Genome Trees and the Nature of Genome Evolution

Berend Snel, Martijn A. Huynen, and Bas E. Dutilh

Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, Nijmegen 6525 ED, The Netherlands; email: snel@cmbi.ru.nl, huynen@cmbi.ru.nl, dutilh@cmbi.ru.nl

Key Words
phylogenomics, phylogeny, comparative genomics, complete genome, horizontal gene transfer

Abstract
Genome trees are a means to capture the overwhelming amount of phylogenetic information that is present in genomes. Different formalisms have been introduced to reconstruct genome trees on the basis of various aspects of the genome. On the basis of these aspects, we separate genome trees into five classes: (a) alignment-free trees based on statistic properties of the genome, (b) gene content trees based on the presence and absence of genes, (c) trees based on chromosomal gene order, (d) trees based on average sequence similarity, and (e) phylogenomics-based genome trees. Despite their recent development, genome tree methods have already had some impact on the phylogenetic classification of bacterial species. However, their main impact so far has been on our understanding of the nature of genome evolution and the role of horizontal gene transfer therein. An ideal genome tree method should be capable of using all gene families, including those containing paralogs, in a phylogenomics framework capitalizing on existing methods in conventional phylogenetic reconstruction. We expect such sophisticated methods to help us resolve the branching order between the main bacterial phyla.
INTRODUCTION

Phylogenies and Genome Trees

Baffled by the variety in life, one of man’s first biological activities has been to classify it. Since Darwin’s theory of evolution, the ultimate goal is to obtain a hierarchical classification that matches the evolutionary relations between species. This makes the construction of phylogenies one of the central activities of biologists, not only to reconstruct the history of life, but also to understand it because “nothing in biology makes sense except in the light of evolution” (16). Traditionally, phylogenies were constructed from phenotypic characteristics, and phenotypic characteristics continue to play a dominant role in the analysis of data such as fossils. However, with the advent of sequencing technologies, it has become possible to construct trees on the basis of nucleotide and amino acid sequences, as foreseen by Zuckerkandl & Pauling (76). Sequence-based trees such as the ribosomal RNA molecules have become the golden standard in areas where phenotypic data are scarce and are at least on equal footing in areas where we have phenotypic data as well as sequence data. Sequence-based analyses have yielded surprising observations, such as the close phylogenetic relationship between archaea and eukaryotes relative to bacteria (27), between fungi and animals relative to plants (3), and the monophyly of the Afrotheria (69). Furthermore, they can be used for organisms for which we do not have phenotypic data or for which we do not even know exist, as in the case of the environmental sampling of ribosomal RNA (5). Yet, the principle of constructing phylogenies on the basis of a single gene has been challenged (17), and for a gene such as ribosomal RNA, many different phylogenetic trees have been published on the basis of different models of sequence evolution (8, 51).

With the availability of complete genome sequences it has become possible to reconstruct phylogenies on the basis of much larger sets of data per species, allowing in principle a more reliable and representative inference of the tree of life. As complete genomes have been available only since 1995 (24), and the methods discussed in this review are all relatively new, there is no consensus on what is the best way of integrating genome data or which genomic data should be used. Furthermore,
the phylogenetic value of genome trees is not as commonly accepted as that of gene trees simply because the different parts of the genome do not necessarily have the same evolutionary history. This observation has led to the question whether it is possible to construct a phylogeny at the level of genomes (17). Given these arguments, it is perhaps best to refer to a clustering of species on the basis of characteristics of complete genomes as a genotype tree rather than a genome phylogeny. Genome trees then are a means to capture and compare the overwhelming amount of information that is present in genomes and to subsequently combine this in a tree that can be interpreted as a phylogeny.

Not only are phylogenies interesting per se, all inferences in comparative biology depend on accurate estimates of evolutionary relationships (39). For example, when we want to investigate how the HOX pathway evolved, we need to know the evolutionary relationships between the species in which it occurs. Similarly, comparing complete genomes using a phylogenetic tree allows researchers to study the evolution of genomic properties such as gene repertoire. Genome trees take an interesting intermediate position in this respect. In addition to being a means to derive genome-wide estimates of evolutionary relationships, they can also serve as a map on which to study the evolution of the genomes themselves. A genome tree is a direct readout of the processes that govern genome evolution, such as the rearrangement of chromosomal gene order. In this review we discuss the various methods used to reconstruct genome trees, the new taxonomic insights they have given with respect to prokaryotic phylogeny, the controversies regarding their construction, and the insights that they have provided into the process of genome evolution.

**Why There Are So Many Ways to Construct Genome Trees**

A plethora of approaches to construct trees from complete genomes have been introduced (9, 14, 23, 28, 31, 34, 44, 52, 56, 62, 67, 73, 74). The reasons for this large variation in genome trees are twofold. First, we cannot simply extend the classical approaches of sequence-based phylogenies to complete genomes. In classical molecular phylogenetics the corresponding homologous characters in a multiple-sequence alignment, nucleotides or amino acids, are the basic elements used to infer the phylogeny. Extending that single-gene phylogeny paradigm by making a long multiple-sequence alignment of genomes is not possible because evolutionary events such as gene order rearrangements, gene loss, and gene duplication occur at such high rates that even genomes from the same species cannot simply be aligned, as in *Escherichia coli*, for which the genome of different strains differs by as much as one megabase (72). Second, and more importantly, there are many more features to complete genomes than to genes. The sheer quantity of data, and types of data from any genome, has inspired researchers to develop new methods to cluster them. On the basis of characters used to cluster genomes, genome trees can be globally divided into five classes (Figure 1): (a) alignment-free genome trees based on statistic properties of the complete genome, (b) gene content trees based on the presence and absence of genes, (c) genome trees based on chromosomal gene order, (d) genome trees based on average sequence similarity, and (e) phylogenomic trees based either on the collection of phylogenetic trees derived from shared gene families or on a concatenated alignment of those families.

In addition to the diversity of the information used to construct genome trees, various methods have been developed to translate the same genomic property into a tree. We classify the myriad genome trees that have appeared, on the basis of the type of genomic information, and discuss the variations in the precise phylogenetic methods that have been applied to each genomic property in the respective sections.
Figure 1
Classification of genome tree reconstruction methods. The genome tree publications are put in the context of the genomic property used to construct the tree. A paper that contains trees constructed with different methods is displayed in all the appropriate contexts. The amount of data available to construct a tree decreases from top (annotated genomes) to bottom (1:1 and ubiquitous families). “1:1 families” means gene families with a single copy in each genome.

FIVE CLASSES OF GENOME TREES, AND COUNTING

Alignment-Free Genome Trees: Quick and Dirty
Several genome tree reconstruction methods use a statistic of the entire genomic DNA, or of all encoded proteins in a genome, to derive a distance between genomes that is then used to cluster them (44, 52, 55, 56). One class of alignment-free tree inference methods relies on word frequency, i.e., oligomers, K-strings, or n-mers in DNA or proteins (71). The K-string method applies such a word-frequency-based method to complete genomes by simply counting the frequencies of all oligomers five or six amino acids in length in all the predicted protein sequences (55, 56). The results are combined in a word-frequency vector, and the angle between two vectors represents the distance between two genomes. A distance-based clustering method is applied to generate the tree. This alignment-free method performs reasonably well. For example, it successfully clusters the proteobacteria as a monophyletic group (56), unlike phylogenies based on single random, nonmarker genes such as the glycolytic enzymes (10).

Another class of alignment-free methods is based on a concept from information theory (71) called shared information, i.e., how much information is needed to obtain genome a, given that we know genome b. For something as complex as a genome, the specific implementations use algorithmic compression, such as Kolmogorov complexity (44) or Lempel-Ziv complexity (52): The distance between two genomes is represented by the length of the shortest computer program to output genome a given the input genome b.
These complexity measures have so far been applied to complete mitochondrial genomes from mammals, for which they accurately reconstruct the known phylogeny.

By not having to decide which genes from species a correspond to which genes from species b, these methods circumvent difficulties in orthology detection that arise from parallel gene loss and ancient gene duplications (Figure 1). They also avoid issues, such as varying rates of evolution (see below), in single-gene tree phylogenetic reconstruction that are inherent to phylogenomic approaches. Furthermore, alignment-free methods are often computationally cheap, and therefore one advantage is that they may provide a quick reference for obtaining the phylogenetic position of a genome or proteome as soon as it becomes available. Last, these are the only methods that really use all the information contained in the genome: The K-strings method uses the information from all protein coding genes, and the algorithmic compression uses the complete DNA sequence. In contrast, homology-based methods use information only from genes that have homologs in other species.

The fact that these alignment-free methods do not incorporate so much standard molecular evolutionary methodology and proven powerful evolutionary concepts raises interesting questions, especially because they perform reasonably well. Why, for example, does the K-string complement of a proteome yield a tree that is similar to sequence-based trees? Is homology let in through the back-door, in the form of well-conserved (i.e., identical) parts of proteins? In any case, further investigation is needed to establish which molecular evolutionary processes enable these methods to perform so well.

**Genome Trees Based on Shared Gene Content**

A natural and convenient way to describe and analyze complete genomes is by their gene repertoire (36). Comparing genomes on the basis of the fraction of genes they share was one of the first comparative genomics activities to be developed with the availability of complete genome sequences (40). This world-view of genomes as bags of genes has allowed for many successful functional/evolutionary analyses, such as differential genomics (35) or phylogenetic profiles, to predict protein function (36, 54).

Genome trees based on gene content were the first type of genome trees of complete, organismal genomes that were published (23, 62, 67). Gene content trees show reasonable correspondence to the known species tree. This might seem trivial given that organisms inherit their genes mostly from their parents, but the concept of gene content trees has been questioned by publications that report “massive” HGT (also called lateral gene transfer), e.g., between archaea and hyperthermophilic bacteria (48). Doolittle (17) argued that a unique organismal phylogeny is not conceivable unless organisms are construed as either less or more than the sum of their genes. In other words, a valid phylogeny may be derived from a gene family or from phenotypic characters, but the true map of organismal evolution cannot be represented by a tree. Rather, it should be represented by a network (4). Gene content trees provide a nice point of reference in this discussion because here a genome is simply treated as the sum of its genes. That we can represent genomes in a tree is of course not an argument that genome evolution is tree-like, as any feature map can be clustered and turned into a tree. That shared gene content between genomes correlates well with evolutionary distance and that a gene content phylogeny is similar to a sequence-based phylogeny is, however, a strong argument that genome evolution is predominantly tree-like.

As the sharing of genes is such a straightforward and logical approach to compare genomes, many different methods to make gene content trees have been introduced. The first difference between the methods is the use of orthology (29, 34, 41, 62, 73) versus homology (23, 67, 74) (Figure 1). Orthology...
is a more fine-grained definition of the sharing of a gene and therefore arguably yields better trees. Today, the use of orthologs is favored over the use of homologs. In the absence of a more sophisticated method based on an explicit model of genome evolution, the tree reconstruction methods for the first gene content trees were distance based (mostly neighbor joining). Now that some consensus on the nature of genome evolution is emerging, more complicated tree reconstruction algorithms have been introduced, such as Dollo parsimony or maximum likelihood distances (29, 34, 73).

A major problem for gene-content-based trees is that in absolute terms large genomes of intermediate evolutionary distance, such as *E. coli* and *Bacillus subtilis*, share more genes than large genomes do with their more closely related but smaller cousins, such as *E. coli* with *Buchnera aphidicola* or *B. subtilis* with *Mycoplasma genitalium*. In fact this genome size effect is one of the strongest signals in shared gene content and thus deserves special attention when developing a distance measure (62, 74). The number of genes that each eubacterium shares with a specific archaeum, such as *Sulfolobus solfataricus*, has a positive relation with very little spread (Figure 2).

Most importantly, the number of shared genes saturates: For small bacterial genomes, their genome size is limiting for the number of shared genes, hence the rise, whereas for bigger genomes the archaeal genome size becomes limiting, hence the plateau. Thus, one way of correcting for this effect is to divide the number of shared genes by the number of genes in the smaller genome, the latter representing the maximum number of genes the two genomes can share. Not properly taking into account the genome size can result in gene content trees that reflect the phylogeny to a lesser extent, as they cluster, for example, small genomes together and the large

Figure 2
The genome size effect. The number of COGs shared between 45 bacterial genomes with three selected archaeal species: *Nanoarchaeum equitans* (323 COGs), *Methanothermobacterthermoautotrophicus* (1127 COGs), and *Sulfolobus solfataricus* (1421 COGs). Genes were assigned to orthologous groups as defined by the COG database (65). Note that there is a saturation in the number of shared COGs at larger bacterial genome sizes.
Another way of handling the genome size effect is to simply leave out the small genomes (23). The genome size effect is intertwined with parallel gene loss, which is a major problem for gene content trees. The gene losses happen independently as well as in a coordinated fashion similar to the loss of many biosynthetic pathways in microbial organisms with a parasitic lifestyle, such as *B. aphidicola* or the mollicutes. This leads to a strong convergent signal, and although distance-based methods have developed tools to manage this, it remains to be seen how well, for example, Dollo parsimony handles it. The application of Dollo parsimony by Wolf et al. (73), and by Snel et al. (unpublished results), clusters the small genomes together. The potential of simpler methods to cope better with the issue of genome size echoes a recent advance in conventional molecular phylogenetics. In the face of highly unequal rates of sequence evolution, parsimony outperforms a more complicated method such as maximum likelihood (39).

**Genome Trees Based on Gene Order**

Gene order, like gene content, correlates fairly well with evolutionary distance, although it does evolve faster (36, 37). For example, *E. coli* and *Haemophilus influenzae* share 78% of their genes, while their gene order is conserved for only 36% (37). As gene order evolves faster than gene content, it is in principle more suited for closely related species and should achieve a higher resolution at close distances. In addition, the rate at which gene order (synteny) evolves varies between taxa. Eukaryotic chromosomal gene order, for example, evolves much faster than prokaryotic gene order (37). Within the prokaryotes, the mollicutes appear to have a relatively well-conserved gene order, possibly because they lack the chromosome rearrangement gene recG (64).

Gene order has been applied only sparsely for the reconstruction of genome trees of microbial genomes (Figure 1). Apart from the high rate of genome rearrangements, which leads to a lack of resolution at large evolutionary distances, this is also due to the fact that, for a large part, gene order depends on gene content. To make a gene order tree, one needs a large-scale definition of orthology. In fact, the two publications on prokaryotic gene order genome trees of which we are aware also contain gene content genome trees (41, 73). In the first publication, the tree based on conserved gene pairs is constructed with Dollo parsimony, but unlike gene content trees based on parsimony, the small genomes in this tree cluster with their big relatives (73). Because the rate of gene order evolution is so much higher than the rate of gene content evolution, there may be many more shared-derived features in the form of lineage-specific gene pairs than there are lineage-specific genes. In the other effort, the gene order and gene content trees were similar to each other (41). The gene order tree showed some improbable higher order affiliations, reflecting a lack of resolution for these longer evolutionary distances in which too many gene rearrangements have occurred. The gene content tree behaved normal for these distances.

In contrast to microbial genomes, gene order has been successfully applied to eukaryotic mitochondrial genomes and specifically to metazoan mitochondrial genomes (6, 7, 59). In fact, trees based on mitochondrial gene order are arguably the first kind of genome trees, predating gene content trees of completely sequenced organismal genomes (59). In this area, real algorithmic progress has been made such as the formal definition of rearrangement distance based on inversions, translocations, and transversions (6). The dense sampling has also revealed cases of extreme rate variation in mitochondrial gene order evolution, such as the accelerated rate of mitochondrial gene order evolution in Echinodermata in contrast to the near stasis of Vertebrata and Hemichordata (12). Moreover, mitochondrial gene order is one of the few areas in which genome
trees are specifically employed by genuine taxonomists to achieve a better picture of the phylogeny of certain species (7).

**Genome Trees Based on Average Sequence Similarity**

The approaches reviewed above do not use sequence information other than for the definition of orthologs. This knowledge is subsequently used to determine the number of shared genes or the extent of gene order conservation, from which a similarity measure is deduced. At the complete opposite of these approaches lies a class of methods sometimes called blastology. Here, a distance matrix is calculated on the basis of the average sequence similarity between genomes or proteomes, explicitly neglecting any knowledge of orthology (Figure 1).

Henz et al. (31) take the most basic approach imaginable. They make BLAST comparisons at the DNA level of 91 complete prokaryotic genomes and use the resulting heat shock proteins to compose a distance matrix (31). This approach then uses the average sequence similarity between two entire genomes as a similarity measure, making no distinction between coding and noncoding regions, although in prokaryotes most heat shock proteins can be expected to fall within the coding regions. A comparable method was introduced earlier by Grishin et al. (28), who used only the coding sequences. Rather than comparing the entire genomic DNA, the authors compare 19 complete proteomes using BLAST (28). They constructed a tree on the basis of the interprotein amino acid substitution rate distribution of all proteins with sufficient similarity (ε < 0.01 in their dataset). Another approach based on complete proteomes was presented by Clarke et al. (13). They built a tree on the basis of the mean normalized BLAST scores for 37 species. Significant hits were normalized by dividing the e-value by the operon reading frame's self-matching score, the average normalized score defining the distance between two species (13). In the genome tree compilation of Wolf et al. (73), the median percent identity of bidirectional best hits between two genomes is used as a similarity measure. The sequence similarities between genomes are transformed logarithmically to obtain a distance matrix, and subsequently, neighbor joining is used to build a tree (73).

Although these methods are straightforward to implement, and although they can be seen as an interesting intermediate between gene-content-based approaches and purely sequence-based approaches, the compilation of genome trees based on average sequence similarity has never had much follow-up. Researchers are reluctant to adopt the method because the approaches appear to combine the problems present in trees based on gene content as well as in trees based on sequence. By using the extra layer of information provided by the orthology assignment, researchers who implement gene content approaches can avoid some of the pitfalls present in naive sequence analysis, such as convergence in nucleotide usage and codon usage. Phylogenomics approaches, on the other hand, use the sequence information in a phylogenetically superior way: They use proper multiple-sequence alignment rather than simply averaging BLAST scores. Furthermore, the average sequence similarity approaches do not allow inclusion of any evolutionary model and prohibit the construction of trees that use maximum parsimony or likelihood methods that could add much value to approaches based on sequence comparison. Comparing homologous genes rather than orthologous genes, as is done in these methods, basically means introducing noise. Optimally, a filter should be applied to reduce the impact of nonorthologous homologs. In fact, the tree from Wolf et al. (73) indeed uses only similarities between bidirectional best hits to include only orthologs, and this improves the topology. In contrast to other average sequence similarity genome trees (28, 31), they successfully retrieve the proteobacteria as a monophyletic clade (73).
Meanwhile, articles that present new genomes often present a list of species for which such a new genome has a large fraction of its best BLAST hits. This practice, which is related to the tree-building method outlined above, provides a fast indication of the taxonomic neighbors of a species.

### Genome Trees Based on Gene Trees: Phylogenomics, Supertrees, and Concatenated Sequences

Because we cannot use traditional sequence alignment tools to compare the sequences of complete genomes, it is a logical step to at least use traditional sequence alignment tools where possible (Figure 1). The advantage is that we can use the entire toolbox of sophisticated phylogenetic reconstruction methods. One approach is to make trees of gene families that are represented in the genomes of interest. The first effort in this direction dates back to 1999, and investigators immediately ran into the issue that trees from different genes have a different topology (66).

To overcome such incongruent gene trees, one can simply concatenate the homologous sequences from the different gene families, as a concatenated alignment automatically yields a single tree (9). This method has had some success but faces difficulty for evolutionary divergent organisms: The concatenated genes not only have to be present in all genomes compared, they should also have a single copy in each genome to make sure that they are indeed orthologous to one another. With the increasing number of sequenced genomes, the number of genes present with exactly one copy in all organisms shrinks dramatically. In closely related species that share many genes this method has been applied successfully, e.g., in a phylogenomic study on the γ-proteobacteria (43). Rather than (or at the same time as) making a concatenated alignment and escaping the issue of what to do with all these different phylogenies, one can also compare them and obtain some consensus, for example, by using approaches comparable to bootstrapping in single-gene trees (43, 58). The advantage of calculating individual gene trees is that one can separate trees that are relatively different from one another, for example, because of unrecognized paralogy or HGT (9). One can even use the orthologous groups that have not been filtered out to construct a new concatenated alignment. There appears to be no straightforward answer to the question whether it is better to concatenate sequences or to integrate individual trees (Figure 3), but the differences can be striking. For example, we find that the concatenated alignment yields the same topology in neighbor joining as in maximum likelihood, while this is not true for the consensus of the phylome (Figure 3). The choice of integration can thus be more influential than the choice of precise phylogenetic method, even with methods as different as neighbor joining and maximum likelihood. On the one hand, concatenation prevents each gene family from being treated with parameters that are specific for that family. The issue with individual gene trees, on the other hand, seems to be that we do not know how to integrate them nicely, other than using a strict consensus. However, as noted, in practice both methods are applied, often in comparison to the same dataset.

The main limitation to the above methods, however, remains that they require one gene per genome per gene family (1:1 family). By relaxing this criterion, one can in principle obtain much more information from the genomes and their phylogenetic position; however, one must implement methods that compare trees with different numbers of species. One such method is the supertree method (14). Although this method still requires a gene family to be present not more than once in each genome (to assure unambiguous orthologous relations), it eases the demand that a gene family should be present in every genome. To handle the different species compositions of the various trees, the authors created a new alignment of co-occurrence of species in all the partitions of each tree. From this new alignment,
a distance matrix is created that is then fed into the neighbor-joining algorithm. This final step may be open to improvement because it seems ad hoc, like many of the gene content genome tree methods. Nevertheless, the resulting tree is of excellent quality (all established prokaryotic taxa such as the Euryarchaeae or the Proteobacteria are monophyletic), possibly because of the aforementioned increase in the amount of data on which it is based.

Leaving out all restrictions on the species distribution of homologs altogether, one can simply create the phylogenies of all the genes from one genome, the phylome (60). Many insights other than purely phylogenetic ones...
can be gained from these collections of trees. One can reconstruct the metabolism of the ancestor of the mitochondria (25) or predict functional relations between genes (57). Nevertheless, these massive phylomes have not yet been integrated into a single hypothesis on the phylogenetic relationships between all genomes.

NEW DEVELOPMENTS IN THE CONSTRUCTION OF GENOME TREES

Filtering for Inconsistent Signals

HGT as well as ancient gene duplications followed by gene loss (unrecognized paralogy) lead to gene phylogenies that are inconsistent with the species phylogeny. For each class of genome trees, methods have therefore been developed to filter out genes with inconsistent histories. Phylogenomic methods have in fact almost without exception applied such filters (9). As it turns out, the incongruence in gene trees rarely has biological reasons like the above but derives mainly from varying rates of evolution and incorrect alignments (15). Nevertheless, in sequence-based trees the filtering is generally reported to make a substantial improvement (9).

Filtering for inconsistent signals has not been so often applied for other types of genome trees. It has been applied once for gene content trees and once for average protein similarity trees (13, 19). Methodologically it is more difficult to define inconsistent signals for both of these types of genomic information than it is for phylogenomics methods. In both approaches the improvements in the quality of the trees were minor. In fact it seems, at least for gene content, that the phyletic distribution of virtually each gene contains at least some phylogenetic information (19). One explanation for this observation is that genes that are horizontally transferred still behave phylogenetically concordant before and after the transfer event.

Modeling Genome Evolution

Phylogenies can in principle be improved by using more information than just the pairwise distances between genes or genomes as is done in clustering methods such as neighbor joining. Not only do maximum likelihood methods include information about the rates of various processes during gene evolution, but more importantly they explicitly calculate the probability that a certain sequence alignment is produced by a specific phylogenetic tree and a specific model of evolution (21). Such an explicit model of evolution can include the rates at which certain point mutations occur. The tree that is most likely to have produced the alignment then, given the model of sequence evolution, is the maximum likelihood tree. In practice, experimentally generated, known phylogenies of bacteriophage T7
have been better reconstructed by maximum likelihood methods than by neighbor joining (13a).

Recently, approaches that incorporate more explicit models of genome evolution have been applied to genome trees. For example, the simulation of gene content evolution in artificial genomes seems to yield a reliable maximum likelihood distance for the reconstruction of gene content trees (29, 34). Such simulations also suggest that a more explicit description of gene content evolution in the form of Dollo parsimony should show excellent performance. Nevertheless, the only implementation so far of Dollo parsimony on real genome data results in a tree that suffers from clustering of unrelated small genomes and paraphyly of established clades such as the γ-proteobacteria (73). Similarly, the application of maximum likelihood on actual genome data results in clustering of small genomes (29).

EVOLUTIONARY INSIGHTS FROM GENOME TREES

New Phylogenetic/Taxonomic Findings

A few new phylogenetic affiliations have been uncovered or were partly resolved with genome trees. One feature almost unanimously supported by genome trees is that Firmicutes (low G + C Gram-positives) and Actinomycetes (high G + C Gram-positives) are polyphyletic. According to the current NCBI taxonomy, these two groups of Gram-positive bacteria are paraphyletic at the root of the Eubacteria, whereas previously they were grouped together in one taxon, based on the 16S rRNA phylogeny. The alternative grouping is now accepted, but mainly because of other discoveries (1). Genome-scale phylogenetic analyses have declared a close evolutionary relationship for various methanobacteria (61, 62), which until now were considered polyphyletic on the basis of 16S rRNA analysis (51). The relegation of Fusobacterium from a separate bacterial division to a member of the Firmicutes is one of the conclusions of genome-scale analysis, which are a part of the original publication of its genome sequence (38, 45). A more hypothetical theme, which nonetheless recurs in many gene content papers from our group and others, is that the hyperthermophilic eubacteria are not primitive; rather Aquifex seems to be affiliated with the proteobacteria and Thermotoga with the Firmicutes (19, 41, 56, 73). The special link of gene order with mitochondrial genome trees is reflected in the contribution of mitochondrial gene order trees to the resolution of the four basal Arthropod lineages (7).

A Limited Role of Horizontal Gene Transfer in Microbial Evolution

One cannot construct genome trees without wondering how we can produce a tree in the presence of HGT. Before the availability of complete genome sequences, HGT was not attributed a quantitatively major role in evolution. The sequencing of complete genomes has drastically changed this view. Publications of genome sequences, studies that specifically targeted HGT occurrence in published genomes and phylogenomic investigations, report massive levels of HGT (48, 50, 66). Although genome trees are not direct assays of the frequency of HGT, papers describing genome trees are almost unanimous in their surprise at how well their tree based on any given genomic property matches the known species phylogeny. Most genome tree papers therefore report HGT as being quantitatively of small importance (14, 62). Part of the argument is semantic: Can we call an alien origin of 12% of the genes in E. coli “massive” (50)? Yet, the frequency of HGT has even been argued to preclude the existence of a species tree, making the contradictory statements more than simply different perspectives on the same data (17).

To resolve this discussion, the investigations into genome evolution have moved
beyond studies that only targeted HGT or studies that only constructed genome trees. First, we can try to estimate the occurrence of processes that affect gene content (such as gene loss, gene duplication, HGT, and the appearance of new gene families), on the basis of the distribution of current gene families in a reliable species phylogeny. Such analyses are challenging because patchy phyletic distributions of genes (in which a gene is sparsely distributed over a taxon) can be explained by HGT as well as by differential gene loss. Most approaches solve this by introducing a cost in terms of gene losses for each HGT event. When comparing different possible explanations for a patchy phyletic pattern, the cost of a HGT event is weighed against the cost of the of gene losses that are no longer necessary if we introduce this HGT event. Such methods allow a broad scope on genome evolution and they reveal substantial continuity in genome evolution: On any given branch, most genes are transmitted vertically even when using low costs for HGT (11, 42, 46, 63).

The second approach studies the extent of aberrant sequence evolution of single genes due to HGT. In a seminal paper by Daubin et al. (15) a comprehensive collection of quartets of unambiguous orthologous genes were tested for their support of the species phylogeny as defined by rRNA. The results showed that few (sometimes even zero) quartets support the two other possible trees in the case of four sequences, implying a limited role for HGT. Interestingly, they also showed that the quartet alignments that do support HGT have long terminal branches compared with the internal branches, suggesting that these might still be the result of errors in tree reconstruction (15). It has been argued that the low level of HGT found in this study, and in a subsequent, more detailed report (43), is the result of using only unambiguously orthologous genes, i.e., no paralogy whatsoever (75). An independent and equally impressive effort applied a similar question to the more loosely defined COGs, which do contain many paralogs (49). Here, the sequence similarities of individual genes were compared with the average sequence similarity between two genomes. Most (70%) of the genes did not show aberrant levels of sequence similarity potentially caused by HGT. Explicit phylogenetic analysis of the remaining 30% indicated that only half of these could be due to HGT, while the other half was due to lineage-specific acceleration of evolution (49).

From both broad-scope views of genome evolution it can be concluded that by far most of the genes, and thus the genome, have evolved by normal vertical transmission. These explicit studies of the evolutionary dynamics of genome evolution have brought together at least some researchers with opposing world views: Gene content tree builders explicitly acknowledge the need for HGT to explain present-day genomic gene repertoires (15, 63) and to filter out its effect (19), while HGT hunters discovered that other tree-like processes such as vertical inheritance and gene invention are, at least quantitatively, more important for genome dynamics than HGT is (42, 49). Disregarding the proposal that the similarity between gene content trees and rRNA trees results from the HGT of the rRNA molecule (26), the outlines of a consensus on the nature of microbial genome evolution thus seem to be emerging from the literature: a quantitatively modest, but qualititatively important role for HGT and a large role for tree-like processes such as gene loss.

CONCLUSIONS

Challenges Ahead: Data and Computation

Apart from large amounts of relatively clean sequence data in the form of complete genome sequences, the fact that DNA sequencing has become much easier has paradoxically also led to a dramatic increase in noisy data. One important development has been the emergence of metagenomics and environmental
sequencing. In these techniques pieces of DNA are sequenced from uncultured samples, such as a drop of ocean water (70) or a sample of liquid from an acid mine (68). The results of such studies are not complete genome data, but they do contain an invaluable amount of phylogenetic information that needs to be classified without using rRNA, because it is unknown which sequenced reads belong with which rRNA. In the first instance, such a sequence read equals the genome and the species. Supertree and other genome tree approaches can thus provide a phylogenetic framework for the sequences in the absence of rRNA.

Another source of growing amounts of noisy sequence data are incomplete genomes. These data are generated for prokaryotic genomes because they provide an easy and cheap method to answer certain microbial questions (53). At the same time, semicomplete eukaryotic genomes are emerging because even with the current relative ease of sequencing many model species [e.g., Gallus gallus (chicken), Fugu rubripes] cannot receive the intense attention that was put into, for example, the human genome project (2, 32). These data are problematic for genome trees based on gene content, because the absence of genes can be explained as easily by not having been sequenced as by a genuine loss from the genome. However, genome tree methods that rely on sequence similarity can still be applied here.

Note that all the metagenomics data and the data from incomplete genomes are deposited in comprehensive sequence databases such as EMBL or GenBank. An interesting study in this light is the effort to build the tree of life from two such databases, namely SwissProt and a subset of GenBank (18). From all these data, a supermatrix is compiled from all groups of phylogenetically informative homologs that are present with a single copy in every genome. This method is a good approach to dealing with these data, although taxonomic labeling of the genes is required, which is exactly what is missing from metagenomics. Nevertheless, this review shows that genome trees can be encompassed in methodologies that integrate ever more data.

Another challenge is computational, especially in light of the speed of DNA sequencing as mentioned above. The computational demands of comparative studies logically always increase faster than the already exponential increase in data. Even given Moore’s law (47), a solution will not simply be found only in faster computers. In addition, the understandable preference for a more phylogenomic approach to genome trees means handling computationally intensive problems such as multiple-sequence alignment and phylogenetic tree reconstruction. Solutions will come in many different shapes and sizes. Some will be algorithmic, such as those already developed for fast and reliable multiple-sequence alignments [e.g., MUSCLE (20)] or maximum likelihood inference of phylogeny [e.g., PHYML (30)]. Another solution is to use the data selectively. A representative and reliable selection of genes or species will be used as references to construct a backbone for the tree of life at higher taxonomic levels. More sequences can be subsequently selected to fill in the details for lower taxonomic levels, which are established from the higher level backbone. Such a procedure is akin to the existing supertree methods that summarize phylogenetic findings from different papers or collections of trees with different species samples (18). As a solution for the all-against-all comparison of sequences, which promises to become a computational nightmare, profile database searches could play an important role. Profile searches do not need to be redone, as their reliability does not depend on database size and profiles also automatically give a reliable (profile anchored) multiple-sequence alignment.

In contrast to the other methods, alignment-free methods will remain computationally inexpensive. They may thereby provide an independent reference. All these efforts are bioinformatic challenges and they
will bring genome trees and phylogenomics closer to a tight integration with existing sequence databases.

**The Future for Genome Trees is Phylogenomics**

Phylogenomics approaches are popular for good reason. They incorporate the best of both worlds: These approaches use sophisticated existing (and continuously developing) conventional molecular phylogenetic methods for the reconstruction of single-gene trees while they apply the power of numbers that is inherent to genome-scale analysis. As orthologous genes are formally defined by the relation of a gene tree to a species tree, phylogenomics is also part of the solution for defining orthology, which is an important challenge in gene content and gene order genome trees. Phylogenomics could thus incorporate these two other types of genome trees. For example, a gene content genome tree that would be based on an orthology derived from single-gene phylogenies is basically nothing more than an alternative to constructing a supertree from a phylogeny.

The continued application of phylogenomics approaches faces certain hurdles. First, there is the computational hurdle mentioned above. Second, most phylogenomics studies require that the genes are present in all genomes under consideration, but with more genomes this will be the case for fewer and fewer genes. This hurdle has already been overcome in part by the supertree approach (14), but the way in which the trees and their different species compositions are integrated is open to improvement (see above). The remaining obstacle might be the requirement of one ortholog/gene per family per genome. This obstacle is related to the integration of trees with different species compositions. One solution could be sought in a more dynamic definition of orthology from the gene tree itself: using subtrees of a gene tree with paralogs to construct a supertree.

**Summarizing Conclusions**

The main result of genome trees, from trees based on word frequencies to trees based on single-gene phylogenies, is that they are all similar to each other and reflect the known species phylogeny regardless of the various specific genomic properties used or the method used to create the phylogeny. Perhaps we find this coherence because these properties, to various extents, depend on each other. Genome trees have yielded the fundamental insight that genome evolution is largely a matter of vertical transmission. Although the dominance of vertical transmission and thus the quantitatively minor role of HGT became a controversial claim, currently, thanks in part to genome trees, the consensus on microbial genome evolution that emerges is that gene repertoires largely follow phylogeny.

Apart from their contribution to a consensus view of genome evolution, genome trees have made some impact on the phylogeny per se. As genome trees are in line with the undisputed parts of the tree of life, they can also be treated as a line of evidence for phylogenetic relationships in inconclusive parts of the single-gene-based species phylogeny. Their impact so far has been on ancient branching points between higher level taxa, such as the position of the Fusobacteria, and the divergence between the high and low G + C Gram-positives. So far, these contributions have been infrequent, as might be expected given the nascent state of this line of research. Yet, the relative branching orders in the bacterial as well as the eukaryotic divisions, and thus relationships between many higher level taxa, still need resolution. Here, the quick succession of divergence events has made resolution difficult, and until now the order of these events has not been resolved with single-gene phylogenies. In principle, genome trees are in a position to contribute to resolving these points. They are the means to use the maximum amount of data available to solve these tough problems and to
thereby solidify the backbone of the bacterial phylogeny.

A number of hurdles remain on the path ahead. Finding solutions to these hurdles will remain a challenge to our creativity as the recognition of the added value of genome phylogenies grows. Thus far, the main contribution has been the recognition that the classic view of genome evolution by vertical inheritance is indeed quantitatively the most important. The promise of more important phylogenetic discoveries will continue to stimulate researchers in this field.

LITERATURE CITED


Contents

Frontispiece
  Georges N. Cohen ................................................................. xiv

Looking Back
  Georges N. Cohen ............................................................... 1

Signaling in the Arbuscular Mycorrhizal Symbiosis
  Maria J. Harrison ........................................................................ 19

Interplay Between DNA Replication and Recombination in
  Prokaryotes
  Kenneth N. Kreuzer ..................................................................... 43

Yersinia Outer Proteins: Role in Modulation of Host Cell Signaling
  Responses and Pathogenesis
  Gloria I. Viboud and James B. Bliska ........................................... 69

Diversity and Evolution of Protein Translocation
  Mechthild Pohlschröder, Enno Hartmann, Nicholas J. Hand, Kieran Dilks,
  and Alex Haddad ........................................................................ 91

Alternative Candida albicans Lifestyles: Growth on Surfaces
  Carol A. Kumamoto and Marcelo D. Vinces ................................... 113

Yeast Evolution and Comparative Genomics
  Gianni Liti and Edward J. Louis ................................................ 135

Biology of Bacteriocyte-Associated Endosymbionts of Plant
  Sap-Sucking Insects
  Paul Baumann ............................................................................. 155

Genome Trees and the Nature of Genome Evolution
  Berend Snel, Martijn A. Huynen, and Bas E. Dutilh ......................... 191

Cellular Functions, Mechanism of Action, and Regulation of FtsH
  Protease
  Koreaki Ito and Yoshinori Akiyama .............................................. 211
Mating in *Candida albicans* and the Search for a Sexual Cycle
*R.J. Bennett and A.D. Johnson* .................................................. 233

Applications of Autofluorescent Proteins for In Situ Studies in Microbial Ecology
*Estibaliz Larraínzar, Fergal O’Gara, and John P. Morrissey* ...................... 257

The Genetics of the Persistent Infection and Demyelinating Disease Caused by Theiler’s Virus
*Michel Brahic, Jean-François Bureau, and Thomas Michiels* .................... 279

Intracellular Compartmentation in Planctomycetes
*John A. Fuerst* .............................................................................. 299

Biogenesis of Inner Membrane Proteins in *Escherichia coli*
*Joen Luirink, Gunnar von Heijne, Edith Houben, and Jan-Willem de Gier* .................................................. 329

Genome-Wide Responses to DNA-Damaging Agents
*Rebecca C. Fry, Thomas J. Begley, and Leona D. Samson* ...................... 357

The Rcs Phosphorelay: A Complex Signal Transduction System
*Nadim Majdalani and Susan Gottesman* ............................................ 379

Translational Regulation of *GCN4* and the General Amino Acid Control of Yeast
*Alan G. Hinnebusch* ....................................................................... 407

Biogenesis, Architecture, and Function of Bacterial Type IV Secretion Systems
*Peter J. Christie, Krishnamohan Atmakuri, Vidhya Krishnamoorthy, Simon Jakubowski, and Eric Cascales* .................................................. 451

Regulation of Bacterial Gene Expression by Riboswitches
*Wade C. Winkler and Ronald R. Breaker* ........................................... 487

Opportunities for Genetic Investigation Afforded by *Acinetobacter baylyi*, A Nutritionally Versatile Bacterial Species that is Highly Competent for Natural Transformation
*David M. Young, Donna Parke, and L. Nicholas Ornston* ...................... 519

The Origins of New Pandemic Viruses: The Acquisition of New Host Ranges by Canine Parvovirus and Influenza A Viruses
*Colin R. Parrish and Yoshibiro Kawaoka* ........................................... 553

Vaccine-Derived Polioviruses and the Endgame Strategy for Global Polio Eradication
*Olen M. Kew, Roland W. Sutter, Esther M. de Gourville, Walter R. Dowdle, and Mark A. Pallansch* .................................................. 587
INDEXES

Subject Index ................................................................. 637
Cumulative Index of Contributing Authors, Volumes 55–59 ...................... 675
Cumulative Index of Chapter Titles, Volumes 55–59 ............................. 678

ERRATA

An online log of corrections to *Annual Review of Microbiology* chapters may be found at http://micro.annualreviews.org/