Genome trees and the Tree of Life

Yuri I. Wolf, Igor B. Rogozin, Nick V. Grishin and Eugene V. Koonin

Genome comparisons indicate that horizontal gene transfer and differential gene loss are major evolutionary phenomena that, at least in prokaryotes, involve a large fraction, if not the majority, of genes. The extent of these events casts doubt on the feasibility of constructing a 'Tree of Life', because the trees for different genes often tell different stories. However, alternative approaches to tree construction that attempt to determine tree topology on the basis of comparisons of complete gene sets seem to reveal a phylogenetic signal that supports the three-domain evolutionary scenario and suggests the possibility of delineation of previously undetected major clades of prokaryotes.
If the validity of these whole-genome approaches to tree building is confirmed by analyses of numerous new genomes, which are currently being sequenced at an increasing rate, it would seem that the concept of a universal 'species' tree is still appropriate. However, this tree should be reinterpreted as a prevailing trend in the evolution of genome-scale gene sets rather than as a complete picture of evolution.

The idea that the evolution of life can be represented as a tree, with leaves corresponding to extant species and nodes to extinct ancestors, came from Charles Darwin and is epitomized in the famous single illustration of his book, *On the Origin of Species by Means of Natural Selection* [1]. The earliest elaborations on this concept are associated with the name of Ernst Haeckel [2]. Haeckel's and other early trees were based on a general idea of a hierarchy of relationships between species and higher taxa. Gradually, quantitative criteria have been developed to measure the degree of morphological difference that was thought to reflect evolutionary distance.

Glossary

Bootstrap analysis: a computational technique for estimating distribution parameters by resampling the original data; used in phylogenetics to estimate the confidence of internal branches of a tree.
Distance method: evolutionary distances are computed for all pairs of analyzed sequences forming a matrix of pairwise distances, and a phylogenetic tree is constructed by analysis of the relationships among these distance values. There are many different distance methods, among the most popular are the neighbor-joining method and the minimum evolution (least squares) method.
Maximum parsimony methods: character states (e.g. amino acids) at each site are analyzed separately. The principle of parsimony approaches is to search for a tree that requires the smallest number of evolutionary changes to explain the differences observed among analyzed sequences.
Maximum likelihood methods: the likelihood value is calculated for the character state configurations among analyzed sequences for each possible tree, and the tree with the maximum likelihood is chosen. The reliability of phylogenetic trees is usually assessed using bootstrap analysis.
Orthologous genes: genes related by vertical descent.
Paralogous genes: genes related by duplication.
Phylogenetic tree construction: The most common methods of phylogenetic analysis can be classified into three major groups: (1) distance methods, (2) maximum parsimony methods and (3) maximum likelihood methods.

Yuri I. Wolf
Igor B. Rogozin
Eugene V. Koonin*
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
*e-mail: koonin@ncbi.nlm.nih.gov

Nick V. Grishin
Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA.

In the early days of molecular phylogenetics, a gene tree was usually equated with the species tree. This view was typified using ribosomal RNA (rRNA) sequences as the principal molecular phylogenetic marker. Phylogenetic analysis of rRNA transformed our understanding of the history of life, resulting in the discovery of a previously unrecognized domain of life, the Archaea, and in a tree topology that has been aptly called the 'standard model' of evolution [3–5]. This model involves the early descent of the bacterial clade from the last universal common ancestor and a subsequent separation of archaea and eukaryotes. In addition to the formulation of the standard model, phylogenetic analysis of rRNAs brought 'the winds of (evolutionary) change' onto taxonomy by revealing, supporting or correcting many major clades among bacteria, archaea and eukaryotes [6]. Thus, this appeared to be the approach of choice to produce the true species tree depicting the history of life. Although it was recognized that the tree topology substantially depends on the method used for tree reconstruction, the actual validity of the rRNA-based approach to species phylogeny was not seriously questioned in the pre-genomic era. All this was to change once comparative genomics yielded more information and multiple complete genome sequences became available for comparison.

Comparative genomics threatens the species tree concept

The first signs that threatened the species tree concept appeared soon after the number of different gene families employed for phylogenetic analysis became more substantial. The problem was that different genes often yielded different trees. This incongruence between tree topologies invaded even the 'sacred of sacred' of phylogenetic taxonomy: the three-domain classification of life. In particular, archaeal genes systematically showed different phylogenetic affinities, with the components of information-processing systems typically affiliating with eukaryotes, whereas metabolic enzymes and structural proteins displayed bacterial connections [7,8].

Systematic comparisons of complete gene sets from sequenced genomes showed beyond reasonable doubt that there is much more to evolution than vertical inheritance. Horizontal gene transfer and lineage-specific gene loss have come to the fore as major evolutionary phenomena, at least in the prokaryotic world [9–14]. The prominence of these events is apparent even without detailed PHYLOGENETIC TREE CONSTRUCTION (see Glossary). Indeed, examination of the phyletic patterns of sets of

http://tig.trends.com 0168-9525/02/$ – see front matter Published by Elsevier Science Ltd. PII: S0168-9525(02)02744-0

ORTHOLOGOUS GENES reveals remarkable patchiness, with most sets represented in only a few lineages, and many having an odd composition, for example, two bacterial and one archaeal species [12,15]. It seems impossible to explain these patterns without invoking massive, lineage-specific gene loss and gain by horizontal gene transfer. A recent quantitative analysis suggested that these processes contributed to the evolution of a substantial majority of orthologous sets of prokaryotic proteins [16].

Phylogenetic tree analysis for multiple gene families sends the same message. A detailed study of 28 protein families (note that, when distant species are involved, phylogenetic analysis is usually carried out with protein rather than DNA sequences) from prokaryotes suggested that, after probable horizontal transfers were removed, there was no reliable phylogenetic signal left in the trees [17]. Similar results were obtained for proteins that comprise the conserved core of archaeal genomes: although they all showed greater conservation within the archaeal domain than outside it, no clear consensus phylogeny within the archaea could be determined [18].

Thus, the study of comparative genomics potentially appeared to undermine the very idea of a ‘Tree of Life’, at least with respect to prokaryotic life (and, since prokaryotes comprise two of the three primary kingdoms, a Tree of Life without them is out of the question) [19,20]. However, genomic information that seems to ‘uproot’ any simple Tree of Life based on a single gene or a small group of genes might also have the potential to offer a way of salvaging the concept itself, at least in a ‘weak’ form. The determination of multiple, complete genome sequences of bacteria, archaea and eukaryotes created the opportunity for a new level of phylogenetic analysis that is based not on a phylogenetic tree for selected molecules (e.g. rRNA) but (ideally) either on the entire body of information contained in the genomes or on a rationally selected, substantial part of this information. Below, we briefly discuss these different genome-based phylogenetic approaches (or, as we call them, ‘genome-tree’ methods), together with the first results in large-scale prokaryotic phylogeny that emerge from the application of these methods.

Genome trees: can comparative genomics help build a consensus?

Methods based on gene content

The most obvious way of comparing genomes is the analysis of gene repertoires. Closely related species share a large proportion of genes; by contrast, distantly related species should have lost a significant fraction of the genes inherited from their last common ancestor, rendering the proportion of shared genes low. If this process continues in a regular fashion (i.e. intergenomic distances based on gene repertoires can be mapped to time scale uniformly across lineages), it could be used for phylogenetic reconstruction.

The latter requirement raises an obvious *a priori* objection. The well-studied plasticity of prokaryotic genomes results in gene content being malleable by selective pressures, both with respect to gene loss (e.g. adaptation to parasitism) and gene acquisition by horizontal transfer (e.g. adaptation to extreme environments). As first presciently noted by Charles Darwin, traits that are subject to strong selection are less suitable for phylogenetic reconstruction than neutral traits because of highly non-uniform rates of changes and the tendency for convergent evolution among the former [1]. This makes gene content comparisons a relatively weak tool to study prokaryote phylogeny *per se*; however, if treated as a means to study similarities and differences between genomes, rather than evolutionary relationships, this approach can produce interesting results.

To use gene content for phylogeny or similarity dendrogram reconstruction, two goals are required: (1) to establish orthology (or, in simpler analysis schemes, homology) relationships between genes; and (2) to choose a method to translate gene presence–absence data into a tree structure. Both goals can be achieved by various means. Snel *et al.* [21,22] and Korbel *et al.* [23] defined orthologs as intergenomic best hits (BeTs) according to Smith–Waterman expectation (*E*)-values, computed the fraction of shared genes for genome pairs, converted it to intergenomic distance and then applied neighbor-joining or least-squares methods to construct trees. Tekaia *et al.* [24] used BLAST *E*-values to establish homology and turned the pairwise matrix of shared gene fraction into factorial space, in which they computed intergenomic distances; these were used to build hierarchical classification trees. Fitz-Gibbon and House [25,26] used intergenomic FASTA *z*-scores followed by single linkage clustering to identify groups of orthologs, then applied PARSIMONY ANALYSIS to reconstruct trees. Lin and Gerstein [27] relied on orthology as defined in the clusters of orthologous groups of proteins (COGs) [15] and built either parsimony or least-squares trees. Natale *et al.* [28] used the coefficient of co-occurrence of genomes in COGs for calculating intergenomic distances, from which neighbor-joining trees were constructed. Wolf *et al.* [29] also employed COGs to construct gene presence–absence (in each genome) matrices, which there then used to construct Dollo parsimony trees. Clarke *et al.* [30] computed the ratio of orthologs (identified as reciprocal BeTs) to the number of genes in the smaller genome and constructed least-squares trees. In addition, Wolf *et al.* [31] and Lin and Gerstein [27] built dendrograms on the basis of the predicted protein fold composition of genomes; the former work used linear correlation coefficient between fold abundance lists and the latter relied on presence–absence data.

The trees produced by different applications of this approach are not directly comparable because of the different species sets and different methods for intergenomic distance calculation used by each group.

Nevertheless, several major trends are recognizable. Enough phylogenetic information is retained in gene repertoires to provide reliable classification on both ends of the phylogenetic distance scale. Gene content trees show good separation between primary kingdoms, and consistently group together closely related species. However, on the intermediate distances, which involve relationships between major lineages, this approach appears to be less suitable for phylogenetic inference. The major factor that determines tree topology appears to be the relative amount of gene loss in different genomes (e.g. the major division in the bacterial branch is between the free-living and parasitic forms), which results in well-defined major lineages (e.g. Proteobacteria) being broken up [29]. This is readily explained by the above-mentioned genome plasticity of prokaryotic genomes, which is manifest in common trends in genome reduction under the selective pressure during the adaptation to parasitism. Attempts to overcome this effect included simple removal of parasites from the analysis [25,26] or normalization of the intergenomic distances by the number of genes of the smaller genome in each pair [21,23]. The latter method results in reasonable phylogenetic reconstructions, with most of the major prokaryotic lineages recovered. The legitimacy of using the fraction of shared genes for phylogenetic reconstructions has been challenged on the grounds that this is a phenetic rather than phylogenetic character [32]. This objection seems to apply to any DISTANCE METHOD of phylogenetic analysis; however, it is well known that, combined with a proper calculation of evolutionary distances, such methods often result in correct trees. More importantly, gene content analysis seems to have less resolution power than some of the other genome-tree approaches (see below).

Figure 1 shows a gene content tree constructed using the latest COG data and the co-occurrence coefficient for calculating intergenomic distances. The separation of free-living and parasitic bacteria is apparent (as discussed above) but, in addition, a potentially phylogenetically meaningful branch unifying cyanobacteria, actinobacteria and deinococcales appears in this tree (see below).

Methods based on gene order

The same logic as above applies to genome comparisons based on gene order. Rearrangements continuously shuffle the genomes, gradually breaking ancestral gene strings. The operonic organization of a prokaryotic genome complicates the kinetics of this process. On the one hand, the selective advantage of physical proximity for co-regulation makes some gene arrays less prone to break-up than others, thus extending the range of evolutionary distances over which gene order comparison is technically possible [33,34]. On the other hand, selective forces acting on operons make them sensitive to the influence of the environmental niche occupied by the organism at any particular time. Furthermore, operons are especially liable to being transferred as a whole, accentuating the effect of lateral transfer on the tree topology [35].

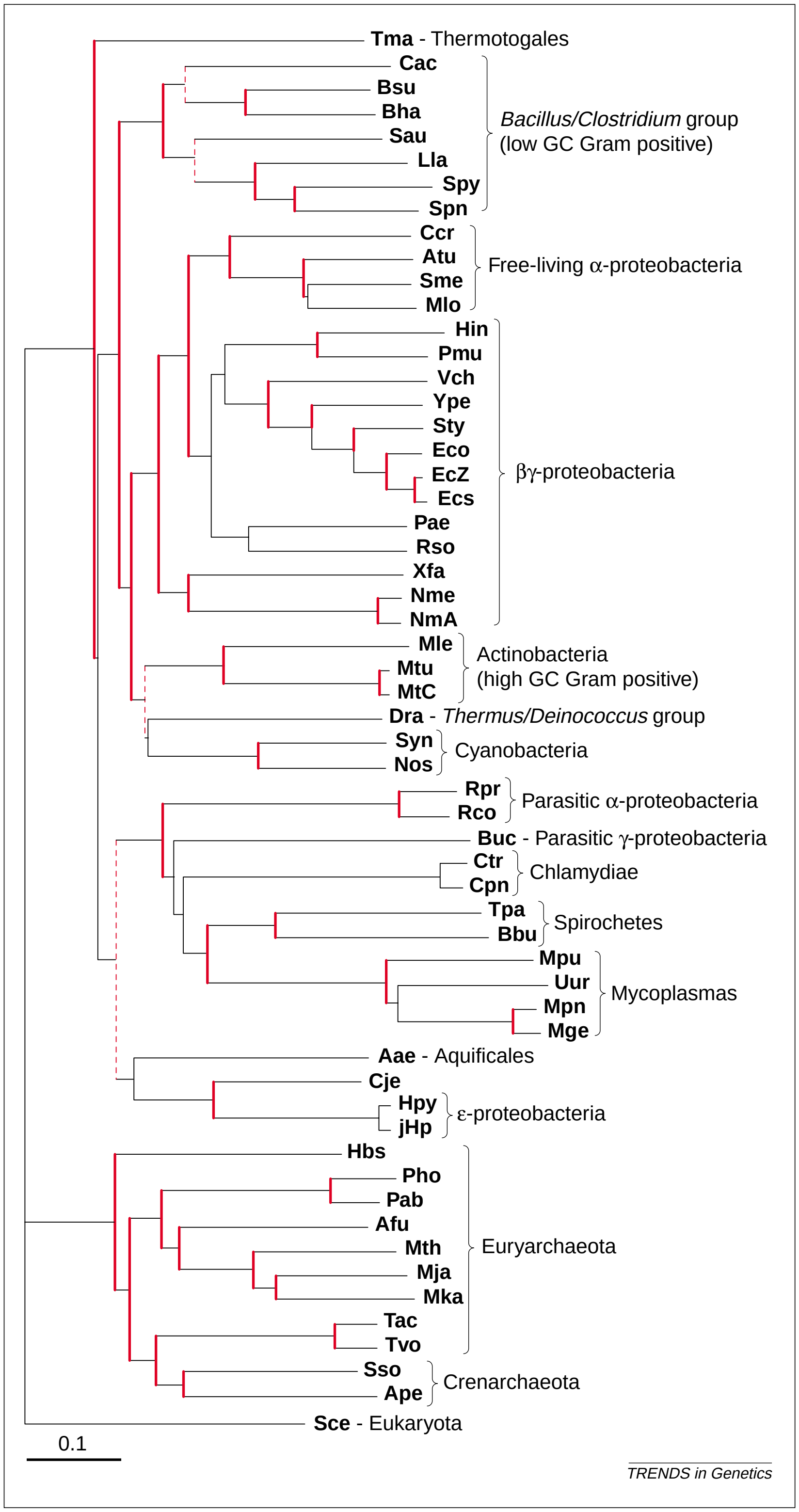
Wolf *et al.* [29] identified pairs of proteins belonging to COGs whose physical proximity is conserved in several genomes. The presence–absence matrices of these pairs were analyzed using Dollo parsimony and neighbor-joining methods, which produced essentially the same topology. Korbel *et al.* [23] counted adjacent pairs of BeT-derived orthologs shared by two genomes, converted the fraction of such pairs to distance and used these distances to construct neighbor-joining or least-squares trees.

Owing to the high rate of intragenomic rearrangements, the gene order trees are (at least in theory) especially suitable to resolving the phylogeny of closely related species [36]. Generally, this approach behaves in a manner similar to the gene content methods, providing a good separation between Archaea and Bacteria, and keeping closely related species together, but offering poor resolution on intermediate distances. Both Wolf *et al.* [29] and Korbel *et al.* [23] described the effect of horizontal gene transfer on the topology of these trees.

Methods based on evolutionary distances between orthologs

Evolutionary distances measured between different pairs of orthologs in a given two genomes show a broad distribution. In theory, this is because of the genuine variability of mean protein evolution rates caused by the variability in the strength of selective constraints acting on functionally different proteins. In practice, several other factors add to the rate variance, including sampling errors, misidentification of orthologs and lateral gene transfer. Nevertheless, if, for the majority of ortholog pairs, the time of divergence coincides with the divergence of species, it seems reasonable to expect that the distribution of the distances retains enough phylogenetic information to be used for tree construction.

Grishin *et al.* [37], Wolf *et al.* [29] and Clarke *et al.* [30] used reciprocal BeTs to identify pairs of likely orthologs for a pair of genomes and to obtain a distribution of distances between orthologs. Grishin *et al.* [37] computed the scaling factor for the best-fitting approximation of the distance distribution function; Wolf *et al.* [29] used the median identity percentage between orthologs transformed into distance; and Clarke *et al.* [30] relied on the median of the relative BLAST score. Least-squares trees were built upon the pairwise distance matrices in all three publications. In addition, Clarke *et al.* [30] employed a statistical test to identify and remove ‘phylogenetically discordant’ sequences (those that displayed an abnormal pattern of similarity to orthologs in several genomes) to reduce the potential effect of lateral gene transfer and the misidentification of orthologs. This procedure resulted in an increased bootstrap support but had little effect on tree topology.



TRENDS in Genetics

Fig. 1. An updated gene content tree of prokaryotes. The intergenome distances were calculated as follows: (1) a set of clusters of orthologous groups of proteins (COGs) represented in each genome was determined; (2) the distance between a pair of genomes (A and B) was calculated as $D_{AB} = 1 - J_{AB}$ where J_{AB} is the Jaccard coefficient, which reflects the similarity between the COG sets A and B and is calculated as $J_{AB} = |A \cap B| / |A \cup B|$ ($J_{AB} \subset [0..1]$); (3) a least-squares tree was calculated from the pairwise distance matrix. Bootstrap analysis was performed by resampling the entire set of 4075 COGs. Solid red lines indicate internal nodes with bootstrap support >90%; dotted red lines indicate internal nodes with bootstrap support between 70% and 90%, and black lines show nodes with <70% support. Species abbreviations: Archaea: Afu, *Archaeoglobus fulgidus*; Hbs, *Halobacterium* sp. NRC-1; Mja, *Methanococcus jannaschii*; Mka, *Methanopyrus kandleri* AV19; Mth, *Methanothermobacter thermautotrophicus*; Tac, *Thermoplasma acidophilum*; Tvo, *Thermoplasma volcanium*; Pho, *Pyrococcus horikoshii*; Pab, *Pyrococcus abyssi*; Ape, *Aeropyrum pernix*; Sso, *Sulfolobus solfataricus*; Sce, *Saccharomyces cerevisiae* (a eukaryotic genome included in addition to the prokaryotes). Bacteria: Aae, *Aquifex aeolicus*; Tma, *Thermotoga maritime*; Dra, *Deinococcus radiodurans*; Cac, *Clostridium acetobutylicum*; Mtu, *Mycobacterium tuberculosis* H37Rv; MtC, *Mycobacterium tuberculosis* CDC1551; Mle, *Mycobacterium leprae*; Lla, *Lactococcus lactis*; Spy, *Streptococcus pyogenes* M1 GAS; Spn, *Streptococcus pneumoniae* TIGR4; Sau, *Staphylococcus aureus* N315; Bsu, *Bacillus subtilis*; Bha, *Bacillus halodurans*; Syn, *Synechocystis* sp.; Nos, *Nostoc* sp. PCC 7120; Eco, *Escherichia coli* K12; EcZ, *Escherichia coli* O157:H7 EDL933; Ecs, *Escherichia coli* O157:H7; Ype, *Yersinia pestis*; Sty, *Salmonella typhimurium* LT2; Buc, *Buchnera* sp. APS; Vch, *Vibrio cholerae*; Pae, *Pseudomonas aeruginosa*; Hin, *Haemophilus influenzae*; Pmu, *Pasteurella multocida*; Xfa, *Xylella fastidiosa* 9a5c; Nme, *Neisseria meningitidis* MC58; NmA, *Neisseria meningitidis* Z2491; Rso, *Ralstonia solanacearum*; Hpy, *Helicobacter pylori* 26695; jHp, *Helicobacter pylori* J99; Cje, *Campylobacter jejuni*; Atu, *Agrobacterium tumefaciens* strain C58 (Cereon); Sme, *Sinorhizobium meliloti*; Mlo, *Mesorhizobium loti*; Ccr, *Caulobacter crescentus*; Rpr, *Rickettsia prowazekii*; Rco, *Rickettsia conorii*; Ctr, *Chlamydia trachomatis*; Cpn, *Chlamydophila pneumoniae*; Tpa, *Treponema pallidum*; Bbu, *Borrelia burgdorferi*; Uur, *Ureaplasma urealyticum*; Mpu, *Mycoplasma pulmonis*; Mpn, *Mycoplasma pneumoniae*; Mge, *Mycoplasma genitalium*.

The trees constructed by Wolf *et al.* [29] and Clarke *et al.* [30] include enough genomes to allow a meaningful comparison and are remarkably similar. Both trees confidently separate Archaea from Bacteria. Both recover most of the known major bacterial lineages: proteobacteria (also providing good separation between α-, βγ- and ε-proteobacteria), low GC Gram-positive bacteria (including mycoplasmas), spirochetes and chlamydiae. Furthermore, both trees suggest a relationship between spirochetes and chlamydiae, and both place bacterial hyperthermophyles (*Thermotoga* and *Aquifex*) near the root of the bacterial subtree, followed by *Deinococcus*, *Mycobacterium* and *Synechocystis*. The trees fail to support the monophyly of euryarchaeota, place *Halobacterium* at the root of the archaeal subtree and unify methanogens with pyrococci.

Methods based on concatenated alignments of orthologous protein sequences

Traditional sequence-based phylogeny relies on gradual sequence change over time. The three main problems with using single genes to determine the relationships between species are the insufficient number of informative sites, the variability of evolutionary rates in different lineages, and the effect of lateral gene transfer. The first two factors add uncertainty to reconstructions; the last factor leads to protein phylogenies being genuinely different from (the hypothetical) species phylogeny. In an attempt to overcome these problems, one can concatenate many sequence alignments into one and use the combined long sequence for tree reconstruction. If there is no systematic bias in the pattern of horizontal transfers involving the concatenated genes, or the likelihood of such transfer is reduced by the careful choice of genes, the trees reconstructed from such an alignment have the potential to provide excellent resolution. However, there are several limitations to this approach. First, all of the concatenated alignments must include exactly the same set of species represented by one protein each, which limits the choice of families to universally represented ones with no PARALOGOUS GENES, or might require a selection of paralogs, which might be a source of bias in itself. Second, concatenation forces a single sequence change model on all proteins (including branch lengths and intra-protein variability of evolutionary rates), which, in general, is not necessarily true.

Teichmann and Mitchison [17] analyzed a concatenated alignment, mostly comprised of translation-related proteins, using the neighbor-joining algorithm. They found that horizontal gene transfer in three families apparently had a major effect on the topology of their trees; after the elimination of probable transfer candidates, the remaining set failed to produce a resolved tree. By contrast, a concatenated alignment, mostly of ribosomal proteins, constructed by Hansmann and Martin [38] resulted in well-supported trees based on MAXIMUM LIKELIHOOD METHODS, especially after the systematic removal of poorly conserved sites. Brown *et al.* made a concatenated set dominated by translation-related proteins to produce parsimony and neighbor-joining trees, which had very similar topologies [39]. Removal of phylogenetically discordant sequences resulted in a well-resolved phylogeny of prokaryotes. Wolf *et al.* used concatenated alignments of universal ribosomal proteins (to minimize the possibility of lateral gene transfer within the dataset) to produce maximum likelihood trees [29]. Brochier *et al.* [40] and Forterre *et al.* [41,42] used, respectively, a diverse set of translation-related proteins from bacteria and the set of archaeal ribosomal proteins for maximum likelihood tree construction; in these studies, phylogenetic tree construction was preceded by principal component analysis to identify and remove phylogenetically discordant (i.e. apparently subject to lateral gene transfer) sequences.

The trees constructed using concatenated alignments share many features with each other and with the trees based on ortholog distance described above. Especially notable is the similarity of the trees produced by three independent groups [29,39,40]. Similar to the ortholog-distance trees, these trees support most of the well-established bacterial clades. In addition, all three show a common clade for spirochetes and chlamydiae. Wolf *et al.* [29] and Brochier *et al.* [40] place bacterial hyperthermophyles on a common branch at the base of the bacterial subtree whereas, in the trees of Brown *et al.* [39], they form separate deep branches. Wolf *et al.* [29] and Brochier *et al.* [40] agree on joining *Deinococcus*, *Mycobacterium* and *Synechocystis* in a single clade. The trees of Brown *et al.* [39] support a *Deinococcus–Synechocystis* clade, but join actinobacteria with green sulfur bacteria (actinobacteria are absent from the work of Wolf *et al.* [29]; in the work of Brochier *et al.* [40], cyanobacteria belong to a distinct branch together with cytophagales. Both Wolf *et al.* [29] and Brown *et al.* [39] support methanogen–pyrococci and *Thermoplasma–Archaeoglobus* clades within the archaea but disagree on the position of crenarchaeota; the phylogeny of concatenated archaeal ribosomal proteins constructed by Forterre *et al.* [41,42] does not reproduce the methanogen–pyrococci clade (nor the monophyly of methanogens themselves) or the *Thermoplasma–Archaeoglobus* clade, but identifies crenarchaeota and euryarchaeota as sister clades.

Approaches based on multiple trees

In principle, combining phylogenetic information contained in multiple, independently reconstructed trees allows one to achieve the same kind of resolution enhancement as with concatenated alignments without the inherent limitations of the latter. However, these approaches face a methodological problem of their own, namely reconciling many trees, possibly including different sets of species and paralogs.

Sicheritz-Ponten and Andersson [43] analyzed over 8000 individual neighbor-joining trees based on proteins from seven prokaryotic species, systematically counting the nearest neighbors of the proteins in question. The concept of 'phylome' (the set of phylogenetic trees including proteins from the given genome) proposed by these researchers, while not producing global trees *per se*, allows one to capture trends in the placement of the given species in the trees. Wolf *et al.* [29] constructed 132 trees representing (nearly) universal protein families with low propensity for lineage-specific duplications. The nearest neighbor census for particular species and/or higher taxa was used to validate hypotheses formulated during the analysis of other types of genome trees. Daubin and Gouy [44] employed a 'supertree' approach by combining matrix representations [45,46] (with weights assigned to internal nodes on the basis of bootstrap probabilities) of individual maximum likelihood trees and analyzing the resulting matrix using the neighbor-joining method. Zhaxybayeva and Gogarten [47] employed maximum likelihood mapping of three topologies possible for a quartet of species in order to poll the trends existing in multiple four-ortholog trees. Two methodologically important studies should also be mentioned, although they did not directly address prokaryote phylogeny. Page [48] used a method that minimizes the total number of 'apparent duplications' in the tree to reconcile many individual trees in an attempt to reconstruct the phylogeny of vertebrates. Bapteste *et al.* [49] analyzed the phylogeny of eukaryotes using the sum of log-likelihoods for competing topologies instead of concatenating alignments, to avoid forcing the same parameters of the evolutionary model on different protein families.

The supertree of Daubin and Gouy [44] is generally consistent with the ortholog-distance trees and concatenated-alignment trees described above. It successfully recovers known taxonomic groups, supports a separate clade for bacterial hyperthermophyles (although it is placed as a sister group to proteobacteria instead of in the bacterial root), joins *Deinococcus*, *Mycobacterium* and *Synechocystis* in another clade, and also groups methanogens and pyrococci. Unlike many other approaches, the supertree places spirochetes and chlamydiae separately in the bacterial root, instead of joining them on a common branch, and recovers crenarchaeota and euryarchaeota as sister groups.

The emerging consensus

Figure 2 shows a synthetic tree of prokaryotes that has taken into account the results of the genome-tree analyses discussed above (specifically, ortholog-distance trees [29,30], concatenated alignment trees [29,39,40–42] and the supertree of Daubin and Gouy [44]) and has attempted to depict the apparent consensus. At least two major new clades are strongly supported by different types of analysis and appear reliable: chlamydiae–spirochetes among bacteria and methanogens–pyrococci among euryarchaeota. In addition, several other major groupings were supported by some but not by other approaches and should be considered tentative for the moment (e.g. the unification of cyanobacteria, deinococcales and actinobacteria, or of aquificales and thermotogales).

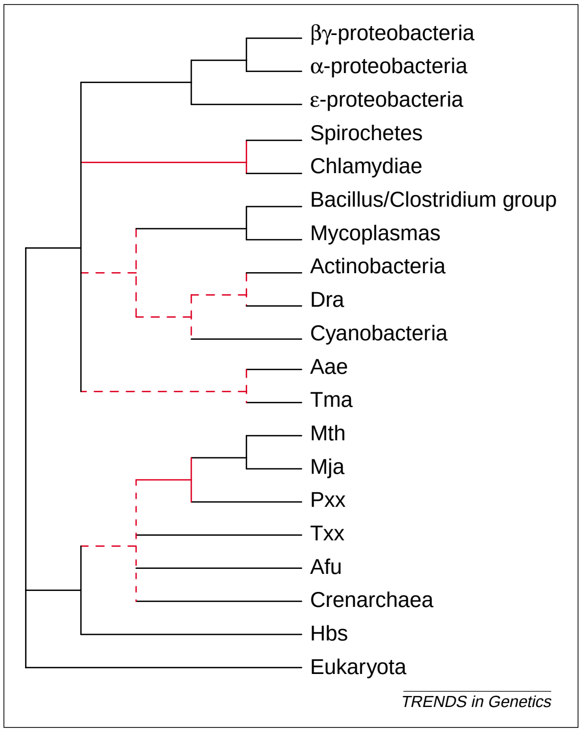


Fig. 2. A tentative consensus of genome trees of prokaryotes. Black branches are conventional clades reproduced by genome-tree approaches; solid red branches are new clades that we consider to be firmly established by genome trees; broken red branches are tentative clades suggested by genome trees; thin lines are unresolved multifurcations. Species/clades abbreviation: Txx, two *Thermoplasma* species (*T. acidophilum* and *T. volcanium*); Pxx, two *Pyrococcus* species (*P. horikoshii* and *P abyssi*); all others, the same as in Fig. 1.

Conclusions and outlook: redefining the concept of the Tree of Life

The results of comparative genomics suggest that the simple notion of a single Tree of Life that would accurately and definitively depict the evolution of all life forms is gone forever. Individual genes, especially those of prokaryotes, follow their unique evolutionary trajectories. This uniqueness stems from the fact that different genes evolved at different points during the history of life and that, in addition to vertical inheritance, evolution of most orthologous families involved multiple instances of lineage-specific gene loss and gene acquisition by horizontal transfer. However, those same comparative genomic studies that have 'uprooted' the Tree of Life give us hope that the concept could be rescued, albeit in a limited sense. Taken together, the results achieved by genome-tree approaches indicate that there is, after all, a phylogenetic signal in the sequences of prokaryotic proteins, but it is weak because of massive gene loss and horizontal transfer and possibly also because of (relatively) rapid divergence of the major lineages in the deep past. To capture this faint signal, analysis of genome-wide protein sets or of carefully selected subsets is required. With all due caution, these approaches seem to converge on some major new taxa, such as the chlamydiae–spirochetes clade, whose existence was not even suspected in the pre-genomic era (Fig. 2). However, the concept of the Tree of Life is bound to change in the post-genomic world. It cannot anymore be thought of as a definitive 'species tree' (something that does not even exist in reality) but only as a central trend in the rich patchwork of evolutionary history, replete with gene loss and horizontal transfer events [50].

Phylogenies based on sequences of single molecules, including rRNA, have helped define the three primary kingdoms and are extremely useful in resolving terminal branches. For the latter task, they will remain the method of choice for the foreseeable future, given that rRNA sequencing is still incomparably easier and faster than genome sequencing. However, these approaches have done little for our understanding of the fundamentally interesting evolutionary events in between. Moreover, such phylogenies can seriously mislead in cases of anomalous evolutionary modes in particular lineages. The recent conundrum around the phylogenetic position of the hyperthemophilic archaeal methanogen *Methanopyrus kandleri* is a case in point. rRNA-based phylogeny placed this organism at the base of the crenarchaeal branch and some analyses suggested an even deeper branching [51]. However, once the genome sequence was determined, genome trees based on concatenated ribosomal proteins, gene content and gene order all unequivocally pointed to a monophyletic clade of archaeal methanogens including *M. kandleri* [52].

What's in store for genome trees? This subfield of molecular phylogenetics is still very young and the most effective methods remain to be developed. Nevertheless, both the initial results and the general notion that using genome-wide information helps enhance the phylogenetic signal suggest that the future belongs to these approaches. The growth of the number of sequenced genomes accentuates the signal but also increases the noise. Furthermore, powerful tree-building algorithms, such as maximum likelihood, tend to rapidly become prohibitively expensive with the increase in the number of analyzed species. Careful selection of slowly evolving species that are most apt for phylogenetic analysis potentially could allow advantage to be taken of the progress of genomics and avoiding these problems. Such analysis should be expected to support or refute some tentative new clades suggested by current genome trees (Fig. 2).

Beyond doubt, major difficulties lie ahead. To mention just one: all alignment-based phylogenetic methods face the problem of site selection, and this becomes all the more pertinent with long concatenated alignments employed in genome-tree analysis. It has been shown that elimination of subsets of aligned residues in concatenated ribosomal protein alignments changes the resulting tree topology, in some cases dramatically [38], but there is no theory that would indicate which subsets tell the right story. This and other problems will keep molecular phylogeneticists and evolutionary genomicists busy for years to come. What is already clear is that genomics has brought an extra layer of complexity to molecular evolution but has also brought the information that is required for generating a new, richer picture of the history of life.

Acknowledgements
We thank Roman Tatusov for providing updated, pre-release COG data, and Mark Ragan and Herve Philippe for providing preprints of their work.

References

1 Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, John Murray
2 Haeckel, E. (1997) *The Wonders of Life: A Popular Study of Biological Philosophy*, De Young Press
3 Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* 51, 221–271
4 Woese, C.R. *et al.* (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4576–4579
5 Doolittle, R.F. and Handy, J. (1998) Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.* 8, 630–636
6 Olsen, G.J. *et al.* (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* 176, 1–6
7 Gupta, R.S. and Golding, G.B. (1996) The origin of the eukaryotic cell. *Trends Biochem. Sci.* 21, 166–171
8 Golding, G.B. and Gupta, R.S. (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol. Biol. Evol.* 12, 1–6
9 Koonin, E.V. *et al.* (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* 25, 619–637
10 Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2129
11 Doolittle, W.F. (1999) Lateral genomics. *Trends Cell Biol.* 9, M5–8
12 Koonin, E.V. *et al.* (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55, 709–742
13 Ragan, M.A. (2001) Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.* 11, 620–626
14 Ragan, M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* 201, 187–191
15 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
16 Snel, B. *et al.* (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12, 17–25
17 Teichmann, S.A. and Mitchison, G. (1999) Is there a phylogenetic signal in prokaryote proteins? *J. Mol. Evol.* 49, 98–107
18 Nesbo, C.L. *et al.* (2001) Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J. Mol. Evol.* 53, 340–350
19 Pennisi, E. (1999) Is it time to uproot the tree of life? *Science* 284, 1305–1307
20 Pennisi, E. (1998) Genome data shake tree of life. *Science* 280, 672–674
21 Snel, B. *et al.* (1999) Genome phylogeny based on gene content. *Nat. Genet.* 21, 108–110
22 Huynen, M.A. *et al.* (1999) Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science* 286, 1443a
23 Korbel, J.O. *et al.* (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* 18, 158–162
24 Tekaia, F. *et al.* (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.* 9, 550–557
25 Fitz-Gibbon, S.T. and House, C.H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* 27, 4218–4222
26 House, C.H. and Fitz-Gibbon, S.T. (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J. Mol. Evol.* 54, 539–547
27 Lin, J. and Gerstein, M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.* 10, 808–818
28 Natale, D.A. *et al.* (2000) Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol.* 1, RESEARCH0009
29 Wolf, Y.I. *et al.* (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* 1, 8
30 Clarke, G.D. *et al.* (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
31 Wolf, Y.I. *et al.* (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res.* 9, 17–26
32 Doolittle, W.F. (1999) Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science* 286, 1443a
33 Lathe, W.C., III *et al.* (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* 25, 474–479
34 Wolf, Y.I. *et al.* (2001) Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. *Genome Res.* 11, 356–372
35 Lawrence, J.G. (1997) Selfish operons and speciation by gene transfer. *Trends Microbiol.* 5, 355–359
36 Suyama, M. and Bork, P. (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* 17, 10–13
37 Grishin, N.V. *et al.* (2000) From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res.* 10, 991–1000
38 Hansmann, S. and Martin, W. (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int. J. Syst. Evol. Microbiol.* 50, 1655–1663
39 Brown, J.R. *et al.* (2001) Universal trees based on large combined protein sequence data sets. *Nat. Genet.* 28, 281–285
40 Brochier, C. *et al.* (2002) Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* 18, 1–5
41 Forterre, P. *et al.* Evolution of the Archaea. *Theor. Popul. Biol.* (in press)
42 Matte-Tailliez, O. *et al.* (2002) Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* 19, 631–639
43 Sicheritz-Ponten, T. and Andersson, S.G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* 29, 545–552
44 Daubin, V. and Gouy, M. (2001) Bacterial molecular phylogeny using supertree approach. *Genome Inform. Ser. Workshop Genome Inform.* 12, 155–164
45 Ragan, M.A. (1992) Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* 1, 53–58
46 Ragan, M.A. (1992) Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *Biosystems* 28, 47–55

47 Zhaxybayeva, O. and Gogarten, J.P. (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. *BMC Genomics* 3, 4

48 Page, R.D. (2000) Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Mol. Phylogenet. Evol.* 14, 89–106

49 Bapteste, E. *et al.* (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1414–1419

50 Woese, C.R. (2000) Interpreting the universal phylogenetic tree. *Proc. Natl. Acad. Sci. U. S. A.* 97, 8392–8396

51 Burggraf, S. *et al.* (1991) *Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens. *Syst. Appl. Microbiol.* 14, 346–351

52 Slesarev, A.I. *et al.* (2002) The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4644–4649

Checking cell size in yeast

Ivan Rupeš

To remain viable, cells have to coordinate cell growth with cell division. In yeast, this occurs at two control points: the boundaries between G1 and S phases, also known as Start, and between G2 and M phases. Theoretically, coordination can be achieved by independent regulation of growth and division, or by participation of surveillance mechanisms in which cell size feeds back into cell-cycle control. This article discusses recent advances in the identification of sizing mechanisms in budding and in fission yeast, and how these mechanisms integrate with environmental stimuli. A comparison of the G1–S and G2–M size-control modules in the two species reveals a degree of conservation higher than previously thought. This reinforces the notion that internal sizing could be a conserved feature of cell-cycle control throughout eukaryotes.

Published online: 25 July 2002

Ivan Rupeš
Dept of Biology, Queen's University, Kingston, Ontario, Canada K7L 3N6.
e-mail: rupesi@biology.queensu.ca

How does a cell control its size? Two schools of thought exist: either a cell divides after it reaches a certain critical size, or cell growth and proliferation are regulated independently, with cell size emerging from a simple correlation of the two [1–5]. The most important difference between the two hypotheses is that cell size feeds back into the cell-cycle regulatory system in the former but not in the latter (Fig. 1a). Thus, the validity of the critical size theory depends on the existence of a 'sizer' – a molecule or set of molecules whose activity correlates with cell size. Nevertheless, the sizer is only one component that determines when cell division occurs; in addition, the extracellular environment influences the timing of the response to the changing activity of the sizer.

Two species of yeast, the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe*, provide genetic models in which to study cell-cycle control. The evolutionary divergence of these yeast is about the same as that between each of them and human [6]. Both yeast share cell-cycle characteristics with higher eukaryotes, such as G1, S, G2 and M phases, cyclin-dependent kinases (CDKs) and checkpoint controls [7–9]. Owing to the uncertain supply of nutrients in the wild, the yeast cell-division rate must be coordinated with widely variable rates of cell growth, otherwise cells would get progressively smaller or larger. Although cell size is reduced in less favorable growth conditions, the range of growth rates exceeds the corresponding range of cell sizes for both yeast. Thus, a relationship between growth and division is a fact of life for yeast cells. The early studies have presented a compelling case for the existence of a critical size that is a prerequisite for progression through the cell cycle. In budding yeast, cell division is asymmetrical and produces cells of unequal size. To compensate for this asymmetry, which becomes more pronounced with increasing nutrient limitation, new daughter cells grow more before division than the mother cells. This additional growth occurs almost entirely in G1, before the reference point known as Start. Once Start is passed, the rest of the cell cycle is relatively constant in length [10] (Fig. 1b).

In fission yeast, G2–M is the primary cell-size control point [9]. The relative length of G2 varies greatly with growth conditions (Fig. 1b). Nitrogen limitation reduces cell size at division, and sudden shifts between different sources of nitrogen generate rapid acceleration or delay of mitosis in cells that are above or below the new cell-size threshold (i.e. the minimum size required for initiation of mitosis) [9]. Even during balanced growth, individual fission yeast cells can compensate for random fluctuations in their size at birth by adjusting their time spent in G2 [11].

Do fission yeast have a size control in G1, and do budding yeast have a control in G2? Start is defined in fission yeast as in budding yeast, although in favorable growth conditions it occurs almost immediately after exit from mitosis. A normally cryptic size-control point at Start is uncovered in mutants in which the G2–M size-control has collapsed and cells enter mitosis prematurely. These cells have an extended G1, suggesting that they initiate S phase only after reaching a certain minimum size [12] (Fig. 1b). Although the two yeast have traditionally been thought of as using different size-control strategies, the aim of this article is to show that they use similar mechanisms. What is different is their emphasis on the size-control points. Because these control mechanisms have been conserved over such long evolutionary distances,

0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(02)02745-2