



## Whole-genome Trees Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels

Jimmy Lin and Mark Gerstein

*Genome Res.* 2000 10: 808-818

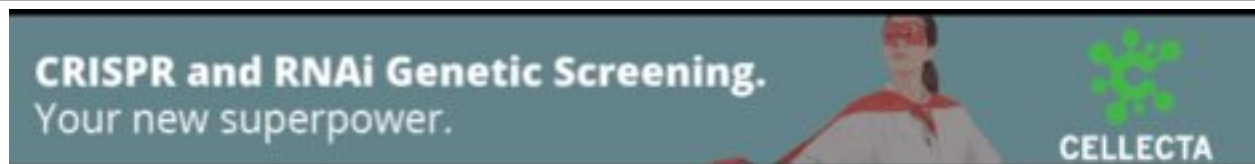
Access the most recent version at doi:[10.1101/gr.10.6.808](https://doi.org/10.1101/gr.10.6.808)

---

**References** This article cites 65 articles, 27 of which can be accessed free at:  
<http://genome.cshlp.org/content/10/6/808.full.html#ref-list-1>

### License

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Cold Spring Harbor Laboratory Press

# Whole-genome Trees Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels

Jimmy Lin<sup>1</sup> and Mark Gerstein<sup>1,2</sup>

*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 USA*

We built whole-genome trees based on the presence or absence of particular molecular features, either orthologs or folds, in the genomes of a number of recently sequenced microorganisms. To put these genomic trees into perspective, we compared them to the traditional ribosomal phylogeny and also to trees based on the sequence similarity of individual orthologous proteins. We found that our genomic trees based on the overall occurrence of orthologs did not agree well with the traditional tree. This discrepancy, however, vanished when one restricted the tree to proteins involved in transcription and translation, not including problematic proteins involved in metabolism. Protein folds unite superficially unrelated sequence families and represent a most fundamental molecular unit described by genomes. We found that our genomic occurrence tree based on folds agreed fairly well with the traditional ribosomal phylogeny. Surprisingly, despite this overall agreement, certain classes of folds, particularly all-beta ones, had a somewhat different phylogenetic distribution. We also compared our occurrence trees to whole-genome clusters based on the composition of amino acids and di-nucleotides. Finally, we analyzed some technical aspects of genomic trees—e.g., comparing parsimony versus distance-based approaches and examining the effects of increasing numbers of organisms. Additional information (e.g. clickable trees) is available from <http://bioinfo.mbb.yale.edu/genome/trees>.

## Introduction: Traditional Single-gene Phylogeny

The sequencing of whole genomes of microbial organisms allows us to reassess how we place organisms into groups and relate them to each other in phylogenetic trees. Traditionally, microorganisms have been grouped together into trees based on the sequence similarity of small subunit ribosomal RNA (SSU rRNA) (Woese et al. 1990; Woese 1987). This approach uses a single important and highly conserved gene, which has complex interactions with many other RNAs and proteins, as a basis of phylogeny. Despite its popularity, there are a number of long-standing difficulties with this approach—for example, long-branch attraction, unresolved tree differences, rate variation among sites, and mutational saturation (Lopez et al. 1999; Doolittle 1999; Lawrence 1999; Jain et al. 1999; Gogarten and Olendzenski 1999). Some researchers have even proposed that rRNA itself can be horizontally transferred between organisms (Nomura 1999; Yap 1999).

Many researchers have also tried building trees based on sequence similarity of individual protein families, such as the cytochromes, ATPases, elongation factors, aminoacyl tRNA synthases, beta-tubulins, or RNA polymerases (Makarova et al. 1999; Teichmann and Mitchison 1999a; Tumbula et al. 1999; Lake et al. 1999; Doolittle 1998; Rivera et al. 1998; Ibba et al.

1999, 1997; Edlind et al. 1996; Baldauf et al. 1996; Brown and Doolittle 1995; Andersson et al. 1998; Tomb et al. 1997; Bult et al. 1996; Lake 1994). These studies have often resulted in a wide range of implied phylogenies. The differences from the accepted ribosomal phylogeny are usually attributed to such factors as horizontal transfer or the existence of ambiguous paralogs. However, sometimes they have been used to argue that the rRNA tree is not representative of the true phylogeny. This has been particularly effective when the protein tree is based on a complex and fundamental protein such as RNA polymerase, or when the protein tree is backed up by extensive natural history evidence (Hirt et al. 1999; Edlind et al. 1996).

## Whole-genome Trees and the Current Controversy

Because SSU rRNA and other individual gene families each correspond to only a tiny fraction of the genomic material in most microorganisms, focusing exclusively on them ignores the bulk of the genetic information in constructing phylogenetic trees. (In particular, the ~1.8 kb of SSU rRNA makes up less than 0.2% of most microbial genomes, which are  $\geq 1$  Mb.) Now with the advent of completely sequenced genomes, it is possible to build trees that encompass much more of the genetic information in an organism. This has led to a profusion of new approaches toward phylogenetic estimation and a heated controversy about the structure of the fundamental tree of life, which has even been featured in the popular press (Pennisi 1998, 1999; Stevens

<sup>2</sup>Corresponding author.  
E-MAIL [Mark.Gerstein@yale.edu](mailto:Mark.Gerstein@yale.edu); FAX (203) 432-5175.

1999). On one extreme, some prefer traditional ribosomal trees. On the other extreme, some argue that trees are not really meaningful given the widespread evidence of horizontal transfer in microorganisms and that nets or more general graph structures should be used instead (Hilario and Gogarten 1993). In between, a third perspective maintains that most microorganisms can be arranged into meaningful trees; however, these trees might not always reflect the branching pattern suggested by rRNA.

To contribute to this debate, we build trees considering progressively more and more of the information in a genome and compare them with the traditionally proposed phylogeny. In particular, we are proposing a number of novel trees based on the occurrence of specific features, either folds or orthologs, throughout the whole genome. We call these *genomic trees* or *whole-genome trees*. Our approach toward genomic trees, which is schematized in Table 1, is similar to the practice in traditional phylogenetic analysis using the presence or absence of morphologic characteristics or heritable traits such as hair or vertebrae to group organisms (Hennig 1965; Maisey 1986).

Our first set of genomic trees is based on the occurrence of orthologous proteins. This work builds upon recent work done by other researchers clustering genomes based on the occurrence of protein families (Tekaiia et al. 1999; Snel et al. 1999; Teichmann and Mitchison 1999a). In particular, Tekaiia et al. (1999) developed a methodology for comparing the whole proteome content of one genome against another and using the loss or acquisition of genes to build trees. We add to this work by dividing the proteome into various classes, building trees based on these, and looking at their consistency.

Our second set of genomic trees is based on protein folds. Folds group together a number of protein

families that may not share sequence similarity but do share the same essential molecular shape. As each protein fold represents a unique three-dimensional shape used by an organism, folds are ideal characteristics for building phylogenetic trees. To build our fold trees, we used a similar approach to that for the ortholog trees, in this case using the presence or absence of folds in particular genomes for tree construction. Our fold tree work is an extension of our previous investigations (Gerstein 1998a). Related work comparing genomes in terms of the occurrence of protein folds has been done by Wolf et al. (1999), who used somewhat different definitions to cluster organisms based on folds.

In a strict sense, our fold occurrence and ortholog occurrence trees are partial proteomes rather than whole-genome trees, because they are based on considering simultaneously a large, but not complete, portion of the protein-coding regions of genomes. The entire genome cannot be used, because not all of the proteins can be classified as part of orthologous groups or can be assigned to fold families. In the last part of our analysis, we look at trees that are based on information from the entire genome, using amino acid and dinucleotide composition. Although these trees represent an unbiased consideration of all the information in the genome, they condense everything to a simple composition vector, discarding much important detail. Our composition tree analysis follows up on our previous work (Gerstein, 1998b) and extensive work done by Karlin and co-workers (Karlin and Burge 1995; Karlin and Mrazek 1997; Campbell et al. 1999). Finally, it is worth pointing out that a number of additional approaches toward whole-genome trees have been advanced beyond those discussed here. In particular, Gupta (1998) used insertions and deletions along with cell-membrane structures for tree reconstruction.

Note that, whereas the occurrence and composi-

**Table 1.** Schematization of the Different Levels

Scale	Levels		Description
Single-gene	Traditional ribosomal	Nucleotide sequence	Based on the small and large subunit ribosomal rRNA
Partial-proteome (multi-gene)	Individual orthologs	Protein sequence	Based on a variety of established orthologous proteins
	Ortholog occurrence	Protein function	Based on the occurrence of specific orthologs in the proteome
Entire genome	Protein fold occurrence	Protein structure	Based on the occurrence of specific protein folds in the proteome
	Amino acid frequency	Proteome composition	Based on frequency of single amino acids in the entire genome
	Dinucleotide frequency	Genome composition	Based on frequency of pairs of adjacent nucleotides in entire genome

This table delineates the three levels on which we compared the microbial genomes. On the single-gene level, both ribosomal RNA and orthologous proteins were used for phylogenetic analysis. These used sequence-based methods for tree construction. The second level uses the presence or absence of orthologous protein groups or protein folds. Because some parts of the genome are assigned neither to a protein fold nor to an ortholog, we call this scale "partial proteome." This scale is based on both protein function and protein structure. The most encompassing scale uses dinucleotide and amino acid composition as a basis for genome comparison. The entire genome is considered.

tion trees have the advantage over the single-gene trees in that they incorporate more genome information, they are not as clearly associated with an evolutionary mechanism. That is, underlying the ribosomal tree there is a specific biologic mechanism, internal to each organism, generating sequence diversity: the mutation of single base pairs, which happens at a rate roughly proportional to time. However, the way a single organism expands or contracts its repertoire of folds or orthologs cannot be explained simply in terms of individual molecular events. If one explains the acquisition of folds with horizontal transfer, the tree is no longer based on ancestral characteristics evolving internally but on the interaction with other organisms. Therefore, as opposed to true evolutionary phylogenies, whole-genome trees would then more appropriately be described as clusterings.

## RESULTS AND DISCUSSION

### Genomes Analyzed and Tree Techniques Used

We focus on the first eight microbial genomes to be sequenced, which include representatives from all three domains of life (Table 2). All our trees were built with the standard programs using both maximum-parsimony and distance-based methods (Felsenstein 1993, 1996; Swofford 1998). In general, we found that distance-based methods gave more reasonable results, probably because of the great divergence of the taxa, which has been remarked upon in other contexts (Swofford et al. 1996). Additional information (clickable trees, plots, etc.) is available on the Internet at <http://bioinfo.mbb.yale.edu/genome/trees>.

### Ortholog and Fold Assignments

Orthologs were selected from the COGs (clusters of orthologous groups) database (Tatusov et al. 1997; Koonin et al. 1998). This database lists groups of ortholog

sequences in eight of the first genomes sequenced based on whether or not they form a mutually consistent sequence of best matches between genomes. This database is in wide use and is accepted as a valid source of functional annotation. It groups orthologs into a hierarchy of functional classes—(e.g. class J, transcription, is part of the “Information Storage and Processing” superclass). The presence or absence of specific proteins for all the COGs for different genomes can be derived from the web site data files.

Folds were assigned to the genome sequences based on a previously described approach (Gerstein 1997, 1998b; Teichmann and Mitchison 1999b; Hegyi and Gerstein 1999). We compared the structure databank (the PDB) against the genome sequences by using both pairwise and multiple-sequence methods and standard thresholds (FASTA and PSI-BLAST, Lipman and Pearson 1985; Pearson 1996; Altschul et al. 1997). We used the SCOP classification to group the domain-level structure matches into different fold families (Murzin et al. 1995). The SCOP (structural classification of proteins) classification is assembled based on expert manual judgement, and we have augmented it with our automatically derived protein-structural alignments (Gerstein and Levitt 1998). Like the COG scheme, the SCOP classification is in wide use and accepted as a reliable classification of a protein's fold.

### Single-gene Trees, a Reference Point

#### *Traditional Ribosomal RNA Trees*

As a reference point, we started our survey by constructing a traditional phylogenetic tree based on the small subunit ribosomal RNA. This established a basis of comparison for the trees generated in this study. The ribosomal tree in Figure 1A resulted in three general clusters corresponding to Archaea, Bacteria, and Eukaryota. The construction of a tree based on the large

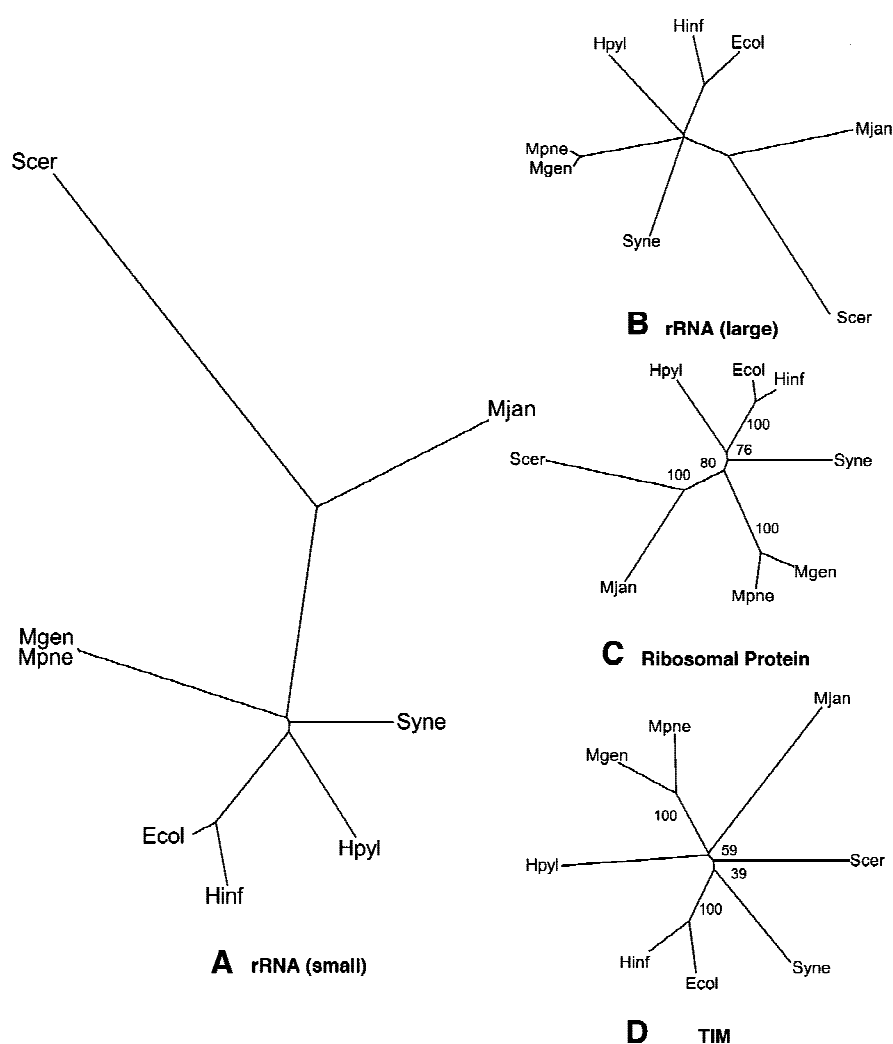
**Table 2.** The Eight Completely Sequenced Organisms Used in this Study

Abbrev.	Organism	Phylogeny			Size (Mb)	Proteins	Reference
Ecol	<i>Escherichia coli</i>	Bacteria	Proteobacteria	gamma subdivision	4.653	4283	Blattner et al. 1997
Hinf	<i>Haemophilus influenzae</i>	Bacteria	Proteobacteria	gamma subdivision	1.830	1703	Fleischmann et al. 1995
Hpyl	<i>Helicobacter pylori</i>	Bacteria	Proteobacteria	epsilon subdivision	1.667	1566	Tomb et al. 1997
Mjan	<i>Methanococcus jannaschii</i>	Archaea	Euryarchaeota	Methanococcales	1.739	1736	Bult et al. 1996
Mgen	<i>Mycoplasma genitalium</i>	Bacteria	Firmicutes	Bacillus/Clostridium	.58	468	Fraser et al. 1995
Mpne	<i>Mycoplasma pneumoniae</i>	Bacteria	Firmicutes	Bacillus/Clostridium	.816	677	Himmelreich et al. 1996
Scer	<i>Saccharomyces cerevisiae</i>	Eukaryota	Fungi	Ascomycota	12.068	5932	Goffeau et al. 1997
Syne	<i>Synechocystis sp.</i>	Bacteria	Cyanobacteria	Chroococcales	3.573	3168	Kaneko et al. 1996

This table lists the currently published microbial genomes, discussed in the text, from the TIGR web site (<http://www.tigr.org/tdb/mdb/mdb.html>). The first column lists the abbreviations that are used for the figures in this paper, corresponding to the genome name in the second column. Column three shows the top three levels of the phylogenetic lineage of the organisms as shown in the NCBI Taxonomy Browser (<http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html>), which is sufficient to locate the taxa on the traditional tree. The size of the complete genome of the particular organism is shown in the fourth column with the total number of proteins in the organism listed on the next column. The final column lists the original publication citation for each of the genomes.

## Whole-genome Trees Based on Folds and Orthologs

**Figure 1** Representative single-gene trees. (A) The traditional small subunit ribosomal phylogenetic tree. This is a tree of eight completely sequenced representative organisms constructed with the SSU rRNA. Trees could be constructed using data from two different sources: the Ribosomal Database Project (RDP, <http://www.cme.msu.edu/RDP>, Maidak et al. 1999) and the rRNA WWW Server (<http://www-rrna.uia.ac.be>, Van de Peer et al. 1999). Although a tree can be abstracted from the RDP, the tree cannot contain both prokaryotes and eukaryotes. Instead, we took sequences from the RDP and the rRNA WWW server and aligned them with Clustal (Thompson et al. 1997). Phy- lip and PAUP were used to construct trees from the aligned sequences using distance and parsimony methods. There was little variation in the resulting trees displayed using TreeView (Page 1996), which was used to show all the trees used in this survey. The PAUP distance-based tree is shown. (B) The large subunit ribosomal tree. Another common method of building phylogenetic trees is the use of the large subunit rRNA (De Rijk et al. 1999). Because of the lack of large subunit rRNA information from the RDP, the sequences were downloaded from the rRNA WWW Server. The same method of tree construction was used as in A. The tree shown in B is the PAUP distance-based tree. Because of the large divergence of the species, the topology of the tree varied slightly when compared to the SSU ribosomal tree in A. The placement of *Synechocystis* was slightly different, as it is placed closer to the eukaryote and Archaea in the large subunit tree. This was relatively less significant when considering the branch lengths of the tree in A. (C, D) Representative trees based on sequence similarity of orthologs. The sequences of proteins for the different organisms were obtained from the COGs web site (<http://www.ncbi.nlm.nih.gov/COG>, Tatusov et al. 1999). Clusters of orthologous groups were chosen that had one protein for each organism in the group. There were eight such COGs with representatives from four different classes. Distance-based trees and parsimony trees were both constructed for each of the orthologous groups. There was great variation in the resulting trees. The tree, which had the highest similarity to the traditional ribosomal tree, is shown in C. In fact, the distance-based tree based on the 30S ribosomal protein S3 (COG92, Class J) in C is exactly the same in topology to the traditional tree. This is not surprising because we expect a ribosomal protein tree to be similar to ribosomal rRNA trees because of their interaction and conservation. For the bootstrap values, all bootstrap replicates grouped *E. coli* with *H. influenzae*, *S. cerevisiae* with *M. jannaschii*, and *M. genitalium* with *M. pneumoniae*. In all, the conserved topology coupled with high bootstrap values shows that phylogenetic trees with even a single protein can exhibit very high fidelity to the traditional ribosomal tree. Besides trees with high similarity to the traditional tree as in C, there were trees that varied significantly from the traditional ribosomal tree. Part D shows a distance-based tree based on the metabolic enzyme triosephosphate isomerase (TIM). In general, there are a lot of differences between this tree and the traditional tree. *M. jannaschii* is grouped with *M. genitalium* and *M. pneumoniae*; *M. jannaschii* is not grouped with *S. cerevisiae* at all. The connectivity of *S. cerevisiae* and *H. pylori* is also different from the traditional tree. The low bootstrap values of 59% and 40% suggest that within the sequence there is great variation and the tree is generated with lower certainty. In general, there were a wide variety of trees produced using sequence similarity of orthologous proteins.



subunit ribosomal RNA in Figure 1B showed some variation in the topology of the tree.

#### Single-gene Ortholog Trees

Next we examined the trees based on individual orthologous proteins chosen from the COGs database. Again this was to establish a reference for comparison

and also to see the variation within single-gene trees. We focused on orthologous groups for which each of our eight organisms had only one representative, to minimize the possible effects of unrecognized paralogy (Teichmann and Mitchison 1999b). As shown in Figure 1, as has been remarked on in previous studies, the clusterings exhibit great variation depending on the

protein chosen. Furthermore, many of the trees have only marginal bootstrap values if we consider 95% to be the cutoff for reasonable confidence (Efron et al. 1996). In Figure 1C, we show a representative tree that agrees well with the traditional ribosomal phylogeny. It is based on the 30S ribosomal protein S3 (COG 92, Class J). It has relatively good bootstrap values compared to other single-gene trees that we constructed and compared to the findings of others (Teichmann and Mitchison 1999a), but not all of the values are above 95%. Figure 1D shows an example of a tree that differs significantly from the traditional phylogeny, that of triosephosphate isomerase (TIM, COG 149, Class C). Perhaps predictably, we found that trees that agreed well with the traditional ribosomal phylogeny tended to be based on proteins involved in transcription and translation, especially those with extensive RNA interactions. [This latter observation is also true for some proteins, such as the signal recognition particle (SRP) GTPase, which are not involved in transcription or translation.] In contrast, soluble enzymes, such as TIM, tended to produce trees with greater variation. These discrepancies provide evidence that trees built on sequence-similarity of individual genes would result in different phylogenies, because different genes have different mutational rates and some are horizontally transferred.

### Genomic Occurrence Trees

Now that we have described the single-gene perspective as a reference, we can progress to the focus of our analysis, constructing genomic trees that consider more than the variation of individual genes. These trees were defined in terms of presence or absence of shared characteristics throughout the whole genome. Broadly, these characteristics potentially could be orthologs, homologs, or folds. We focused on orthologs and folds. We used both distance-based and parsimony methods for tree construction. For the distance-based methods, we defined the distance between two organisms with a normalized Hamming distance, which was expressed as the fraction of unshared characters divided by the total number of characters in the genomes—that is  $(A + B - 2S)/(A + B)$ , where  $A$  and  $B$  are the characters in the first and second genomes, respectively, and  $S$  is the number of shared characters between  $A$  and  $B$ .

We had hoped that parsimony would produce reasonable trees, because this method would automatically propose ancestral organisms that had the intermediate configurations of orthologs or folds. However, we found that, in general, distance-based methods resulted in trees closer to the traditional phylogeny. This may be because of the great divergence of the organisms studied. We give an example of the superiority of distance-based methods in Figure 3, which compares

fold trees based on distances and parsimony. Also, because of the divergence of the organisms, a number of our trees may show some evidence of long-branch attraction. This arises when there are differing rates of variation among different sites in a gene, resulting in the clustering of organisms with higher rates of sequence change (Felsenstein 1996). However, we feel long-branch attraction affects the ribosomal tree as much as, if not more than, the whole-genome trees, as evident in its longer relative branch lengths and more star-shaped appearance (i.e., with less well-resolved branching). Furthermore, it is known that distance-based methods are less sensitive to long-branch attraction than parsimony, perhaps suggesting why we found the distance-based trees more reasonable.

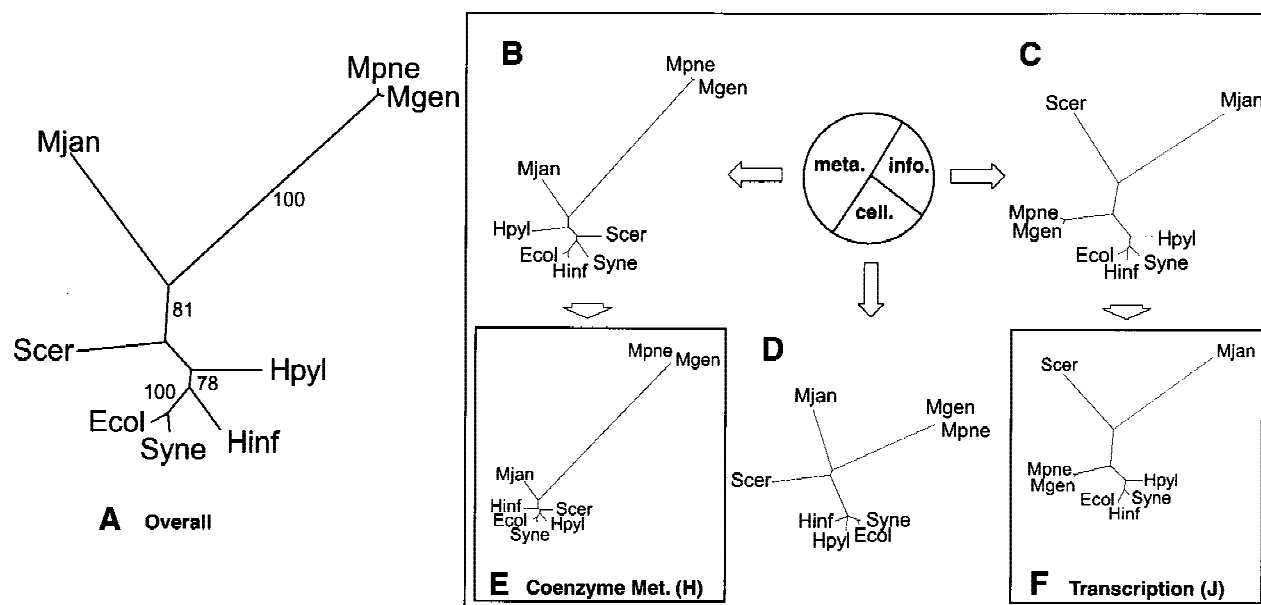
### Genomic Tree Derived from Occurrence of Orthologs

Figure 2 shows the trees built in terms of the overall occurrence of ortholog families in the eight genomes. We used the COGs database to determine whether a genome had a particular orthologous group. The overall clustering based on all the orthologs in the COGs database is notably different from the traditional ribosomal tree. The *Mycoplasma pneumoniae* and *Mycoplasma genitalium* cluster is placed with the *Methanococcus jannaschii*, and the generally conserved grouping of *Escherichia coli* and *Haemophilus influenzae* does not occur.

### Subdivision by Functional Class

The COGs database contains three main functional subdivisions: metabolism, information storage and processing, and cellular processes. For convenience, we will refer to these as the *metabolic*, *information*, and *cellular* subdivisions. More than half of the total ortholog groups and half of the signal in the overall tree come from proteins in the metabolic subset. If we remove the metabolic subset, we get a tree much more consistent with the traditional phylogeny. Figure 2C shows a genomic tree based on the occurrence of orthologs just in the information set. Its topology is almost identical to the traditional tree, the only difference being the placement of *Synechocystis* and *Helicobacter pylori*, which are reversed. This difference is minor as the divergence between these two organisms is very small even in the traditional tree. One sees, furthermore, that the traditional topology is preserved even when one selects a subset of the information set, namely the orthologs involved in transcription (class J) in Figure 2F.

In contrast to the information-subset tree, the metabolic-subset tree in Figure 2B corresponds closely to the topology of the overall ortholog-occurrence tree, dominating it and giving rise to its nontraditional topology. This unusual tree topology is accentuated even more when we look at a subset of the metabolic pro-



**Figure 2** Genomic trees based on the occurrence of orthologs. (A) Distance-based genomic tree based on the overall occurrence of orthologous proteins in the complete genome. One of the alternative methods we used for phylogenetic analysis involved building trees based on the presence or absence of orthologs in the complete genome, using the information from the original COGs web site with eight genomes (<http://www.ncbi.nlm.nih.gov/COG>, Tatusov et al. 1999). For each of the microbial organisms, the occurrence of proteins in each of the clusters of orthologous groups was tabulated with 1 for present and 0 for absent. With the parsed data, a distance matrix was then calculated using the normalized Hamming distance, as described in the text. The trees were subsequently constructed using the kitsch program in the PHYLIP package, which allowed for easy automation. For the bootstrap values, we used PAUP. The resulting tree shown is a distance-based tree using the information of the occurrence of all the COGs in the genomes. As expected, the *M. pneumoniae* and *M. genitalium* are grouped with bootstrap values of 100%. However, interestingly, *E. coli* and *Synechocystis* are also clustered with this bootstrap value—a grouping that is not in the traditional tree. Also, *M. jannaschii* is clustered with *M. pneumoniae* and *M. genitalium* with a bootstrap value of 81%. Furthermore, the eukaryote, *S. cerevisiae*, is placed among the bacteria. (B, C, and D) Ortholog occurrence genomic trees based on a three-way partition of the whole ortholog set. As described in the text, the total COGs were divided into three large subsets, the information, cellular, and metabolic subsets. The pie chart in Figure 3 shows the number of COGs in each group as percentages. The metabolic subset dominated the total group with 362 COGs, approximately half of all the COGs. The information subset has 190 COGs, just above one quarter, while the cellular subset has 132 COGs, just less than a quarter. For each of the subsets, distance-based trees were generated using the same methods described in A. Because of the smaller sizes of these subsets, the bootstrap values were often ill-defined. The largest subset was the metabolic partition shown in B. There was a high correlation between the trees in A and B. Aside from the different placement of *H. pylori* and *S. cerevisiae*, the trees are nearly identical, even having similar branch lengths. The second largest partition was the information subset shown in C. Surprisingly, this subset produced a tree almost identical to the traditional ribosomal tree. The only difference is the switch in the placement of *H. pylori* and *Synechocystis*. This shows that although using the entire group of COGs may produce trees much different from the traditional tree, using a smaller subset may in fact produce a tree that is closer to the traditional topology. Part D shows the smallest partition, the cellular subset. (E and F) Representative genomic trees of ortholog occurrence based on specific functional classes J and H. Using the functional classes obtained from the COGs web site, the metabolic, information, and cellular partitions were subdivided further, into specific functional classes. For each of the different functional classes, there was a range of trees produced. Two representatives were chosen to show this variety. Class J (translation, ribosomal structure, and biogenesis), which has 108 clusters of orthologous proteins, is a further subdivision of the information subset. It has a tree very similar to the traditional ribosomal tree in Figure 1A. Class H (Coenzyme metabolism), which has 77 clusters of orthologous proteins, is a further subdivision of the metabolism subset. It produced a tree that did not correspond well to the traditional phylogeny.

teins corresponding to less essential functions that are less evenly maintained across organisms. This is shown in Figure 2E, which shows a tree with 65 COGs involved in coenzyme metabolism (class H).

The topology of the metabolic subset, in fact, seems skewed by the number of orthologous proteins that each genome in this subset has. The two genomes with the largest number of orthologous proteins in the subset, *E. coli* (350) and *Synechocystis* (330), were grouped together. *Haemophilus influenzae* (264), having the third highest number of clusters of orthologous

metabolic proteins, branches off from this cluster, followed by *S. cerevisiae* (262), *H. pylori* (226), *M. jannaschii* (215), and the mycoplasmas (81 and 75). We tried a variety of alternative distance measures to correct for this effect (e.g., by dividing the number of shared orthologous groups over the number of groups only present in one) but were unsuccessful in getting the traditional topology from the metabolic subset.

Thus, our results suggest that the occurrence of proteins associated with transcription and translation is closer to the traditional rRNA phylogeny than that

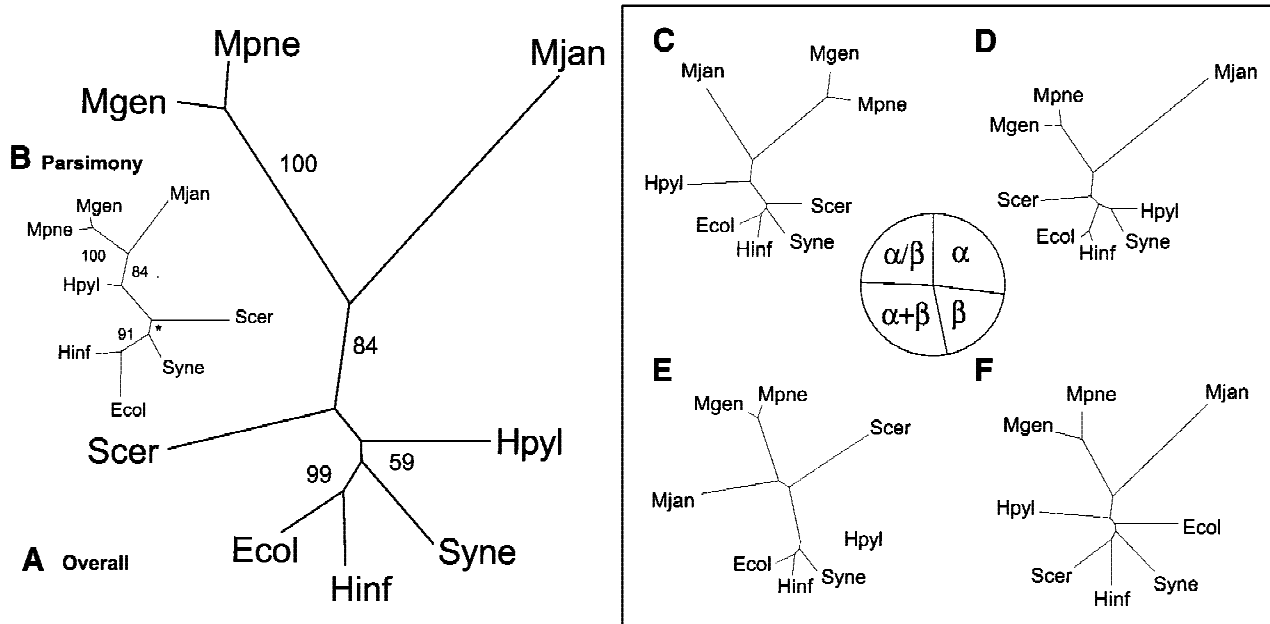
associated with metabolism. These outcomes are reasonable and in consonance with the results that show that trees based on individual proteins involved in metabolism are usually farther from the established phylogeny than those based on proteins involved with transcription (see references above).

### Fold Trees

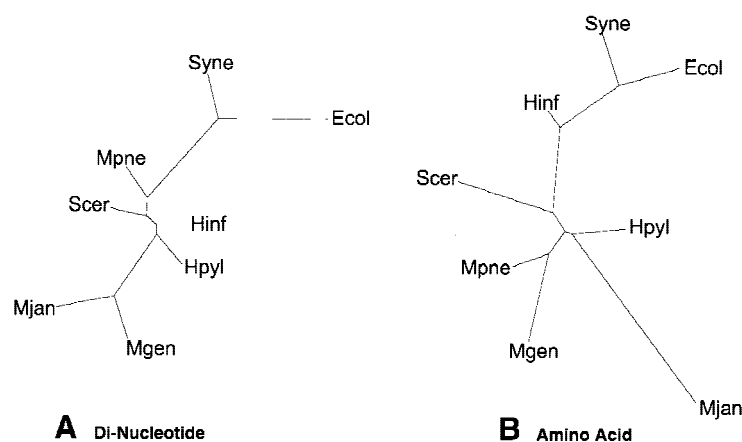
In Figure 3 we show trees based on the occurrence of protein folds throughout the genome. Folds unite protein families that share the same basic architecture but which might not have any appreciable sequence similarity. They provide an ideal type of character to use in the construction of occurrence trees, since it is believed that there are only a very limited number of protein folds (Chothia 1992). The occurrence of a particular fold within the genome represents the organism select-

ing a particular three-dimensional shape from the overall master parts list found in nature. Fold trees have the advantage over ortholog trees in that the assignment of a particular ORF to a fold can be done fairly automatically and objectively, whereas the assignment of ORFs to various orthologous groups is often more ambiguous and requires considerable manual intervention.

Our overall fold tree is shown in Figure 3A. It has a remarkably similar topology to the traditional ribosomal tree, especially when one considers how radically different the criteria are for building the trees. The tree shown in Figure 3A is built with distance-based methods, using a normalized Hamming distance (the number of different folds between genomes as a fraction of their total). As a contrast, in Figure 3B we show a fold tree built on the basis of parsimony. The parsimony tree clearly differs from the distance-based tree.



**Figure 3** Genomic trees based on the occurrence of folds. (A) Genomic tree based on the overall occurrence of folds in the genomes, generated by a distance-based method. For each of the microbial organisms the presence or absence of folds was marked with 1 or 0, respectively. The folds that were not present in any of the genomes were excluded, because this does not provide any distinguishing information. Similar to the ortholog occurrence, a distance matrix was generated with the Hamming distance. (B) Genomic tree based on the overall occurrence of folds in the genomes, generated by parsimony. Instead of generating a distance matrix, parsimony can be used for tree construction. For this task, PAUP was used and the resulting tree is mostly similar to the distance-based tree. However, the locations of *S. cerevisiae* and *H. pylori* are switched. Also, in contrast to the traditional ribosomal tree, *S. cerevisiae* is placed closer to *M. jannaschii*, whereas *H. pylori* is placed with the other bacteria. Therefore, the distance-based method as described in A seems to be better. For all the trees presented here, both distance-based and parsimony trees were generated; in general, as observed in this instance, the distance-based tree is closer to the ribosomal tree. The star shown in the bootstrap value represents a node where the bootstrap consensus tree results in a star decomposition and cannot be resolved. (C, D, E, and F) Distance-based genomic trees based on occurrence of folds in particular fold classes. In this analysis, instead of dividing the COGs into functional classes, the folds are fractionated into classes: all-alpha, all-beta, alpha+beta, and alpha/beta. As seen in the pie chart, the distribution of folds among the different classes is rather equal; each has approximately one quarter of the total. Of the four divisions of folds, the alpha+beta group is most similar to the overall tree, having the exact same topology. It also has the largest number of folds (81, 29% of the total). The all-alpha fold group has 27% (75) of the total folds and has almost the exact topology of the overall tree, except that *H. pylori* and *Synechocystis* are grouped together instead of just being close to each other. The alpha/beta group has 24% (68) of the total folds and is also very similar to the overall fold tree. The most surprising tree is that of the all-beta group. This is based on the smallest number of folds, which is 20% (55) of the total folds.



**Figure 4** Trees based on overall composition. (A) Dinucleotide composition tree. We counted the relative frequency of the dinucleotides for the complete genomes of the eight organisms. Distance between two species of dinucleotides is the distance between the 16-dimensional vectors, with each axis representing a dinucleotide pair. PAUP then generated trees using the distance matrix. Figure 4A shows the resulting dinucleotide composition tree, which has almost no resemblance to the traditional ribosomal tree in Figure 1A. Even the *M. genitalium* and *M. pneumoniae* clustering, which is conserved throughout the survey, does not appear. This suggests that the dinucleotide method is not very accurate in the production of phylogenetic trees. Although it encompasses entire genomes, it reduces them to a 16-dimensional vector, losing much information. (B) Amino acid composition tree. This shows the tree generated from amino acid composition. Again, the relative frequencies of the amino acids were counted and a similar distance measure as in A was used. The distances between the genomes are calculated using 20-dimensional vectors, one for each amino acid. The resulting distance matrices were used to generate trees using PAUP. Interestingly, although this tree is still significantly different from the traditional tree in Figure 1A, it indeed is a great improvement upon the dinucleotide composition tree. Relatively, the organisms are closer in position to the traditional tree.

Although these two trees were still similar, the parsimony tree alternates the position of *H. pylori* and *S. cerevisiae*, placing the latter much closer to the *E. coli*, *H. influenzae*, *Synechocystis* bacterial cluster than to the Archaea, *M. jannaschii*.

#### Subdivision by Fold Class

As with the orthologous protein trees, we analyzed the composition of the total tree through various subdivisions. We can subdivide folds into four major structural classes, all-alpha, all-beta, alpha/beta, and alpha+beta (Levitt and Chothia 1976). One might not expect much variation between the classes, since unlike the orthologous proteins whose subsets had definite functional and evolutionary implications, the fold class subdivisions are not clearly related to any of these aspects. However, the trees are, in fact, different among the fold structural class subdivisions, as seen in Figure 3.

While the alpha+beta and all-alpha subdivisions look most similar to the overall fold tree and the traditional phylogeny, the all-beta class has an unusual clustering. Specifically, *E. coli* is not placed with *H. in-*

*fluenzae*, and *S. cerevisiae* is placed deep within the bacterial cluster. Perhaps this reflects the less even distribution of all-beta proteins and their selective proliferation in various taxa. It has been suggested before, based on bulk structure prediction, that there seems to be a different distribution of all-beta proteins in eukaryotes than prokaryotes (Gerstein 1997).

#### Genome Composition Trees

Because we cannot assign every single ORF in a genome to a known fold or ortholog family, the genomic occurrence trees based on these characters are in a strict sense partial proteome trees. Consequently, they suffer from potential biases because the orthologs or folds that are selected may not represent a truly random sampling of proteins in the genomes. One simple way of building a tree that takes into account all the information within the genome in unbiased fashion is using the overall nucleotide composition. To complete our analysis, we built trees based on dinucleotide and amino acid (essentially trinucleotide) composition. We constructed vectors representing the normalized composition of the genome, denoted by  $\mathbf{f}$ , and then took the difference of these vectors to define our tree. We thus used the following relation for the distance between genomes  $i$  and  $j$ :

where  $\mathbf{f}(i,k)$  represents the composition and the sum over  $k$  runs from 1 to  $M=16$  or  $M=20$ , depending on whether the tree is for

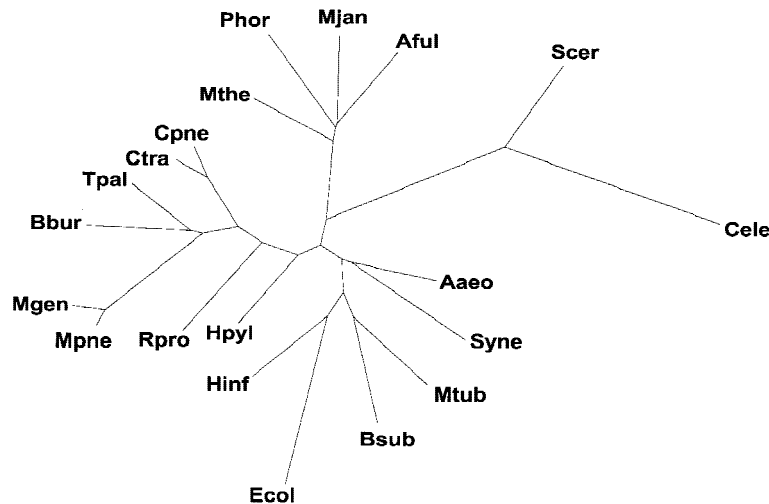
dinucleotide composition or amino acid. Clearly, in the computation of composition, while we are broadly considering the entire genome, we are discarding much information by reducing the entire genome into a single composition vector.

#### Dinucleotide Composition

Dinucleotide composition results in a tree that is very different from the traditional tree. As seen in Figure 4A, the clustering did not show any of the patterns observed in most of the trees in this paper. Even *M. pneumoniae* and *M. genitalium* were not clustered together.

#### Amino Acid Composition

The amino acid composition tree contained great similarity to other trees presented in this survey. The mycoplasmas are clustered together, as are *E. coli*, *Synechocystis*, and *H. influenzae*; *M. jannaschii* is far from the main clusters. As can be seen in Figure 4B, simple amino acid composition, for even the entire sequence, cannot generate anything like the traditional tree; it is noteworthy that by changing from two nucleotides in the dinucleotide analysis to three in the amino acid



**Figure 5** Prospects for the future. The figure shows a 20-genome tree based on the occurrence of folds. This is similar to Figure 3A. The unit in the SCOP classification that was used was the structural superfamily rather than the fold. For eight genome occurrence trees there is no difference between one made at the fold or superfamily level. However, for the 20-genome tree this distinction matters. The additional species names in the 20-genome tree are: Aaao (*Aquifex aeolicus*), Aful (*Archaeoglobus fulgidus*), Bsub (*Bacillus subtilis*), Bbur (*Borrelia burgdorferi*), Cpne (*Chlamydia pneumoniae*), Ctra (*Chlamydia trachomatis*), Ecol (*Escherichia coli*), Hinf (*Haemophilus influenzae*), Hpyl (*Helicobacter pylori*), Mthe (*Methanobacterium thermoautotrophicum*), Mjan (*Methanococcus jannaschii*), Mtub (*Mycobacterium tuberculosis*), Mgen (*Mycoplasma genitalium*), Mpne (*Mycoplasma pneumoniae*), Phor (*Pyrococcus horikoshii*), Rpro (*Rickettsia prowazekii*), Scer (*Saccharomyces cerevisiae*), Syne (*Synechocystis sp.*), and Tpal (*Treponema pallidum*).

$$D(i,j) = |\mathbf{f}(i) - \mathbf{f}(j)| = (I/M) \sum_{k=1,M} (\mathbf{f}(i,k) - \mathbf{f}(j,k))^2,$$

analysis (and taking into consideration reading frames) much more information is revealed.

## CONCLUSION

We built trees grouping organisms based on the overall occurrence of molecular features throughout their genomes. Most broadly, these characteristics could be orthologs, homologs, or folds. We focused on orthologs and folds. For folds we found that the overall genomic tree agreed surprisingly well with the traditional ribosomal tree. However, the distribution of all-beta folds was somewhat different. For orthologs we found that an occurrence tree based just on proteins involved in transcription and translation also agreed quite well with the traditional phylogeny. However, one built based on metabolic proteins had a rather skewed topology, and as the metabolic subset comprised most of orthologs, the overall ortholog tree also shared this appearance. This implies that adding more features does not necessarily increase the accuracy of the tree, an observation that, of course, has been made numerous times in relation to traditional phylogeny (Swofford et al. 1996). We compared our occurrence trees with

many other possible trees, ranging from single-gene trees based on sequence similarity of individual orthologous proteins to entire-genome composition trees. We found that many of these alternate trees had rather unusual topologies, providing a good context for appreciating how remarkable was the agreement between whole-genome occurrence trees, particularly the fold tree, and the traditional tree.

## Prospects: Trees Based on More Genomes

The number of genomes being sequenced is increasing at a fast rate and obviously the eight-genome scale of analysis presented here will be soon out of date. However, the real question is whether our approach of comparing genomes in terms of the occurrence of orthologs or folds will scale with more and more organisms. We believe it will. Recently, we have built whole-genome occurrence trees based on more than eight genomes and found that they had quite a reasonable topology. As an illustration, in Figure 5, we show a fold tree based on 20 genomes.

## ACKNOWLEDGMENTS

We thank H. Hegyi for help with the fold assignments and G. Naylor for discussions on phylogeny. M.G. thanks the NIH and the Donaghy Foundation for support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

## REFERENCES

- Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
- Andersson, S.G., A. Zomorodipour, J.O. Andersson, T. Sicheritz-Ponten, U.C. Alsmark, R.M. Podowski, A.K. Naslund, A.S. Eriksson, H.H. Winkler, and C.G. Kurland. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**: 133–140.
- Baldauf, S.L., J.D. Palmer, and W.F. Doolittle. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* **93**: 7749–7754.
- Blattner, F.R., G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**: 1453–1462.
- Brown, J.R. and W.F. Doolittle. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92**: 2441–2445.
- Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne, et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**: 1058–1073.
- Campbell, A., J. Mrazek, and S. Karlin. 1999. Genome signature

- comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **96**: 9184–9189.
- Chothia, C. 1992. Proteins—1000 families for the molecular biologist. *Nature* **357**: 543–544.
- De Rijk, P., E. Robbrecht, S. de Hoog, A. Caers, Y. Van de Peer, and R. De Wachter. 1999. Database on the structure of large subunit ribosomal RNA. *Nucleic Acids Res.* **27**: 174–178.
- Doolittle, R.F. 1998. Microbial genomes opened up. *Nature* **392**: 339–342.
- Doolittle, W.F. 1999. Phylogenetic classification and the universal tree [see comments]. *Science* **284**: 2124–2129.
- Edlind, T.D., J. Li, G.S. Visvesvara, M.H. Vodkin, G.L. McLaughlin, and S.K. Katiyar. 1996. Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa. *Mol. Phylogenet. Evol.* **5**: 359–367.
- Efron B, E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **96**: 13,429–13,434.
- Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.
- . 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Meth. Enzymol.* **266**: 418–427.
- Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J. Merrick, et al. 1995. Whole-genome random sequencing and assembly of *haemophilus influenzae rd*. *Science* **269**: 496–512.
- Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**: 397–403.
- Gerstein, M. and M. Levitt. 1998. Comprehensive assessment of automatic structural alignment against a manual standard, the Scop classification of proteins. *Protein Sci.* **7**: 445–456.
- Gerstein, M. 1997. A structural census of genomes: Comparing eukaryotic, bacterial and Archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**: 562–576.
- Gerstein, M. 1998a. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des.* **3**: 497–512.
- Gerstein, M. 1998b. Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census. *Proteins* **33**: 518–534.
- Goffeau, A., R. Aert, M.L. Agostini-Carbone, A. Ahmed et al. 1997. The yeast genome directory. *Nature* **387(Suppl)**: 5–105.
- Gogarten, J.P. and L. Olendzenski. 1999. Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9**: 630–636.
- Gupta, R.S. 1998. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.* **62**: 1435–1491.
- Hegyi, H. and M. Gerstein. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**: 147–164.
- Hennig, W. 1965. Phylogenetic systematics. *Ann. Rev. Entomol.* **10**: 97–116.
- Hilario, E. and J.P. Gogarten. 1993. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. *Biosystems* **31**: 111–119.
- Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**: 4420–4449.
- Hirt, R.P., J.M. Logsdon, Jr., B. Healy, M.W. Dorey, W.F. Doolittle, and T.M. Embley. 1999. Microsporidia are related to fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci. USA* **96**: 580–585.
- Ibba, M., H.C. Losey, Y. Kawarabayasi, H. Kikuchi, S. Bunjun, and D. Soll. 1999. Substrate recognition by class I lysyl-tRNA synthetases: A molecular basis for gene displacement. *Proc. Natl. Acad. Sci. USA* **96**: 418–423.
- Ibba, M., S. Morgan, A.W. Curnow, D.R. Pridmore, U.C. Vothknecht, W. Gardner, W. Lin, C.R. Woese, and D. Soll. 1997. A euryarchaeal lysyl-tRNA synthetase: Resemblance to class I synthetases. *Science* **278**: 1119–1122.
- Jain, R., M.C. Rivera, and J.A. Lake. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. USA* **96**: 3801–3806.
- Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirotsawa, M. Sugiura, S. Sasamoto, et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). *DNA Res.* **3**: 185–209.
- Karlin, S. and C. Burge. 1995. Dinucleotide relative abundance extremes: A genomic signature. [Review]. *Trends Genetics* **11**: 283–290.
- Karlin, S. and J. Mrazek. 1997. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* **94**: 10,227–10,232.
- Koonin, E.V., R.L. Tatusov, and M.Y. Galperin. 1998. Beyond complete genomes: From sequence to structure and function. *Curr. Opin. Struct. Biol.* **8**: 355–363.
- Lake, J.A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA* **91**: 1455–1459.
- Lake, J.A., R. Jain, and M.C. Rivera. 1999. Mix and match in the tree of life. *Science* **283**: 2027–2028.
- Lawrence, J.G. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol.* **2**: 519–523.
- Levitt, M. and C. Chothia. 1976. Structural patterns in globular proteins. *Nature* **261**: 552–558.
- Lipman, D.J. and W.R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* **227**: 1435–1441.
- Lopez, P., P. Forterre, and H. Philippe. 1999. The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* **49**: 496–508.
- Maidak, B.L., J.R. Cole, C.T. Parker, Jr., G.M. Garrity, N. Larsen, B. Li, T.G. Lilburn, M.J. McCaughey, G.J. Olsen, R. Overbeek, R., et al. 1999. A new version of the RDP (Ribosomal Database Project). *Nucleic Acids Res.* **27**: 171–173.
- Maisey, J.G. 1986. Heads and tails: A chordate phylogeny. *Cladistics* **2**: 201–256.
- Makarova, K.S., L. Aravind, M.Y. Galperin, N.V. Grishin, R.L. Tatusov, Y.I. Wolf, and E.V. Koonin. 1999. Comparative genomics of the Archaea (Euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* **9**: 608–628.
- Murzin, A., S.E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: A structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536–540.
- Nomura, M. 1999. Engineering of bacterial ribosomes: Replacement of all seven *Escherichia coli* rRNA operons by a single plasmid-encoded operon [see comments]. *Proc. Natl. Acad. Sci. USA* **96**: 1820–1822.
- Page, R. D. 1996. TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**: 357–358.
- Pearson, W.R. 1994. Using the FASTA program to search protein and DNA sequence databases. *Meth. Mol. Biol.* **24**: 307–331.
- . 1996. Effective protein sequence comparison. *Meth. Enzymol.* **266**: 227–258.
- Pennisi, E. 1998. Genome data shake tree of life [news]. *Science* **280**: 672–674.
- . 1999. Is it time to uproot the tree of life? [news]. *Science* **284**: 1305–1307.
- Rivera, M.C., R. Jain, J.E. Moore, and J.A. Lake. 1998. Genomic evidence for two functionally distinct gene classes. *Genetics* **11**: 6239–6244.
- Snel, B., P. Bork, and M.A. Huynen. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21**: 108–110.
- Stevens, W.K. 1999. Rearranging the branches on a new tree of life. *The New York Times*: Aug. 31 1999: F1.

- Swofford, D.L. 1998. *PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods). Version 4*. Sinauer Associates, Sunderland, MA.
- Swofford, D., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic Inference. In *Molecular Systematics*, pp. 407–514. Sinauer Associates, Sunderland, MA.
- Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278**: 631–637.
- Teichmann, S.A. and G. Mitchison. 1999a. Is there a phylogenetic signal in prokaryote proteins? *J. Mol. Evol.* **49**: 98–107.
- . 1999b. Making family trees from gene families. *Nat. Genet.* **21**: 66–67.
- Tekaia, F., A. Lazcano, and B. Dujon. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9**: 550–557.
- Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL\_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**: 4876–4882.
- Tomb, J.-F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty, et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**: 539–547.
- Tumbula, D., U.C. Vothknecht, H. Kim, M. Ibba, B. Min, T. Li, J. Pelaschier, C. Stathopoulos, H. Becker, and D. Soll. 1999. Archaeal aminoacyl-tRNA synthesis. Diversity replaces dogma [In Process Citation]. *Genetics* **152**: 1269–1276.
- Van de Peer, Y., E. Robbrecht, S. de Hoog, A. Caers, P. De Rijk, and R. De Wachter. 1999. Database on the structure of small subunit ribosomal RNA. *Nucleic Acids Res.* **27**: 179–183.
- Woese, C.R. 1987. Bacterial evolution. *Microbiol Rev.* **51**: 221–271.
- Woese, C.R., O. Kandler, and M.L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**: 4576–4579.
- Wolf, Y.I., S.E. Brenner, P.A. Bash, and E.V. Koonin. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9**: 17–26.
- Yap, W.H., Z. Zhang, and Y. Wang. 1999. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J. Bacteriol.* **181**: 5201–5209.

Received October 21, 1999; accepted in revised form April 5, 2000.