The molecular signal for the adaptation to cold temperature during early life on Earth

Mathieu Groussin, Bastien Boussau, Sandrine Charles, Samuel Blanquart, Manolo Gouy

The phylogenetic signal for a non-hyperthermophilic Last Universal Common Ancestor

Mathieu Groussin[1,*], Bastien Boussau[2], Sandrine Charles[1], Samuel Blanquart[3] and Manolo Gouy[1]

1: Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France

2: Department of Integrative Biology, University of California, Berkeley, United States of America

3: Inria Lille Nord Europe, LIFL UMR 8022 (CNRS Université de Lille 1), Villeneuve d'Ascq, France

Keywords: Nonhomogeneous substitution model, Ancestral Sequence Reconstruction, Optimal Growth Temperature, Last Universal Common Ancestor.

*: Email: mathieu.groussin@univ-lyon1.fr

Abstract

Several lines of evidence such as the basal location of thermophilic lineages in large scale phylogenetic trees and the ancestral sequence reconstruction of single enzymes or large protein concatenations support the conclusion that the ancestors of the bacterial and archaeal domains were thermophilic organisms which were adapted to hot environments during the early stages of the Earth. A parsimonious reasoning would therefore suggest that the Last Universal Common Ancestor (LUCA) was also thermophilic. Various authors have used branch-wise non-homogeneous evolutionary models that better capture the variation of molecular compositions among lineages to accurately reconstruct the ancestral G+C contents of ribosomal RNAs and the ancestral amino acid composition of highly conserved proteins. They confirmed the thermophilic nature of the ancestors of Bacteria and Archaea, but concluded that LUCA, their last common ancestor, was a mesophilic organism having a moderate optimal growth temperature. In this letter, we investigate the unknown nature of the phylogenetic signal that informs ancestral sequence reconstruction to support this non parsimonious scenario. We find that rate variation across sites of molecular sequences provides information at different time scales by recording the oldest adaptation to temperature in slow-evolving regions and subsequent adaptations in fast-evolving ones.

Introduction

Several lines of evidence support the hypothesis that, during early stages of the evolution, life was adapted to high temperatures that may have prevailed on the surface of the early earth. For instance, early studies discovered that the deepest branching lineages within the bacterial and archaeal domains are thermophilic [1]. This scenario was also supported by the resurrection of ancestral elongation factor Tu sequences that appear more and more thermostable when going back in time [2], and by an estimation of the amino acid composition of ancestral proteomes that appear more similar to the composition of extant thermophiles than of mesophiles [3].

Boussau et al. [4] concluded that molecular sequence data confirm the hypothesis of high temperature adaptation during early stages of life, namely for the ancestors of the bacterial and archaeal domains. However, these authors reported strong evidence for a scenario in which the Last Universal Common Ancestor (LUCA) itself, living at a still earlier stage of the history of life, was a mesophilic organism. This non parsimonious scenario was obtained exploiting the tight relation that exists between either the G+C content in ribosomal RNAs or the amino-acid contents in proteins and the Optimal Growth Temperature (OGT) of Bacteria and Archaea. Such correlations between molecular compositions and temperature may be explained by structural adaptation to increase RNA and protein thermostability [5,6] and are likely to remain constant over evolutionary time. They allow the construction of molecular thermometers [7] that can provide estimates of ancestral environmental temperatures if one obtains ancestral base and amino acid compositions through ancestral sequence reconstruction.

Through a number of control experiments, Boussau et al. [4] have shown that the use of nonhomogeneous models, which are capable of capturing the variation of composition among lineages, are key to accurately estimate ancestral base and amino acid compositions, and therefore, ancestral temperatures. But these authors have not identified the nature of the signal present in extant molecular sequences that informs evolutionary models to support such a non parsimonious scenario. In this letter, we aim at solving this issue.

Material and Methods

Data sets and nonhomogeneous models

Boussau et al. [4] built a concatenate of small- and large-subunit rRNAs from 456 organisms (2,239 sites) and used the sites restricted to stem regions (1,043 sites) to infer the ancestral G+C contents over the tree of life. From these alignments, we selected 125 species covering a broad taxonomic diversity without redundancy in the taxonomic sampling. Regarding the concatenation of proteins, the 56 gene families and 30 species considered in Boussau et al. [4] were used here, completed up to 38 species, especially with Archaea species, which were poorly represented in the first set of species. We reconstructed ML phylogenetic trees for rRNAs (on the 2,239 sites dataset) and proteins with PhyML [9]. A three-domain tree was obtained and the root was placed on the bacterial branch. As in [12] and [4], the branch-wise equilibrium frequencies were estimated along these universal phylogenetic trees. The stem dataset was analyzed with the BppML program [11] assuming a discrete Gamma distribution with 8 categories to model rate variation among sites and the nonhomogeneous Galtier and Gouy (GG) substitution model [10]. The GG model specifies branch-wise equilibrium G+C contents, as well as an independent G+C content on the root. For proteins, we used a new branch-wise nonhomogeneous model implemented in the Maximum Likelihood (ML) framework, named COaLA [13] that we recently designed. See supplementary Material for a description of the COaLA model and an evaluation of the fit to data of the nonhomogeneous models in comparison with homogeneous models.

Molecular thermometers

OGT highly correlates with the G+C content of the stem regions of rRNAs (ρ=0.76, *p-value<0.001*, Supplementary Figure 2) and with the second axis of the COA computed on amino-acid compositions of the protein dataset restricted to prokaryotic species (ρ=0.88, *p-value<0.001*, Supplementary Figure 3). We controled for phylogenetic inertia with the phylogenetic independent

contrast approach [14] using the R package ape [15] and observed that these correlations are still strongly significant. Linear regressions between OGTs and compositions were then computed to obtain the molecular thermometers.

Inference of ancestral compositions and OGTs

The ancestral sequences were inferred with BppAncestor [11] using the evolutionary parameters estimated by BppML. For each node of the tree, 100 ancestral sequences were generated by drawing amino-acids from the posterior distributions of probabilities. The average composition of these ancestral sequences was calculated and the corresponding ancestral temperatures were deduced from the molecular thermometers (See Supplementary Material for the confidence intervals computation).

Results and Discussion

We first confirm results obtained in [4] with the present rRNA and protein datasets and the nonhomogeneous GG [10] and COaLA [13] substitution models in Maximum Likelihood (ML). Figures 1 and 2 show that LUCA is estimated to have lived in colder environments than the ancestors of Bacteria and Archaea (wilcoxon test, $p\text{-value}<0.001$), which were hyperthermophiles. Supp. Fig. 1 shows that this pattern is also recovered when an alternative tree topology is used, in which Eukaryotes branch within Archaea (Eocyte hypothesis [17]) but is less pronounced with the homogeneous LG model, which infers a thermophilic LUCA.

The phylogenetic signal that informs a non-hyperthermophilic LUCA and yet two hyperthermophilic descendants is currently unknown. However, several points suggest that the variation in evolutionary rate among sites plays a role. First, Fournier and Gogarten [18] highlighted that amino acids that are found in higher proportions in hyperthermophilic species are rarer at slow-

evolving sites. Such amino-acids notably include charged residues [7]. Second, the signal for a parallel adaptation to high temperatures is partially lost when COaLA is employed without a Gamma distribution to model the variation in rate among sites (Supp Fig 1).

To highlight the influence of rate variation among sites in the differential recording of ancestral compositions, we partitioned the rRNA and protein datasets according to the site evolutionary rates. Figure 2a and 2b show that with slow-evolving sites, all ancestors are inferred to be mesophilic organisms, LUCA being adapted to lower temperatures than its two descendants. The ancestral compositions of fast-evolving sites tend to favor hotter ancestral environments, even for LUCA with proteins. But LUCA is still inferred to live at lower temperatures than the ancestors of Bacteria and Archaea. As expected, the quantitative estimates of past temperatures inferred by both slow- and fast-evolving sites are different from those obtained with the complete dataset. Indeed, although slow-evolving sites conserved reliable signals for ancestral compositions, they carry less phylogenetic information for the early parallel adaptation to high temperature, which explains why this pattern is less pronounced than with the complete dataset. However, both the G+C content (rRNAs) and the third axis of a correspondence analysis (proteins) computed from the slow-evolving sites of extant sequences correlate with OGT (ρ=0.72, *p-value<0.001* and ρ=0.43, *p-value<0.05*, respectively), adding support to the idea that slow-evolving sites can respond to temperature and can represent accurate fossils of ancestral adaptation to temperature. Fast-evolving sites contain a stronger signal for this parallel adaptation but necessarily less reliable information for ancestral compositions and so, ancestral temperatures.

All these results suggest the presence of a genuine signal in molecular sequences indicating a mesophilic LUCA. This signal was recorded thanks to a combination of compositional variation in time and rate variation in site, such that slow-evolving sites more accurately reflect older temperatures, while fast evolving sites partially erased this oldest signal in favor of subsequent adaptations to higher temperatures.

Gowri-Shankar and Rattray [8] showed that there is an intrinsic correlation between

evolutionary rates across sites and base composition in rRNAs. Therefore nucleotide composition varies across the sites of a rRNA alignment. These authors showed that branch-wise non-homogeneous models, which account for the variation of composition in time but assume across-site homogeneity, may infer biased ancestral sequence compositions for sequences generated by a time-homogeneous process in which evolutionary rate and base compositions are correlated. The inference bias is directed towards the composition of slow-evolving sites which are, in the case of full-length rRNAs including both stem and loop regions, GC-poor. One could therefore wonder whether such an inference bias would be responsible for the low G+C content inferred for LUCA compared to the higher G+C contents of its first descendants. We reject this bias with two points. First, as in the present study, Boussau et al. [4] applied the molecular thermometers on rRNAs to only the stem regions of the molecule. Supp Fig 4 shows that, for these regions, the correlation found by Gowri-Shankar and Rattray [8] is in the opposite direction, although non-significant, with G+C-enriched slow-evolving sites. Second, we simulated data in a context where the bias would apply, assuming only heterogeneity among sites and no heterogeneity among branches and verified if the correlation between site evolutionary rates and site compositions incorrectly informs the non-homogeneous model to estimate a lower G+C content of LUCA than for its descendants. We partitioned rRNA alignment sites in 8 categories according to their evolutionary rate. For each rate-specific category, we simulated DNA sequences with a homogeneous Tamura92 model and the G+C equilibrium frequency fixed to the observed G+C frequency of the category and then concatenated the 8 simulated sets. We repeated this procedure 100 times and reconstructed ancestral G+C contents with the non-homogeneous GG model on each concatenated simulated alignment. Supp Fig 5 shows that the pattern of parallel increase in G+C content since LUCA found on real data is not recovered. Instead, LUCA has a higher G+C content than its two descendants. As slow-evolving sites of stem regions have globally higher G+C contents than fast-evolving ones (Supp Fig 4), this simulation result is in agreement with the bias of Gowri-Shankar and Rattray [8]. It further suggests that, if the non-homogeneous model applied on real data is affected by the bias as it is when applied

to simulations, the true G+C content of LUCA may so far have been overestimated.

All these results indicates that non-homogeneous models can capture a genuine time-wise variation in composition and that the pattern of parallel increase to high temperatures does not result from a bias due to a correlation between site-specific rates and site-specific compositions [8], but emerges in spite of this bias.

Acknowledgements

The authors thank Nicolas Lartillot, Vincent Daubin and the members of the Bioinformatics and Evolutionary Genomics team for suggestions and fruitful discussions.

This work was supported by the French Agence Nationale de la Recherche (ANR) and is a contribution to the Ancestrome project (ANR-10-BINF-01-01).

References

[1] Stetter KO. 2006. Hyperthermophiles in the history of life. *Phil. Trans. R. Soc. Lond. B* 361:1837-1843.

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

[3] Brooks DJ, Fresco JR and Singh M. 2004. A novel method for estimating ancestral amino acid composition and its application to proteins of the last universal ancestor. *Bioinformatics* 20:2251-2257.

[4] Boussau B, Blanquart S, Necsulea A, Lartillot N, and Gouy M. 2008. Parallel Adaptation to High Temperature in the Archaean Eon. *Nature* 456:942-945.

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

[6] Zeldovich KB, Berezovsky IN, and Shakhnovich EI. 2007. Protein and DNA Sequence Determinants of Thermophilic Adaptation. *Plos Comput Biol* 3:e5.

[7] Groussin M and Gouy M. 2011. Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea. *Mol Biol Evol* 28:2661-2674.

[8] Gowri-Shankar V and Rattray M. 2006. On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol Biol Evol* 23:352-64.

[9] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307-321.

[10] Galtier N and 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.

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

[12] Galtier N, Tourasse N, and Gouy M. 1999. A Nonhyperthermophilic Common Ancestor to Extant Life Forms. *Science* 283:220-221.

[13] Groussin M, Boussau B, and Gouy M. 2013. A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. *Syst Biol* 62:523-538.

[14] Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat* 125:1-15.

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

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

[18] Fournier GP and Gogarten JP. 2007. Signature of a Primitive Genetic Code in Ancient Protein Lineages. *J Mol Evol* 65:425-436.

Figure legends

Figure 1

Evolution of OGT along the universal tree of life obtained with the protein dataset.

Branches have been colored according to temperature estimates at nodes, following a linear interpolation from node to node. OGTs for Eukaryotes are not available, their branches are therefore grey colored. The branch length scale is in substitution/site. The color scale is in °C. Mean estimates of temperature for LUCA and the ancestors of major domains are given above branches. Confidence intervals (95%) for estimates of ancestral OGTs are given between square brackets.

Figure 2

The nonhomogeneous models recover the signal for a parallel adaptation to high temperatures within the across-site rate variation.

a) rRNA dataset. b) Protein dataset. Ancestral temperatures for domain ancestors and for LUCA were estimated from ancestral compositions inferred with nonhomogeneous models, either on all sites of the datasets (Complete dataset) or on slow-evolving or fast-evolving sites only. ***: *p-value<0.001*. NS: non-significant.