doi:10.1038/nature07393 nature

LETTERS

Parallel adaptations to high temperatures in the Archaean eon

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

Fossils of organisms dating from the origin and diversification of cellular life are scant and difficult to interpret[1], for this reason 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 the effects of temperature on ancestral organisms left 'genetic footprints' that 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]. As the archaeal ancestor is also thought to have been thermophilic[5], the 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 environmental-temperature-related phases during the evolutionary history of 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 the LUCA and leading to the ancestors of Bacteria and Archaea–Eukaryota convergently adapted to high temperatures, possibly in response to a climate change of the early Earth[1,8,9], and/or aided by the transition from an RNA genome in the 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.

Investigations into whether the LUCA was a hyperthermophilic (optimal growth temperature (OGT) $\geq$80 °C), thermophilic (OGT 50–80 °C), or mesophilic (OGT $\leq$50 °C) organism have relied on correlations between the species' OGT and the composition of their macromolecular sequences. In extant prokaryotic species, the G+C content of rRNA stems (that is, double-stranded parts) has been shown to correlate with OGT[12]. Exploiting this correlation, support was obtained for a non-hyperthermophilic LUCA[2]. In contrast, studies based on correlations between the composition of the 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 evolutionary models used.

These previous 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. Two of these studies[13,14] relied on homogeneous models of evolution which make the simplifying hypothesis that substitutions occur with constant probabilities over time and across all lineages. If genomes and proteins had 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. Another approach[2] has been to use 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], which 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[18] mixture model to account for 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 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 the 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 data set of concatenated small- and large-subunit rRNAs—restricted to double-stranded regions—from 456 organisms (1,043 sites), and an amino acid data set of 56 concatenated nearly universal proteins from 30 organisms (3,336 sites), were assembled, each data set sampling all forms of cellular life. Correspondence analyses of the protein data set 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 and 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 I, V, Y, W, R, E and L (hereafter referred to as 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

[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. †Present address: Département de Biochimie, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal QC H3C3J7, Canada.
*These authors contributed equally to this work.