J Mol Evol (2004) 58:527–539
DOI: 10.1007/s00239-003-2575-6

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York, LLC 2004

The Consistent Phylogenetic Signal in Genome Trees Revealed by Reducing the Impact of Noise

Bas E. Dutilh,[1] Martijn A. Huynen,[1] William J. Bruno,[2] Berend Snel[1]

[1] Center for Molecular and Biomolecular Informatics/Nijmegen Center for Molecular Life Sciences, University of Nijmegen, Nijmegen, The Netherlands
[2] Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received: 27 June 2003 / Accepted: 12 November 2003

Abstract. Phylogenetic trees based on gene repertoires are remarkably similar to the current consensus of life history. Yet it has been argued that shared gene content is unreliable for phylogenetic reconstruction because of convergence in gene content due to horizontal gene transfer and parallel gene loss. Here we test this argument, by filtering out as noise those orthologous groups that have an inconsistent phylogenetic distribution, using two independent methods. The resulting phylogenies do indeed contain small but significant improvements. More importantly, we find that the majority of orthologous groups contain some phylogenetic signal and that the resulting phylogeny is the only detectable signal present in the gene distribution across genomes. Horizontal gene transfer or parallel gene loss does not cause systematic biases in the gene content tree.

Key words: Genome phylogeny — Horizontal gene transfer — Gene loss — Genome evolution — Character weighting — Thermophilic Bacteria

Introduction

With the availability of complete genome sequences, it has become possible to use the information contained in whole genomes to infer phylogenies (for a review see Wolf et al. 2002). Genome trees are created in an attempt to combine all the phylogenetic messages in all the genes. The main idea is that one can obtain a more representative phylogeny by averaging out the confounding signals in single gene trees. It has been argued that the gene repertoire is a phenetic character (Doolittle 1999a; Gogarten et al. 2002) and that gene content can undergo convergence through selective pressures. Thus, some of the processes that impair single gene trees, such as horizontal transfer (Doolittle 1999b) and parallel loss of related genes (Snel et al. 2002; Wolf et al. 2002), can, when frequent enough, also affect genome trees. For example, some phenotypic characteristics, such as a parasitic lifestyle, are reflected in a similarity in the functional classes of genes in the genome (Zomorodipour and Andersson 1999).

It is true that gene content is a more phenotypic character than gene sequence. After all, the gene repertoire determines the phenotype of an organism. We have argued that gene content phylogenies take a position intermediate to phylogenies based on single genes and phylogenies based on phenotypic characteristics (Snel et al. 1999). However sequence evolution can also reflect the phenotype, e.g., thermophily is reflected in the amino acid content of a genome (Cambillau and Claverie 2000; Kreil and Ouzounis 2001; Suhre and Claverie 2003), and in general sequence-based phylogenetics can suffer from homoplastic events. Fast-evolving positions create a problem when inferring ancient phylogenetic relationships, adding noise rather than signal to the data

Correspondence to: Bas E. Dutilh, Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands; *email:* dutilh@cmbi.kun.nl

(Gribaldo and Philippe 2002). Unless parallel gene loss and horizontal gene transfer occur along demarcated transfer routes, these processes will also only add noise. Sequence analysis has developed tools to identify and remove this noise (Bruno et al. 2000; Goldstein and Pollock 1994). Because the gene presence/absence profile is a binary sequence in all organisms, we can use similar tools to remove noise from genome phylogenies (Brown et al. 2001; Clarke et al. 2002). Clarke et al. (2002) suggested an implementation in which they rid the genome of phylogenetically discordant signals (PDSs) by applying a filter that identified horizontal transfers as sequences with an irregular ranking of the BLAST expectancy values of their orthologs. Removing these PDSs did improve bootstrap support for basal nodes in the phylogeny but, aside from that, altered hardly any topological features.

In the current investigations, we reduce the impact of noise in gene content phylogenies by two schemes that treat the presence/absence profiles as sequence alignments. As we identify PDSs by examining their species distribution, we avoid the pitfalls inherent in sequence analysis, unlike Clarke et al. (2002), who reverted to sequence comparison for identification of the PDSs. By using both orthology and sequence information, Clarke et al. try to combine possibly inconsistent sources of information. This approach can be expected to erroneously identify sequences in rapidly evolving lineages as phylogenetically discordant. Our approach should be less sensitive to this long-branch artifact, as the orthology assignment (Tatusov et al. 2001; von Mering et al. 2003) suffers from this problem only to a small extent. We identify as PDSs instances of horizontal gene transfer, and contrary to Clarke et al. (2002), our schemes also identify parallel gene loss as PDS. As we take orthologous groups as the starting material, orthologous gene displacement within an orthologous group is not identified as a discordant signal.

The coding of genomes as binary sequences allows our approaches to deal with noise from fast-evolving positions. First, we use a method that finds PDSs in a reconstructed genome phylogeny and removes them from the data set. To properly incorporate changes, construction of phylogenies, and identification and removal of PDSs are repeated iteratively until the trees converge. To further test whether the phylogenetic signal in gene content is the only dominant signal, we also used this approach to determine to which topology our trees converge from 100 different random initial topologies.

Second, we use an adapted method that was originally developed for assessing amino acid sequences (Bruno 1996). Assigning high weights to the clade specific genes, and low weights to genes that evolve rapidly, we were able to scale down the impact of noisy signals and infer a filtered phylogeny.

Methods

Orthology

To be able to compare genomes based on their gene content, it is first necessary to identify which genes are shared between genomes, i.e., which genes are orthologs. Orthologs are genes in different species that are directly related by vertical inheritance (Fitch 1970). Paralogs are genes within a species that are derived from gene duplication. If a group of paralogs in a certain species has dispersed after the latest speciation event, all these genes will have the same orthology relationship with their relatives in the sister species. Thus, groups of orthologs will best represent the ancestral relationships of a collection of genes in a set of species.

Inferring orthology relationships is far from trivial, especially because orthology has been defined for the comparison between two species (Fitch 1970). It is not unusual in comparative genomics to define as orthologs those homologs that have a BLAST expectation value lower than a certain threshold (e.g., Bansal and Meyer 2002; Fitz-Gibbon and House 1999). Another operational definition of orthology that is often used is that of reciprocal best BLAST matches (called BeTs [Tatusov et al. 1997], BBHs [Tamames 2001], or RBMs [Clarke et al. 2002]). Although this definition will be a closer approximation of the evolutionary definition of orthology than the close homologs method, it does not give us directly a group orthology that is best suited for our study. A very suitable database of groups of orthologous genes is the manually curated COG database (Clusters of Orthologous Groups of Proteins; NCBI; see www.ncbi.nlm.nih.gov/COG [Tatusov et al. 1997]). Within each of the 3166 COGs, the proteins are assumed to have evolved from the same ancestral gene, and if present, the COG is represented by an individual protein or a group of paralogs within a certain species. We use this database, extended by von Mering et al. (2003) to a current total of 19,433 orthologous groups (OGs) in 89 completely sequenced genomes (for more information see www.bork.embl-heidelberg.de/STRING), to compare these organisms on the basis of their gene content.

Distance Measure

For each OG, a binary profile was created, indicating its presence (1) or absence (0) in the 89 genomes considered (see Fig. 1). Using these profiles as a similarity measure, a matrix was made containing the distances between all species according to Eq. (1) (Korbel et al. 2002).

$$dist(A,B) = 1 - \frac{shared\_OGs(A,B)}{(\sqrt{2} * size\_A * size\_B)/(\sqrt{size\_A^2 + size\_B^2})} \quad (1)$$

As larger genomes can share more genes, we normalize the number of shared OGs by dividing by the weighted average genome size (see Eq. (1) [Korbel et al. 2002]), where the genome size is defined as the number of considered OGs in the genome. Other approaches for normalization such as division by the smallest of the two genomes, or by the geometric average of the genome sizes, show an inferior fit to the relation between genome size and the number of shared genes (not shown). The distance is calculated by subtracting the resulting similarity fraction from 1 (see Eq. [1]).

Iterative Removal of Phylogenetically Discordant Signals (PDSs)

The idea of the iterative method is to compare the presence/absence profile of every OG to the phylogeny to determine to what extent it can be considered discordant. Those OGs that are discordant ac-

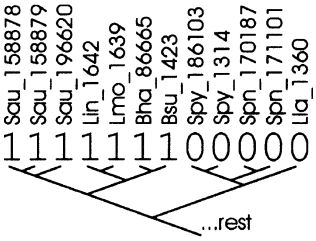


Fig. 1. Example of an OG profile that shows its presence in 7 species (the OG is absent from the rest of the 89 species; not displayed). The profile covers the maximum number of subtrees in this phylogeny (i.e., 7 leaf nodes + 6 internal nodes = 13). The species abbreviations are explained in the legend to Fig. 3.

cording to a certain threshold are then removed, and a new phylogeny is inferred from the remaining profiles. This means that we need a first instance phylogeny to identify the first PDSs and start the iterations. In the standard runs, this was done by using the distance matrix calculated from all the OGs to construct a first instance neighbor-joining tree (Saitou and Nei 1987) using Neighbor (Felsenstein 1989). We also started 100 runs from randomized initial topologies.

The profile of every OG was then compared to the tree to determine to what extent its distribution was monophyletic. To do so, we counted the number of subtrees in which all the leaves contain the OG in question (i.e., all species in this partition have a 1 in the presence/absence profile; see Fig. 1). The number of completely covered subtrees was used to calculate a score for how monophyletic the distribution is. For a given number of species, the score lay between the average coverage of 1000 randomly generated profiles in the first instance neighbor-joining tree (lower bound, set to 0) and the maximum number of partitions possibly covered by this number of species (upper bound, set to 1). The maximum number of tree partitions covered by a profile depends on the number of species in which the OG is present, according to Eq. (2). Equation (2) is based on the number of partitions in a rooted tree, as every bipartition defines a rooted subtree in the entire phylogeny (cf. Fig. 1).

$$max\_covered\_partitions = 2 \cdot species - 1 \qquad (2)$$

The resulting coverage score, which can be compared between OGs present in any number of species, allows us to choose a threshold. OGs that scored below this threshold were removed from the data set, and a new distance matrix and neighbor-joining tree were computed based on the remaining profiles. For every threshold score, this procedure was iterated until convergence was established. Note that convergence to a limit cycle of phylogenies is possible as OGs are allowed to return to the data set if, in the new tree, their profile does cover sufficient branches. After each convergence, we increased the threshold in a simulated annealing-like approach (Kirkpatrick et al. 1983). In the work presented here, we chose 10 annealing steps of 0.1 each (the horizontal lines in Fig. 2). Taking smaller annealing steps (e.g., 50 steps of 0.02 each) did not result in different phylogenies (not shown).

Weighting Method

As an alternative to the above method based on counting subtrees that share a gene, we employed a method that was originally developed to address the sequence weighting problem in amino acid multiple sequence alignments. The Rind program (Bruno 1996) uses a simple maximum likelihood model to estimate the frequency of characters on the tree and corrects for phylogenetic correlations. The Rind frequency gives an estimate of the number of times a character appeared *de novo* in evolution (Bruno 1996). If a character appears throughout a clade consisting of short branches, it is assiged a lower frequency than a gene that appears throughout a clade of the same number of taxa but is made of long branches. If the monophyly of a clade is disrupted by taxon with an inconsistent character, this will have a smaller effect if the branch length of that taxon is longer.

As the presence/absence profiles of the OGs in all species can be seen as a multiple sequence alignment, we were able to run the Rind program on these binary sequences. Genes or columns that have a low Rind frequency, but are relatively abundant according to the raw data, are very clade specific. Thus, to get a score for the monophyly of each character, we divided the raw gene frequency by the Rind frequency. We scaled these scores so that the lowest received a weight of 0 and the highest got a weight of 1, and inferred a neighbor-joining tree as explained above, using the scores to assign weights to each OG.

Assessing Tree Quality

To determine how well the distances in the distance matrix were represented in the neighbor-joining tree, a new distance matrix was derived from the tree, by measuring the distances along the branches between all species pairs. The total difference between all the corresponding values in the two distance matrices was calculated and is expressed as a fraction of the average total distance in the trees. This gives a measure for how well the neighbor-joining tree represents the distance matrix.

We assessed the reliability of the genome tree by counting how often of the partitions occurred in 100 phylogenies constructed by resampling 100% of the OGs with replacement (bootstrapping).

Reference Trees

For reference, we used a SSU rRNA tree and an (unresolved) reference phylogeny from the NCBI taxonomy database (www.ncbi.nlm.nih.gov/Taxonomy [Wheeler et al. 2000]). The rRNA tree is based on a database of expert aligned SSU rRNA sequences of all the species present in the current investigations (www.rna.icmb.utexas.edu [Cannone et al. 2002]). If the correct species was not available, a SSU rRNA sequence from a closely related organism was chosen; if multiple sequences per species were available, the longest and most reliable was selected. We used Clustal to construct a simple neighbor-joining tree based on this alignment (Thompson et al. 1994).

Results

The Choice for a Distance-Based Phylogeny

At first glance, the genome size effect and the concomitant parallel loss of genes should be represented by the Dollo parsimony (Farris 1977). This method is based on the idea that in evolution it is harder to gain a complex feature than to lose it, and we assert that a gene or orthologous group (OG) is such a complex feature that can only be independently gained by horizontal gene transfer. In a given phylogenetic tree, the Dollo algorithm explains the distribution of a character by allowing one origin (i.e., a change from 0 to 1) and as many reversions (1 to 0) as are necessary to explain the pattern of states seen. It then searches for the tree that minimizes the number of 1-to-0 re-

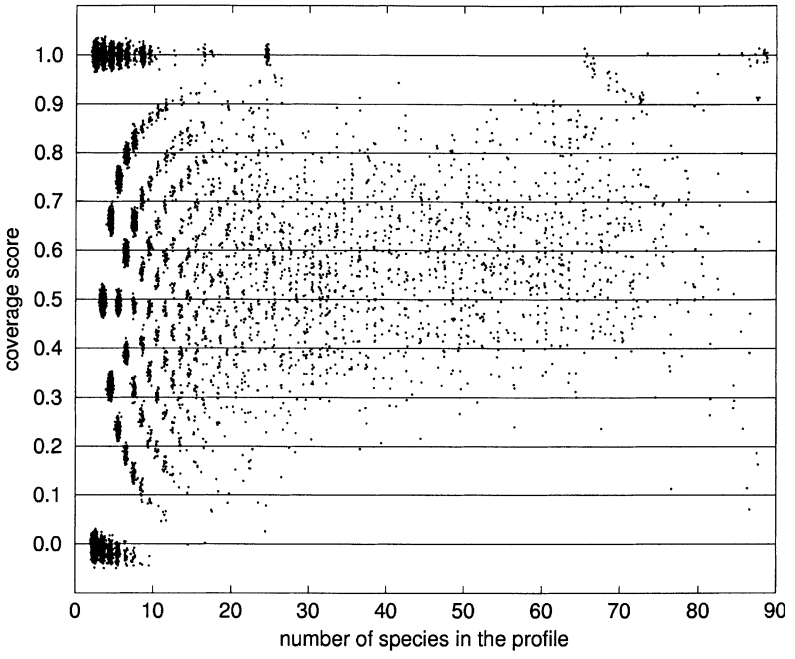


Fig. 2. The coverage score of all the OGs in the first iteration neighbor-joining tree is plotted against the total number of species that contain this OG (number of species in the profile). Note that the coordinates have been scattered (by adding a random number from a normal distribution) to get better insight into the density. The horizontal lines are the simulated annealing steps going from the average of the random distribution (score 0; bottom line) to the maximum possible number of completely covered partitions (score 1).

versions. Although this approach performs slightly better than standard parsimony (not shown), the resulting phylogeny still contains many errors including the clustering of small genomes. Likewise a maximum likelihood approach, such as implemented in MrBayes (Huelsenbeck and Ronquist 2001), in which the presence/absence of OGs was treated as the presence/absence of phenotypic characteristics, did not result in the clustering of the small parasitic genomes with their close relatives with large genomes.

In general, the main caveat of off-the-shelf parsimony or maximum likelihood methods is that they treat the evolution of each character independently. A model for genome evolution has to take variations in the number of genes present in a genome explicitly into account, as has been done for distance-based gene content phylogenies (Korbel et al. 2002). It is not the aim of this work to build a model but rather to develop methods to identify and filter out phylogenetic noise based on the presence/absence pattern of genes, and for that distance-based gene content phylogenies suffice.

The Signal in Gene Content

Phylogenetic Signal in Most of the OGs. Prior to iterating, we establish which OGs behave discordantly, based on the genome tree of the complete data set. This comparison already reveals how well the tree represents the data, as most of the OGs (13,375 of 19,433 = 69%; see Table 1) have a presence/absence profile that is (to a certain extent) consistent with the initial phylogeny based on all the OGs. Profiles that are present in only a few species are more likely to be either perfect or worse than random; the 31% of the OGs that had a negative coverage score contained an average of only 2.8 species. The rest of the OGs cover more subtrees than random profiles would and have a positive coverage score (see Fig. 2). All these profiles are consistent with the genome tree of the complete data set (lowest simulated annealing threshold). No fewer than 5320 of the OGs (27%) are even completely in accordance with the first iteration neighbor-joining tree (they have a coverage score of 1). Many of these "perfect" OGs are present in the same few species; e.g., large groups of over 300 OGs with the same profiles occur between two species like the Ascomycota (352), the Cyanobacteria (341), the Methanosarcinales (364), the Sulfolobaceae (335), or the Xanthomonadaceae (305), but also between a three-species Mammalia–*Drosophila* group (579), and between the four Metazoa (890) included in this data set. All the OGs in these large groups are nonsupervised orthologous groups (NOGs) from the extended data set of von Mering et al. (2003). The other 2154 "perfect" OGs are distributed over only 90 different profiles, among which there are profiles specific for groups such as the 16 Archaea, the 24 Archaea (16) plus Eukaryotes (8), and the 8 α-proteobacteria. Some of the larger groups can be seen as clusters on the line $y = 1$ in Fig. 2.

Results from Iterations. The first instance tree (Fig. 3) is already quite similar to the SSU rRNA reference phylogeny and the NCBI taxonomy (see Table 1). This confirms that gene content contains a strong phylogenetic signal (Fitz-Gibbon and House 1999; Snel et al. 1999; Tekaia et al. 1999). Throughout the iterations, this signal is shown be persistent in the evolving genome tree, and the phylogeny inferred

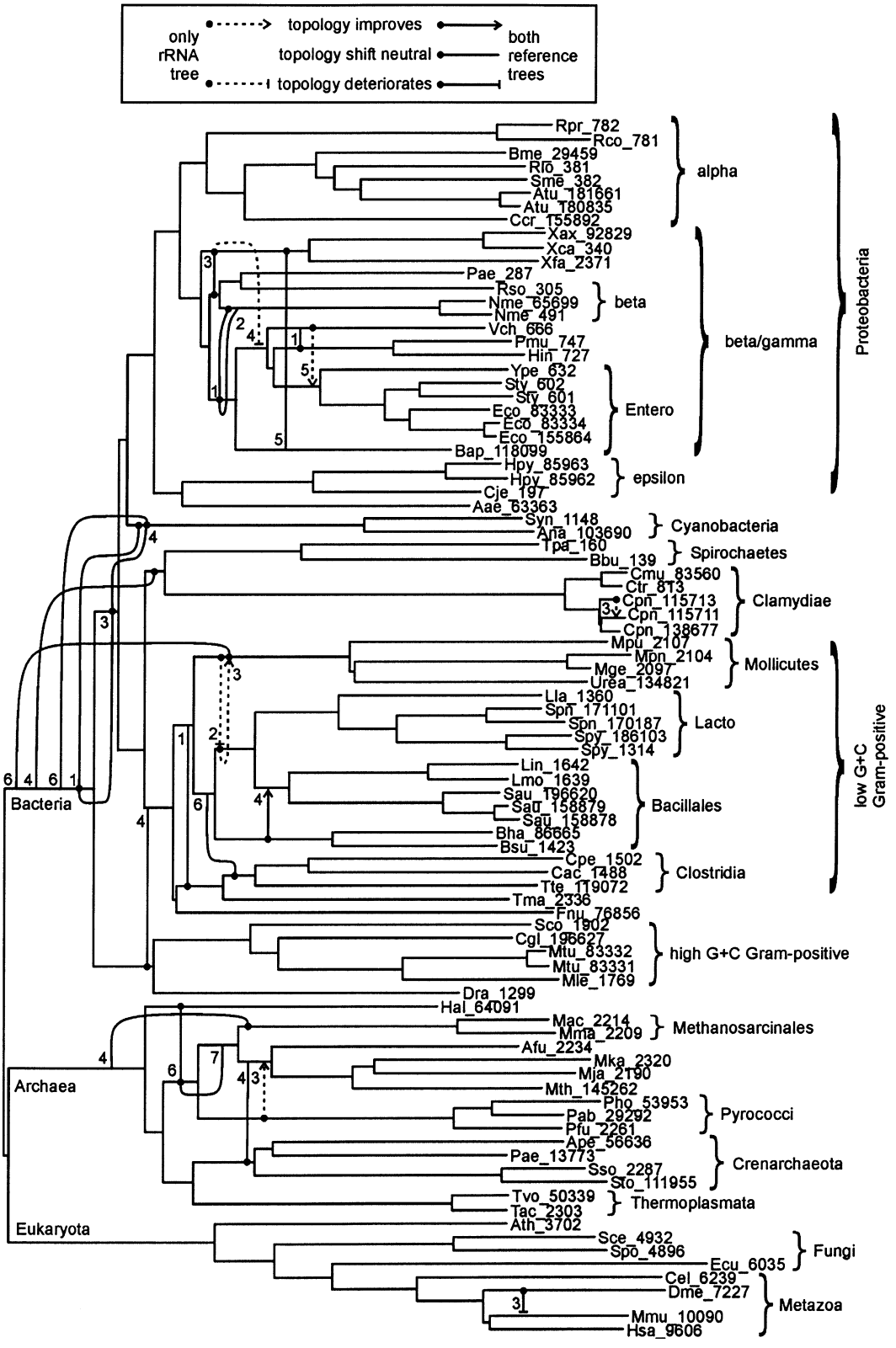
Table 1. Statistics on the evolving phylogeny under a scheme that eliminates discordant OGs from the data set with an increasing stringency

Topology number (1)	Score threshold (2)	OGs (3)	Average score (4)	Branches, rRNA (5)	Branches, NCBI (6)	Matrix vs. NJ tree (7)	Average bootstrap (8)
0	0.0	19,433	0.494	0.628	0.818	0.237	0.881
	0.1	13,375	0.720			0.272	0.868
1	0.2	13,350	0.721	0.628	0.818	0.273	0.880
	0.3	13,152	0.729			0.278	0.880
		12,769	0.745			0.283	0.878
2	0.4	12,777	0.744	0.616	0.818	0.283	0.888
	0.5	11,737	0.780			0.302	0.871
3		9,239	0.856	0.640	0.800	0.338	0.877
		9,379	0.863			0.339	0.849
4	0.6	8,449	0.896	0.640	0.818	0.362	0.855
5		8,428	0.898	0.651	0.818	0.364	0.864
		8,468	0.897			0.361	0.827
6	0.7	7,046	0.946	0.651	0.818	0.388	0.819
7		7,074	0.945	0.651	0.818	0.396	0.829
	0.8	7,081	0.945			0.397	0.759
8		5,930	0.980	0.628	0.818	0.429	0.798
	0.9	5,975	0.981			0.430	0.627
9		5,651	0.987	0.581	0.727	0.452	0.645
		5,644	0.989			0.449	0.568
10	1.0	5,564	0.990	0.570	0.709	0.421	0.563
11		5,563	0.990	0.570	0.709	0.421	0.573
12		5,565	0.990	0.570	0.709	0.419	0.569

Note: Similarity to the rRNA and NCBI reference trees reaches a maximum after seven simulated annealing steps. The simulated annealing threshold in the iterations is given in column 2. Note that the simulated annealing threshold scores are raised after convergence of the topology, so more topologies may be visited in a single threshold score step. The number of OG profiles used to construct each tree is given in column 3. The average scores of the OG profiles used to reconstruct the phylogeny are given in column 4. The fraction of branches shared with the SSU rRNA reference tree and in the taxonomy from NCBI is shown in columns 5 and 6 (the value for the unresolved NCBI taxonomy is higher because it contains fewer branches and will automatically share a larger fraction of its partitions). The difference between the distance matrix and the neighbor-joining tree is shown in column 7, and column 8 contains the average bootstrap value of all the partitions. The topology shifts above the bold line are shown in detail in Fig. 3.

from the restricted gene repertoire even improves. The improvements with respect to the reference trees (see Table 1; SSU rRNA, column 5; and NCBI taxonomy, column 6) are only minor, largely because the first instance tree already shows a considerable resemblance. As the threshold increases (column 2), the tree is based on a decreasing fraction of the OG profiles (column 3), which are selected to cover a maximum number of subtrees. At a certain point, the threshold becomes too high, and we start to exclude false negatives. The phylogeny then breaks down because too many OGs are removed that contain a phylogenetic signal.

This breakdown is evident in Table 1: the difference between the matrix and the neighbor-joining tree increases, and after topology number 7, the overlap with the reference trees shows a sharp drop. Topology number 8 decreases the average bootstrap value of the partitions from 80 to 63%. To illustrate the types of changes that occur in the evolving tree, Fig. 3



shows the shifts leading from the initial phylogeny to topology number 7. Up to the point of this breakdown, where 64% of the discordant OGs were excluded, the reconstructed phylogenies change little and remain very close to the reference trees. The phylogeny contains almost 82% of the branches of the (unresolved) NCBI taxonomy and just over 65% of the branches of the SSU rRNA tree. This result shows that the phylogenetic signal in gene content, as present in the first instance tree, is the dominant

Fig. 3. Initial phylogeny inferred from all the gene presence/absence profiles. The branches that shift in the tree during the iterations are indicated. The number given is the topology number (cf. Table 1) at which the shift occurs. Note that there are no shifts that improve or deteriorate the topology relative to the NCBI taxonomy alone: because the NCBI taxonomy is not completely resolved, changes relative to this tree always also change the fraction of branches shared with the SSU rRNA tree. The phylogeny we inferred was unrooted; we chose to display the Archaea and the Eukaryota as sister taxa, as that is the most commonly accepted view (though there are many papers from the Philippe group to combat this [e.g., Philippe and Forterre 1999]). There is no legend for the branch lengths, as they are only informative in the first instance tree (in two cases they have been very slightly altered to fit the arrows indicating the shifts into the figure). The species abbreviations are the first letter of the family name and the first two letters of the species name, followed by the taxonomic identifier (in alphabetical order): Aae_63363, *Aquifex aeolicus*; Afu_2234, *Archaeoglobus fulgidus*; Ana_103690, *Anabaena* sp.; Ape_56636, *Aeropyrum pernix*; Ath_3702, *Arabidopsis thaliana*; Atu_181661, *Agrobacterium tumefaciens* C58/ATCC 33970 (Cereon); Atu_180835, *Agrobacterium tumefaciens* C58/ATCC 33970 (U. Washington); Bap_118099, *Buchnera aphidicola*; Bbu_139, *Borrelia burgdorferi*; Bha_86665, *Bacillus halodurans*; Bme_29459, *Brucella melitensis*; Bsu_1423, *Bacillus subtilis*; Cac_1488, *Clostridium acetobutylicum*; Ccr_155892, *Caulobacter crescentus*; Cel_6239, *Caenorhabditis elegans*; Cgl_196627, *Corynebacterium glutamicum*; Cje_197, *Campylobacter jejuni*; Cmu_83560, *Chlamydia muridarum*; Cpe_1502, *Clostridium perfringens*; Cpn_115711, *Chlamydia pneumoniae* AR37; Cpn_115713, *Chlamydia pneumoniae* CWL029; Cpn_138677, *Chlamydia pneumoniae* J138; Ctr_813, *Chlamydia trachomatis*; Dme_7227, *Drosophila melanogaster*; Dra_1299, *Deinococcus radiodurans*; Eco_155864, *Escherichia coli* O157:H7 EDL933; Eco_83333, *Escherichia coli* K-12 MG1655; Eco_83334, *Escherichia coli* O157:H7 substr. RIMD 0509952; Ecu_6035, *Encephalitozoon cuniculi*; Fnu_76856, *Fusobacterium nucleatum*; Hal_64091, *Halobacterium* sp.; Hin_727, *Haemophilus influenzae*; Hpy_85962, *Helicobacter pylori* 26695; Hpy_85963, *Helicobacter pylori* J99; Hsa_9606, *Homo sapiens*; Lin_1642, *Listeria innocua*; Lla_1360, *Lactococcus lactis* subsp. *lactis*; Lmo_1639, *Listeria monocytogenes*; Mac_2214, *Methanosarcina acetivorans*; Mge_2097, *Mycoplasma genitalium*; Mja_2190, *Methanococcus jannaschii*; Mka_2320, *Methanopyrus kandleri*; Mle_1769, *Mycobacterium leprae*; Mma_2209, *Methanosarcina mazei*; Mmu_10090, *Mus musculus*; Mpn_2104, *Mycoplasma pneumoniae*; Mpu_2107, *Mycoplasma pulmonis*; Mth_145262, *Methanobacterium thermoautotrophicum*; Mtu_83331, *Mycobacterium tuberculosis* CDC1551; Mtu_83332, *Mycobacterium tuberculosis* H37Rv; Nme_491 *Neisseria meningitidis*; Nme_65699, *Neisseria meningitidis*; Pab_29292, *Pyrococcus abyssi*; Pae_287, *Pseudomonas aeruginosa*; Pae_13773, *Pyrobaculum aerophilum*; Pfu_2261, *Pyrococcus furiosus*; Pho_53953, *Pyrococcus horikoshii*; Pmu_747, *Pasteurella multocida*; Rco_781, *Rickettsia conorii*; Rlo_381, *Rhizobium loti*; Rme_382, *Rhizobium meliloti*; Rpr_782, *Rickettsia prowazekii*; Rso_305, *Ralstonia solanacearum*; Sau_158878, *Staphylococcus aureus* subsp. *aureus* Mu50; Sau_158879, *Staphylococcus aureus* subsp. *aureus* N315; Sau_196620, *Staphylococcus aureus* subsp. *aureus* MW2; Sce_4932, *Saccharomyces cerevisiae*; Sco_1902, *Streptomyces coelicolor*; Spn_170187, *Streptococcus pneumoniae* TIGR4; Spn_171101, *Streptococcus pneumoniae* R6; Spo_4896, *Schizosaccharomyces pombe*; Spy_1314, *Streptococcus pyogenes*; Spy_186103, *Streptococcus pyogenes*; Sso_2287, *Sulfolobus solfataricus*; Sto_111955, *Sulfolobus tokodaii*; Sty_601, *Salmonella typhi*; Sty_602, *Salmonella typhimurium*; Syn_1148, *Synechocystis* sp.; Tac_2303, *Thermoplasma acidophilum*; Tma_2336, *Thermotoga maritima*; Tpa_160, *Treponema pallidum*; Tte_119072, *Thermoanaerobacter tengcongensis*; Tvo_50339, *Thermoplasma volcanium*; Upa_134821, *Ureaplasma parvum*; Vch_666, *Vibrio cholerae*; Xax_92829, *Xanthomonas* axonopodis; Xca_340, *Xanthomonas campestris*; Xfa2371, *Xylella fastidiosa*; and Ype_632, *Yersinia pestis*. The strain is not specified unless more instances of the same species make this necessary.

signal. More importantly, the shifts in the tree do not specifically affect organisms with shared phenotypic characters, e.g., parasites or hyperthermophilic species. As we do not see the effect of phenotype in the tree, such phenotypic convergence does not appear to be the cause or the result of large, systematic biases in the horizontal transfers.

Random Initializations. To investigate whether the quality of the reconstructed gene content trees throughout the iterations depended on the good first instance phylogeny, we repeated the experiments, starting from random initial topologies. The 100 random initial topologies, though completely different (they shared an average of 1% of their branches), rapidly converged. Based on the random first instance phylogenies, an average of 90% of the OGs was deleted. The second iteration phylogenies, composed of those OGs that were not discordant in the random initial trees (lowest simulated annealing threshold), already shared an average of 70% of their branches. The rapid convergence of the topology over the iterations illustrates how consistently this single phylogenetic signal is present in the gene repertoire data. To analyze the topological paths the phylogenies took after these random initializations, we looked in more detail at those trees with the highest resemblance to the reference phylogenies (the rRNA tree and the NCBI taxonomy) and to a selected topology from the standard initialization (topology 7; cf. Table 1). Of the 100 random initial topologies, a large group of 68 paths converged to one topology, which shared 97% of its branches with the standard initialization. Abundant though this topology was, it contained some improbable shifts compared to the phylogeny from the standard procedure. The position of *Halobacterium* was closer to the archaeal root, and *Thermotoga* was placed next to *Thermoanaerobacter* rather than at the root of the low-G+C Gram-positives. Seven of the paths converged exactly to the topology of the standard genome tree. The other 25 paths converged to six other phylogeny, sharing an average of 94% of the partitions with the phylogeny from the standard initialization.

A Worst-Case Scenario. An often-discussed case of "massive" horizontal gene transfer is that from the Archeae to the hyperthermophilic Bacteria *Aquifex aeolicus* and *Thermotoga maritima* (Aravind et al. 1998). We tested whether starting our iterations with

an edited phylogeny, in which we grouped *A. aeolicus* and *T. maritima* at the root of the Archaea, would result in the selection of those horizontally transferred genes, and a convergence of the tree to one in which the hyperthermophilic Bacteria would cluster with the Archaea. In the first iteration the tree converged to the same tree as the one that was started with the unedited tree. This illustrates the point that there may be cases of large-scale horizontal between some species, but their signal is not strong enough to cause systematic biases in the tree based on gene content, even when biasing the selection of genes for the phylogeny toward a set involved in horizontal transfer.

Weighted Tree. The tree obtained from weighing fast-evolving positions is very similar to the rRNA phylogeny (Fig. 4) and successfully improves relative to the unweighted first instance tree. The tree shares over 85% of the branches with the (unresolved) NCBI taxonomy and just over 65% of the branches with the SSU rRNA reference tree. The weighting procedure reinforces especially strongly the separation into three kingdoms, as the internal branches separating the three kingdoms have become longer. When comparing the genome phylogenies that result from the two approaches for filtering the OGs in detail, it becomes apparent that the iterative removal method has a bigger impact on the topology of the tree. As the threshold increases, there are many more topological shifts than in the weighting method. Topology number 3 looks most like the weighted tree (they share 90% of the branches), and all topologies that follow move farther away from the initial phylogeny. Nonetheless, both methods accomplish comparable improvements relative to the unfiltered tree. The advantage of the weighted tree relative to the iterated tree is that the former does not require arguably subjective criteria, like the breakdown of the bootstrap values, to determine when to stop increasing the score threshold.

Phylogenetic Implications

Shifts in the Archaea. In the archaeal phylogeny, the Crenarchaeota remain monophyletic, but they appear to be derived from the Euryarchaeota, making the Euryarchaeota a paraphyletic taxon. This is inconsistent with the rRNA tree but often found in genome trees (Wolf et al. 2002). The current approach does manage to shift *Halobacterium* away from its (erroneous) ancestral position, into the Euryarchaeota. Instead, the Methanosarcinales move to the archaeal root, next to *Halobacterium*, followed later by the Thermoplasmata. Cavalier-Smith already proposed to join the Methanosarcinales and the Halobacteria in the phylum Halomebacteria. This was based on the fact that many of the differences between the two can be attributed to the loss of ancestral proteins by the Methanobacteria, whereas many similarities in RNA polymerases, antibiotic sensitivities, and rRNAs can be found (Cavalier-Smith 1986, 2002). Slesarev and co-workers (2002) showed that genome trees based on gene content or on conserved gene pairs group all methanogenic Archaea. Indeed, this is true for the methanogens sequenced at the time of that research, but we show here that the Methanosarcinales are not part of the otherwise strongly supported methanogenic subtree. As in the gene content tree presented by Slesarev et al. (2002), we find the position of *Archaeoglobus fulgidus* to be stable at the root of the methanogens.

Hyperthermophilic Bacteria. Gene content phylogenies are especially interesting for those clades where rRNA trees might fail. The phylogenetic position of thermophilic Bacteria is such a point (Cavalier-Smith 2002). The inference that the thermophilic Bacteria are primitive, based on rRNA trees, has been doubted because this placement might be an artifact from long-branch attraction (Gribaldo and Philippe 2002) and selection for high G + C content in hyperthermophilic rRNA (Galtier and Lobry 1997). Recently it has indeed been shown that this artifact can be circumvented by considering only the slowly evolving nucleotides in the rRNA sequence. This places the hyperthermophilic Bacteria as a division whose relation to other divisions remains unclear (Brochier and Philippe 2002). Interestingly our results indicate a consistent (i.e., throughout the iterations) affiliation of *Thermotoga* with the Firmicutes and of *Aquifex* with the (δ/ε)-Proteobacteria. They stay there even after removal of possible phylogenetically discordant signals, such as the abundant horizontal transfers of these species with the Archaea (Nelson et al. 1999).

The hypothesis of independent origins of eubacterial (hyper)thermophily finds strong support in the work of Forterre et al. (2000). They show that reverse gyrase, an enzyme that is crucial for stabilizing the DNA in hyperthermophilic organisms, in *Aquifex* and in *Thermotoga* was independently obtained by two seperate horizontal transfer events. Our iterative approach discards reverse gyrase (COG1110) as a discordant signal at a threshold score of 0.5 and in the weighting approach it is assigned a weight of 0.16. When combined with results from other genome tree-like approaches and other independent evidence, the position of *Aquifex* with the Proteobacteria, as well as *Thermotoga* with the Gram-positive Bacteria, is supported.

Structurally, the outer membrane of *Aquifex* has been shown to contain lipopolysaccharide (Plotz et al. 2000), like the Proteobacteria, but unlike Gram-pos-

Fig. 4. Phylogeny inferred from gene presence/absence profiles with weighted characters. The species abbreviations are explained in the legend to Fig. 3.

itive Bacteria. Klenk et al. (1999) made phylogenetic analyses of the two largest subunits of bacterial RNA polymerases and placed *Aquifex* with the Proteobacteria. This position is also supported by a supertree composed of the phylogenies of hundreds of orthologous gene families (Daubin et al. 2001), gene content trees (Wolf et al. 2001), and gene order trees (Wolf et al. 2001) Analysis from rare genomic events, such as conserved insertions and deletions in several proteins, also shows that *Aquifex* should be placed next to the Proteobacteria (Gupta and Griffiths 2002).

The position of *Thermotoga* as an evolutionary neighbor to the Gram-positives is supported by the same insertions and deletions study (Gupta and Griffiths 2002). Both Tiboni et al. (1993) and Pesole et al. (1995) show that glutamine synthetase I trees group *Thermotoga* with the low-G + C Gram-positive Bacteria. Gribaldo et al. (1999) show a deletion in the sequence of HSP70, shared by *Thermotoga* and the Gram-positive Bacteria, and though the phylogenies inferred from the protein sequence do not cluster these groups, this may be artifactual and the result of

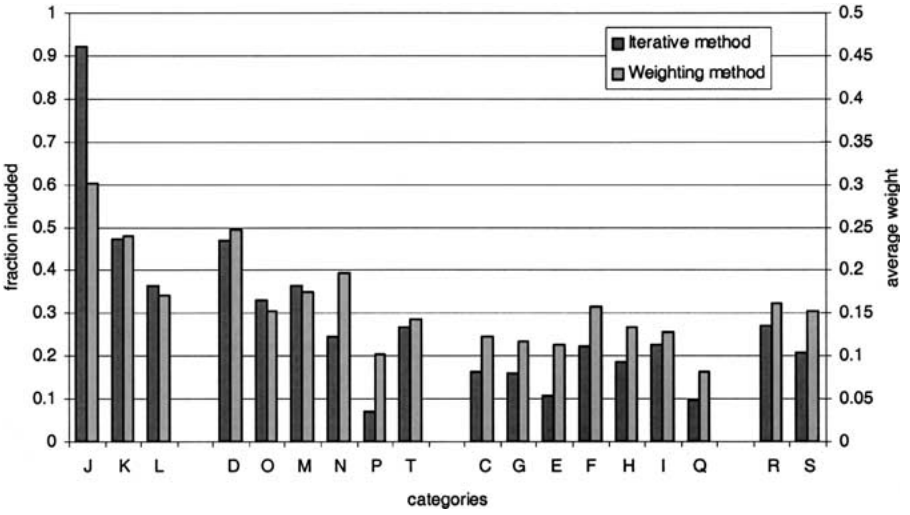


Fig. 5. Functional categories of the contributing COGs for each of the two methods. The dark gray bars (left) are the fractions that were not removed in the seventh topology (coverage score, 0.8; cf. Table 1) of the iterative method. The light gray bars (right) are the average weights assigned in the weighting method. Both methods identify the same functional categories as discordant. The categories are grouped in the four main COG classes ''information storage and processing'' (translation, ribosomal structure, and biogenesis [J], transcription [K], and DNA replication, recombination, and repair [L]), ''cellular processes'' (cell division and chromosome partitioning [D], posttranslational modification, protein turnover, chaperones [O], cell envelope biogenesis, outer membrane [M], cell motility and secretion [N], inorganic ion transport and metabolism [P], and signal transduction mechanisms [T]), ''metabolism'' (energy production and conversion [C], carbohydrate transport and metabolism [G], amino acid transport and metabolism [E], nucleotide transport and metabolism [F], coenzyme metabolism [H], lipid metabolism [I], and secondary metabolite biosynthesis, transport, and catabolism [Q]), and ''poorly characterized'' (general function prediction only [R] and function unknown [S]) (Tatusov et al. 1997).

convergence within the hyperthermophilic sequences (Cambillau and Claverie 2000; Kreil and Ouzounis 2001; Suhre and Claverie 2003).

Problems. In the iterative method, the eukaryotic subtree is very stable. The only topological change is for the worse: *Drosophila melanogaster* is placed in between the mammals (see Fig. 3). This results from an artifact of the definition of NOGs by von Mering et al. (2003), which unites the mammals to form a single clade, thus disallowing the formation of any NOGs shared only by these two species. The weighting method does not show this shift.

Two groups of Bacteria with exceptionally small genomes are pushed to the root during the iterations. Though we have corrected for genome size in Eq. (1), the Chlamydiae/Spirochaetes group and the Mollicutes are still problematic cases, though more so in the iterative than in the weighting approach. This size effect is the result of the fact that small genomes can share only a certain maximum number of genes. This is a known problem in gene content phylogenetics (Wolf et al. 2002), and though it has been addressed (Korbel et al. 2002), a definitive solution has still not been found.

Which Genes Are Discordant?

It has been proposed that metabolic genes undergo more horizontal gene transfer than informational genes. Here we obtain detailed information on this hypothesis, by determining which types of genes were discordant, i.e., which OGs were filtered out in our procedures. To summarize this, we look at the COG functional classes (NOGs are not functionally classified). Figure 5 shows the extent to which the different COG functional categories were allowed to remain in the data set. For the weighting method, the average assigned weight of all the genes in the functional category is plotted. For the iterative removal method, the fraction of genes that remained in the data set of topology number 7 is indicated. This topology, where the coverage score threshold of 0.8 excluded 64% of the OGs, was selected to maintain consistency with Fig. 3.

The results for both schemes investigated in this research are remarkably similar. The ''translation, ribosomal structure, and biogenesis'' category (J) is the least discordant; less than 8% of the COGs from this category are removed in the iterative procedure, and the average weight assigned to the COGs in this category was 0.30. Of the ''inorganic ion transport and metabolism'' category (P), less than 8% remained in the data set, and it can be considered the fastest-evolving category of genes with respect to gene content. The category where the lowest weights were assigned was ''secondary metabolite biosynthesis, transport, and catabolism'' (Q), where the COGs received an average weight of 0.08. In general, ''metabolism'' COGs are filtered out most in our

procedure, whereas "information storage and processing" COGs are relatively stable in evolution. This supports the complexity hypothesis (Jain et al. 1999) that, generally, operational genes are transferred more readily throughout evolution than informational genes, which are more often involved in complex networks of interactions. However, our study reveals a more detailed picture: apart from the inorganic ion transport and metabolism category (P), the other "cellular processes" categories, such as "cell division and chromosome partitioning" (D) and "cell envelope biogenesis, outer membrane" (M), are intermediately discordant with the COGs from the "metabolism" and "information storage and processing" classes.

Discussion

If we correct for genome size, a very good gene content phylogeny, subject to some caveats, can already be inferred. This means that the noise, which results from processes like horizontal transfer or convergence through parallel gene loss and may confound a genome phylogeny, can be effectively averaged out by considering genome scale data. In this initial tree, improvements can be made by reducing the impact of the noise, which is shown in the current paper using two independent approaches. This result is in contrast with Clarke et al. (2002), who did not find any improvements in their tree when filtering for discordant genes. This is probably due to the fact that the filtering scheme used by these authors is not strong enough. The topological improvement of their phylogeny may also be restrained by their choice to use one source (orthology) for the reconstruction of the tree and another, albeit related, source (sequence) for filtering. Here we show that the topology will change during the iterative removal of the noise, as well as in a scheme that selectively downweighs the noise. This is not to say that the genes designated as noise are biologically irrelevant. Genes that have a nonphylogenetic distribution often have functional significance, such as shared pathogenicity factors between *Helicobacter pylori* and *Haemophilus influenzae* (Huynen et al. 1998) or reverse gyrase in the hyperthermophiles (Forterre et al. 2000). But these qualitative, phenetic, patterns in shared gene content apparently play a quantitatively minor role relative to the phylogenetic signal and can be considered noise when constructing genome trees.

In the iterative procedure, we have shown that there is a consistent phylogenetic signal in the majority of OGs: throughout the iterations, the phylogeny shows few changes until the fraction signal over noise that is removed becomes too high. This result is also supported by the converging trajectories starting from random initial phylogenies. Being too strict in removing discordant OGs leads to a breakdown of the phylogenetic pattern, leaving too little signal for a reliable tree topology. The phylogenetic signal is thus the only detectable signal in the gene content. The rest is noise. A recent investigation of the relation between horizontal transfer and phylogenetic incongruence in gene trees revealed that, in most cases, alternate topologies represent construction artifacts rather than the accumulation of horizontal transfer events with time (Daubin et al. 2003).

In the current paper, we have implemented two methods, based on the same ideas, and both give comparable results in terms of improvements in the phylogeny and in the types of functions that are considered discordant. Improvement for the current approaches may be achieved by the implementation of a better measure for the discordance of a signal in the phylogeny, but we do not expect major changes in the results given the similarity in outcome from the two procedures. The main improvements for both the iterative and the weighting method may be expected from a better, i.e., more fine-grained, definition of orthology, which will allow more detail and thus better-defined relationships between the species.

Other improvements might come from maximum likelihood or Bayesian approaches, which can include explicit statistical models of genome evolution. Full Bayesian methods are already available for gene/species tree reconciliation (Arvestad et al. 2003). This specific development, and that of Bayesian inference in general, opens up several lines along which gene content phylogenies can be improved. First, their model of gene content evolution can be used for the likelihood of a species phylogeny, incorporating all genome sizes and the distribution of the OGs over the species. Second, a more complicated approach could be implemented that does not treat the OGs as a binary distribution but as a gene tree. This makes it possible to directly use the methodology from Arvestad et al. (2003), but with the extension that the species tree is one of the parameters that are to be determined using the likelihood algorithm. The biggest drawbacks are expected to be the computational time needed to construct reliable gene trees for all OGs, computing the likelihood for all the trees, and the great increase in computational time needed for the Monte Carlo Markov chain to simultaneously and sufficiently sample tree space.

Acknowledgments. This work was supported in part by European Union Contract QLTR-2000-01676 and by a grant from The Netherlands Organization for Scientific Research (NWO).

References

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet 14:442–444

Arvestad L, Berglund AC, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19 (Suppl 1):I7–I15

Bansal AK, Meyer TE (2002) Evolutionary analysis by whole-genome comparisons. J Bacteriol 184:2260–2272

Brochier C, Philippe H (2002) Phylogeny: A non-hyperthermophilic ancestor for bacteria. Nature 417:244

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nat Genet 28:281–285

Bruno WJ (1996) Modeling residue usage in aligned protein sequences via maximum likelihood. Mol Biol Evol 13:1368–1374

Bruno WJ, Socci ND, Halpern AL (2000) Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol 17:189–197

Cambillau C, Claverie JM (2000) Structural and genomic correlates of hyperthermostability. J Biol Chem 275:32383–32386

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, uuml, Iler KM, Pande N, Shang Z, Yu N, Gutell RR (2002) The Comparative RNA Web (CRW) Site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3:2

Cavalier-Smith T (1986) The kingdoms of organisms. Nature 324: 416–417

Cavalier-Smith T (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol 52:7–76

Clarke GD, Beiko RG, Ragan MA, Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J Bacteriol 184:2072–2080

Daubin V, Gouy M, Perriere G (2001) Bacterial molecular phylogeny using supertree approach. Genome Inform Ser Workshop Genome Inform 12:155–164

Daubin V, Moran NA, Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes. Science 301:829–832

Doolittle WF (1999a) Lateral gene transfer, genome surveys, and the phylogeny of Prokaryotes. Science 286:1443a

Doolittle WF (1999b) Phylogenetic classification and the universal tree. Science 284:2124–2129

Farris RJ (1977) Phylogenetic analysis under Dollo's law. Syst Zool 26:77–88

Felsenstein J (1989) PHYLIP—Phylogeny inference package (version 3.2). Cladistics 5:164–166

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 27:4218–4222

Forterre P, Bouthier De La Tour C, Philippe H, Duguet M (2000) Reverse gyrase from hyperthermophiles: Probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet 16:152–154

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

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–2238

Goldstein DB, Pollock DD (1994) Least squares estimation of molecular distance—Noise abatement in phylogenetic reconstruction. Theor Popul Biol 45:219–226

Gribaldo S, Philippe H (2002) Ancient phylogenetic relationships. Theor Popul Biol 61:391–408

Gribaldo S, Lumia V, Creti R, de Macario EC, Sanangelantoni A, Cammarano P (1999) Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol 181:434–443

Gupta RS, Griffiths E (2002) Critical issues in bacterial phylogeny. Theor Popul Biol 61:423–434

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS Lett 426:1–5

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806

Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. Science 220:671–680

Klenk HP, Meier TD, Durovic P, Schwass V, Lottspeich F, Dennis PP, Zillig W (1999) RNA polymerase of Aquifex pyrophilus: Implications for the evolution of the bacterial rpoBC operon and extremely thermophilic bacteria. J Mol Evol 48:528–541

Korbel JO, Snel B, Huynen MA, Bork P (2002) SHOT: A web server for the construction of genome phylogenies. Trends Genet 18:158–162

Kreil DP, Ouzounis CA (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 29:1608–1615

Nelson KE, Clayton RA, Gill SR, et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399:323–329

Pesole G, Gissi C, Lanave C, Saccone C (1995) Glutamine synthetase gene evolution in bacteria. Mol Biol Evol 12:189–197

Philippe H, Forterre P (1999) The rooting of the universal tree of life is not reliable. J Mol Evol 49:509–523

Plotz BM, Lindner B, Stetter KO, Holst O (2000) Characterization of a novel lipid A containing D-galacturonic acid that replaces phosphate residues. The structure of the lipid a of the lipopolysaccharide from the hyperthermophilic bacterium Aquifex pyrophilus. J Biol Chem 275:11222–11228

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, Tatusov RL, Wolf YI, Stetter KO, Malykh AG, Koonin EV, Kozyavkin SA (2002) The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. Proc Natl Acad Sci USA 99:4644–4649

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21:108–110

Snel B, Bork P, Huynen MA (2002) Genomes in flux: The evolution of archaeal and proteobacterial gene content. Genome Res 12:17–25

Suhre K, Claverie JM (2003) Genomic correlates of hyperthermostability: An update. J Biol Chem 278:17198–17202

Tamames J (2001) Evolution of gene order conservation in prokaryotes. Genome Biol 2:RESEARCH0020

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: New developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28

Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res 9:550–557

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tiboni O, Cammarano P, Sanangelantoni AM (1993) Cloning and sequencing of the gene encoding glutamine synthetase I from the archaeum Pyrococcus woesei: Anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase I sequences. J Bacteriol 175:2961–2969

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: A database of predicted functional associations between proteins. Nucleic Acids Res 31:258–261

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA (2000) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 28:10–14

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1:8

Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. Trends Genet 18:472–479

Zomorodipour A, Andersson SG (1999) Obligate intracellular parasites: Rickettsia prowazekii and Chlamydia trachomatis. FEBS Lett 452:11–15