Research



# Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox

Céline Brochier*, Patrick Forterre† and Simonetta Gribaldo†

Addresses: *Equipe Phylogénomique, Université Aix-Marseille I, Centre Saint-Charles, 13331 Marseille Cedex 3, France. †Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris-Sud, 91405 Orsay, France.

Correspondence: Céline Brochier. E-mail: celine.brochier@up.univ-mrs.fr

Published: 26 February 2004

*Genome* ***Biology*** 2004, **5:**R17

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/3/R17

Received: 14 November 2003
Revised: 5 January 2004
Accepted: 21 January 2004



## Abstract

**Background:** Phylogenetic analysis of the Archaea has been mainly established by 16S rRNA sequence comparison. With the accumulation of completely sequenced genomes, it is now possible to test alternative approaches by using large sequence datasets. We analyzed archaeal phylogeny using two concatenated datasets consisting of 14 proteins involved in transcription and 53 ribosomal proteins (3,275 and 6,377 positions, respectively).

**Results:** Important relationships were confirmed, notably the dichotomy of the archaeal domain as represented by the Crenarchaeota and Euryarchaeota, the sister grouping of Sulfolobales and *Aeropyrum pernix*, and the monophyly of a large group comprising Thermoplasmatales, *Archaeoglobus fulgidus*, Methanosarcinales and Halobacteriales, with the latter two orders forming a robust cluster. The main difference concerned the position of *Methanopyrus kandleri*, which grouped with Methanococcales and Methanobacteriales in the translation tree, whereas it emerged at the base of the euryarchaeotes in the transcription tree. The incongruent placement of *M. kandleri* is likely to be the result of a reconstruction artifact due to the high evolutionary rates displayed by the components of its transcription apparatus.

**Conclusions:** We show that two informational systems, transcription and translation, provide a largely congruent signal for archaeal phylogeny. In particular, our analyses support the appearance of methanogenesis after the divergence of the Thermococcales and a late emergence of aerobic respiration from within methanogenic ancestors. We discuss the possible link between the evolutionary acceleration of the transcription machinery in *M. kandleri* and several unique features of this archaeon, in particular the absence of the elongation transcription factor TFS.

## Background

Deciphering the evolutionary history of the Archaea, the third domain of life [1,2], is essential to resolve a number of important issues, such as the dissection of their many eukaryote-like molecular mechanisms, understanding the adaptation of life to extreme environments, and the exploration of novel metabolic abilities (for recent reviews on the Archaea, see [3,4]). Until recently, the phylogeny of the Archaea was mainly based on 16S small ribosomal RNA (16S rRNA) sequence comparisons [5]. Such analyses, which included

environmental samples, suggest a diversity comparable to that of the Bacteria [3,6], with cultured lineages falling into two main phyla, the Euryarchaeota and the Crenarchaeota [2]. 16S rRNA trees suggest a specific order of emergence and mutual relationships among archaeal lineages that have important implications for understanding the evolution of many archaeal features, as well as the very nature of the archaeal ancestor. For example, the early emergence of Methanopyrales suggests that methanogenesis (methane production from $H_2$ and $CO_2$) is an ancestral character [7], whereas the sister grouping of Methanomicrobiales/Methanosarcinales and Halobacteriales would imply a late emergence of aerobic respiration in archaea.

New phylogenetic approaches that exploit the expanding database of completely sequenced archaeal genomes have recently challenged some of these conclusions. In particular, a consensus of a number of whole-genome trees based on gene-content comparison among all archaeal genomes does not recover the monophyly of Euryarchaeota, as Halobacteriales are at the base of the archaeal tree (see [8] and references therein). Moreover, whole-genome trees, whether based on gene content or on the conservation of gene order, pair-group *Methanopyrus kandleri* with Methanobacteriales and Methanococcales [9], contradicting the early branching of this archaeon in the 16S rRNA tree. Phylogenies based on whole-genome analyses may, however, be biased by the abundant lateral gene transfer (LGT) events that have occurred between archaea and bacteria, as well as between archaeal lineages [10-14]. For example, the early branching of Halobacteriales in whole-genome trees may reflect the fact that Halobacteriales contain a high number of genes of bacterial origin [15,16]. Similarly, the grouping of *M. kandleri* with other thermophilic methanogens may be explained by extensive LGT across different lineages of methanogens sharing the same biotopes.

One possible way to bypass the problem of LGT is to focus on informational proteins, as their genes are supposed to be less frequently transferred [17]. In general, the use of large datasets of concatenated sequences (that is, fusions) has proved very useful in increasing tree resolution, especially if procedures are used to remove from the analysis proteins that have been affected by LGT [18-21]. Our recent analyses of bacterial and archaeal phylogenies based on ribosomal proteins showed a minimal occurrence of transfers, suggesting that the phylogenetic signal carried by the components of the translation apparatus is not biased by LGT and can provide a *bona fide* species tree [20,21]. In archaeal trees based on a concatenated dataset of 53 ribosomal proteins from 14 taxa, the dichotomy Euryarchaeota/Crenarchaeota was recovered, with Halobacteriales being a sister group of Methanosarcinales, as in the 16S rRNA tree [21]. At that time, the position of *M. kandleri* could not be tested, as its genome was not yet available. A more recent tree based on a fusion dataset of ribosomal proteins has shown that *M. kandleri* groups with Methanobacteriales and Methanococcales [9], as in whole-genome trees [8]. Surprisingly, however, this analysis showed Halobacteriales at the base of the archaeal tree [9]. To further investigate archaeal phylogeny with components of informational systems, we updated our ribosomal protein concatenation by including newly available genome sequences, and we performed a similar analysis with proteins of the transcription apparatus. Previous analyses based on large subunits of archaeal RNA polymerases have indeed suggested that transcription proteins may be good phylogenetic markers for the archaeal domain [22].

## Results

### Sequence retrieval

By surveying proteins involved in transcription in 20 complete, or nearly complete, archaeal genomes we retrieved and constructed 15 sequence alignment datasets corresponding to 12 subunits of RNA polymerase and three transcription factors (see Materials and methods). Several of the archaeal RNA polymerase subunits do not have any homologs in bacteria, and all of them can be only partially aligned over their eukaryotic homologs (dramatically shortening the number of positions for analysis and increasing the risk of reconstruction artifacts). Consequently, as in Matte-Tailliez *et al.* [21], we decided not to include any bacterial/eukaryote outgroup in our analysis. To compare the results obtained with transcription proteins with those obtained with ribosomal proteins, our previous alignment dataset of ribosomal proteins [21] was updated by including four additional taxa (*Sulfolobus tokodaii, Thermoplasma volcanium, Methanopyrus kandleri, Methanococcus maripaludis*).

### Detection of LGT and dataset construction

Phylogenetic analyses were carried out on the 15 single datasets of transcription proteins in order to identify possible LGT events. Undisputed groups such as Thermoplasmatales, Halobacteriales, Sulfolobales, Thermococcales, Methanosarcinales and Methanococcales were recovered in the majority of the single trees (data not shown). However, other relationships were largely unresolved in several trees as a result of the small size of the datasets. The only case of putative LGT was detected in the phylogeny based on RNA polymerase subunit H, as Thermoplasmatales were robustly grouped with *M. kandleri* (83% Boostrap proportion (BP)) (Figure 1). This surprising grouping (never observed in other phylogenies), was also strongly supported by a well-conserved insert of five or six amino acids shared only by the RNA polymerase subunits H from *M. kandleri* and Thermoplasmatales (Figure 1). The proximity of Halobacteriales suggests that *M. kandleri* acquired its subunit H gene from Thermoplasmatales and not the other way round. RNA polymerase subunit H was thus excluded from further analysis in order to limit the introduction of a possible bias. The remaining 11 RNA polymerase subunits (A', A", B, D, E', E", F, K, L, N, P), and the transcription factors NusA, NusG and TFS, were then concatenated

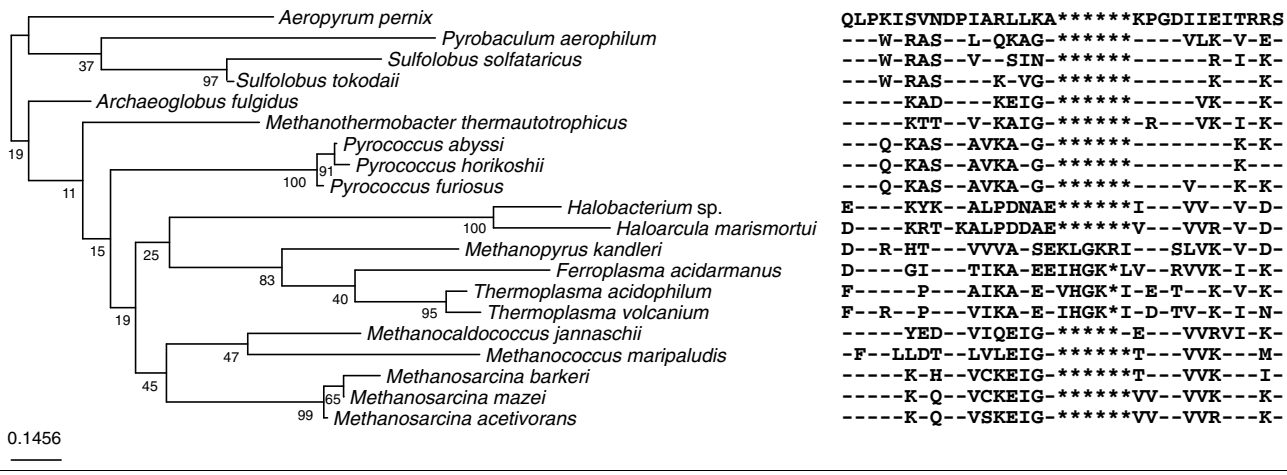


**Figure 1**
Unrooted neighbor-joining phylogenetic tree of the RNA polymerase subunit H computed from a Γ-corrected matrix of distances. Numbers close to nodes are bootstrap proportions. The scale bar represents the number of changes per position per unit branch length. For each taxon, the portion of the alignment from positions 57 to 83 is displayed. For clarity, identical amino acids shared by the current taxa and the first taxon (*Aeropyrum pernix*) are indicated by dashes, whereas stars correspond to missing amino acids.

into a large fusion of 3,275 amino acids. A previous analysis on 53 ribosomal proteins showed a minimal occurrence of LGT [21]. We did not observe any new case of LGT in our updated datasets with the four additional taxa. The 53 ribosomal proteins were thus concatenated into a large fusion containing 6,377 positions.

### Phylogenetic analyses

The trees resulting from the transcription and translation datasets (hereafter referred to as the 'transcription tree' and the 'translation tree') are shown in Figure 2a and 2b, respectively. The same topologies were recovered with the three methods used for phylogenetic reconstruction, but with little variation in bootstrap values (data not shown). The transcription and the translation trees presented interesting similarities, such as the Crenarchaeota/Euryarchaeota dichotomy (100% BP), the sister grouping of Sulfolobales and *Aeropyrum pernix* (84% and 100% BP) and the monophyly of a large group comprising Thermoplasmatales, *Archaeoglobus fulgidus*, Methanosarcinales and Halobacteriales (96% and 100% BP), with the latter two orders forming a well-sustained cluster (100% BP). However, the transcription tree strongly supported *A. fulgidus* as the sister group of the Methanosarcinales/Halobacteriales clade (100% BP), whereas in the translation tree *A. fulgidus* grouped, albeit with weak confidence (41% BP), with Thermoplasmatales. Moreover, the transcription tree recovered a robust monophyly (80% BP) of three methanogens (*Methanothermobacter thermoautotrophicum, Methanocaldococcus jannaschii*, and *Methanococcus maripaludis*), while in the translation tree these taxa were paraphyletic with a moderate support (BP 62%). The apparent incongruence between the two trees concerning the positions of *A. fulgidus* and of the three methanogens most probably reflects a lack of phylogenetic signal rather than LGT or long-branch attraction. Future analyses including more positions and a wider taxonomic sampling will help in resolving these nodes better. The two phylogenies differed remarkably concerning the base of the Euryarchaeota. The transcription tree showed *M. kandleri* as the first offshoot (100% BP) just before Thermococcales, whereas in the translation tree Thermococcales represented the most basal branch, with *M. kandleri* grouping paraphyletically with Methanococcales and Methanobacteriales (88% BP).

Interestingly, *M. kandleri* displayed a very long branch in the transcription tree (Figure 2a), a peculiarity not observed in the translation tree (Figure 2b), suggesting an acceleration of evolution of *M. kandleri* transcription proteins. We tested the possibility that this acceleration was due to a composition bias by removing aspartate and glutamate from the transcription dataset, as the proteome of *M. kandleri* displays an unusually high content of negatively charged amino acids [9], possibly as an adaptation to the very high intracellular salinity (1 M of cyclic 2,3-diphosphoglycerate) [23]. The resulting phylogeny was very similar to the transcription tree of Figure 2a, with *M. kandleri* emerging at the base with a very long branch (data not shown).

The comparison of the percentages of amino-acid differences in transcription and translation fusion datasets for each pair of species is shown in Figure 3. A strong correlation between the percentages of amino-acid differences in the two datasets could be observed for each pair of species (R = 0.88). For *M. kandleri*, however, this correlation was less strong, reflecting

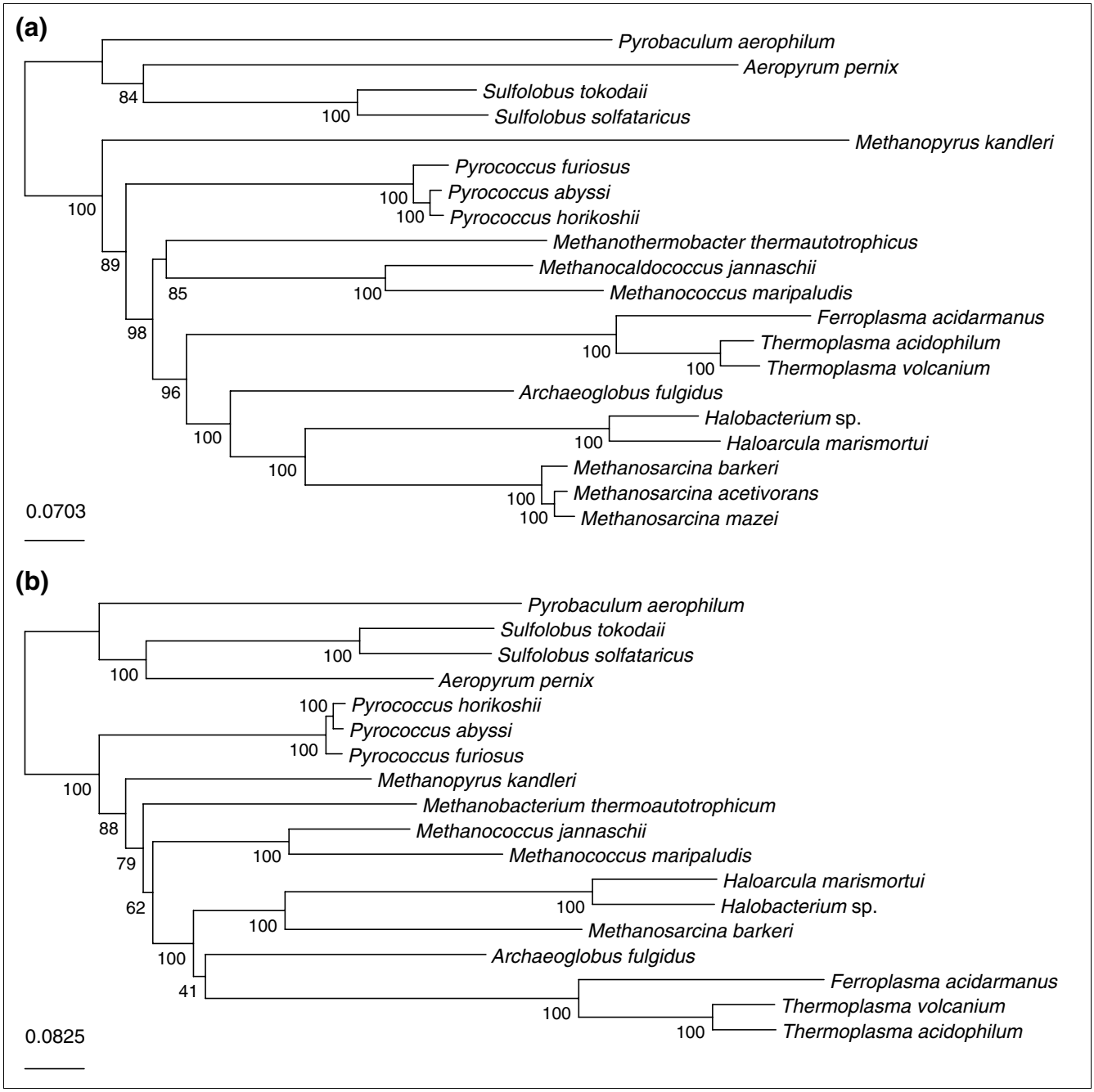


**Figure 2**
Unrooted maximum likelihood (ML) phylogenetic trees obtained from the transcription and translation datasets. **(a)** Transcription; **(b)** translation. The best tree and the branch lengths were calculated using the program PUZZLE with a Γ-law correction. Numbers at the nodes are ML bootstrap supports computed with the RELL method using the MOLPHY program without correction for among-site variation. The scale bars represent the number of changes per position per unit branch length.

the fact that the transcription dataset displayed much higher evolutionary rates compared to the translation dataset (see legend to Figure 3).

We then tested the possibility that the basal placement of *M. kandleri* in the transcription tree might be due to a biased phylogenetic signal specifically contributed by one or more RNA polymerase subunits. Indeed, we found that *M. kandleri* displayed a strongly supported basal position associated with a long branch in single trees based on RNA polymerase large subunits A' and A" (Figure 4a and 4b, respectively), whereas it was grouped with the two other thermophilic methanogens

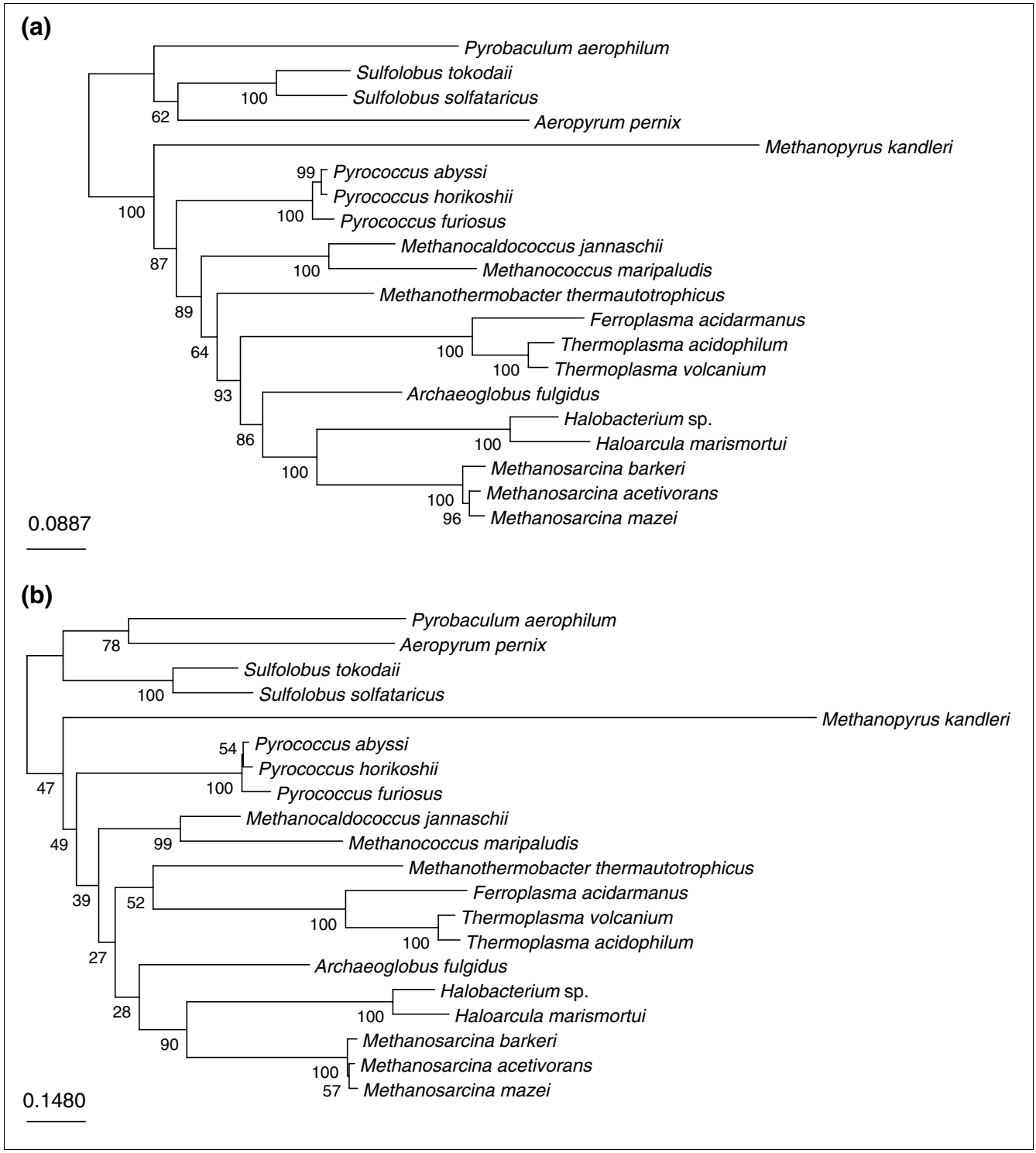


**Figure 3**
Unrooted neighbor-joining phylogenetic tree of the RNA polymerase subunits A' and A" computed from a Γ-corrected matrix of distances. **(a)** Polymerase A'; **(b)** polymerase A". Numbers close to nodes are bootstrap proportions. The scale bars represent the number of changes per position per unit branch length.

in a tree based on RNA polymerase large subunit B (Figure 5). This indicates that subunits A' and A" may be largely responsible for the basal placement of *M. kandleri* in the transcription dataset. This was not very surprising, as RNA

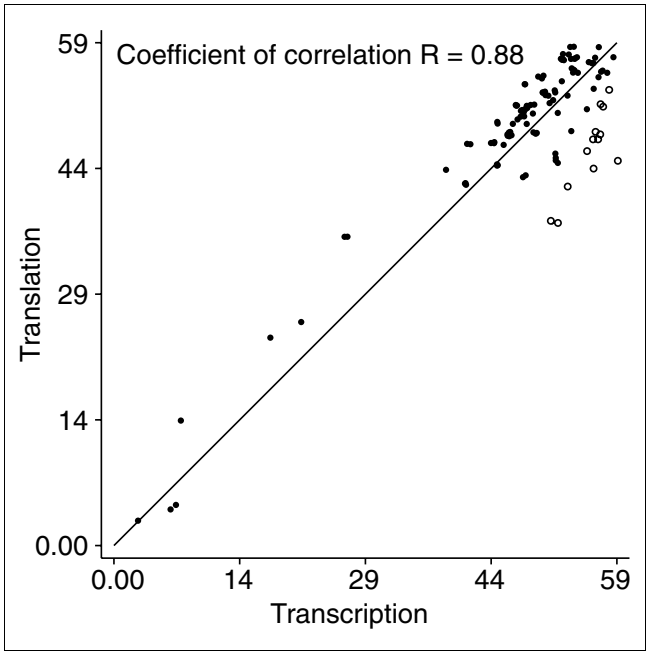


**Figure 4**
Comparison between the percentage of differences observed in the transcription and ribosomal datasets for each couple of taxa. The *x*-axis represents the percentage of amino-acid differences observed between two taxa for the concatenated transcription dataset. The *y*-axis represents the percentage of amino-acid differences observed between two taxa for the concatenated ribosomal dataset. Circles show for each pair of taxa the comparison between the observed percentage of differences for the concatenated transcription and ribosomal datasets. The majority of circles are localized close to the diagonal indicating a strong correlation (R = 0.88) between the differences observed into the two concatenated datasets. White circles represent the comparisons of *Methanopyrus kandleri* with other taxa.

polymerase A' and A" represents about 30% of the fusion sites (812 and 360 sites, respectively). However, as *M. kandleri* still emerged first when these subunits were removed from the dataset (data not shown), other factors may be involved. Interestingly, *M. kandleri* emerged with a relatively long branch at the base of the euryarchaeal part of a RNA polymerase subunit B tree reconstructed without correction for variation of evolutionary rates among sites (data not shown). When a Γ-law is taken into account, this basal placement disappears (Figure 4), strongly suggesting that long-branch attraction artifact could affect the *M. kandleri* placement.

### Rare evolutionary events

To gain further insight into the nodes showing contradictory placements between the transcription and translation trees, we searched for rare evolutionary events that may be used as synapomorphies for clade identification. We first analyzed the genomic context to look for possible signatures that support some nodes in our phylogenies. The genes encoding RNA polymerase subunits are clustered in several 'operon-like structures' in all archaeal genomes, together with genes encoding NusA, TFS, and several ribosomal proteins (data not shown). Unfortunately, we could not infer any possible grouping based on the structure of these operons, except for the confirmation of closely related species.

An interesting rare character in the transcription dataset was the split/fusion of the RNA polymerase B subunit [21,24]. This subunit is encoded by a single gene (*rpoB*) in crenarchaeotes, Thermococcales and Thermoplasmatales, and by two genes (*rpoB'* and *rpoB"*) in all other euryarchaeotes. The split of the B-subunit gene has taken place at the same position in all archaeal species, suggesting that it occurred only once in the archaeal domain. Consistently with both the *rpoB* tree (Figure 4) and translation trees (Figure 2b), the most parsimonious scenario that may explain the distribution of this character is the occurrence of a single *rpoB* gene split soon after the divergence of Thermococcales, followed by a gene fusion event in the lineage leading to Thermoplasmatales [21]. Importantly, this scenario supports the emergence of *M. kandleri* after Thermococcales.

Finally, we focused on large insertions/deletions (indels), as these events are less prone to convergence than amino-acid substitutions and may be potentially good phylogenetic characters [25]. Indels were looked for in all individual transcription protein datasets. Unfortunately, no indel-sharing indicative of phylogenetic relationship among groups could be found. Intriguingly, the proteins from the *M. kandleri* transcription set harbored a greater number of indels than observed in any other archaeal species; 27 of these indels were specific to this species, whereas the average number of indels specific to other archaeal lineages was between one and eight (Table 1). In addition, the specific indel regions in *M. kandleri* are frequently flanked by very highly divergent regions (Figure 6). The presence of such a high proportion of indels in the *M. kandleri* transcription dataset is consistent with an accelerated evolution of transcriptional proteins in this taxon with respect to any other archaeal lineage included in the present analysis.

## Discussion

The availability of completely sequenced genomes offers new opportunities to determine inter-species evolutionary relationships. It was suggested for some time that this task would be hopeless for prokaryotes because of the extent of LGT between domains and phyla [26,27]. However, it has subsequently been shown that a universal tree of life roughly similar to the 16S rRNA tree (with the tripartite division of cellular organisms) could be recovered by different whole-genome approaches, indicating that a *bona fide* phylogenetic signal may still be present in contemporary organisms [8,28,29]. Nevertheless, whole-genome trees are highly sensitive to LGT, which can produce misleading placements of specific lineages [30]. As an alternative approach, several authors have used sets of concatenated protein sequences to increase tree resolution [9,18-21,31]. These approaches are based on

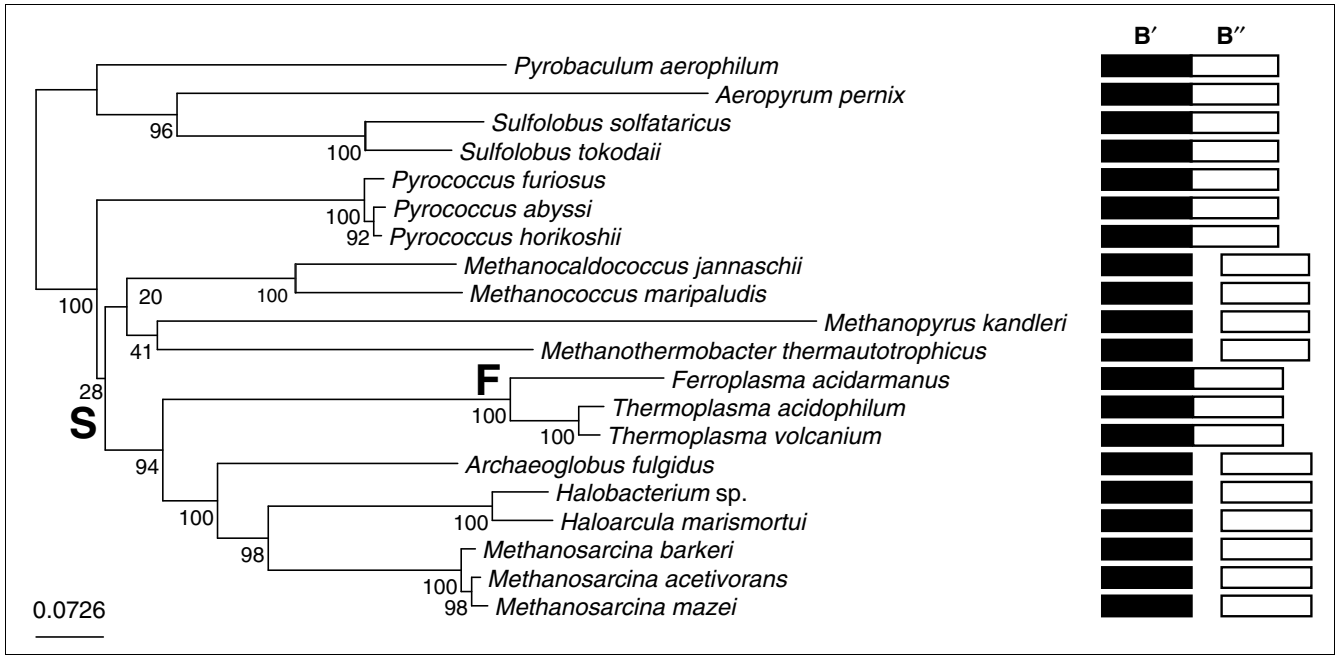


**Figure 5**
Unrooted neighbor-joining phylogenetic tree of the RNA polymerase subunit B computed from a Γ-corrected distance matrix. Numbers close to nodes are bootstrap proportions. The scale bar represents the number of changes per position for a unit branch length. In *Methanococcus maripaludis*, *Methanocaldococcus jannaschii*, *Methanopyrus kandleri*, *Methanothermobacter thermoautotrophicus*, *Archaeoglobus fulgidus*, Thermoplasmatales, Methanosarcinales and Halobacteriales genomes, the gene for the RNA polymerase subunit B is split in two parts: B' and B". The black and white boxes correspond to the B' and B" parts of the gene, respectively. S and F represent the split and fusion event hypotheses of the B' and B" parts of the gene.

the idea that a core of proteins (mostly informational proteins) has evolved mainly through vertical inheritance and can thus be used to retrace a genuine species phylogeny. Furthermore, by focusing on relatively small groups of proteins, it is possible to identify and remove proteins affected by LGT by performing single phylogenetic analyses. We have previously applied such a strategy to a dataset of ribosomal proteins used to retrace the phylogeny of the Bacteria [20] and the Archaea [21]. These analyses showed that LTG events involving ribosomal proteins are rare, and that these rare transfers affect the resulting phylogenies only slightly [20,21]. The similar analysis presented in this paper revealed no new case of LGT in our updated ribosomal protein dataset and a single case in the transcription dataset (Figure 1). This confirms that a large fraction of informational genes belong to a core of genes refractory to frequent transfers and they may therefore be used to retrace a genuine organismal phylogeny [17,32]. An alternative explanation may be that the genes involved in transcription and in translation are systematically transferred together. However, this hypothesis would imply the co-transfer and replacement of more than 70 genes localized in different regions of the genome.

The likely displacement of the original RNA polymerase subunit H of *M. kandleri* by the orthologous subunit from Thermoplasmatales indicates that orthologous displacement is nevertheless possible 'at the heart of the transcription machinery', at least across euryarchaeal lineages. The likely location of subunit H on the outside of the archaeal RNA polymerase, as in eukaryotic RNA polymerase [33,34], might facilitate its replacement. Interestingly, this gene replacement occurred *in situ*, that is without disruption of gene arrangement, as the phylogenies obtained from the nearest neighbors of the gene encoding subunit H in *M. kandleri* (subunits B, A' and A") did not indicate any specific affiliation of this species with Thermoplasmatales. Several such precise homologous gene displacements have recently been reported [35,36], and may be explained by a high rate of LGT and intrachromosomal recombination, followed by purifying selection for the maintenance of operon structure [36].

The phylogenies based on the transcription and translation datasets shared a number of nodes. In particular, a robust cluster comprising Thermoplasmatales, *A. fulgidus*, and a Halobacteriales/Methanosarcinales clade strengthens the notion of a late emergence of aerobic respiration in archaea from within methanogenic ancestors. This result is in agreement both with the classical rooted 16S rRNA trees [5] and with a recent whole-genome tree obtained by Daubin *et al.* [37]. Furthermore, the hypothesis of a late emergence of aerobic respiration in Halobacteriales is in line with the finding that enzymes involved in this process in *Halobacterium* were probably recruited by LGT from bacteria [16]. Our results thus strengthen the hypothesis that the early emergence of

**Table 1**

**Indels in the 12 subunits of RNA polymerase**

| | Total number of indels | Number of specific indels | Percentage of specific indels |
|---|---|---|---|
| *Aeropyrum pernix* | 38 | 4 | 10.53 |
| *Pyrobaculum aerophilum* | 33 | 5 | 15.15 |
| *Sulfolobus solfataricus* | 28 | 1 | 3.57 |
| *Sulfolobus tokodaii* | 30 | 3 | 10 |
| *Archaeoglobus fulgidus* | 17 | 2 | 11.76 |
| *Halobacterium* sp. | 23 | 2 | 13.04 |
| *Haloarcula marismortui* | 24 | 3 | 12.50 |
| *Methanocaldococcus jannaschii* | 24 | 7 | 29.17 |
| *Methanococcus maripaludis* | 22 | 6 | 27.27 |
| *Methanopyrus kandleri* | 57 | 27 | 47.37 |
| Methanosarcinales | 10 | 2 | 20 |
| *Methanothermobacter thermautotrophicus* | 17 | 2 | 11.76 |
| Thermococcales | 19 | 2 | 10.53 |
| Thermoplasmatales | 36 | 8 | 22.22 |

For each species, regions containing insertions/deletions (indels) have been counted for the 12 RNA polymerase subunits (A', A", B, D, E', E", F, H, K, L, N, P), TFS, NusA and NusG. We use 'indel region' terms because if two species exhibit indels in the same region, even if they are different sizes, we count this region as a shared indel region. For each species, the number and percentage of specific regions containing indels (that is, the indel region is exclusive to that species and is not shared by any other species) are indicated. As they share exactly the same indels, the three *Pyrococcus* species, the three *Methanosarcina* species and the two *Thermoplasma* species plus *Ferroplasma* are grouped in Thermococcales, Methanosarcinales and Thermoplasmatales respectively. Consequently, the specific indels are those specific to the group.

*Halobacterium* species in some whole-genome trees might be due to the high proportion of genes of bacterial origin in *Halobacterium* [8,15]. The early branching of halobacteria in the ribosomal protein tree published by Slesarev *et al.* [9] may be explained by an artifact caused by the inclusion of a bacterial outgroup, as archaeal ribosomal proteins are difficult to align over their bacterial homologs.

We were particularly interested in clarifying the controversial position of *M. kandleri*, as this is relevant to the important issue of the origin of methanogenesis [7]. The emergence of *M. kandleri* at the base of the euryarchaeal phylum in the 16S rRNA tree would point to a methanogenic (and hyperthermophilic) ancestor for euryarchaeotes, and possibly for all the Archaea. Accordingly, some specific features of *M. kandleri* have been interpreted as ancient characters. An example is the presence of an unsaturated terpenoid, considered to be a precursor of normal archaeal lipids, as the major membrane component [38]. However, following the recently published genome of *M. kandleri*, whole-genomes trees constructed by different methods, as well as ribosomal protein trees, have challenged the supposed ancestral character of this lineage, suggesting instead that *M. kandleri* should be included with other methanogens in a monophyletic group [9]. Our translation tree was in agreement with Slesarev *et al.*, showing a placement of *M. kandleri* just after Thermococcales and close to Methanobacteriales and Methanococcales (Figure 2b), thus further supporting a relatively late emergence of methanogenesis in the Archaea. The emergence of *M. kandleri* at the base of the Euryarchaeota (that is, before Thermococcales) in the transcription tree (Figure 2a) was reminiscent of that observed (albeit with lower support) in the 16S rRNA tree [3]. However, the long branch of *M. kandleri* suggests that this basal placement in the transcription tree may be due to a tree-reconstruction artifact, possibly magnified by a misleading phylogenetic signal contributed by the large RNA polymerase subunits A'/A" (Figure 4a and 4b). Consequently, the late emergence of this species observed in the translation tree (Figure 2b), which is not likely to be biased by tree-reconstruction artifacts, is probably the correct one.

Moreover, a late placement of *M. kandleri* is congruent with our analysis of the split/fusion of RNA polymerase B subunit (Figure 5), as an early emergence of this taxon would imply a less parsimonious scenario involving an additional split event for the *rpoB* gene. Importantly, the inclusion of Methanosarcinales in our analysis clearly indicates that methanogens are not monophyletic, as the common ancestor of all methanogens is also the ancestor of non-methanogenic organisms (Thermoplasmatales, Halobacteriales and Archaeoglobales). The presence in this group of non-methanogenic lineages would be due to secondary loss, as is indeed suggested by the presence of relics of the methanogenic pathway in *A. fulgidus* [9,39].

| | |
|---|---|
| *Sulfolobus tokodaii* | KKVEELIKQYNE**GTLELIP********GRTAEESLEDHILETLDQLRKVAGDIATKY |
| *Sulfolobus solfataricus* | VE-DN--QK-KN**-E--P--*********---L-----NY--D---K--ST-----S-- |
| *Pyrobaculum aerophilum* | GR-YSI-EE-RK**----P--*********--SL-----IK-M-V--E---RVQEV-SNN |
| *Aeropyrum pernix* | E--DK--DDFRS**-H--AM-*********-F-V--TF-NKVT-I-SKV-ED-AVVVE-- |
| *Pyrococcus abyssi* | ER-QK--EA-KR**-E--PL-*********-K-L--T--SK-MAV-AEA-DN--S--E-- |
| *Pyrococcus horikoshii* | ER-QR--EA-KR**-E--PL-*********-KSL--T--SK-MAV-AEA-DN--SV-E-- |
| *Pyrococcus furiosus* | ER-NK--EA-KR**-E--PL-*********-KSL-DT--SL-MAV-AEA-DN--AV-E-- |
| *Methanocaldococcus jannaschii* | E--K-I-EK-ER**-E---L-*********-LNL---R-AY-SNV-REA-DK--A--ER- |
| *Methanococcus maripaludis* | QD--DIVEK-EN**----SL-*********--GV---R-AY-MQI-GKA-DQ--NV-E-- |
| *Methanothermobacter thermautotrophicus* | AR-DQ--EA-EN**-E--PL-*********--SL--T--MK-MQV-GEA-DKS-E--ES- |
| *Methanosarcina barkeri* | DA-QN---S-QN**KE--PL-*********---LD-TI-MS-MQK-GKA-DET-G--DSH |
| *Methanosarcina mazei* | DA-QN---S-IN**KE-DPL-*********---LD-TI-MS-MQK-GKA-DKT-N--DSH |
| *Methanosarcina acetivorans* | DA-QN---S-QN**KE--PL-*********---LD-TI-MS-MQK-GKA-DET-N--DSH |
| *Archaeoglobus fulgidus* | NE-NR--EA-RR**-D--PM-*********--SI--T--MR-MQV-GRA-DR--K--QRH |
| *Halobacterium* sp. | DR-Q---ET-EN**-D--SL-*********---VD-T--MK-MQ--GKA-DS---V-EEN |
| *Haloarcula marismortui* | DR-----ET-DR**-E--SL-*********---VD-T--MK-MQ--GKA-DS-----EDH |
| *Thermoplasma acidophilum* | DRINK--ET-KR**-E-QPA-*********--SV-DT--IE--SEAGVV-DES-K--SS- |
| *Thermoplasma volcanium* | DRINK--ETFRR**-E-QPA-*********--SV-DT--ME--SEAGVV-DES-K--SS- |
| *Ferroplasma acidarmanus* | E-I-K-VDAFRS**-Q-QPL-*********--SV-DT--VE--SSTGGV-DES-K--SQ- |
| *Methanopyrus kandleri* | --AK-I-ERGERRLQE--EHETCNRSRIER-EML-RNI-SEVMAI-N-P-VETERLLK-H |

**Figure 6**
An example of an indel being flanked by divergent regions in *Methanopyrus kandleri*. The portion of the alignment corresponds to positions 1,281 to 1,340 in our RNA polymerase subunit A' dataset. For clarity, identical amino acids shared by each taxon and the first taxon (*Sulfolobus tokodaii*) are indicated by dashes, whereas stars correspond to missing amino acids.

In the present study we show that *M. kandleri* displays higher evolutionary rates in its transcriptional proteins (Figure 3) compared with the other archaeal species analyzed, consistently with a surprisingly high number of specific indels (Table 1). We have identified two new specific features in the molecular biology of *M. kandleri* that may explain such evolutionary acceleration: the displacement of RNA polymerase subunit H by a homologous protein from a distantly related archaeal lineage, and the loss of the transcription factor TFS. As both proteins contact the RNA polymerase core [33,34,40,41], their replacement or loss may have led to the overall release of evolutionary constraints in core RNA polymerase subunits. This phenomenon was possibly further amplified by an extremely low diversity of signaling systems in the genome of *M. kandleri*, and an unusual under-representation of DNA-binding proteins generally implicated in transcriptional regulation of specific operons in archaea [9].

The absence of transcription elongation factor TFS in *M. kandleri* is especially intriguing. Archaeal TFSs are homologous to both eukaryotic RNA polymerase subunit M and to the carboxy-terminal domain of the eukaryotic transcription elongation factor TFIIS [42]. However, biochemical experiments have shown that archaeal TFS is not part of the RNA polymerase core and displays an activity more consistent with the function of eukaryotic TFIIS [43,44]. Eukaryotic TFIIS has the ability to strongly enhance the weak intrinsic nuclease activity of RNA polymerase II (PolII), allowing it to bypass template-arrest sites by activating the cleavage reaction of nascent RNAs and releasing stalled RNA polymerase complexes [45]. Bacteria have no homolog of TFIIS, but two functional analogs, GreA and GreB, which perform exactly the same reaction *in vitro* and interact with the RNA polymerase core in a very similar fashion [40,41]. The ubiquitous distribution of TFIIS in eukaryotes and GreA/GreB in bacteria underlines the extremely important role of these proteins, which is probably similar for archaeal TFS (for reviews, see [46,47]). To our knowledge, *M. kandleri* is the only cellular organism whose genome has been completely sequenced that lacks a homolog of either TFS or GreA/GreB. Given the high evolutionary rates of the transcriptional machinery in *M. kandleri*, the absence of TFS may be tolerated because of specific mutations in the sequence of large subunits that would either increase the intrinsic RNA polymerase nuclease activity, or render stalled elongation complexes less stable, leading to the dispensability of TFS-mediated dissociation [48]. Alternatively, TFS function may be replaced in *M. kandleri* by a non-homologous enzyme yet to be discovered.

It is tempting to speculate that these peculiarities in the transcription apparatus of *M. kandleri* may explain a number of unique features of this species by the effects of some alteration in this machinery on the evolution of this organism. Indeed, in addition to the presence of unusual lipids in its membranes, *M. kandleri* displays specific features not observed in other archaea. This is the case for its reverse gyrase, for example. In all other hyperthermophilic archaeal taxa reverse gyrase is a monomer formed by the fusion of a

helicase and a topoisomerase, but in *M. kandleri* it is composed of two proteins, one corresponding to the helicase module and the amino terminus of the topoisomerase module and the other to the carboxy terminus of the topoisomerase module [49]. Another peculiarity of *M. kandleri* is its histone protein, formed by the fusion of two monomers into a single polypeptide containing two tandemly repeated histone folds [50]. Interestingly, the recent sequencing of the *M. kandleri* genome has identified several other cases of unique protein fusions [9]. Also, *M. kandleri* contains the largest proportion of orphan genes found in any prokaryotic genome [51]. This is reminiscent of the presence in *M. kandleri* of a unique DNA topoisomerase, Topo V, which is exclusive to this archaeon [52]. All these observations suggest an unusually high level of gene loss, gene capture and intramolecular recombination (producing gene fusions and formation of indels) in this archaeon.

We hypothesize that the loss of TFS in *M. kandleri* may be directly linked to all these oddities. In fact, as TFIIS, as well as GreA/GreB, is involved in the release of stalled elongation complexes [45] and transcription fidelity [53,54], an appealing hypothesis is that the absence of TFS in *M. kandleri* may induce some transcriptional mutagenesis. For instance, absence of TFS may possibly allow transcriptional bypass of DNA lesions that would normally trigger transcription-coupled repair systems. Also, the lack of TFS may prevent dissociation of stalled complexes and consequently increase the number of replication fork disruptions due to collision between the replication and transcription machineries. This situation may mobilize mutagenic DNA repair systems to promote replication restart via homologous recombination. Of course, one cannot exclude the possibility that all the idiosyncrasies of *M. kandleri* may be due to another as-yet undetermined feature of this organism, such as the one that triggered the initial evolutionary acceleration of RNA polymerase subunits that may have facilitated the loss of TFS. Nevertheless, the hypothesis of a direct effect of the loss of a TFIIS-like transcription elongation factor on the rate of genome evolution is fascinating and should be readily testable using the *TFIIS* and *greA greB* mutants already available. If this hypothesis turns out to be correct, this would imply a strong correlation, previously unnoticed, between transcription and the rate of genome evolution.

## Materials and methods

### Sequence retrieval and dataset construction

All proteins annotated as implicated in transcription in the genome of *Pyrococcus abyssi* [55] were used as seeds for BLASTP and PSI-BLAST searches [56] on 20 complete or near-complete archaeal genomes (*Pyrobaculum aerophylum*; *Aeropyrum pernix*; the two Sulfolobales - *Sulfolobus solfataricus* and *S. tokodaii*; the three Thermococcales - *Pyrococcus furiosus*, *P. horikoshii* and *P. abyssi*; the two Methanococcales - *Methanococcus maripaludis* and *Methanocaldococcus jannaschii*; the Methanobacteriales *Methanothermobacter thermoautotrophicus*; the Methanopyrales *Methanopyrus kandleri*; the three Thermoplasmatales - *Ferroplasma acidarmanus*, *Thermoplasma acidophilum* and *T. volcanium*; the Archaeoglobales *A. fulgidus*; the three Methanosarcinales - *Methanosarcina barkeri*, *M. mazei* and *M. acetivorans*; and the two Halobacteriales *Halobacterium* species and *Haloarcula marismortui*). The protein sequences retrieved were: rpoA' (PAB0424), rpoA" (PAB0425), rpoB (PAB0423), rpoD (PAB2410), rpoE' (PAB1105), rpoE" (PAB7428), rpoF (PAB0732), rpoH (PAB7151), rpoK (PAB7132), rpoL (PAB2316), rpoM/TFS (PAB1464), rpoN (PAB7131), rpoP (PAB3072), NusA (PAB0426), NusG (PAB2352), TPB (PAB1726), TFB (PAB1912), TFE (PAB0950), TFIIH (PAB2385), TIP49 (PAB2107). BLAST searches were performed at the National Center for Biotechnology Information (NCBI) [57] for published sequences, and locally for two unfinished genomes *Haloarcula marismortui* ([58] and S. DasSarma, personal communication) and *Methanococcus maripaludis* strain LL [59].

For some proteins of small size, additional TBLASTN searches were performed, as they were not annotated or their sequences were partial (for example, the complete sequence of the RNA polymerase subunit K from *Ferroplasma acidarmanus* was retrieved by this approach, as the annotated sequence was partial as a result of misdetection of the initial methionine). Single protein datasets were aligned by CLUSTALW [60], manually refined by the use of the program ED from the MUST package [61].

We retained only the proteins which were present in a single copy in each genome and which were missing in not more than one species. The majority of transcription factors (*bona fide* or putative) were discarded, as they were present in multiple copies (TBP, TFB) or had a scattered distribution (for example, TFE, TFIIH, TIP49), which prevented their reliable use as phylogenetic markers. We thus kept only the putative transcription factors NusA, NusG, and TFS (also annotated as RNA polymerase subunit M). Although present in two copies in *Halobacterium* sp. and *Haloarcula marismortui*, TFS was retained because phylogenetic analysis indicated a recent duplication event specific to Halobacteriales (data not shown). Surprisingly, no TFS homolog was found in the complete genome of *M. kandleri*. We also gathered 12 proteins annotated as RNA polymerase subunits (A', A", B, D, E", E", F, H, K, L N, P). Subunits E" and P were not found in *Ferroplasma acidarmanus*, possibly because the genome sequence of this species is still incomplete. Finally, 15 aligned datasets were kept for transcription proteins (NusA, NusG, TFS, and 12 RNA polymerase subunits).

Previous datasets of archaeal ribosomal proteins [21] were updated to include four additional taxa (*Sulfolobus tokodaii*, *Methanopyrus kandleri*, *Thermoplasma volcanium*,

*Methanococcus maripaludis*). The 53 datasets presenting a sufficient taxonomic sampling and no evidence of multiple paralogies and/or LGT were retained and concatenated into two large fusions, one consisting of 14 proteins of the transcription apparatus (3,275 amino-acid positions), the other consisting of 53 ribosomal proteins (6,377 positions).

### Phylogenetic analyses

Phylogenetic analyses were performed on both single and concatenated datasets by neighbor-joining [62] from gamma-corrected JTT-F distance matrices calculated by PUZZLE 4.0 [63]. Bootstrap values were calculated on 1,000 replicates of the original alignments [61]. Maximum likelihood (ML) analyses were performed by ProtML of the MOLPHY 2.3 package [64]. ML trees were selected among the 2,000 top-ranking trees resulting from heuristic searches using the JTT-F model of amino-acid substitution [65]. ML bootstrap proportions were computed using the RELL (ReEstimation of Log likelihood) method [66]. To conclude, exhaustive topology searches were also performed on a partially constrained starting tree where a few undisputed nodes were chosen according to preliminary analyses {{(*S. solfataricus, S. tokodaii*), *Aeropyrum pernix, Pyrobaculum aerophylum*}, *M. kandleri*, ((*P. abyssi, P. horikoshii*), *P. furiosus*), (*Ferroplasma acidarmanus*, (*T. acidophilum, T. volcanium*)), *A. fulgidus*, ((*M. acetivorans, M. mazei*), *M. barkeri*), (*Halobacterium* species, *Haloarcula marismortui*), *Methanocaldococcus jannaschii, Methanococcus maripaludis, Methanothermobacter thermautotrophicus*}. ML values were computed for each topology by PUZZLE 4.0 [63] by taking into account the rate among site variations by a gamma correction (eight rates). All individual and concatenated alignments and the corresponding phylogenetic trees are available online [67].

## Acknowledgements

S.G. was the recipient of a postdoctoral fellowship from the Association de la Recherche contre le Cancer (ARC).

## References

1. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74:**5088-5090.
2. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proc Natl Acad Sci USA* 1990, **87:**4576-4579.
3. Forterre P, Brochier C, Philippe H: **Evolution of the Archaea.** *Theor Popul Biol* 2002, **61:**409-422.
4. Makarova KS, Koonin EV: **Comparative genomics of archaea: how much have we learned in six years, and what's next?** *Genome Biol* 2003, **4:**115.
5. Woese CR: **Bacterial evolution.** *Microbiol Rev* 1987, **51:**221-271.
6. Delong EF, Pace NR: **Environmental diversity of bacteria and archaea.** *Syst Biol* 2001, **50:**470-478.
7. Burggraf S, Stetter KO, Rouviere P, Woese CR: ***Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens.** *Syst Appl Microbiol* 1991, **14:**346-351.
8. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life.** *Trends Genet* 2002, **18:**472-479.
9. Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, *et al.*: **The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens.** *Proc Natl Acad Sci USA* 2002, **99:**4644-4649.
10. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: **Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles.** *Trends Genet* 1998, **14:**442-444.
11. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, *et al.*: **Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*.** *Nature* 1999, **399:**323-329.
12. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304.
13. Gogarten JP, Doolittle WF, Lawrence JG: **Prokaryotic evolution in light of gene transfer.** *Mol Biol Evol* 2002, **19:**2226-2238.
14. Jain R, Rivera MC, Moore JE, Lake JA: **Horizontal gene transfer in microbial genome evolution.** *Theor Popul Biol* 2002, **61:**489-495.
15. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, *et al.*: **Genome sequence of *Halobacterium* species NRC-1.** *Proc Natl Acad Sci USA* 2000, **97:**12176-12181.
16. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S: **Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence.** *Genome Res* 2001, **11:**1641-1650.
17. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96:**3801-3806.
18. Moreira D, Le Guyader H, Philippe H: **The origin of red algae and the evolution of chloroplasts.** *Nature* 2000, **405:**69-72.
19. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Muller M, Philippe H: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*.** *Proc Natl Acad Sci USA* 2002, **99:**1414-1419.
20. Brochier C, Bapteste E, Moreira D, Philippe H: **Eubacterial phylogeny based on translational apparatus proteins.** *Trends Genet* 2002, **18:**1-5.
21. Matte-Tailliez O, Brochier C, Forterre P, Philippe H: **Archaeal phylogeny based on ribosomal proteins.** *Mol Biol Evol* 2002, **19:**631-639.
22. Klenk HP, Palm P, Lottspeich F, Zillig W: **Component H of the DNA-dependent RNA polymerases of Archaea is homologous to a subunit shared by the three eucaryal nuclear RNA polymerases.** *Proc Natl Acad Sci USA* 1992, **89:**407-410.
23. Shima S, Herault DA, Berkessel A, Thauer RK: **Activation and thermostabilization effects of cyclic 2, 3-diphosphoglycerate on enzymes from the hyperthermophilic *Methanopyrus kandleri*.** *Arch Microbiol* 1998, **170:**469-472.
24. Puhler G, Leffers H, Gropp F, Palm P, Klenk HP, Lottspeich F, Garrett RA, Zillig W: **Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome.** *Proc Natl Acad Sci USA* 1989, **86:**4569-4573.
25. Gribaldo S, Philippe H: **Ancient phylogenetic relationships.** *Theor Popul Biol* 2002, **61:**391-408.
26. Doolittle WF: **Lateral genomics.** *Trends Cell Biol* 1999, **9:**M5-M8.
27. Pennisi E: **Is it time to uproot the tree of life?** *Science* 1999, **284:**1305-1307.
28. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21:**108-110.
29. Fitz-Gibbon ST, House CH: **Whole genome-based phylogenetic analysis of free-living microorganisms.** *Nucleic Acids Res* 1999, **27:**4218-4222.
30. Korbel JO, Snel B, Huynen MA, Bork P: **SHOT: a web server for the construction of genome phylogenies.** *Trends Genet* 2002, **18:**158-162.
31. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: **Universal trees based on large combined protein sequence data sets.** *Nat Genet* 2001, **28:**281-285.
32. Daubin V, Moran NA, Ochman H: **Phylogenetics and the cohesion of bacterial genomes.** *Science* 2003, **301:**829-832.
33. Bushnell DA, Kornberg RD: **Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription.** *Proc Natl Acad Sci USA* 2003, **100:**6969-6973.
34. Armache KJ, Kettenberger H, Cramer P: **Architecture of initiation-competent 12-subunit RNA polymerase II.** *Proc Natl Acad Sci USA* 2003, **100:**6964-6968.

35. Brochier C, Philippe H, Moreira D: **The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome.** *Trends Genet* 2000, **16:**529-533.
36. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV: **Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ*.** *Genome Biol* 2003, **4:**R55.
37. Daubin V, Gouy M, Perriere G: **A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history.** *Genome Res* 2002, **12:**1080-1090.
38. Hafenbradl D, Keller M, Thiericke R, Stetter KO: **A novel unsaturated archaeal ether core lipid from the hyperthermophile *Methanopyrus kandleri*.** *Syst Appl Microbiol* 1993, **16:**165-169.
39. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, *et al.*: **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*.** *Nature* 1997, **390:**364-370.
40. Kettenberger H, Armache Kj, Cramer P: **Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage.** *Cell* 2003, **114:**347-357.
41. Opalka N, Chlenov M, Chacon P, Rice WJ, Wriggers W, Darst SA: **Structure and function of the transcription elongation factor GreB bound to bacterial RNA polymerase.** *Cell* 2003, **114:**335-345.
42. Kaine BP, Mehr IJ, Woese CR: **The sequence, and its evolutionary implications, of a *Thermococcus celer* protein associated with transcription.** *Proc Natl Acad Sci USA* 1994, **91:**3854-3856.
43. Hausner W, Lange U, Musfeldt M: **Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase.** *J Biol Chem* 2000, **275:**12393-12399.
44. Best AA, Olsen GJ: **Similar subunit architecture of archaeal and eukaryal RNA polymerases.** *FEMS Microbiol Lett* 2001, **195:**85-90.
45. Awrey DE, Weilbaecher RG, Hemming SA, Orlicky SM, Kane CM, Edwards AM: **Transcription elongation through DNA arrest sites. A multistep process involving both RNA polymerase II subunit RPB9 and TFIIS.** *J Biol Chem* 1997, **272:**14747-14754.
46. Wind M, Reines D: **Transcription elongation factor SII.** *BioEssays* 2000, **22:**327-336.
47. Fish RN, Kane CM: **Promoting elongation with transcript cleavage stimulatory factors.** *Biochim Biophys Acta* 2002, **1577:**287-307.
48. Trautinger BW, Lloyd RG: **Modulation of DNA repair by mutations flanking the DNA channel through RNA polymerase.** *EMBO J* 2002, **21:**6944-6953.
49. Krah R, Kozyavkin SA, Slesarev AI, Gellert M: **A two-subunit type I DNA topoisomerase (reverse gyrase) from an extreme hyperthermophile.** *Proc Natl Acad Sci USA* 1996, **93:**106-110.
50. Slesarev AI, Belova GI, Kozyavkin SA, Lake JA: **Evidence for an early prokaryotic origin of histones H2A and H4 prior to the emergence of eukaryotes.** *Nucleic Acids Res* 1998, **26:**427-430.
51. Jensen LJ, Skovgaard M, Sicheritz-Ponten T, Jorgensen MK, Lundegaard C, Pedersen CC, Petersen N, Ussery D: **Analysis of two large functionally uncharacterized regions in the *Methanopyrus kandleri* AV19 genome.** *BMC Genomics* 2003, **4:**12.
52. Slesarev AI, Belova GI, Lake JA, Kozyavkin SA: **Topoisomerase V from *Methanopyrus kandleri*.** *Methods Enzymol* 2001, **334:**179-192.
53. Erie DA, Hajiseyedjavadi O, Young MC, Von Hippel PH: **Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription.** *Science* 1993, **262:**867-873.
54. Thomas MJ, Platas AA, Hawley DK: **Transcriptional fidelity and proofreading by RNA polymerase II.** *Cell* 1998, **93:**627-637.
55. Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Querellou J, Ripp R, *et al.*: **An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*.** *Mol Microbiol* 2003, **47:**1495-1512.
56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
57. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/BLAST]
58. **Halophile Genome Project homepage** [http://zdna2.umbi.umd.edu]
59. **University of Washington Genome Center** [http://www.genome.washington.edu]
60. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
61. Philippe H: **MUST, a computer package of Management Utilities for Sequences and Trees.** *Nucleic Acids Res* 1993, **21:**5264-5272.
62. Felsenstein J: **An alternating least squares approach to inferring phylogenies from pairwise distances.** *Syst Biol* 1997, **46:**101-111.
63. Strimmer K, Von Haeseler A: **PUZZLE.** *Mol Biol Evol* 1996, **13:**964-969.
64. Adachi J, Hasegawa M: **MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood.** *Comput Sci Monogr* 1996, **28:**1-150.
65. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8:**275-282.
66. Kishino H, Miyata T, Hasegawa M: **Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts.** *J Mol Evol* 1990, **31:**151-160.
67. **Additional data for this manuscript** [http://www.up.univ-mrs.fr/evol/phylogenomics-lab/celine/celine.html]