BMC Genomics

BioMed Central

Research article

Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses

Olga Zhaxybayeva and J Peter Gogarten*

Address: Department of Molecular and Cell Biology University of Connecticut 75 North Eagleville Road Storrs, CT 06269-3044 USA

E-mail: Olga Zhaxybayeva - olga@carrot.mcb.uconn.edu; J Peter Gogarten* - gogarten@uconn.edu

*Corresponding author

Published: 5 February 2002

BMC Genomics 2002, **3**:4

Received: 4 December 2001

Accepted: 5 February 2002

This article is available from: http://www.biomedcentral.com/1471-2164/3/4

Abstract

Background: Horizontal gene transfer (HGT) played an important role in shaping microbial genomes. In addition to genes under sporadic selection, HGT also affects housekeeping genes and those involved in information processing, even ribosomal RNA encoding genes. Here we describe tools that provide an assessment and graphic illustration of the mosaic nature of microbial genomes.

Results: We adapted the Maximum Likelihood (ML) mapping to the analyses of all detected quartets of orthologous genes found in four genomes. We have automated the assembly and analyses of these quartets of orthologs given the selection of four genomes. We compared the ML-mapping approach to more rigorous Bayesian probability and Bootstrap mapping techniques. The latter two approaches appear to be more conservative than the ML-mapping approach, but qualitatively all three approaches give equivalent results. All three tools were tested on mitochondrial genomes, which presumably were inherited as a single linkage group.

Conclusions: In some instances of interphylum relationships we find nearly equal numbers of quartets strongly supporting the three possible topologies. In contrast, our analyses of genome quartets containing the cyanobacterium *Synechocystis* sp. indicate that a large part of the cyanobacterial genome is related to that of low GC Gram positives. Other groups that had been suggested as sister groups to the cyanobacteria contain many fewer genes that group with the *Synechocystis* orthologs. Interdomain comparisons of genome quartets containing the archaeon *Halobacterium* sp. revealed that *Halobacterium* sp. shares more genes with Bacteria that live in the same environment than with Bacteria that are more closely related based on rRNA phylogeny . Many of these genes encode proteins involved in substrate transport and metabolism and in information storage and processing. The performed analyses demonstrate that relationships among prokaryotes cannot be accurately depicted by or inferred from the tree-like evolution of a core of rarely transferred genes; rather prokaryotic genomes are mosaics in which different parts have different evolutionary histories. Probability mapping is a valuable tool to explore the mosaic nature of genomes.

Background

The introduction of small subunit ribosomal RNA as a tool in microbial taxonomy by Carl Woese and George Fox [1] led most microbiologists to assume that the concepts of animal and plant taxonomy could be extended to the realm of prokaryotes. In particular, it was assumed

that a natural taxonomic system for microorganisms was feasible [2]. The goal of a natural taxonomic system is the formation of taxonomic groups that are defined by shared ancestry [3]. By definition, an ancestor that defines a monophyletic group can only give rise to members of this group. No organism outside this group has a lineage that traces back to the same ancestor (paraphyletic group); however, there might be earlier ancestors that define more inclusive monophyletic groups. The metaphor for organismal evolution that underlies a natural taxonomic system is a strictly bifurcating tree of species. A decade ago ribosomal RNA promised that one day it might be possible to place every extant organism on a universal tree of life, and the hope was that more genomic sequences would make this placement more accurate.

However, the analyses of completely sequenced genomes initiated a reassessment of concepts in microbial evolution [4]. While some molecular markers were found to agree with one another e.g., [5], others do not [6–12]. Transfer of genetic information between divergent organisms has turned the tree of life into a net or web [13], and genomes into mosaics. Different parts of genomes have different histories, and representing the history of genome evolution as a single tree appears inconsistent with the data. Nevertheless, the assumption of a tree-like process still underlies many approaches. Genome content trees have been calculated based on the presence and absence of genes [14–16] or types of protein folds [17]. While there is limited agreement between genome and rRNA phylogeny, at present it remains unclear whether this similarity is based on shared ancestry of part of a less frequently exchanging genome core [18], or if the apparent congruence is itself the result of horizontal gene transfer [19].

Overall genome content is not best represented on a single tree. Fig. 1 gives an example of an alternative depiction, where thickness of a line reflects percentage of genes shared between two genomes. The coherence among the three domains of life (Bacteria, Archaea, Eucarya [20]), is clearly reflected in genome content; i.e., Archaea share more genes with other Archaea than with Bacteria, but many features are incompatible with representing the relationships between different genomes as a tree. For example, the mesophilic euryarchaeon *Halobacterium* sp. has more genes in common with the mesophilic Bacteria than does the thermophilic crenarchaeote *Aeropyrum pernix*. However, the extremophilic euryarcheote *Archaeoglobus fulgidus* shares many more genes with the extremophilic bacteria, *Aquifex aeolicus* and *Thermotoga maritima* than does *Halobacterium*. While this example illustrates the web-like relationships among genomes, recent phylogenetic reconstructions from molecular data have explored only few alternatives to the tree-paradigm (e.g. [21,22]).

One obvious drawback of the star-like representation in Fig. 1 is that it utilizes BLAST search results only. Any phylogenetic information retained in the sequences is not utilized beyond the presence or absence decision based on a single expectation value cut-off. Because of recombination, individual genes themselves might be mosaic [23]; however, within-gene recombination of protein coding genes occurs mostly between closely related organisms. The redundancy of the genetic code greatly reduces recombination between divergent proteins. Even if a region is 100% conserved on the amino acid level, the encoding DNA can be so different as to allow the mismatch repair system to prevent recombination. For studies of single divergent orthologous protein encoding genes the assumption of a tree-like evolutionary history remains a reasonable expectation. In this manuscript we focus on methods that utilize the phylogenetic information that is retained in molecular sequence data, while not presuming that genomes as a whole evolved in a tree-like fashion.

In an elegant approach Korbinian Strimmer and Arndt von Haeseler [24] utilized Bayesian posterior probabilities to assess the phylogenetic information contained in an alignment of four homologous sequences. With four sequences there are only three possible tree topologies, and thus the three posterior probabilities corresponding to these three trees must sum up to one. Utilizing a barycentric coordinate system, the resulting probability vector is represented as a point in an equilateral triangle (Fig. 2), where the distances of the point P to the three sides represent the three probabilities. Strimmer and von Haeseler applied this approach to depict the phylogenetic information content present in a multiple sequence alignment. They plot the results from the analyses of all possible quartets, where the four sequences are selected from a single multiple sequence alignment in the same coordinate system. If there is a lot of phylogenetic information in the alignment, then most probability vectors will fall close to one of the corners; conversely datasets containing little phylogenetic information will mainly result in vectors falling into the center of the triangle. Here we explore the application of this and similar approaches in comparative genome analyses. In particular, we compare different approaches to calculate Bayesian posterior probabilities, and we compare these probabilities to the more widely used bootstrap support values. We assess the reliability of the different probability mapping approaches through their application to mitochondrial genomes, and we illustrate their usefulness by mapping selected interphylum and interdomain relationships.

Results and Discussion

Overview of data flow in probability mapping

An outline of our approach to genome probability mapping is given in Figure 3. Using SEALS [25] and MySQL we

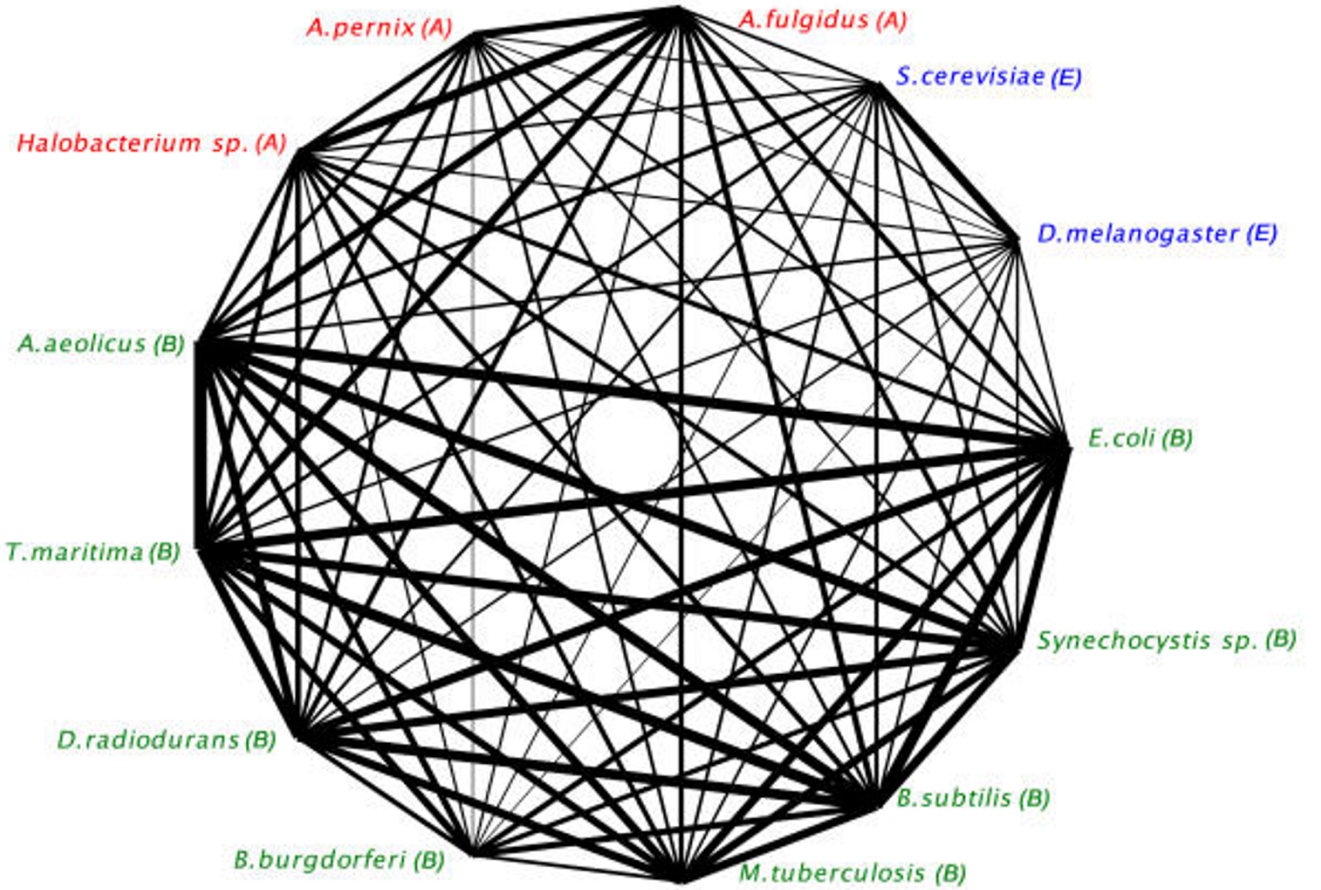

Figure 1
Star Like Representation of Genome Relationships. The diagram depicts pairwise comparisons among thirteen genomes. Every genome is represented as a point on the perimeter of a circle. The thickness of the line connecting two genomes reflects the percentage of shared genes between the genomes. The thickest line connecting *Aquifex aeolicus* and *Thermotoga maritima* corresponds to 51% shared genes, and the thinnest line connecting *Aeropyrum pernix* and *Borrelia burgdorferi* corresponds to 9% shared genes. A gene is considered shared when it had a BLAST hit in the other genome with an E-value below 10^{-8}. The percentage of genes shared between genomes A and B is calculated as ((#of genes in A shared with B/total # of genes in A)+(# of genes in B shared with A/total # of genes in B))/2. Bacteria are depicted in green, Archaea in red and Eukaryotes in blue. The domain affiliation is also indicated by a letter following the species name (A: Archaea, B: Bacteria, and E: Eukaryotes).

developed scripts that identify and retrieve **quart**ets of **o**rthologous **p**rotein-encoding open reading frames (QuartOPs) from four selected genomes. We use the term genome to denote the collection of all ORFs identified in a genome. (In the case of genomes that are not well annotated, it is feasible to use a very wide definition of ORF, e.g, all amino acid sequences encoded between two stop codons in any of the six possible reading frames. As long as one of the genomes included in the analyses is properly annotated, only those identified ORFs that are actually homologous to an identified ORF will become part of a quartet of orthologs.) We utilize an operational definition of an ortholog: two open reading frames are considered orthologous, if and only if they are each other's top scoring BLAST hit when one is used as a query to search the other genome. A QuartOP is formed when each of the open reading frames picks the other members of the quartet as the top scoring hit in searches of the respective genomes. QuartOPs are similar to the clusters of orthologous groups (COGs) maintained by the NCBI [26–28], but differ in that COGs require only unidirectional, circular best hit relationships for three of the reference genomes, whereas we require the reciprocal top hit relationship for the four genomes included in a quartet, and we do not limit our identification of QuartOPs to a number of reference genomes. Montague and Hutchison

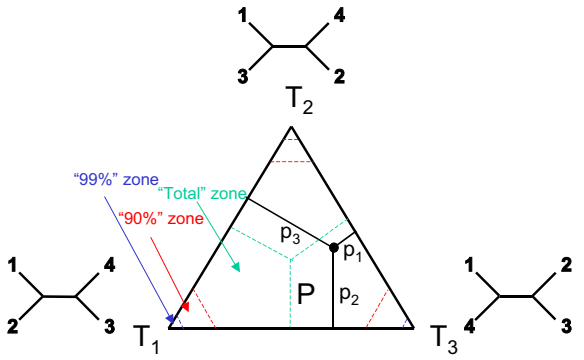

Figure 2
Mapping of the probability vector onto an equilateral triangle. Each QuartOP is represented as a probability vector P inside an equilateral triangle. The position of P is determined by the barycentric coordinates (p_1, p_2, p_3), which correspond to the posterior probabilities or bootstrap support values of the three possible tree topologies. The vertices of the triangle T_1, T_2 and T_3 represent the three possible unrooted tree topologies. Geometrically, each of the coordinates (p_1, p_2, p_3) equals the distance between P and the side of the triangle opposite the corresponding vertex. Points closer to a vertex T_i have a larger corresponding probability p_i and represent a more probable tree topology than the two alternatives. All the points are classified by their position in one of three zones: "total" zone, "90%" zone and "99%" zone, which are depicted schematically and not drawn to scale. In this diagram, point P corresponds to a dataset which has highest probability for the topology T_3, but the probability is below 90%, so the point P is located in the "total" zone, but not in the 99% or 90% zone. Figure adapted from [24].

utilized a comparable approach in their definition of congruent COGs [29]. So far we have analyzed 68 genome quartets (see supplementary material). The number of QuartOPs identified per genome quartet ranges from 82 (for genome quartet #6: *Deinococcus radiodurans, Treponema pallidum, Escherichia coli,* and *Halobacterium sp.*) to 1182 (for genome quartet #63: *Agrobacterium tumifaciens, Sinorhizobium meliloti, Mezorhizobium loti* and *Caulobacter crescentus*).

Each of the aligned QuartOPs from a genome quartet was analyzed with respect to the posterior probability of the three possible tree topologies given the aligned QuartOP. Routinely we calculated these probabilities using Strimmer's and von Haeseler's approach [24]: Using each of the three topologies as a usertree, we calculated the maximum likelihood for each of the three topologies given the data. We then use the three maximum likelihoods to calculate the probability for topology *i* according to the formula: $P_i = L_i/(L_1+L_2+L_3)$, where L_i is the likelihood for the best tree given topology *i*. Other types of reliability measures used to evaluate QuartOPs were bootstrap support values and Bayesian posterior probabilities estimated using MrBayes program (see below).

An example for the comparison of four genomes from different phyla is given in Fig. 4A. Surprisingly, each of the three tree topologies is strongly supported by more than 40 QuartOPs, and most of the QuartOPs appear to strongly support one of the trees. None of the three possibilities has majority support. Figure 5 lists the functional categories of those QuartOPs that strongly support the different tree topologies. None of the categories shows a preference for a particular tree topology. For each tree topology more than 50% of the strongly supporting QuartOPs belong to the category "information storage and processing", while this category contains only about 1/3 of the genes present in the genomes. While the genes in this category appear more conserved and phylogenetically informative, the strong support that the genes in this category provide is nearly evenly split between the three possibilities.

Impact of model parameters and sequence conservation

To test if ill-aligned sequences might have had an impact on the analyses, we repeated the analysis of genome quartet # 8 (see Figure 4) using only QuartOPs that contained very similar sequences. By default we only excluded top hits with an E-value larger than 10^{-4}. We repeated the example given in Fig. 4A with a cut-off of 10^{-20}, i.e., we not only required the sequences in a QuartOP to be each others top hit, but in addition we asked for a high similarity between the two sequences. As a result the support for the three topologies in genome quartet #8 dropped to 54 (44, 38), 51 (45, 32), 39 (29, 28) (the numbers in parenthesis are the numbers of quartets that support the topology

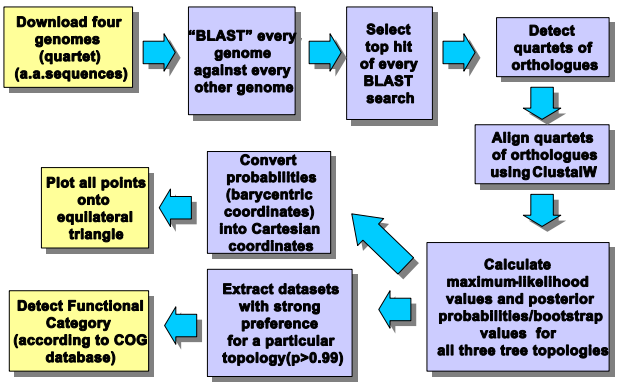

Figure 3
Data flow for the genome quartet analysis. See Materials and Methods for details.

with posterior probability larger than 90% and 99%, respectively). To access the level of sequence conservation within the QuartOPs' sequences, we calculated the average percentage of pairwise identity per QuartOP. It varied from 40.53 10.54% to 43.84 ± 9.7% when the E-value cutoff was varied between 10^{-2} and 10^{-20} (see supplementary material for the summary table). While pairwise sequence identity is not a universally dependable measure of phylogenetic information content, these values illustrates that the sequences within a QuartOP are neither identical to one another, nor so divergent as to be saturated with substitutions and of questionable homology [30]. Using only the most conserved QuartOPs does not change the qualitative result: each of the three possible tree topologies is supported by about an equal number of QuartOPs (see supplementary material).

We recalculated the likelihoods for all QuartOPs in genome quartet #8 using a model that incorporates **a**mong **s**ite **r**ate **v**ariation (ASRV). The posterior probabilities calculated according to Strimmer and von Haeseler did not change dramatically and each of the three tree topologies is still supported by roughly equal number of QuartOPs. The maps for this analysis are available in the supplementary material.

Estimating Bayesian Posterior Probabilities

The formula used by Strimmer and von Haeseler [24] to calculate posterior probabilities (i.e. the probability that tree topology T_i is true given an aligned set of four sequences) considers only three trees (i.e. branch lengths and topology), each with the same prior probability. These three trees are those that have the highest likelihood for the three possible topologies. However, there are infinitely many other trees that differ from the three chosen ones only by differences in branch lengths. What is the effect on the calculated posterior probability of using only the single best tree as a representative of all the trees with the same topology? There is no *a priori* reason to exclude the other trees that have slightly lower likelihoods.

A different approach that does not make these assumptions is the use of Markov Chain Monte Carlo methods to explore tree space. We used the program MrBayes written by Huelsenbeck and Ronquist [31]. Using a QuartOP with posterior probabilities of .76, .10 and .13 we explored different parameter choices for the biased random walk through tree-space. We chose two chains with 5,000 burn-in cycles, and 25,000 cycles with sampling after every cycle as a compromise between increased precision of the probability estimate and computation time (see Materials and Methods for more details).

The result of calculating the posterior probabilities of all QuartOPs in genome quartet #8 is given in figure 4B.

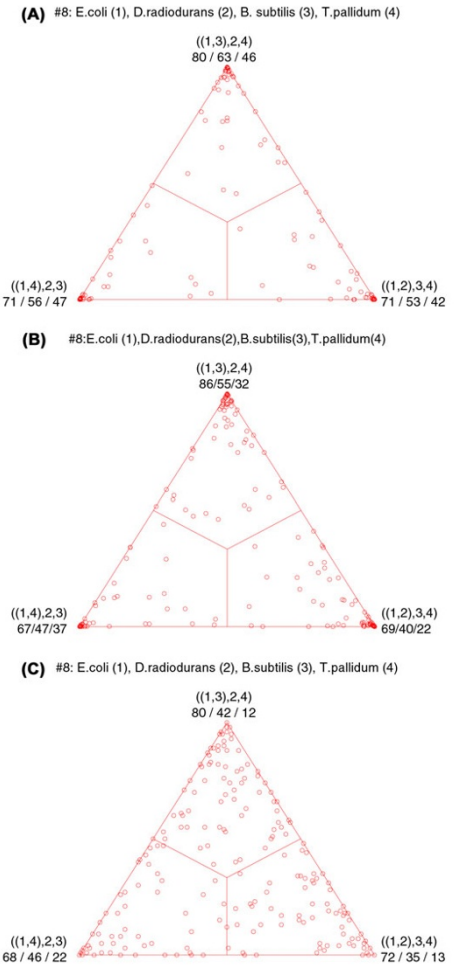

Figure 4

Maps of a genome quartet with organisms from four different bacterial phyla: *Escherichia coli* (Gram negative), *Deinococcus radiodurans* (Deinococcales), *Bacillus subtilis* (Gram positive) and *Treponema pallidum* (spirochete). Tree topologies assigned to the vertices are depicted in New Hampshire tree format near the corresponding vertex of the triangle and they are equivalent to the unrooted tree topologies as depicted in Figure 2. The three numbers associated with each tree topology indicate how many QuartOPs fall into each of the three zones: "total", 90% and 99% respectively. For definition of zones see figure 2. **A)** Probabilities are calculated according to Strimmer and von Haeseler [24]. There is no single topology that is supported by the majority of the QuartOPs and all three possible tree topologies are supported by roughly equal number of QuartOPs at the different probability levels. **B)** Probabilities are calculated with MrBayes program [31]. **C)** Bootstrap support values are plotted. For this case the zones are "total", 70% and 90% support, respectively. Bootstrapping appears to provide a more conservative reliability estimate than the posterior probabilities used in cases A and B. Nevertheless, each tree topology is still supported by a roughly equal number of bootstrapped datasets.

		#8		
Functional Categories of COGs:		**1**	**2**	**3**
Information storage and processing		23	28	25
J	Translation, ribosomal structure and biogenesis	15	22	15
K	Transcription	0	0	4
L	DNA replication, recombination and repair	8	6	6
Cellular processes		8	8	11
D	Cell division and chromosome partitioning	0	2	0
O	Posttranslational modification, protein turnover, chaperones	4	2	4
M	Cell envelope biogenesis, outer membrane	3	3	1
N	Cell motility and secretion	1	1	5
P	Inorganic ion transport and metabolism	0	0	1
T	Signal transduction mechanisms	0	0	0
Metabolism		7	8	7
C	Energy production and conversion	1	1	0
G	Carbohydrate transport and metabolism	2	2	3
E	Amino acid transport and metabolism	2	1	1
F	Nucleotide transport and metabolism	0	2	1
H	Coenzyme metabolism	2	1	2
I	Lipid metabolism	0	1	0
Poorly characterized		5	3	6
R	General function prediction only	5	3	3
S	Function unknown	0	0	3

Figure 5
Distribution among different functional categories for those datasets that support one of the three topologies with better than 99% posterior probability. Tree topologies are indicated by column numbers 1, 2 and 3. Column 1 corresponds to topology ((1,4),2,3), columns 2 and 3 correspond to topologies ((1,3),2,4) and ((1,2),3,4) respectively. Divisions into functional categories are adopted from the COG database [27]. Functional categories are aggregated into four broad functional meta-categories. Distributions of datasets among the meta-categories are plotted as pie charts for each tree topology. In this case all three topologies are supported by roughly equal number of datasets from each meta-category.

Again all of the three tree topologies are strongly supported by some QuartOPs. When we repeated this analysis using the same settings, none of the probabilities changed by more than a few percent. The support for the three tree topologies changed from 67/47/37, 69/40/22 and 86/55/32 (the three numbers indicate total support and QuartOPs that supported a topology with more than .90 and .99 respectively) in the first run to 67/47/37, 70/41/22 and 85/55/32 in the second, indicating that the chosen parameters provided satisfactory reproducibility. Plots of both analyses are available in the supplementary material. Comparing figure 4B with 4A it is clear that in this case the Bayesian posterior probabilities estimated with MrBayes are more conservative assessments of reliability than the ones calculated according to [24]. The 99% support level calculated according to [24] approximately corresponds to the 90% support level calculated with MrBayes.

Bootstrap support values versus posterior probabilities

To facilitate comparison of Bayesian posterior probabilities with a more widely used confidence measure, we generated 100 bootstrapped samples [32] from each QuartOP in case #8. Each of the bootstrapped samples was analyzed using maximum likelihood with the same model of substitution as before. Each of the bootstrapped samples supports one of the three possible topologies, thus the sum of the bootstrap support values for the three topologies adds up to 100%, and the percentage of bootstrapped samples for each QuartOP that best supported each tree was again plotted in a barycentric coordinate system (Fig. 4C). Many more QuartOPs map into the central region of the triangle as compared to Figure 4A and 4B. Clearly, for this test bootstrap support values are more conservative measures of support than either of the posterior probabilities calculated above. Nevertheless, there are still several QuartOPs that strongly support each of the three tree topologies; however, there are 22 QuartOPs that support grouping *E. coli* with *Treponema pallidum* with better than 90% bootstrap support, whereas the alternatives are supported by only 12 and 13 QuartOPs, respectively. Comparing Figures 4A and 4C it appears that in analyzing quartets 70% bootstrap support is comparable to .99 posterior probability calculated according to [24].

Comparison of the different reliability assessment tools

ML-mapping according to [24] is the least conservative of the tools explored. For the test cases analyzed a posterior probability of .99 according to [24], corresponds to a Bayesian posterior probability of .90 calculating using a Markov chain exploration of tree-space using [31] and about 70% bootstrap support. We did not find a strong dependence of the results on the substitution models used in calculating likelihoods and separate runs indicated satisfactory precision of the calculated probabilities and bootstrap values. Given that we only analyzed about 300 QuartOPs using all three approaches it would be premature to generalize our findings; however, other analyses that utilized both bootstrapping and Bayesian posterior probabilities also found bootstrapping to be more conservative than posterior probabilities calculated using Bayesian methods with Markov chain Monte Carlo sampling (e.g., [33–35]).

Mitochondrial genomes

While gene transfer into the mitochondrial genomes has been inferred [36–39], mitochondrial genomes are expected to have undergone many fewer legitimate and illegitimate recombination events than free-living prokaryotes. Clearly, if probability mapping is to be considered a reliable approach, we expect that when analyzing quartets of mitochondrial genomes, the different genes should all support the same tree topology.

In most instances, this expectation is fulfilled (see Table 1), even though we selected instances in which the splits could be expected to be ill resolved, e.g., echinoderm, mammal, insect, mollusk (m4), or protist, fungus, animal, plant (m7). The only exception was an ORF in quartet m7 that encodes the cytochrome oxidase subunit II. This ORF did not support grouping the animal with the

Table 1: Results of analyses for the control mitochondrial genome quartets #m1-m7.

#	Genome 1	Genome 2	Genome 3	Genome 4	((1,2),3,4)			((1,3),2,4)			((1,4),2,3)		
					Tot.	A	B	Tot.	A	B	Tot.	A	B
m1	**Drosophila melanogaster**	**Drosophila yakuba**	Ceratitis capitata	Apis mellifera ligustica	**9**	**8**	**8**	0	0	0	1	0	0
					10	**8**	**8**	0	0	0	0	0	0
					9	**7**	**5**	0	0	0	1	1	0
m2	**Alligator**	Opossum	**Stork**	Donkey	0	0	0	**11**	**11**	**11**	0	0	0
					0	0	0	**11**	**11**	**11**	0	0	0
					0	0	0	**11**	**11**	**11**	0	0	0
m3	**Turtle**	Opossum	**Stork**	Donkey	0	0	0	**12**	**12**	**12**	0	0	0
					0	0	0	**12**	**12**	**12**	0	0	0
					0	0	0	**12**	**12**	**12**	0	0	0
m4	**Starfish**	**Donkey**	Fruit Fly	Doorsnail	**8**	**7**	**7**	2	1	1	0	0	0
					8	**7**	**7**	1	0	0	1	0	0
					8	**7**	**5**	1	1	0	1	0	0
m6	**Reclinomonas americana**	Saccharomyces cerevisiae	**Arabidopsis thaliana**	Homo sapiens	0	0	0	**4**	**3**	**3**	0	0	0
					1	0	0	**3**	**3**	**3**	0	0	0
					0	0	0	**4**	**3**	**2**	0	0	0
m7	**Cafeteria roenbergensis**	Saccharomyces cerevisiae	**Arabidopsis thaliana**	Homo sapiens	0	0	0	**3**	**2**	**2**	1	1	1
					1	0	0	**2**	**2**	**1**	1	1	1
					1	0	0	**2**	**1**	**0**	1	1	1

The groupings corresponding to the expected organismal phylogenies are given in bold. The three numbers in each table cell correspond to the three approaches used. The top number corresponds to results obtained using Strimmer and von Haeseler's approach [24], the middle number corresponds to results obtained using MrBayes program [31], and the bottom number corresponds to results of bootstrap support values calculation. Column "Tot." lists the number of QuartOPs from "total" zone, column A lists the number of QuartOPs from "90%" zone (70% for bootstrap support), and column B lists the number of QuartOPs from "99%" zone (90% for bootstrap support). For definition of zones see Fig. 2. With the exception of the one dataset for quartet #m7, the analyses proved to be consistent with organismal tree topologies. The alignment for the exceptional dataset is presented in Fig. 6. The common names for the organisms listed correspond to the following scientific names: alligator corresponds to *Alligator mississippiensis*, opossum to *Didelphis virginiana*, stork to *Ciconia ciconia*, donkey to *Equus asinus*, turtle to *Chelonia mydas*, starfish to *Asterina pectinifera*, doorsnail to *Albinaria caerulia*, fruit fly to *Drosophila melanogaster*.

fungal homolog as expected; rather it grouped the protist and animal homologs together (posterior probability according to [24] was t 0.99). Inspection of the aligned sequences (Fig. 6) revealed that there are more residues shared between the homologs from *Cafeteria roenbergensis* and *Homo sapiens* thanbetween the homologs from *Cafeteria roenbergensis* and *Arabidopsis thaliana*. No artifact that could be responsible for this unexpected grouping was detected. The same high support for this unexpected grouping is also recovered in bootstrap analysis and in posterior probabilities calculated with MrBayes [31].

The finding of a QuartOP in a mitochondrial genome quartet that supports a non-traditional grouping could either reflect a rare recombination event, selection pressures that led to convergent evolution in two lineages, or a chance event – if one looks at enough samples one will find some that (considered by themselves) appear significant. At present it is not possible to decide between these three possible explanations. Our analysis of mitochondrial genomes shows that in most instances the calculated probabilities (ML-mapping, Bayesian posterior probabilities, or bootstrap values) support the expected tree topologies, albeit with surprisingly strong support values. Rarely, unexpected groupings can be recovered and support for these probably erroneous groupings can be high. In most instances the ML-mapping approach accurately revealed the expected relationships between the mitochondrial genomes. This confirms the suitability of this approach in genome analyses.

Interphylum genome quartets

Here we focus on examples that illustrate the utility of the probability mapping approach. Focusing on the relation-

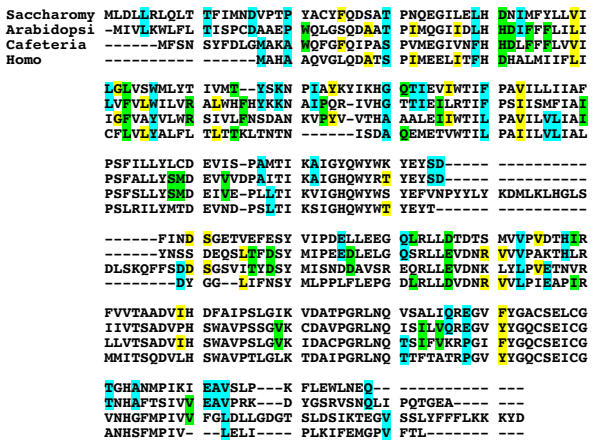

Figure 6
Alignment of mitochondrial cytochrome oxidase subunit II. The alignment for the control mitochondrial quartet m7 (see Table 1) that supports the unexpected ((Homo sapiens, Cafeteria), Saccharomyces, Arabidopsis) topology. The exact matches for each tree topology are colored in three different colors. Blue corresponds to the ((Homo sapiens, Cafeteria), Saccharomyces, Arabidopsis), yellow corresponds to the ((Homo sapiens, Arabidopsis), Saccharomyces, Cafeteria) and green corresponds to the ((Homo sapiens, Saccharomyces), Arabidopsis, Cafeteria) tree topology. As can be seen, the majority of the matches are in favor of ((Homo sapiens, Cafeteria), Saccharomyces, Arabidopsis) tree topology. There are nine parsimony informative positions favoring the latter topology, and only three for each of the other two topologies.

ships between the cyanobacteria with other bacterial phyla we calculated several genome quartets that include the Synechocystis sp. genome and three members each of other bacterial phyla (see Table 2). In all cases that included both Bacillus subtilis as a representative of the low GC Gram positives, and Synechocystis sp., the majority of QuartOPs supported the topologies that grouped these two organisms together. The alternative topologies were significantly supported by some QuartOPs, but the number of strongly supporting QuartOPs was lower than for the B. subtilis – Synechocystis grouping. This also was true when one of the other two genomes was from a high GC Gram positive (genome quartets #53, #54, #55 and #68). Only when two low GC Gram positives were included in the same quartet, was the intra-phylum grouping of low GC Gram positives supported by many more QuartOPs than the grouping of B. subtilis with Synechocystis sp. (genome quartets #51 and #52).

Previous analyses based on a limited number of proteins and signature insertions and deletions had suggested different bacterial groups as closest relatives to cyanobacteria. Among the suggested sister groups were the Deinococcales [40] and spirochetes [41,42]. Our analyses do not support these earlier claims, but are in agreement with the recent analyses of genes involved in chlorophyll biosynthesis [43], which indicated that the low GC Gram positive heliobacteria are closest to the last common ancestor of all oxygenic photosynthetic lineages. The analyses summarized in Table 2 also illustrate that interphylum HGT, while turning genomes into mosaics, has not eroded all associations between bacterial phyla. In the case of cyanobacteria, a close association between low GC Gram positives and the cyanobacteria is supported by the majority of conserved genes. Similar observations of reproducible associations between phyla based on genome wide comparisons were recently published [44–47]. However, at present it cannot be decided to what extent these closer associations reflects shared ancestry or are due to preferred HGT [19].

Interdomain genome quartets

In our search for the "sister-phylum" to the cyanobacteria we also analyzed a few quartets including Archaea. One noteworthy finding was that in the genome quartet including *Synechocystis* sp., *Halobacterium* sp., *Aquifex aeolicus* and *Thermotoga maritima* the grouping of *Halobacterium* sp. with *Synechocystis* sp. was recovered by many more QuartOPs (56 with p > .99) than the grouping that would be expected following 16S rRNA phylogeny (12 QuartOPs with p > .99; see Table 3). To test if this association was specific for *Synechocystis* sp., we repeated the analyses replacing *Synechocystis* sp. with *Bacillus subtilis*. The result was qualitatively the same: at the p > .99 level 53 QuartOPs supported grouping *Bacillus subtilis* with *Halobacterium* sp., and only 27 supported grouping *Aquifex aeolicus* with *Halobacterium* sp. (Fig. 7).

Clearly, there are many artifacts possible in analyzing divergent sequences. For many QuartOPs the ortholog from *Halobacterium* sp. is expected to be the longest branch. To test for the possibility that long branch attraction [48] might be the reason for the strong support of *Halobacterium* sp. grouping with *Synechocystis* sp., we repeated the analysis replacing the *Halobacterium* sp. genome with that from *Archaeoglobus fulgidus*, another archaeon. Gratifyingly, many more QuartOPs supported the grouping of the thermophilic archaeon *Archaeoglobus* with the thermophilic bacteria *Aquifex* and *Thermotoga*. The different interdomain genome quartets that include a meso- or thermophilic archaeon are summarized in Table 3.

An analysis of the putative functional assignments of the QuartOPs that grouped *Halobacterium* sp. with the mesophilic bacteria is given in Table 4. To assess which of these categories have an increased percentage of QuartOPs as compared to distribution of ORFs within the ge-

Table 2: Summary of the genome quartets that include *Synechocystis* sp., *Bacillus subtilis*, and two bacterial genomes from other phyla.

#	Genome 1	Genome 2	Genome 3	Genome 4	((1,2),3,4)			((1,3),2,4)			((1,4),2,3)		
					Tot	0.9	0.99	Tot	0.9	0.99	Tot	0.9	0.99
9	*Synechocystis*	*P. aeruginosa*	*D. radiodurans*	*B. subtilis*	94	76	63	101	73	*57*	186	158	**126**
10	*Synechocystis*	*P. aeruginosa*	*T. pallidum*	*B. subtilis*	69	54	50	51	33	*28*	102	80	**67**
14	*Synechocystis*	*R. sphaeroides*	*B. subtilis*	*E. coli*	65	53	44	286	263	**248**	28	25	17
15	*Synechocystis*	*R. sphaeroides*	*B. subtilis*	*D. radiodurans*	95	72	60	201	173	**149**	73	60	47
16	*Synechocystis*	*D. radiodurans*	*B. subtilis*	*T. pallidum*	63	50	*40*	94	74	**63**	60	46	*35*
17	*Synechocystis*	*B. subtilis*	*T. pallidum*	*E. coli*	93	72	**66**	55	43	*34*	66	49	39
18	*Synechocystis*	*D. radiodurans*	*T. pallidum*	*E. coli*	86	68	54	65	49	38	75	54	47
19	*Synechocystis*	*D. radiodurans*	*B. subtilis*	*E. coli*	129	105	*86*	156	131	**104**	98	76	62
50	*Synechocystis*	*B. subtilis*	*E. coli*	*M. loti*	276	255	**228**	44	31	21	54	43	38
53	*Synechocystis*	*B. subtilis*	*E. coli*	*M. leprae*	125	104	**82**	101	84	65	119	19	66
54	*Synechocystis*	*B. subtilis*	*E. coli*	*M. tuberculosis*	141	114	**97**	101	82	69	128	101	89
55	*Synechocystis*	*B. subtilis*	*M. loti*	*M. tuberculosis*	189	164	**139**	92	74	59	80	58	44
64	*Synechocystis*	*C. trachomatis*	*B. subtilis*	*T. pallidum*	47	31	24	108	96	**86**	36	25	18
67	*Synechocystis*	*C. trachomatis*	*B. subtilis*	*M. loti*	64	48	32	116	104	**84**	72	51	44
68	*Synechocystis*	*C. trachomatis*	*B. subtilis*	*M. tuberculosis*	77	55	45	94	80	**62**	68	52	38
51	*Synechocystis*	*B. subtilis*	*E. coli*	*S. aureus*	33	19	15	361	349	333	15	7	5
52	*Synechocystis*	*B. subtilis*	*E. coli*	*S. pyogenes*	34	22	18	259	249	227	24	17	9

The # column refers to the unique number assigned to the genome quartets analyzed. Columns "((1,2),3,4)", "((1,3),2,4)" and "((1,4),2,3)" refer to the three possible tree topologies. Numbers in columns "Tot", ".90" and ".99" give the number of QuartOPs that support the indicated tree topology with a posterior probability higher than the other two posterior probabilities, or with 90% or 99% probability, respectively. The numbers in bold indicate the number of orthologs supporting the grouping of *Synechocystis* sp. and *Bacillus subtilis* in the absence of another low GC gram-positive in the genome quartets. Note that those numbers are the largest of the three numbers, a finding that supports the recent analyses by [43]. In the presence of another low GC Gram-positive in addition to *Bacillus subtilis*, the largest number of QuartOPs support grouping of low GC Gram-positives with each other (underlined). Other groupings that involve putative sister groups to the cyanobacteria (Deinococaceae and spirochetes) that had been suggested by others (e.g., [40,41]) are indicated in italics.

nome, we also calculated the distributions of ORFs among functional categories in the *Halobacterium* sp. and *A. fulgidus* genomes. Open reading frames within a genome are distributed almost evenly among the four meta-categories (see columns labeled "H" and "A"). However, the QuartOPs that group the halobacterial orthologs with those from mesophilic Bacteria are distributed differentially among the meta-categories. Most of the QuartOPs are in the "Metabolism" and "Information Storage and Processing" meta-categories. These are also the categories in which *Halobacterium* sp. shows many more OuartOPs in support of topology 3 than *A. fulgidus*.

The analyses described in this section reconfirm that genes have been transferred across domain boundaries [6–12]. Not surprisingly, these transfers appear to occur preferentially between organisms living in the same or similar environment. The genome of the mesophilic *Halobacterium* sp. contains many genes that group with the orthologs from mesophilic bacteria, whereas the majority of genes from the thermophilic archaeon *Archaeoglobus fulgidus* group with the orthologs from the extremely thermophilic bacteria. The majority of QuartOPs that group the halobacterial orthologs with the ortholog from the mesophilic bactria belong to two of four meta-categories: "Information Storage and Processing" and "Metabolism". QuartOPs in Information Storage and Processing meta-category that support the grouping of *Halobacterium* sp. with *Synechocystis/Bacillus* are listed in the Table 5. A complete listing is available in the supplementary material. As expected, this list includes several tRNA synthetases, which were previously found to be frequently transferred [6–8], and enzymes involved in DNA repair (*cf.* [9]). More surprisingly, this list also includes translation initiation factors and several ribosomal proteins. The latter were assumed to be infrequently transferred, but recent analyses reported them to be horizontally transferred among bacterial lineages [10,11]. The initiation factor IF-2 in *Halobacterium* sp. was previously shown to have strong similarity to the initiation factor IF-2 from Bacteria [49]. Most of the genes that group *Halobacterium* with the mesophilic bacteria encode functions that were postulated to be frequently exchanged [50]. While no meta-category ap-

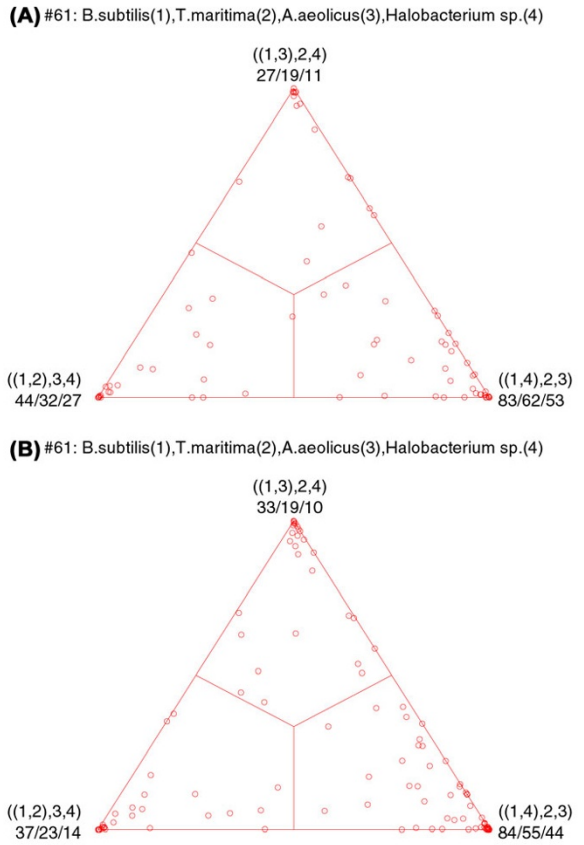

Figure 7
ML map of the quartet representing *Bacillus subtilis*, the deep branching bacteria *T. maritima* and *A. aeolicus*, and the salt-loving archaeon *Halobacterium* sp.. The majority of the orthologous datasets support the grouping of the *Halobacterium* with *Bacillus subtilis*. The topology that corresponds to the 16S rRNA topology (lower left vertex) is supported by the least number of orthologous datasets. The result stayed qualitatively the same when *B. subtilis* was replacedwith the cyanobacterium *Synechocystis* sp. (see results for quartet #11 in Table 3). For details on the figure notations see legend for Figure 4. **A.** Probabilities calculated according to Strimmer and von Haeseler [24]. **B.** Probabilities calculated with the MrBayes program [31].

pears exempt from HGT, some functions appear to be more often transferred than others (cf. Table 4).

Conclusions

Maximum likelihood mapping is a useful tool for analyzing and depicting the mosaic nature of genomes. ML-mapping is much less conservative than other approaches of estimating Bayesian posterior probabilities. If ML-mapping is used as the only probability mapping tool, the overestimation of supporting probabilities has to be taken into consideration. A posterior probability of .99 calculated with ML-mapping often corresponds to a posterior probability of only .90.

Many relationships among prokaryotes cannot be depicted by a tree-like pattern reflecting a core of rarely transferred genes. Rather prokaryotic genomes are mosaics where different parts have different evolutionary histories. However, HGT between divergent organisms has not erased all patterns of interphylum relationship. For example, the majority of QuartOPs group the cyanobacteria with the low GC Gram positives as sister phyla.

Due to horizontal gene transfer even organisms from different domains living in the same or similar environments share more genes with each other than organisms with a similar degree of divergence that live in different environments. These interdomain horizontal transfers mainly concern proteins involved in nucleotide, carbohydrate and amino acid transport and metabolism; however, proteins that are part of the translation machinery or are involved in DNA repair appear to be transferred across domain boundaries as well.

Materials and Methods

Genome Data

Completed genomes were retrieved from the NCBI's FTP site [ftp://ncbi.nlm.nih.gov/genbank/genomes/] in the form of amino acid sequences encoded by open reading frames (ORFs) as identified in the annotated genomes. Mitochondrial genomes were obtained from the Organelle Genomes Page at NCBI [http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/euk_o.html] . The genomes were formatted using the *formatdb* program from the stand-alone BLAST package, initially of version 2.0.11 and later of versions 2.1.2 and 2.2.1 as they were released [51]. All analyses were performed locally.

Data Flow in Quartet Analyses

For each set of four genomes, BLAST [51,52] searches of every ORF in one genome against the other three genomes were performed using the *blastp* program. The E-value cutoff for the BLAST searches was set to 10^{-4} (in one test case an E-value cutoff of 10^{-20} was used). For every BLAST search the GI number of the top hit (if it was below the cutoff) was saved along with the GI number of the query sequence forming a GI pair. This resulted in twelve lists of GI pairs for each of the twelve possible pairwise genome comparisons. This information was further used to identify quartets of orthologous proteins (QuartOPs). Following Tatusov et al. [27] we defined QuartOPs as those sets of genes that mutually pick each other as the top scoring hit in the BLAST comparisons. The detection of the QuartOPs was performed using the MySQL database software [http://www.mysql.com] . The lists of GI pairs were entered into twelve tables of a database. The tables were

Table 3: Summary of the genome quartets that include the mesophilic archaeon *Halobacterium* sp. or the thermophilic archaeon *Archaeoglobus fulgidus*, deep-branching bacteria *Thermotoga maritima* and *Aquifex aeolicus*, and bacteria *Synechocystis* sp. or *Bacillus subtilis*.

#	Genome 1	Genome 2	Genome 3	Genome 4	*((1,2),3,4)*			*((1,3),2,4)*			*((1,4),2,3)*		
					Tot.	**0.9**	**0.99**	**Tot.**	**0.9**	**0.99**	**Tot.**	**0.9**	**0.99**
11	*Synechocystis sp.*	*Thermotoga maritima*	*Aquifex aeolicus*	*Halobacterium sp.*	29	20	12	45	34	27	86	69	56
13	*Synechocystis sp.*	*Thermotoga maritima*	*Aquifex aeolicus*	*Archaeoglobu fulgidus*	47	36	30	63	51	44	50	34	25
61	*Bacillus subtilis*	*Thermotoga maritima*	*Aquifex aeolicus*	*Halobacterium sp.*	44	32	27	27	19	11	83	62	53
62	*Bacillus subtilis*	*Thermotoga maritima*	*Aquifex aeolicus*	*Archaeoglobus fulgidus*	64	50	40	50	35	30	41	27	23

For table notations see legend for Table 2. Quartets #11 and 61 indicate that the majority of the QuartOPs group *Halobacterium* sp. together with *Synechocystis* sp. and with *Bacillus subtilis* respectively, which is in disagreement with 16S rRNA topology. In two control quartets (#13 and #62) *Halobacterium* was substituted with *Archaeoglobus fulgidus*, and in these cases the majority of QuartOPs support the topology that is in agreement with SSU rRNA topology.

joined into one table under conditions that satisfy the definition of the QuartOPs (see above). This resulted in a table with four columns of GI numbers for QuartOPs. The amino acid sequences for each QuartOP were retrieved from GenBank at NCBI and were aligned using ClustalW 1.8 [53]. QuartOPs were analyzed using the ML-mapping approach according to Strimmer and von Haeseler, Bayesian probabilities mapping and bootstrap support values mapping techniques (see details below).

Posterior probabilities according to Strimmer and von Haeseler

For all three possible unrooted tree topologies maximum-likelihood values and posterior probabilities were calculated using in-house JAVA programs that were written utilizing classes from the Phylogenetic Analysis Library version 1.0 [54] and parts of Vanilla package version 1.0 [54]. If not indicated otherwise, likelihood values were estimated using the automatically selected suitable substitution model (chosen from BLOSUM62, CPREV, Dayhoff, JTT, MTREV24, VT and WAG) with no ASRV. The maximum-likelihood mapping approach was further used to visualize support for each tree topology [24], i.e. the posterior probability vector for each QuartOP was plotted into an equilateral triangle. Maximum-likelihood maps were generated using GNUPlot v. 3.7 [http://www.gnuplot.info/] .

Posterior Probabilities calculated with MrBayes program

Posterior probabilities were also calculated with MrBayes version 2.01 [31]. Each QuartOP was analyzed with two simultaneous Markov chains for 25,000 cycles under the JTT substitution model [55] without ASRV. One chain was heated with the temperature set equal to the default value of 0.2. Samples were taken at each cycle. The "burn in" option was set to 5,000 cycles. The remaining 20,000 cycles were used to calculate posterior probabilities for each of the three tree topologies. The posterior probabilities were plotted to equilateral triangle as described above. For discussion of the choice of the parameters see below.

Bootstrap Support Values

As an alternative to posterior probability vectors, bootstrap support values were calculated and plotted. Each QuartOP was bootstrapped 100 times and the proportion of bootstrapped datasets supporting each tree topology was recorded as a bootstrap probability vector. The bootstrap probability vectors were plotted into an equilateral triangle with the zones changed to "total", "70%" and "90%" (see Fig. 2).

Empirical Search for Optimal MrBayes Parameters

To find parameters that will return consistent posterior probabilities within reasonable computation time, one QuartOP from mitochondrial genome quartet #m1 was analyzed multiple times with different parameters. According to Strimmer and von Haeseler's approach [24] this QuartOP has posterior probabilities of 0.76, 0.10 and 0.13. In all runs samples were taken at each cycle; two chains and the JTT substitution model [55] without ASRV were used.

First, we analyzed the dataset with 250,000 cycles. We tried different "burn in" options in the range of 1,000–20,000. The posterior probability values changed by less than 0.3% from case to case. We selected a "burn in" of

Table 4: Distribution of the datasets that strongly support (with 99% posterior probability) one of the three topologies among different functional categories.

	#11			#13			#61			#62			H	A
Functional Categories of COGs:	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**		
Information storage and processing	5	7	20	7	7	10	11	4	17	12	4	0	24	17
J Translation, ribosomal structure and biogenesis	4	6	14	5	6	9	9	3	11	10	3	0	31	44
K Transcription	1	1	0	1	1	0	1	1	0	1	1	0	30	30
L DNA replication, recombination and repair	0	0	6	1	0	1	1	0	6	1	0	0	39	26
Cellular processes	1	5	4	5	7	1	5	1	6	6	3	1	21	16
D Cell division and chromosome partitioning	0	0	1	2	0	1	3	0	0	3	0	0	7	5
O Posttranslational mod., protein turnover, chaperones	1	0	1	1	1	0	1	0	3	2	0	0	22	18
M Cell envelope biogenesis, outer membrane	0	3	0	1	1	0	0	1	0	1	0	0	13	14
N Cell motility and secretion	0	0	1	1	1	0	1	0	0	0	0	0	15	10
P Inorganic ion transport and metabolism	0	2	1	0	4	0	0	0	3	0	3	1	30	31
T Signal transduction mechanisms	0	0	0	0	0	0	0	0	0	0	0	0	14	22
Metabolism	6	16	29	18	30	12	9	6	30	22	28	14	30	37
C Energy production and conversion	2	3	1	1	4	1	0	0	0	3	0	0	23	31
G Carbohydrate transport and metabolism	1	2	1	2	2	0	3	0	2	2	1	1	12	8
E Amino acid transport and metabolism	1	5	16	12	10	7	2	2	14	10	12	6	27	24
F Nucleotide transport and metabolism	2	4	10	1	9	4	2	2	11	4	10	7	10	7
H Coenzyme metabolism	0	2	1	2	5	0	2	2	3	3	5	0	19	16
I Lipid metabolism	0	0	0	0	0	0	0	0	0	0	0	0	9	14
Poorly characterized	1	1	3	1	1	0	2	1	2	1	1	1	24	30
R General function prediction only	1	1	3	1	1	0	2	1	2	1	1	1	64	58
S Function unknown	0	0	0	0	0	0	0	0	0	0	0	0	36	42

The distribution corresponds to the genome quartets listed in Table 3. Functional categories are as designated in Fig. 5. Columns 1, 2 and 3 correspond to the three possible unrooted topologies for each genome quartet (see Table 3). Column entries indicate the number of QuartOPs in each functional category. The last two columns represent the distribution of ORFs in *Halobacterium* sp. (H) and *Archaeoglobus fulgidus* (A) genomes among different functional categories. For these two columns, numbers in the rows corresponding to the meta-categories give the percentage of proteins in each meta category relative to the total number of classifiable proteins and numbers in the rows for each functional category indicate the percent distribution of the proteins within the corresponding meta-category.

5,000 cycles in further analyses. Second, we tried different numbers of cycles to calculate posterior probabilities. The probabilities were calculated using 10,000–240,000 cycles with increment of 10,000 cycles. Again, the posterior probability values did not change significantly from case to case. Third, we raised the "temperature" parameter *T* to 2.0 for the second, heated chain. This did not result in changes of the estimated posterior probabilities. Fourth, we used 25,000 cycles and repeated the analysis 10 times, calculating average and standard deviation of all runs. For all three probabilities the standard deviation was less than 0.01. Based on these analyses we selected 25,000 cycles with a "burn in" of 5,000 as a compromise between precision of probability estimation and computational time spent. As a final test, we performed the analysis of the quartet #8 twice with selected parameters. This did not result in significantly different maps. Graphs and tables depicting the results of these analyses are given in the supplementary material.

Mapping taking ASRV into account

For the genome quartet #8 we calculated posterior probabilities under the model which takes ASRV into account with Strimmer and von Haeseler's [24] approach and with the MrBayes program version 2.01 [31]. TREE-PUZZLE 5.0 [56] was used to calculate posterior probabilities according to Strimmer and von Haeseler [24]. A discrete approximation of the gamma distribution [57] was used to

Table 5: List of genes putatively horizontally transferred between *Halobacterium* sp. (*H.* sp.) and the mesophilic Bacteria *Synechocystis* sp. and *Bacillus subtilis* ("Information Storage and Processing" meta-category only).

Protein Name	*H. sp. GI number*
tRNA synthetases for serine, valine, methionine, cysteine, arginine, proline	10581491, 10581937,10579953, 10580644,10584349, 10580016
phenylalanyl-tRNA synthetase subunit alpha	10581896
Glu-tRNA amidotransferase subunits A, B	10580435, 10579969
tRNA-pseudouridine synthase	10581191
dimethyladenosine transferase	10580702
DNA gyrase subunits A, B	[10580453, 10580452]
DNA helicase	10580995
excision nuclease ABC chains A, B, C (involved in DNA repair)	10582016, 10581796, 10581790
endonuclease V (involved in DNA repair)	10579981
DNA mismatch repair protein	10579807
Putative translation factor SUA5	10581723
Translation initiation factor eIF-2B subunit alpha	10581299
Initiation factor IF2	10581429
ribosomal proteins L1, L11, L3, S4	[10580652, 10580653], 10581159, 10580672

GI numbers in brackets correspond to genes in operons. This list is derived analyses of genome quartets #11 and #61. A complete list of all GI numbers for each QuartOP as well as the four definition lines is available in the supplementary material.

describe ASRV. Eight rate categories were used in TREE-PUZZLE [56], and four rate categories were used in MrBayes [31]. The maps are available in the supplementary material. Due to the amount of time required for calculations, the analyses were not performed for other genome quartets.

Functional assignments using the COG database

Datasets for QuartOPs with strong preference for a particular tree topology (i.e. with posterior probability above 99% for that particular topology, or in other words the QuartOPs located in the very corners of the equilateral triangle) were extracted. For each of those QuartOPs the COG functional category [27] was identified. In order to detect the functional categories, the COG database was downloaded from NCBI's FTP site (initially the year 2000 release and later the year 2001 release). The COG database was formatted using the *formatdb* program of BLAST package. Every QuartOP was compared to the COG database using the *blastp* program. The category of the each sequence in the QuartOP was assigned according to the category of the top hit of each BLAST search. The numbers of QuartOPs in each functional category were calculated for each of the three tree topologies.

Distribution of ORFs among COG categories for complete genomes of Halobacterium sp. and Archaeoglobus fulgidus

Every predicted ORF in a genome was compared to the COG database (release of year 2001) using the *blastp* program with E-value cutoff 10^{-4}. The category of each ORF was set to be equal to the category of the top hit of the corresponding BLAST search. Category Q was dropped from the results, because the corresponding genome quartets were analyzed with the previous release of the COG database (release of year 2000) that did not contain the Q category.

Data Analysis Automation

The repetitive tasks of analyses were automated using the SEALS package version 0.824 [25]. The tasks that were not available through SEALS package were programmed in PERL v. 5.005. The PERL scripts and JAVA programs are available upon request.

Mitochondrial Genome Quartets Analyses

Seven mitochondrial genome quartets were used as controls and were analyzed with the three approaches for genome quartet analysis described above. For calculation of posterior probabilities with MrBayes at least 25,000 cycles were used.

List of Abbreviations

HGT horizontal gene transfer

COG cluster of orthologous groups

ML maximum likelihood

rRNA ribosomal ribonucleic acid

sp. species

BLAST Basic Local Alignment Search Tool

QuartOP quartet of orthologous proteins

SEALS System for Easy Analysis of Lots of Sequences

NCBI National Center for Biotechnology Information

ASRV Among Site Rate Variation

Supplementary Material

Supplementary material is located at the QuartOP web page [http://carrot.mcb.uconn.edu/quartets/] . This web page includes the summary of all genome quartets analyzed (with maps), the results of control analyses, and a form to request the scripts described in this article. ML maps are available in postscript and PDF formats. An offline version of the QuartOP web page is available as a compressed archive named supp_material.zip and as a self-extracting archive supp_material.exe for Microsoft Windows users. The archive can be expanded using WinZip [http://www.winzip.com/] for Windows, StuffIt for Macintosh [http://www.stuffit.com/] , or unzip utility for Unix. The uncompress utilities have to be run with the option to preserve the subdirectory structure inside the archive. To access the information in the archive, the file index.html has to be opened using an Internet browser. This index.html file is located in the root directory named "offline_quartops". All the files in the archive are hyperlinked and accessible through the index.html file.

Additional material

Additional file 1

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-3-4-S1.zip]

Additional file 2

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-3-4-S2.exe]

Acknowledgements

We thank Paul Lewis and Lorraine Olendzenski for many stimulating discussions and for critically reading the manuscript. The work was supported through the NASA Exobiology Program and through the NASA Astrobiology Institute at Arizona State University.

References

1. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci U S A* 1977, **74:**5088-5090
2. Woese CR: **Bacterial evolution.** *Microbiol Rev* 1987, **51:**221-271
3. Hennig W: **Phylogenetic systematics.** *Urbana, University of Illinois Press* 1966
4. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284:**2124-2129
5. Ludwig W, Strunk O, Klugbauer S, Klugbauer N, Weizenegger M, Neumaier J, Bachleitner M, Schleifer KH: **Bacterial phylogeny based on comparative sequence analysis.** *Electrophoresis* 1998, **19:**554-568
6. Doolittle RF, Handy J: **Evolutionary anomalies among the aminoacyl-tRNA synthetases.** *Curr Opin Genet Dev* 1998, **8:**630-636
7. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25:**619-637
8. Olendzenski L, Liu L, Zhaxybayeva O, Murphey R, Shin DG, Gogarten JP: **Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches.** *J Mol Evol* 2000, **51:**587-599
9. Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, Sunjevaric I, Rothstein R, Elion J, Taddei F, *et al*: **Evolutionary implications of the frequent horizontal transfer of mismatch repair genes.** *Cell* 2000, **103:**711-721
10. Makarova KS, Ponomarev VA, Koonin EV: **Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins.** *Genome Biol* 2001, **2:**research0033.1-0033.14
11. 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
12. Wolf YI, Aravind L, Grishin NV, Koonin EV: **Evolution of aminoacyl-tRNA synthetases–analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events.** *Genome Res* 1999, **9:**689-710
13. Gogarten JP: **The early evolution of cellular life.** *Trends in Ecology and Evolution* 1995, **10:**147-151
14. Tekaia F, Lazcano A, Dujon B: **The genomic tree as revealed from whole proteome comparisons.** *Genome Res* 1999, **9:**550-557
15. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content [see comments].** *Nat Genet* 1999, **21:**108-110
16. Fitz-Gibbon ST, House CH: **Whole genome-based phylogenetic analysis of free-living microorganisms.** *Nucleic Acids Res* 1999, **27:**4218-4222
17. Lin J, Gerstein M: **Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels.** *Genome Res* 2000, **10:**808-818
18. Graham DE, Overbeek R, Olsen GJ, Woese CR: **An archaeal genomic signature.** *Proc Natl Acad Sci U S A* 2000, **97:**3304-3308
19. Olendzenski L, Zhaxybayeva O, Gogarten JP: **Horizontal gene transfer: A new taxonomic principle?** *In: Horizontal Gene Transfer,*
20. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proc Natl Acad Sci U S A* 1990, **87:**4576-4579
21. Huson DH: **SplitsTree: analyzing and visualizing evolutionary data.** *Bioinformatics* 1998, **14:**68-73
22. Ribeiro S, Golding GB: **The mosaic nature of the eukaryotic nucleus.** *Mol Biol Evol* 1998, **15:**779-788
23. Gogarten JP, Olendzenski L: **Orthologs, paralogs and genome comparisons.** *Curr Opin Genet Dev* 1999, **9:**630-636
24. Strimmer K, von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** *Proc Natl Acad Sci U S A* 1997, **94:**6815-6819
25. Walker DR, Koonin EV: **SEALS: a system for easy analysis of lots of sequences.** *ISMB* 1997, **5:**333-339
26. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637
27. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28:**33-36
28. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28
29. Montague MG, Hutchison CA 3rd: **Gene content phylogeny of herpesviruses.** *Proc Natl Acad Sci U S A* 2000, **97:**5334-5339

30. Pearson WR: **Effective protein sequence comparison.** *Methods Enzymol* 1996, **266:**227-258
31. Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755
32. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39:**783-791
33. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, *et al*: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294:**2348-2351
34. Rannala B, Yang Z: **Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference.** *J Mol Evol* 1996, **43:**304-311
35. Karol KG, McCourt RM, Cimino MT, Delwiche CF: **The closest living relatives of land plants.** *Science* 2001, **294:**2351-2353
36. Goddard MR, Burt A: **Recurrent invasion and extinction of a selfish gene.** *Proc Natl Acad Sci U S A* 1999, **96:**13880-13885
37. Rousvoal S, Oudot M, Fontaine J, Kloareg B, Goer SL: **Witnessing the evolution of transcription in mitochondria: the mitochondrial genome of the primitive brown alga Pylaiella littoralis (L.) Kjellm. Encodes a T7-like RNA polymerase.** *J Mol Biol* 1998, **277:**1047-1057
38. Cermakian N, Ikeda TM, Miramontes P, Lang BF, Gray MW, Cedergren R: **On the evolution of the single-subunit RNA polymerases.** *J Mol Evol* 1997, **45:**671-681
39. Schinkel AH, Tabak HF: **Mitochondrial RNA polymerase: dual role in transcription and replication.** *Trends Genet* 1989, **5:**149-154
40. Gupta RS, Johari V: **Signature sequences in diverse proteins provide evidence of a close evolutionary relationship between the Deinococcus-thermus group and cyanobacteria.** *J Mol Evol* 1998, **46:**716-720
41. Gupta RS, Mukhtar T, Singh B: **Evolutionary relationships among photosynthetic prokaryotes (Heliobacterium chlorum, Chloroflexus aurantiacus, cyanobacteria, Chlorobium tepidum and proteobacteria): implications regarding the origin of photosynthesis.** *Mol Microbiol* 1999, **32:**893-906
42. Gupta RS: **The natural evolutionary relationships among prokaryotes.** *Crit Rev Microbiol* 2000, **26:**111-131
43. Xiong J, Fischer WM, Inoue K, Nakahara M, Bauer CE: **Molecular evidence for the early evolution of photosynthesis.** *Science* 2000, **289:**1724-1730
44. Sicheritz-Ponten T, Andersson SG: **A phylogenomic approach to microbial evolution.** *Nucleic Acids Res* 2001, **29:**545-552
45. Brochier C, Bapteste E, Moreira D, Philippe H: **Eubacterial phylogeny based on translational apparatus proteins.** *Trends Genet* 2002, **18:**1-5
46. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evol Biol* 2001, **1:**8
47. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification.** *Annu Rev Microbiol* 2001, **55:**709-742
48. Felsenstein J: **Cases in which parsimony and compatibility methods will be positively misleading.** *Syst. Zool.* 1978, **27:**401-410
49. Hasegawa Y, Sawaoka N, Kado N, Ochi M, Itoh T: **Cloning and sequencing of the homologues of both the bacterial and eukaryotic initiation factor genes (hIF-2 and heIF-2 gamma) from archaeal Halobacterium halobium.** *Biochem Mol Biol Int* 1998, **46:**495-507
50. Lawrence JG, Roth JR: **Selfish operons: horizontal transfer may drive the evolution of gene clusters.** *Genetics* 1996, **143:**1843-1860
51. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410
53. 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
54. Drummond A, Strimmer K: **PAL: an object-oriented programming library for molecular evolution and phylogenetics.** *Bioinformatics* 2001, **17:**662-663
55. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *CABIOS* 1992, **8:**275-282
56. Strimmer K, von Haeseler A: **Quartet puzzling: quartet A maximum-likelihood method for reconstructing tree topologies.** *Molecular Biology and Evolution* 1996, 964-969
57. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39:**306-314