship between OGT and molecular compositions in rRNAs and proteins initially observed was not solely due to the nonindependence of data points.

Evolutionary model parameters initially estimated by BppML (e.g., branch lengths, substitution model parameters, gamma law parameter) were used by BppAncestor (Dutheil and Boussau 2008) to reconstruct rRNA and protein ancestral sequences, using the same topology. One hundred putative ancestral sequences were estimated for both data sets at each node of the tree. The G+C contents and amino acid compositions of these ancestral

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

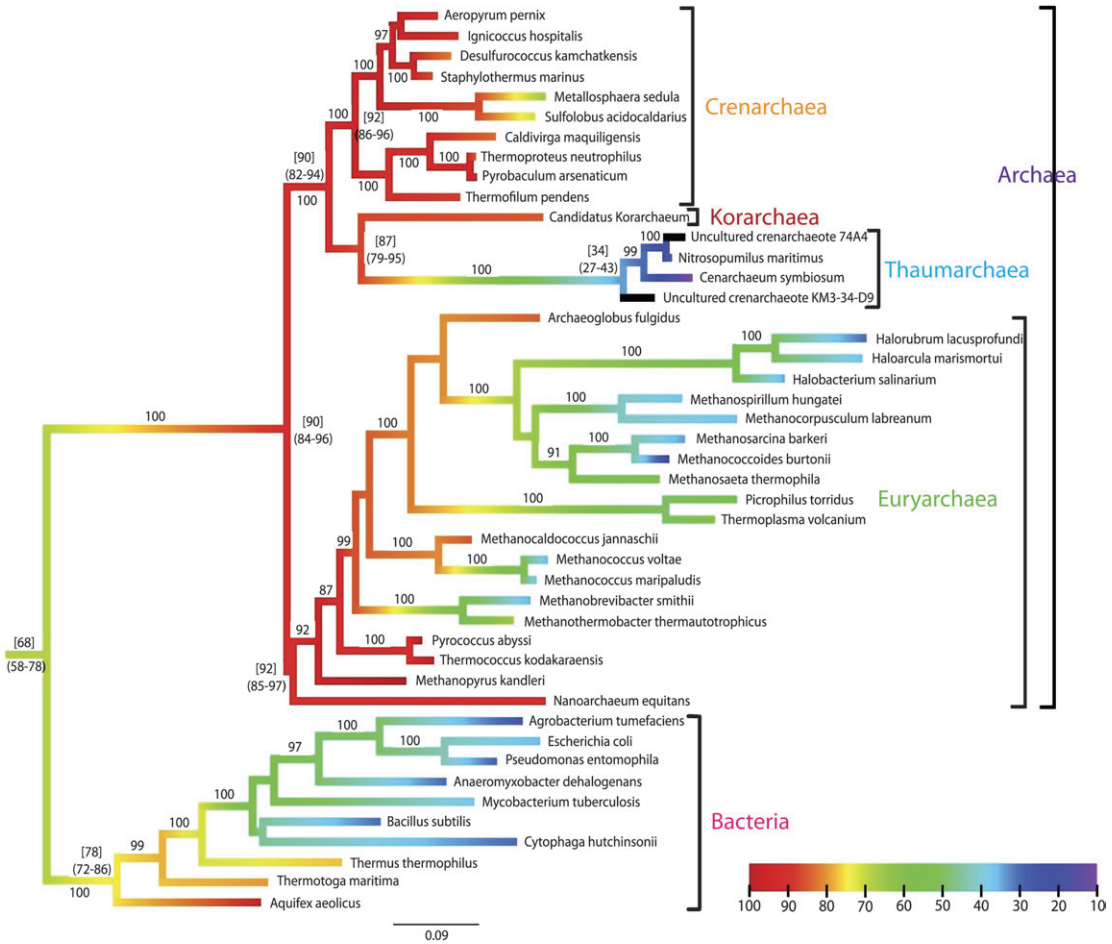

Fig. 2. Evolution of OGT from a hyperthermophilic ancestral state over the rRNA archaeal tree. Branch lengths have been colored according to temperature estimates at nodes. A linear gradient of color has been drawn between nodes. No evolution of OGT is represented in the vertical tree lines. As OGTs for uncultured Thaumarchaea are not available, their branches are black colored. The branch length scale is in substitution per site. The color scale is in degree Celsius. Mean estimates of temperature at key nodes are given between square brackets. Confidence intervals (95%) for estimates of ancestral OGTs are given between round brackets. Bootstrap values higher than 85% are represented. The concatenation of small and large rRNA subunits provided an alignment of 3,719 positions.

sequences were then computed. For each ancestral node, the mean of the distribution of G+C content values was determined and projected in the previously established correlation. Concerning proteins, the amino acid composition of each ancestral sequence was added to the correspondence analysis to get its projection on the second factor. Finally, the mean of the distribution of second factor values was used to infer OGT.

Figures 2 and 3 show that there is a parallel adaptation to high temperatures from a common ancestor of Archaea and Bacteria to a common ancestor of each domain. The last archaeal common ancestor is predicted to be hyperthermophilic and will be named below the HACA (Hot Archaeal Common Ancestor). From the HACA, whose OGT is estimated around 82 °C by proteins (90 °C by rRNAs), there is a slight increase of OGT until ancestors of Euryarchaea on one side (83 °C) and of Crenarchaea, Thaumarchaea, and Korarchaea on the other side (85 °C), but this increase is not statistically significant if confidence intervals are taken into account. Among Crenarchaea, OGT seems to increase along the tree until extant species such as *Aeropyrum pernix* (95 °C).

A progressive adaptation to lower temperatures is observed along the Euryarchaeal clade, similarly to what Boussau et al. (2008) observed within the bacterial domain. The euryarchaeal ancestor is predicted to be hyperthermophilic (83 °C and 92 °C for proteins and rRNAs, respectively) and deep-branching species are also adapted to these high temperatures. An adaptation of Euryarchaea to lower temperatures can then be observed with the exception of *Archaeoglobus fulgidus* and *Methanocaldococcus jannaschii* which may have readaptated to higher temperatures. The OGTs inferred for HACA are markedly higher in the present study (74–89 °C) than in the Boussau et al. (2008) study (59–73 °C). We investigated the reason(s) why the credibility intervals were not overlapping between the two studies. Three hypotheses were tested: the influence of the taxon sampling, the model of sequence evolution, and the gene sampling. We ruled out the taxon sampling and the model of sequence evolution hypotheses.

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

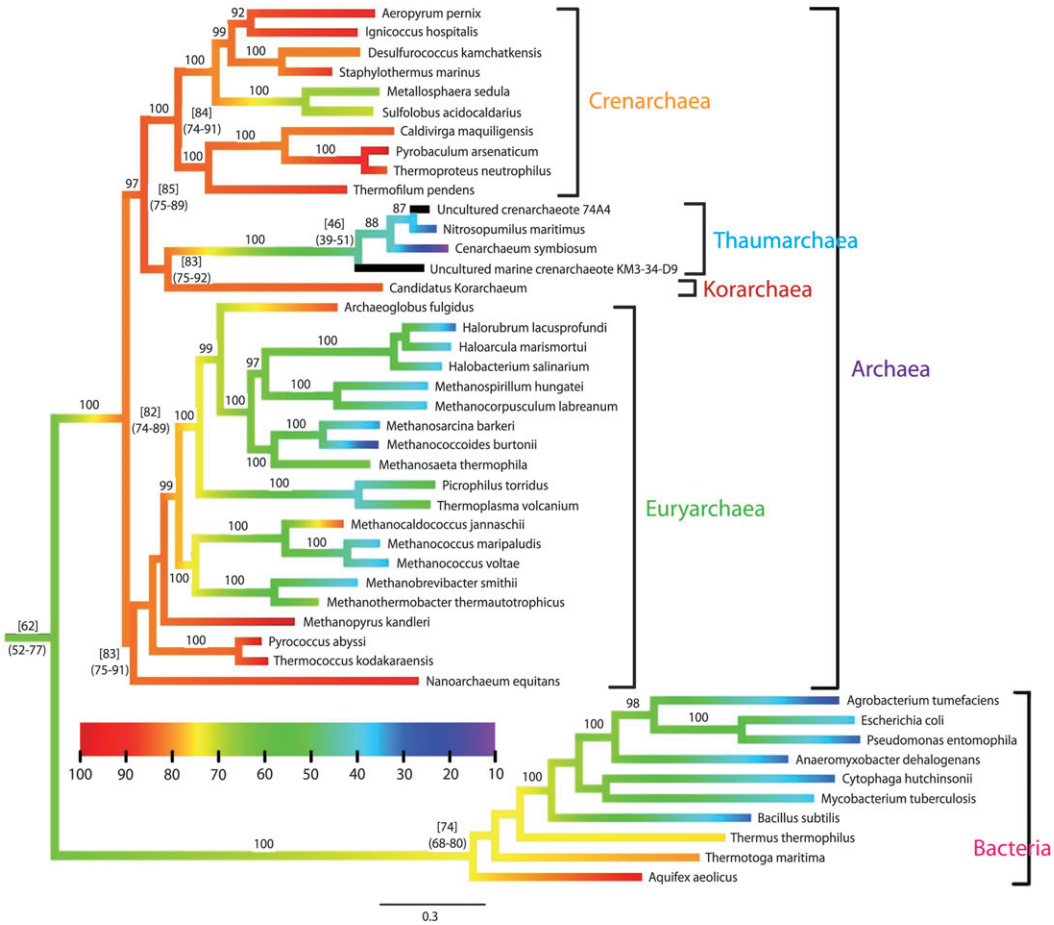

Fig. 3. Evolution of OGT from a hyperthermophilic ancestral state over the protein archaeal tree. See legend of figure 2 for details. The phylogenetic reconstruction is based on 72 genes and on a 9,799 amino acid long alignment.

Indeed, when the seven archaeal species used by Boussau et al. (2008) are analyzed with our data set, the ancestral OGT for HACA remains 82 °C, as with our 35 archaeal species. Furthermore, we analyzed the Boussau et al. (2008) data set with our model of sequence evolution, whereas in Boussau et al., the data were analyzed using a Bayesian approach and the CAT-BP (Blanquart and Lartillot 2008) model. The authors inferred that HACA lived around 66 °C. Here, using maximum likelihood and an a priori attribution of equilibrium frequencies along the tree to relax the homogeneity property, we obtained results very similar to Boussau et al.'s: the mean OGT inferred for HACA is 69 °C. We conclude that the difference between the two ancestral temperatures is a consequence of two interacting factors: 1) gene sampling and 2) uncertainty of the molecular thermometers. Among the 56 protein families selected by Boussau et al. (2008), only sites with less than 5% gaps were conserved. We observed that among these 56 families, only 24 contain archaeal sequences and so contributed to their final alignment. Among these 24 families, 22 are present in our 72 protein families data set. Thus, we split our data set in two parts: one with the proteins also analyzed by Boussau et al. (2008) and one with the remaining protein families. We observed that whereas the added families still predict a temperature around 83 °C, the families that are common with the Boussau et al. data set predict a temperature of 78 °C. Second, the molecular thermometer used by Boussau et al. was inferred from their own data set, using data from the bacterial domain. Here, the molecular thermometer was inferred from archaeal species only. In both approaches, the uncertainty in the regression was not taken into account and would increase the final credibility intervals.

Evaluation of the Two Thermometers

Remarkably, rRNAs and proteins converge to quite similar estimated OGTs (figs. 2 and 3). However, the differences between rRNA and protein-based OGT predictions could result from a different signal between rRNAs and proteins. For most internal nodes, rRNAs tend to predict lower OGTs for low temperatures (<65 °C) and higher OGTs for high temperatures (>65 °C) than proteins. Thus, a negative correlation ($r = -0.79$) exists between the differences of prediction (Protein − rRNA) and the rRNA predictions (supplementary fig. 5A, Supplementary Material online). Interestingly, the same profile occurs with extant molecules: If the molecular thermometers are used to estimate OGTs based on the compositions of extant rRNAs and proteins, a similar negative correlation ($r = -0.53$) is found (supplementary fig. 5B, Supplementary Material online).

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

Consequently, one can reasonably assume that the differences of OGT prediction between rRNAs and proteins for internal nodes result from the differences of precision of the thermometers between each other and not from a different signal that these two molecules intrinsically carry. Finally, the precision of each thermometer was investigated. The OGTs predicted from rRNAs and proteins for extant species were compared with the reference OGTs found in the databases (supplementary fig. 6, Supplementary Material online). For both rRNAs and proteins, there is a strong positive correlation between the two variables ($r = 0.95$ and $r = 0.8$, respectively), which proves that both thermometers are reliable. However, the sum of the squares of deviations to the $y = x$ line is much lower for rRNAs than for proteins (3,298 against 11,202), which indicates that the rRNA thermometer tends to be more precise than the protein thermometer.

The Influence of the Input Topology

The estimations of ancestral OGTs are not sensitive to the initial topology. Indeed, OGTs have been estimated at each node for the 15 topologies. The variance is small for all OGTs, and the HACA, euryarchaeal, and crenarchaeal ancestors are always predicted to be hyperthermophilic (supplementary table 4, Supplementary Material online). In general, for all ancestral nodes of each phylum, the pattern observed with the topology no. 14 stays unchanged. The same conclusion can be drawn concerning the possible influence of Eukarya. Supplementary table 5 (Supplementary Material online) shows that inferences of ancestral OGTs remain roughly the same for crucial nodes (Crenarchaea, Thaumarchaea, Korarchaea, and Euryarchaea) of the four archaeal domains of the eocyte tree. In particular, the common ancestor of all archaeal groups and Eukarya is still inferred to be hyperthermophilic.

Temperature Is the Major Selective Pressure Governing Sequence Evolution in Archaea

Several hypotheses have been developed to explain the adaptation of mesophilic archaea to low OGTs (López-García et al. 2004). A classical scenario posits that mesophilic species have adapted from their thermophilic ancestor to cold temperatures through extensive HGTs from mesophilic bacteria or other archaea. Such a hypothesis could explain why some species were able to colonize other ecological niches and became adapted to cold environments. Recently, Drake showed that thermophilic species display very low mutation rates in comparison to mesophilic species (Drake 2009). So, once the process of adaptation to colder temperatures begun, one can propose that the lineages were subjected to a relaxation of negative selective pressures and to an increase of mutational rates. Our results strongly support this hypothesis. Indeed, there is a strong positive correlation ($r = 0.82$, P value < 0.001) between deviations in rRNA G+C content and evolutionary distances (branch lengths) from the HACA to leaves (fig. 4A). Mesophilic species tend to have long branches associated with a strong deviation of rRNA G+C content from the HACA. Concerning proteins, the correlation between the deviation of second factor values of the correspondence analysis and branch lengths is also statistically significant, but weaker ($r = -0.57$) (data not shown). However, it has been reported that amino acid compositions of thermo- or hyperthermophiles are strongly linked to OGT. Thus, Hickey and Singer (2004) and Tekaia et al. (2002) showed that proteins of thermophilic species are slightly enriched in charged residues (Glu, Arg, Lys), whereas being depleted in polar uncharged (Asn, Gln, Ser, Thr) and in thermolabile residues (His, Gln, Thr). So, the evolution of the protein compositions for these amino acids was studied. A high correlation between deviations of ERK content (Glu, Arg, Lys) from the HACA to species and branch lengths exists ($r = 0.74$) (fig. 4C) and so does between HQT (His, Gln, Thr) content ($r = -0.55$, P value $= 0.001$) or NSTQ (Asn, Gln, Ser, Thr) content ($r = -0.69$, P value < 0.001) and branch lengths (supplementary fig. 3A and C, Supplementary Material online). As discussed above, the regressions of figure 4A and C may be potentially biased by the nonindependence of the data points. Thus, the PIC approach was also employed here to measure the statistical significance of the relationship between evolutionary rates and variations of composition. Figure 4B and D compares PICs in the evolutionary rates and in GC contents for rRNAs and ERK contents for proteins, respectively. The correlation coefficients equal 0.76 and 0.67, respectively, and are highly significant (P value < 0.001). Furthermore, similar conclusions can be drawn for the HQT and NSTQ contents (supplementary fig. 3B and D, Supplementary Material online), with correlation coefficients of -0.35 (P value < 0.05) and -0.65 (P value < 0.001), respectively. Consequently, intrinsic molecular evolution of rRNAs and proteins co-occurred with the continuous adaptation of mesophilic species to colder environments.

Control for a Putative Bias in the Nonhomogeneous Approach

In the archaeal domain, short branches have a small variation of G+C content, and long branches have a high variation. One could argue that a bias exists in our nonhomogeneous estimation of evolutionary parameters, which would systematically lead to this pattern. Of course, short branches exclude high G+C content variations between the two branch extremities, but long branches could exist with and without extensive base composition variation. Thus, a simulation experiment was carried out for the rRNA data set, where the association between branch lengths and variation of G+C content is the highest (figs. 2 and 4A). A model tree with the topology of the archaeal domain tree was used. Random branch lengths and random G+C equilibrium frequencies were attributed to each branch of this tree. Branch lengths were randomly extracted from 95% of a Poisson distribution with a mean chosen to preserve the total variance of branch lengths of the archaeal tree. The G+C equilibrium frequencies were extracted

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

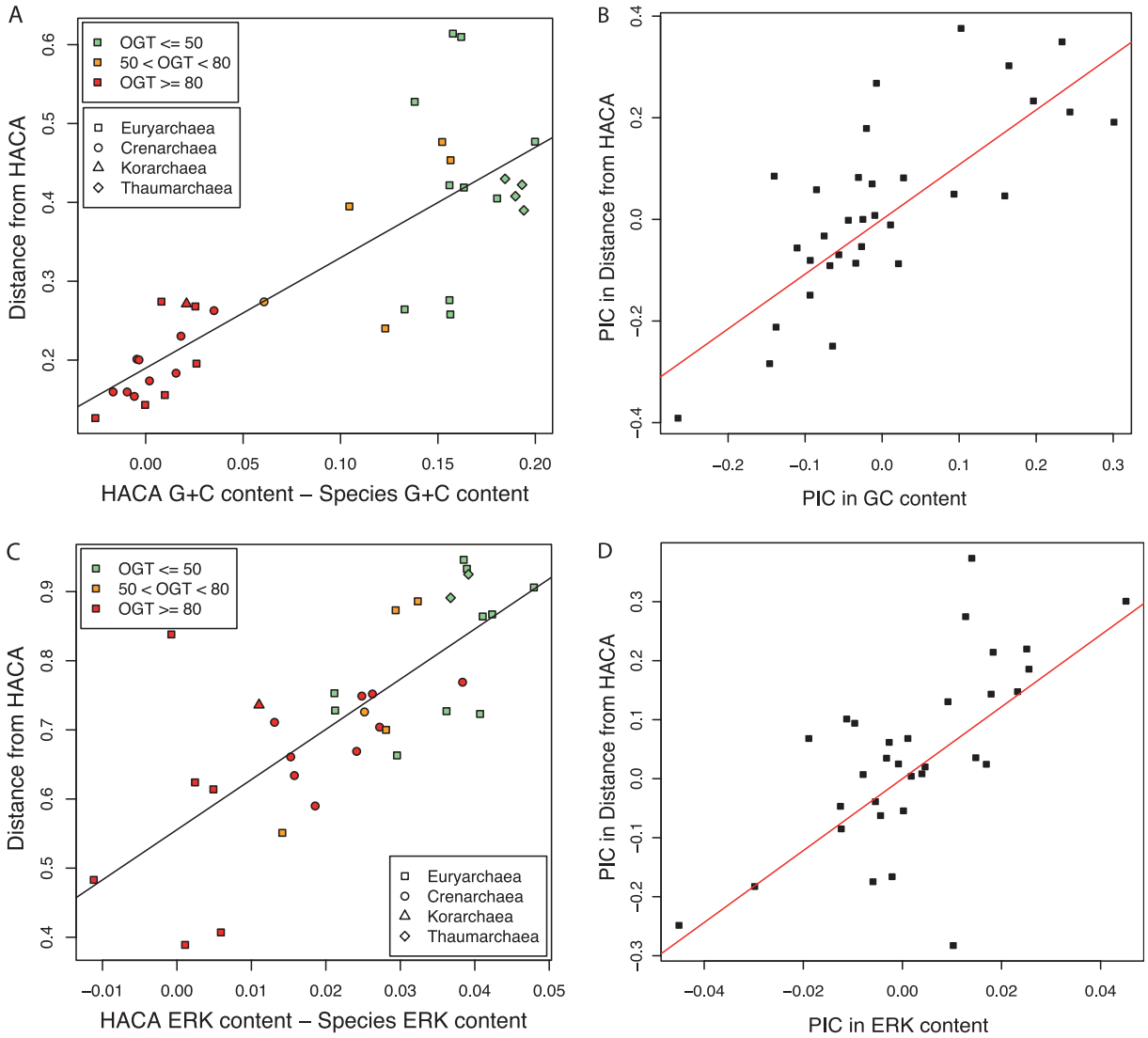

Fig. 4. A nonclock behavior of archaeal rRNAs and proteins and its relation with environmental temperature. In figure 4*A* and *C*, correlations between the raw evolutionary distances from HACA (total of branch lengths between the HACA and the extant species) and raw deviations of specific base (G+C%, fig. 4*A*) and amino acid (E+R+K%, fig. 4*C*) contents between the HACA and extant species are represented. E, R, and K amino acids represent charged residues. Mesophilic species are colored in green, thermophilic in orange, and hyperthermophilic in red. The four major groups of Archaea are plotted with different symbols. The linear correlation coefficient is 0.82 (P value < 0.001) for rRNAs and 0.74 for proteins (P value < 0.001). In figure 4*B* for rRNAs and figure 4*D* for proteins, the raw values have been corrected using the PICs method. The correlation coefficients are 0.76 (P value < 0.001) and 0.67 (P value < 0.001), respectively.

from 95% of a normal distribution with mean and variance equal to those of the archaeal tree. One hundred trees were constructed in this way. Each simulated tree was used to reconstruct simulated alignments with BppSeqGen, which then were used as input alignments to run the HKY85θpb model with BppML. Using simulated sequences, we reestimated the ancestral G+C content of the HACA and computed correlations between differences in G+C content and evolutionary distances from the HACA to leaves. The resulting distribution of correlation coefficients (supplementary fig. 4,Supplementary Material online) has a mean value of 0.23 and a maximum value of 0.69, far from the correlation coefficient of the real data ($r = 0.82$). This result strongly suggests that the observed pattern is a real signal and not a bias of the nonhomogeneous parameter estimation protocol.

The Node-Density Artifact

Mesophilic species could have longer branch lengths because of the node-density artifact highlighted by Webster et al. (2003). This phenomenon could be particularly true in the Euryarchaeal domain, where more bifurcations exist, but does not apply to the long branch leading to thaumarchaeal species, which is poor in internal nodes. To rule out this possible bias in the euryarchaeal domain, eight euryarchaeal species were removed from the analysis and a nonhomogeneous experiment with the HKY85θpb model was carried out. The resulting tree (data not shown)

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

still displays longer branch lengths for euryarchaeal mesophilic species than for thermophilic species, which disproves the node-density artifact.

Discussion

The main goal of this study was to investigate the evolution of the adaptation to OGT in Archaea over evolutionary times. We do not claim that the archaeal topology (figs. 2 and 3) used to infer ancestral compositions and OGTs reflects the true evolutionary history. Indeed, many protein families used here are likely to have been affected by HGT. However, the results above show that the chosen archaeal tree does not strongly influence the OGT estimates of the ancestors of the major archaeal phyla.

The nonhomogeneous models employed better fit the data than homogenous ones and allow for a more realistic description of the evolutionary process. Moreover, many archaea-specific gene families that do not have members in Bacteria were used. This further increased the signal to estimate ancestral compositions in the archaeal domain in comparison to the work reported by Boussau et al. (2008). Here, as Galtier and Gouy (1998) already highlighted with their nonhomogeneous model (one T92 substitution model per branch with κ [transversion/transition ratio] shared in the whole tree), we confirmed that the crucial parameter which characterizes the evolution of rRNAs is θ (G+C content). However, the use of one HKY85 model per branch with κ, θ_1, and θ_2 shared in the whole tree, allowed to significantly improve phylogenetic estimations: the two additional parameters (θ_1 and θ_2) remove the constraint of equal base frequencies assumed in the T92 model and enhance the model fit to the data. The Bowker's test used in this study was implemented for DNA or RNA. It suggests that the HKY85θpb model was suitable to fit the heterogeneity of the rRNA data set. A rather similar approach that could be applied to the protein data set was proposed by Foster (2004).

The evolution of OGT along the archaeal tree presented here is in line with previous studies that used phylogenetics to infer ancestral conditions of life (Galtier et al. 1999; Boussau et al. 2008; Gaucher et al. 2008). Nevertheless, this is the first one that reaches such a level of accuracy for one particular domain of life. Our results show that the archaeal domain, from an ancestral state adapted to high temperatures, progressively colonized colder environments on Earth in the euryarchaeal phylum. This evolution of OGT is very similar to the one reconstructed for the bacterial domain (Boussau et al. 2008; Gaucher et al. 2008). In Thaumarchaea, partial genomic sequences of *Uncultured crenarchaeote 74A4* and *Uncultured crenarchaeote KM3-34-D9* were used to infer the ancestral sequences and OGTs (only one gene of each organism was present in the protein alignment). This did not cause any bias in ancestral estimations of OGTs because no major inconsistency was observed at internal nodes of this phylum in comparison with the rRNA tree.

Remarkably, the global pattern of OGT predictions is qualitatively similar between the two data sets. Even if some discrepancies exist between rRNAs and protein-based inferences, the results presented here suggest that these differences do not result from different evolutionary signals carried by rRNAs and proteins but originate from the specific prediction bias of each thermometer.

It is worthwhile to note that the difference between the present study and that by Boussau et al. (2008) concerning the OGT of HACA reveals the uncertainty in the current approach regarding the estimation of ancestral OGTs. We have ruled out that this difference in inferred OGTs resulted from the differences in taxon sampling or in evolutionary models between the two studies and have shown that the difference mostly results from different protein gene sets. In future approaches, the uncertainties of the molecular thermometers should be incorporated in the inferences of temperatures, allowing to improve the modeling of the evolution of protein sequences, without constraining it to a single dimension (here the regression line).

How did organisms that were adapted to life at high temperatures acquire the ability to colonize colder environments? An attractive hypothesis would be to bring into play intensive HGTs between archaeal species that live in hot environments and other species (bacterial or archaeal) that live in colder ones. In their analysis of partial genomic sequence data from mesophilic crenarchaea, Lopez-Garcia et al. (2004) proposed that HGTs could have been crucial in the adaptation of Thaumarchaea to cold environments. They mentioned the case of the HSP70 chaperone, present in mesophilic euryarchaea and thaumarchaea but not in hyperthermophilic crenarchaea. They supposed that the gene could have been acquired by HGT and could have facilitated the adaptation of thaumarchaea to lower temperatures. At present, with more completely sequenced genomes, this assumption remains valid because no hyperthermophilic crenarchaeon possesses this gene (data not shown). Moreover, with a deeper analysis of more newly sequenced thaumarchaeal fosmids, these authors showed that chromosomal rearrangements in the region of the rRNA genes occurred during the evolution of Thaumarchaea, more than in other lineages, and that many HGTs from bacterial and mesophilic euryarchaeal lineages can be highlighted (Brochier-Armanet C, personal communication).

How did archaeal genomes and proteomes evolve during the transition from thermophilic to mesophilic environments? Concerning rRNAs, it has been shown that thermo- or hyperthermophilic organisms display especially high values of G+C%. As rRNAs possess a large fraction of double-stranded regions, a high G+C% could provide a higher stability at high temperatures (Galtier and Lobry 1997). Concerning proteins from organisms living in hot temperatures, they are very stable from both a thermodynamic and a kinetic point of view (Sterner and Liebl 2001). Some estimations focusing on the comparison between the two bacteria *Escherichia coli* and *Thermus thermophilus*

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

for the RNAse H have shown that the melting temperature of the thermophilic protein is 20 °C higher than that of the mesophilic species (Hollien and Marqusee 1999). The literature mentions a lot of characters to explain why thermophilic proteins are so thermostable. Sterner and Liebl (2001) proposed a nonexhaustive list of several characters that could avoid chemical degradation of the polypeptide chain, such as an increase of hydrogen bonds, improved electrostatic interactions, and increased compactness. The results of Tekaia et al. (2002) and Hickey and Singer (2004) are good evidence to support these trends: proteins of thermophilic species are slightly enriched in charged residues (Glu, Arg, Lys), whereas being impoverished in polar uncharged (Asn, Gln, Ser, Thr) and in thermolabile residues (His, Gln, Thr).

Thus, the highly correlated evolutionary changes observed with the PIC method between branch lengths and the variation of GC (for rRNAs), ERK, NSTQ, and HQT (for proteins) contents between the HACA and extant species are informative. This phenomenon offers an explanation of the deviation from molecular clock in the archaeal domain and highlights the critical role played by environmental temperature on the archaeal molecules. An exception to this evolutionary scenario concerns *N. equitans* and its very long branch. Indeed, this organism developed a parasitic life at the surface of its hyperthermophilic crenarchaeal host, *I. hospitalis* (Hubert et al. 2002). This way of life could explain the increase of evolutionary rates unrelated to temperature in this lineage.

The use of nucleotide and amino acid sequences to estimate the timing of the history of life on earth was proposed early on (Zuckerkandl and Pauling 1965) in the history of molecular biology. The molecular clock hypothesis assumed that molecules evolved at constant rates over time, which allowed the inference of divergence times between species from molecular sequence data. However, it was later clearly demonstrated that evolutionary rates are not constant over time, either between lineages or within a particular lineage. The branches of the present archaeal domain phylogenetic trees clearly differ extensively in length, which contradicts the molecular clock model. Nowadays, relaxed molecular clock methods are developed to take these variations of evolutionary rates into account (Thorne et al. 1998; Rannala and Yang 2007).

Recent studies (Friedman et al. 2004; Drake 2009) seem to confirm that increased temperature imposes increased constraints on genetic innovation. Species adapted to high temperatures exhibit very low mutational rates. The possibility that neutral or slightly deleterious mutations in cold environments may become highly deleterious in hot temperatures, especially concerning protein folding, could explain this phenomenon. Thus, the increase of evolutionary rates during the colonization of cold environments is partly explained by increased possibilities to explore the substitutional space, without fitness impact. However, if only a neutral process of evolution were involved, we would expect to observe a broader range of G+C content and second factor values among extant mesophilic species rRNAs in figure 1A. All mesophilic species rRNAs are G+C poor and are characterized by low values of the second factor of the correspondence analysis. Therefore, natural selection forces archaeal organisms to have low substitutional rates in hot temperatures but, even if mesophilic species can have higher mutational and substitutional rates and are freer to explore more genetic combinations, environmental temperature continuously constrains the base and amino acid equilibrium frequencies.

In conclusion, mesophilic species have adjusted their molecular compositions during an adaptation process to colonize new mesophilic environments. The results obtained by Cherry (2010) support this view. The author showed in five eukaryotes and one mesophilic bacterium (*E. coli*) that highly expressed and slowly evolving proteins have similar compositions to those of proteins from thermophilic organisms. Because the amino acid composition of thermophilic proteins has been shown to increase the stability of the folded state at high temperatures (Singer and Hickey 2003), Cherry (2010) proposed that there is a strong selection against protein misfolding. This selection would be higher for highly expressed proteins and would be the reason of their low evolutionary rates (Pál et al. 2001; Drummond et al. 2005). In Eukaryotes (Cherry 2010) and mesophilic bacteria (Rocha and Danchin 2004) (and probably mesophilic archaea), this selection is higher for highly expressed genes. For thermophilic species (Archaea and probably Bacteria), environmental temperature appears to be a major selective factor at the whole proteome level, explaining the decrease of evolutionary rates in thermophilic proteins.

Supplementary Material

Supplementary figures 1–6 and tables 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/)

Acknowledgments

The authors would like to thank four anonymous referees as well as the Associate Editor for their constructive comments, which allowed to significantly improve the present manuscript. The authors are particularly grateful to Céline Brochier-Armanet, Bastien Boussau, and Vincent Daubin for their help, suggestions, and fruitful discussions. Finally, the authors sincerely thank Julien Dutheil for all his help with the Bio++ libraries.

References

Ababneh F, Jermiin LS, Ma C, Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics* 22:1225–1231.

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol*. 25:842–858.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456:942–945.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol*. 55:756–768.

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P. 2006. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol*. 6:R42.

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. 2008. Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol*. 6:245–252.

Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biol Lett*. 5:401–404.

Bromham L, Rambaut A, Harvey PH. 1996. Determinants of rate variation in mammalian DNA sequence evolution. *J Mol Evol*. 43:610–621.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540–552.

Cherry JL. 2010. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol*. 27:735–741.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci U S A*. 105:20356–20361.

DeLong EF. 1992. Archaea in coastal marine environments. *Proc Natl Acad Sci U S A*. 89:5685–5689.

Denamur E, Matic I. 2006. Evolution of mutation rates in bacteria. *Mol Microbiol*. 60:820–827.

Drake JW. 2009. Avoiding dangerous missense: thermophiles display especially low mutation rates. *PloS Genet*. 5(6):e1000520.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102:14338–14343.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol*. 8:255.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5:113.

Elkins JG, Podar M, Graham DE, et al. (20 co-authors). 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci U S A*. 105:8102–8107.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat*. 125:1–15.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Forterre P, Gribaldo S, Brochier-Armanet C. 2009. Happy together: genomic insights into the unique Nanoarchaeum/Ignicoccus association. *J Biol*. 8:7.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol*. 53:485–495.

Foster PG, Cox CJ, Embley TM. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos Trans R Soc Lond B Biol Sci*. 364:2197–2207.

Foster PL. 2007. Stress-induced mutagenesis in bacteria. *Crit Rev Biochem Mol Biol*. 42:373–397.

Friedman R, Drake JW, Hughes AL. 2004. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. *Genetics* 167:1507–1512.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol*. 15:871–879.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*. 44:632–636.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* 283:220–221.

Gaucher EA, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451:704–707.

Gowri-Shankar V, Rattray M. 2006. On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol Biol Evol*. 23:352–364.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*. 27:221–224.

Gribaldo S, Brochier-Armanet C. 2006. The origin and evolution of Archaea: a state of the art. *Philos Trans R Soc Lond B Biol Sci*. 361:1007–1022.

Grogan DW, Carver GT, Drake JW. 2001. Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius. *Proc Natl Acad Sci U S A*. 98:7928–7933.

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

Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.

He Q, Sanford RA. 2003. Characterization of Fe(III) reduction by chlororespiring *Anaeromyxobacter dehalogenans*. *Appl. Environ. Microbiol*. 69:2712–2718.

Hegan PS, Mermall V, Tilney LG, Mooseker MS. 2007. Roles for *Drosophila melanogaster* Myosin IB in maintenance of enterocyte brush-border structure and resistance to the bacterial pathogen *Pseudomonas entomophila*. *Mol Biol Cell*. 18:4625–4636.

Hickey D, Singer G. 2004. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol*. 5:117.1–117.7.

Hollien J, Marqusee S. 1999. A thermodynamic comparison of mesophilic and thermophilic ribonucleases H. *Biochemistry* 38:3831–3836.

Hubert H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 417:63–67.

Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol*. 4:18.

Lanfear R, Welch JJ, Bromham L. 2010. Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol Evol*. 25:395–503.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. 7:S4.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21:1095–1109.

Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci U S A*. 81:3786–3790.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol*. 25:1307–1320.

Le SQ, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Lopez-Garcia P, Brochier C, Moreira D, Rodriguez-Valera F. 2004. Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ Microbiol*. 6:19–34.

Mackwan RR, Carver GT, Drake JW, Grogan DW. 2007. An unusual pattern of spontaneous mutations recovered in the halophilic archaeon *Haloferax volcanii*. *Genetics* 176:697–702.

Mackwan RR, Carver GT, Kissling GE, Drake JW, Grogan DW. 2008. The rate and character of spontaneous mutation in *Thermus thermophilus*. *Genetics* 180:17–25.

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022

Nazar RN. 1980. A 5.8 S rRNA-like sequence in prokaryotic 23 S rRNA. *FEBS Lett*. 119:212–214.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perrière G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*. 10:S3.

Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol*. 56:453–466.

Ripplinger J, Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? *Syst Biol*. 57:76–85.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol*. 21:108–116.

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

Singer G, Hickey D. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 317:39–47.

Sterner R, Liebl W. 2001. Thermophilic adaptation of proteins. *Crit Rev Biochem Mol Biol*. 36:39–106.

Stetter KO. 2006. Hyperthermophiles in the history of life. *Philos Trans R Soc B Biol Sci*. 361:1837–1843.

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc B Biol Sci*. 269:137–142.

Tekaia F, Yeramian E, Dujon B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* 297:51–60.

Thioulouse J, Chessel D, Dolédec S, Olivier J. 1997. ADE-4: a multivariate analysis and graphical display software. *Stat Comput*. 7:75–83.

Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*. Chapter 2:Unit 2.3.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*. 15:1647–1657.

Valentine DL. 2007. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat Rev Microbiol*. 5:316–323.

Vetriani C, Maeder DL, Tolliday N, Yip KS, Stillman TJ, Britton KL, Rice DW, Klump HH, Robb FT. 1998. Protein thermostability above 100 degreesC: a key role for ionic interactions. *Proc Natl Acad Sci U S A*. 95:12300–12305.

Webster AJ, Payne RJ, Pagel M. 2003. Molecular phylogenies link rates of evolution and speciation. *Science* 301:478.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p. 97–166.

Downloaded from https://academic.oup.com/mbe/article/28/9/2661/1013956 by Universite Claude Bernard (Lyon 1) user on 22 September 2022