Assessing evolutionary relationships among microbes from whole-genome analysis

Jonathan A Eisen

The determination and analysis of complete genome sequences have recently enabled many major advances to be made in the area of microbial evolutionary biology. These include the determination of the first genome of a Crenarchaeota, the suggestion that horizontal gene transfer may be the rule rather than the exception, and revelations about how genomes evolve on short timescales.

Addresses
The Institute for Genomic Research, 9712 Medical Center Drive,
Rockville, Maryland 20850, USA; e-mail: jeisen@tigr.org

Current Opinion in Microbiology 2000, **3**:475–480

1369-5274/00/$ – see front matter

Abbreviation
rRNA ribosomal RNA

Introduction

In 1965, Zuckerkandl and Pauling [1] started a revolution in evolutionary biology by showing how the evolutionary history of species could be inferred from comparisons of gene sequences. Subsequently, Woese and colleagues [2,3] used molecular systematic analysis of ribosomal RNA (rRNA) molecules to provide the first useful evolutionary classification of microbes as well as a universal tree of life. Although the rRNA tree of life was and continues to be incredibly useful, it has also been the subject of much controversy. Perhaps the most controversy has involved whether any single gene tree can represent the evolution of species. The use of a single tree of life assumes that species are related through vertical descent; however, not all genes follow the rules of vertical descent. For example, some genes can be transferred between lineages, a phenomenon known as horizontal or lateral gene transfer. Horizontal transfer complicates evolutionary reconstruction because it means that some species are chimeric, with several histories for different parts of the genome.

Before the genomics era, although some cases of horizontal gene transfer were generally accepted (e.g., transfers between organelles and the nucleus), the extent of the phenomenon was not clear. With great hope and some trepidation, many microbial evolutionary biologists looked forward to the ‘clarification’ that complete genome sequences would bring to this and other issues concerning microbial evolution. Fortunately for those who work on microbial evolution, the picture that is emerging from analysis of complete genome sequences, while clarified, is far from simple. Each lineage appears to have its own complex combination of vertical descent, gene transfer, gene and genome duplication, gene invention, gene loss and degradation, recombination, convergence and selection. Here I summarize recent work on using complete genome sequences to study evolution of microbes in light of these complications.

The available genome sequences are not representative of evolutionary diversity

Inferring evolutionary history from genomes is of course dependent on which genomes are analyzed. During the period of this review, the complete genomes of eight new microbial species have been published including representatives of a major new Archaeal lineage (Crenarchaeota [4]) and three new major bacterial lineages (Thermotogales [5••], the *Deinococcus-Thermus* phyla [6•] and the β-Proteobacteria [7,8]). Although genomes from representatives of many of the major microbial groups have now been sequenced (Figure 1), the diversity of species represented is still poor (Figure 2). Because species sampling can greatly affect results in evolutionary analysis, analysis based on these genomes should be considered with a note of caution.

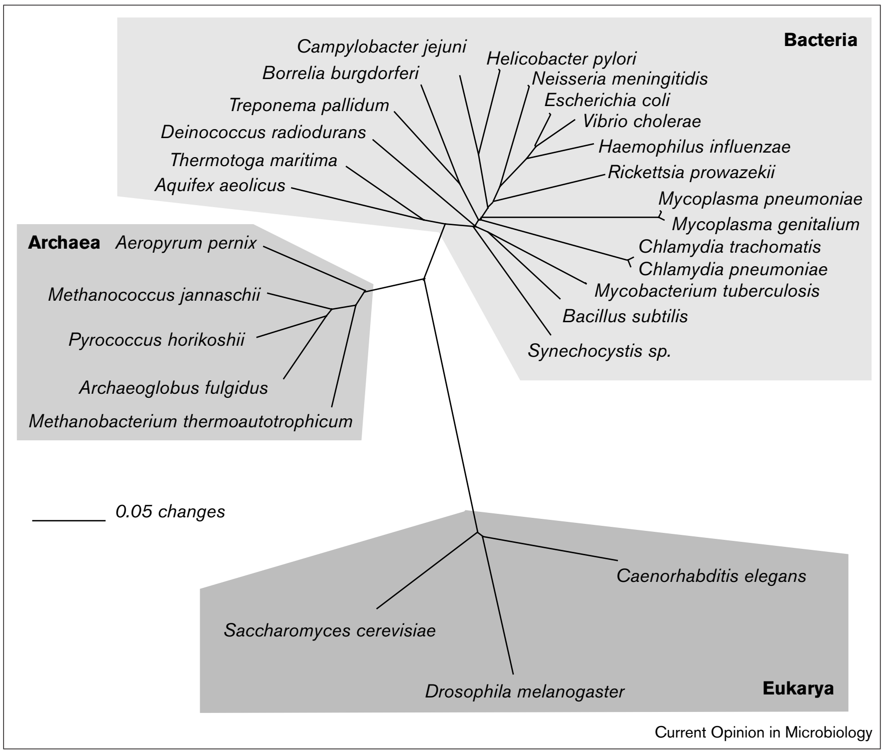
Methods for assessing evolutionary relationships among species from genome sequences

Comparing phylogenetic trees of universal genes

In theory, the best way to determine the evolutionary history of different genomes is to generate, and then compare and contrast, phylogenetic trees for every gene in every genome. A chief limitation of this approach is that phylogenetic tree reconstruction is not reliably automated as yet. To generate accurate trees, sequence alignments should be examined carefully, and regions of the alignments that are ambiguous or hypervariable should be excluded from the final phylogenetic analysis. Therefore, evolutionary studies of complete genome sequences using phylogenetic trees have focused on limited sets of genes. Ideally, the set of genes analyzed should be found in all of the species being studied. This removes the possibility that differences in the phylogenetic trees of different genes is due to different species sampling, which can profoundly effect phylogenetic reconstruction [9].

A few recent studies have reported phylogenetic analysis of sets of universal genes from the same complete genome sequences. These included a set of 33 genes found in yeast, four Archaea and 20 Bacteria [5••], and larger sets of genes from four Bacteria and two Archaea [10] or from four Archaea [11•]. The results of these studies are informative but not conclusive. Species from the same major microbial group in the rRNA tree (e.g. Archaea, Spirochetes) group together in the trees of most universal genes [5••]. This suggests that horizontal gene transfers, if they occurred, were either between members of the same group, or before the divergence of the major microbial lineages. Within the Euryarchaeota, however, trees of conserved genes are quite variable, which suggests that within-group transfers may be

Figure 1



A phylogenetic tree of the species for which complete genome sequences are available. The tree was constructed from rRNA sequences using a distance-based method (neighbour-joining with distances calculated using a maximum-likelihood measurement as implemented by the PAUP* program). rRNA sequences were downloaded from the Ribosome Database Project site (available at www.cme.msu.edu/RDP/html/index.html).

common [11•]. Relationships between major microbial groups are variable [5••,10], but these differences are not significant enough and too few species were analyzed to warrant any significant conclusions. As more genomes become available, it is likely that phylogenetic trees of universal genes will become more valuable and informative.

The trees generated from universal genes are not necessarily accurate. Factors such as convergence, mis-identification of orthologs, gene conversion and ambiguities in sequence alignments can lead to inaccurate trees. Therefore, a more careful analysis is needed to determine which genes are informative. For example, careful analysis of amino-acid usage can determine which proteins are prone to convergence and are therefore less useful for evolutionary analysis. In addition, trees of universal genes cannot reveal the evolution of whole genomes. Thus phylogenetic studies of non-universal genes and other methods that address the evolution of whole genomes are still of great use.

Patterns of best matches to different species

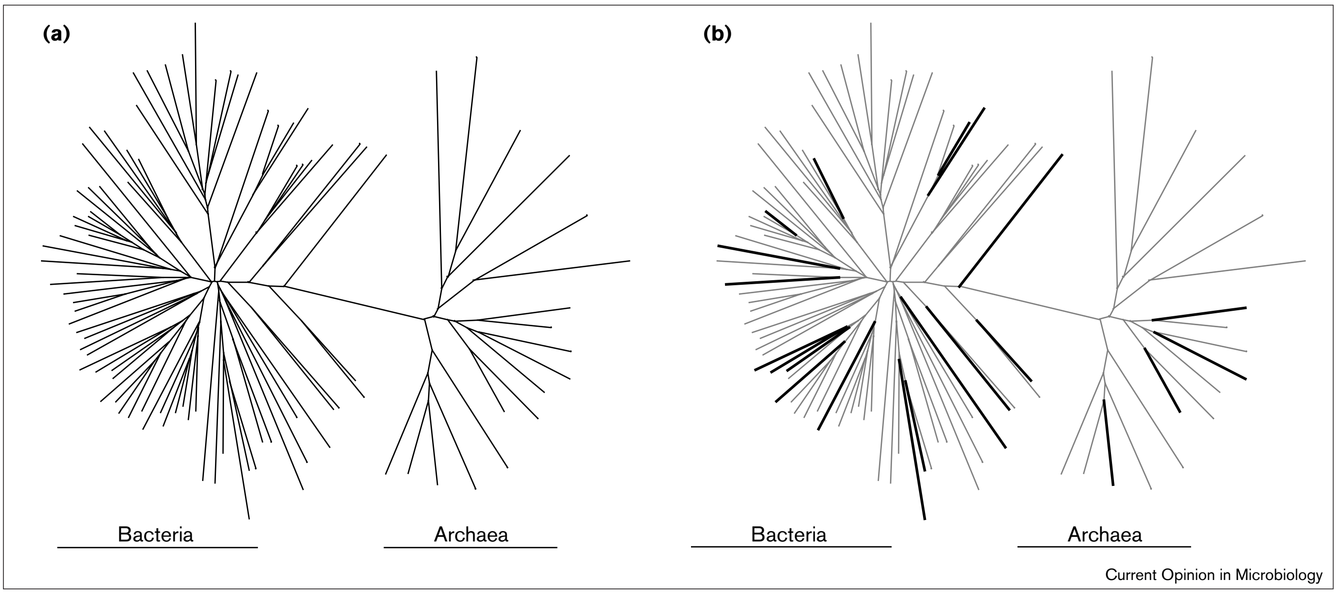
A common alternative to generating phylogenetic trees for all genes is to assess the degree of similarity of genes between genomes. Frequently, for analysis of a particular genome, the proportion of genes that show best matches to those of a different species is used as a measure of relatedness to the genome of that species. This ‘best-match’ approach led to one of the biggest stories in the study of microbial evolution in recent years — the suggestion of extensive horizontal gene transfers between thermophilic Bacteria and Archaea [5••,12•]. This suggestion was based on the finding that a large percentage of the genes in the bacteria *Aquifex aeolicus* and *Thermotoga maritima* (~20–25%) were most similar to genes from Archaea rather than from Bacteria. Best-match methods have also been used to suggest distinct origins for different genetic elements within species such as *Deinococcus radiodurans* and *Vibrio cholerae* [6•,13] and to identify likely organellar genes in nuclear genomes of eukaryotes [14].

The main advantage of best-match methods is their speed. In addition, they can readily be adapted to other types of analysis (such as the identification of recent gene duplications [6•]). Best match approaches are limited, however, because sequence similarity is not a perfect indicator of evolutionary relatedness [15]. For example, the high level of similarity between thermophilic Bacteria and Archaea could be due to an increased rate of evolution or gene loss in mesophilic Bacteria [16•]. Thus, sequence similarity should only be used as a screening method to identify genes of interest to pursue with other methods.

Clustering species by oligonucleotide relative abundance

All species have biases in the nucleotide composition of their genomes (e.g. in codon usage, GC nucleotide content, dinucleotide frequency) [17,18]. On a global level, these biases are relatively uniform within any particular genome. This led Karlin and co-workers [17,18] to use the similarity of relative abundances of di-nucleotides

Figure 2



Current Opinion in Microbiology

The diversity of bacterial and Archaeal species for which complete genomes are available is still poor. **(a)** Phylogenetic tree, based on rRNA sequences, of representatives of many major Bacterial and Archaeal lineages. **(b)** Lineages for which complete genomes are available are highlighted. rRNA tree for representative species was downloaded from the Ribosome Database Project (available at www.cme.msu.edu/RDP/html/index.html).

(referred to as genomic signature) as a measure of evolutionary relatedness. However, their own analysis suggests that genomic signature is too prone to convergence to be a useful evolutionary marker. For example, the genomic signatures of many mitochondria are most similar to those of *Clostridium* and *Sulfolobus* species [19]. The evidence that mitochondria are derived from microbes in the α-Proteobacterial lineage is simply overwhelming [20••], the similarities of genomic signature can be due only to convergence and not to mitochondria being derived from a *Sulfolobus–Clostridia* fusion.

Identifying regions of the genome with unusual compositions

Nucleotide composition is relatively constant within species, and this has been used to identify possible cases of horizontal gene transfer. It takes time for genes that have been transferred from other species to become adapted to the signature of their new host genome. Thus, it is possible to detect 'foreign' genes in a genome through the identification of unusual nucleotide composition [21,22]. An advantage of this approach is that it only requires the genome sequence of one species. A disadvantage is that, although recently transferred genes may have unusual compositions, unusual compositions can also be caused by other factors such as selection and mutation bias. In addition, transfers between species with similar signatures will not be detected by this method. Finally, this approach usually does not allow the determination of the particular history of any gene, just whether or not it may be foreign.

Using patterns of shared homologous proteins to build whole genome trees

The number of genes that are shared between genomes has been used to build 'whole genome trees'. Snel *et al.* [23••] showed that a distance tree based on number of genes shared between genomes is remarkably similar to the tree based on rRNA sequences for those same species. Subsequently, similar results were obtained using parsimony analysis (an evolutionary reconstruction method that is based on identifying the tree with the smallest number of evolutionary changes) [24] and a hierarchical classification method [25] based on the presence and absence of genes. It is important to note that these gene-content trees represent averages of patterns produced by phylogeny, gene duplication and loss, and horizontal transfer and are not real phylogenetic trees. The fact that these 'trees' are very similar to phylogenetic trees of single genes (such as those for rRNA or RecA) suggests that there may be an average phylogenetic history to species even if individual genes do not show this history. Thus, if horizontal gene transfers are occurring they either represent a small portion of the genome or are more frequent among similar organisms. This then allows members of major lineages in the rRNA tree to share many derived features [11•,26].

The number of shared genes between species is greatly affected by genome size in addition to evolutionary relatedness [7]. The gene content of distantly related organisms can become more similar to each other through the loss of genes that they do not share in common or through the duplication of shared genes. Closely related species can become dissimilar through the same forces (i.e. by loss of common genes or amplification of unshared genes). Therefore, analysis of shared gene content must be done with great care.

The big picture

Horizontal transfer versus vertical inheritance

Horizontal gene transfer can be very difficult to prove because each piece of evidence in support of it can be explained by other forces. For example, unusual nucleotide composition in a genome could be due to selection not transfer (see above). Thus, several lines of evidence should be used to identify horizontal transfer. For example, the suggestion of transfer between *T. maritima* and Archaea (based on best matches) has been supported by the finding that the Archaeal-like genes are clustered in the genome, show significantly biased nucleotide composition, and branch in evolutionary trees next to Archaeal genes [5••,27]. Although some cases of gene transfer will no doubt turn out to be artifacts, overall, the emerging picture is that horizontal gene transfer is not a rare exception, but may actually be quite common [28•,29•]. However, as mentioned above, it appears that the extent of gene transfer may be constrained by phylogenetic relationships.

The emphasis now needs to turn to determining the rules that govern transfers. For example, on the basis of their phylogenetic analysis of universal genes in six genomes (see above), Jain *et al.* [10] conclude that genes that interact with many other genes (informational genes) are less prone to gene transfer than those with fewer interactions (operational genes). A similar conclusion has been reached from the analysis of the four completely sequenced Archaeal genomes [11•]. It is still not clear, however, whether the complexity of interactions is the most important factor here. For example, experimental studies show that 'informational' genes from one species can work well in another. This is even true for rRNA [30]. The high conservation in sequence of some 'informational' genes may allow them to be transferred between species even though they interact with many other genes [31•].

Relationships among species and deep phylogeny

As discussed above, studies of complete genome sequences indicate that the average nature of most genomes can be predicted from the rRNA tree of life. For example, if an organism is classified as a Spirochete in the rRNA tree, its proteome is similar to that of other Spirochetes. As more genomes are completely sequenced, we will be able to determine whether this general pattern applies to all microbial phylogenetic groups or just the major taxa. The patterns of relationships among the different microbial groups is still ambiguous, however, partly because of the occurrence of gene transfers but also simply because of the difficulty of inferring events that occurred a long time ago. This is most apparent in attempts to infer the root of the tree of life [32,33••]. Based, in part, on analysis of complete genome sequences, Philippe and coworkers [32,33••] suggest that the previous rooting of the tree of life in which Archaea and Eukaryotes are considered sister groups is due to an artifact of phylogenetic reconstruction called 'long branch attraction'. In this phenomenon, unrelated lineages with long evolutionary branches artificially group together in trees. They propose that Bacteria and Archaea may share a common ancestor [32,33••]. This hypothesis seems reasonable, but it does not sufficiently address the problem of horizontal gene transfer. If gene exchanges have really occurred across the main domains of life as extensively as it seems, then there is no real way to root the tree of life. Until we know more about horizontal gene transfer (such as which genes are most prone to it), attempts to infer a root of the tree of life should be viewed with some skepticism.

Genome reduction and degradation

Complete genome sequences are helping reveal a great deal about genome reduction and degradation. For example, analysis of the *Ricketssia prowazekii* genome is providing great insight into gene loss, gene transfer and genome degradation in the evolution of mitochondria and endosymbionts [14,34,35,36••,37]. Interestingly, one analysis suggests that *R. prowazekii* may not be a perfect model for studying mitochondrial evolution because it has apparently lost different genes than have mitochondria [37]. Many of the species for which genomes are available have undergone large genome reductions relatively recently (at least in terms of the history of life). Thus, careful comparison of these genomes with those of other species will help to provide insight into the general processes of gene loss and degradation.

The value of comparing genomes from closely related species

The evolutionary distance separating the species for which genome sequences are available determines the types of questions that can be addressed using these data. The first genomes to be published were from distantly related species; therefore, inferences could only be made about events that occurred either long ago or very rarely. However, sets of genomes from closely related species or strains are now becoming available [7,38•,39•,40,41]. The comparison of such genome sequences will yield information on recent or frequent events, such as genome rearrangement and mutation processes. For example, comparison of the genomes of four Chlamydial species has revealed the occurrence of frequent tandem gene duplication and gene loss, as well as large chromosomal inversions [39•]. The identification of differences between strains of the same species will also be of great use in studying the population genetics of those species.

Conclusions

The analysis of complete genome sequences is showing that a genome can be considered almost a living breathing entity. Genomes grow by duplication, shrink by deletion and degradation, consume through horizontal exchange, reproduce and die. Genomes are even ‘breathing’ while the strains are being grown to prepare DNA for genome sequencing [42•]. In order to represent the complicated patterns of genome evolution, we now need new paradigms and models. In addition, we need to move on from simply detecting events, to understanding the rules that govern their occurrence. For future studies, it is important to recognize that a complete genome sequence is much more valuable than an incomplete one. For example, an understanding of genome evolution is dependent, in part, on knowing the relative position of genes and not just their presence and absence [39•,43••]. It is also important to realize that evolutionary analysis is not just for inferring the history of species. It is also useful for identifying gene duplication events [15], pathogenicity genes [21], organellar genes [14]; for predicting gene functions [44••–46••,47]; and for identifying vaccine candidates [7]. Thus, even if the picture of the relationships among species remains cloudy, evolutionary studies of genomes will continue to be of great value.

Acknowledgements

I thank all the people at The Institute for Genomic Research for support, assistance, and helpful discussions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Zuckerkandl E, Pauling L: **Molecules as documents of evolutionary history.** *J Theor Biol* 1965, **8**:357-366.

2. Woese C: **Bacterial evolution.** *Microbiol Rev* 1987, **51**:221-271.

3. Woese C, Fox G: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74**:5088-5090.

4. Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A *et al.*: **Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3.** *DNA Res* 1998, **5**:55-76.

5. •• Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA *et al.*: **Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*.** *Nature* 1999, **399**:323-329.

Presents several lines of evidence suggesting horizontal gene transfers have occurred between thermophilic Bacteria and Archaea. Over 20% of the genes in the genome have best matches to Archaeal genes.

6. • White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL *et al.*: **Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1.** *Science* 1999, **286**:1571-1577.

This paper shows how evolutionary analysis can be helpful in understanding the radiation resistance of *Deinococcus radiodurans* by identifying recent gene duplications, by predicting gene functions, and by showing how closely related this species is to *Thermus* species.

7. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ *et al.*: **Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58.** *Science* 2000, **287**:1809-1815.

8. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T *et al.*: **Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491.** *Nature* 2000, **404**:502-506.

9. Eisen JA: **The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16s rRNAs from the same species.** *J Mol Evol* 1995, **41**:1105-1123.

10. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96**:3801-3806.

11. • Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV: **Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9**:608-628.

A detailed comparative study of four complete Euryarchaeal genomes. The authors identify a set of proteins found in all four of these species. Phylogenetic trees of these are variable suggesting lateral transfers have occurred among the Euryarchaeotal lineages. The authors also present evidence that different types of genes may be more prone to horizontal transfer.

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

The first publication suggesting that extensive gene transfers have occurred between thermophilic Bacteria and Archaea. The conclusions are based on the analysis of best matches of proteins in the *Aquifex aeolicus* genome, as well as gene clustering, nucleotide composition, and evolutionary analysis.

13. Heidelberg JF, Eisen JA, Nelson WC, Clayton RC, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L *et al.*: **The genome sequence of *Vibrio cholerae*, the etiologic agent of cholera.** *Nature* 2000, **406**:477-484.

14. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M *et al.*: **Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*.** *Nature* 1999, **402**:761-768.

15. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.** *Genome Res* 1998, **8**:163-167.

16. • Kyrpides NC, Olsen GJ: **Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry?** *Trends Genet* 1999, **15**:298-299.

A brief but relatively thorough discussion of the problems with relying on sequence similarity alone as an indicator of evolutionary relatedness.

17. Campbell A, Mrazek J, Karlin S: **Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA.** *Proc Natl Acad Sci USA* 1999, **96**:9184-9189.

18. Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature.** *Trends Genet* 1995, **11**:283-290.

19. Karlin S, Brocchieri L, Mrazek J, Campbell AM, Spormann AM: **A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes.** *Proc Natl Acad Sci USA* 1999, **96**:9190-9195.

20. •• Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG: **The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria.** *Nature* 1998, **396**:133-140.

The first complete genome sequence from a species in the α-subgroup of the Proteobacteria. The authors' analysis provides remarkable insight into the evolution of mitochondria.

21. Lawrence JG, Ochman H: **Molecular archaeology of the *Escherichia coli* genome.** *Proc Natl Acad Sci USA* 1998, **95**:9413-9417.

22. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44**:383-397.

23. •• Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21**:108-110.

The first attempt to use gene presence and absence in different genomes to build a ‘whole-genome tree’. The authors show that this tree is very similar to the rRNA tree for the same species, suggesting that there is an average phylogeny of a genome that is accurately reflected in the rRNA of life.

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

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

26. Graham DE, Overbeek R, Olsen GJ, Woese CR: **An archaeal genomic signature.** *Proc Natl Acad Sci USA* 2000, **97**:3304-3308.

27. Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW: **Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*.** *Nucleic Acids Res* 2000, **28**:706-709.

28. Martin W: **Mosaic bacterial chromosomes: a challenge en route to**
• **a tree of genomes.** *BioEssays* 1999, **21**:99-104.
An excellent review of horizontal gene transfer in microorganisms and the difficulties extensive gene transfer presents to attempts to study the evolution of species. The authors outline the need to search for principles governing genetic exchange.

29. Doolittle WF: **Phylogenetic classification and the universal tree.**
• *Science* 1999, **284**:2124-2129.
A discussion of how the concept of species classification of species may need to be revised in the light of extensive horizontal gene transfer.

30. Asai T, Zaporojets D, Squires C, Squires CL: **An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria.** *Proc Natl Acad Sci USA* 1999, **96**:1971-1976.

31. Logsdon JM, Faguy DM: **Thermotoga heats up lateral gene**
• **transfer.** *Curr Biol* 1999, **9**:R747-R751.
Commentary on the problems of inferring evolutionary information from whole genome data.

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

33. Brinkmann H, Philippe H: **Archaea sister group of Bacteria?**
•• **Indications from tree reconstruction artifacts in ancient phylogenies.** *Mol Biol Evol* 1999, **16**:817-825.
An excellent discussion of the problems that long branch attraction and noise cause in attempts to infer the rooting of the tree of life. This paper presents an alternative in which the rooting of the tree of life is on the Eukaryotic branch.

34. Andersson JO, Andersson SG: **Insights into the evolutionary process of genome degradation.** *Curr Opin Genet Dev* 1999, **9**:664-671.

35. Andersson SG, Kurland CG: **Origins of mitochondria and hydrogenosomes.** *Curr Opin Microbiol* 1999, **2**:535-541.

36. Andersson JO, Andersson SG: **Genome degradation is an ongoing**
•• **process in *Rickettsia*.** *Mol Biol Evol* 1999, **16**:1178-1191.
An experimental study comparing the sequence of the *metK* gene in different *Rickettsia* species, showing that mutations such as deletions and point mutations are very frequent.

37. Muller M, Martin W: **The genome of *Rickettsia prowazekii* and some thoughts on the origin of mitochondria and hydrogenosomes.** *Bioessays* 1999, **21**:377-381.

38. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR,
• Noonan B, Guild BC, deJonge BL *et al.*: **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*.** *Nature* 1999, **397**:176-180.
The first whole-genome comparison of closely related genomes shows how different types of information can be inferred from this type of comparison than from comparison of distantly related genomes.

39. Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O,
• Hickey EK, Peterson J, Utterback T, Berry K *et al.*: **Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39.** *Nucleic Acids Res* 2000, **28**:1397-1406.
A detailed comparison of four closely related genomes (including two new genome sequences and two published genome sequences). This paper shows how comparisons of closely related species can be used to identify recent events such as inversions, duplications and deletions.

40. Maeder DL, Weiss RB, Dunn DM, Cherry JL, Gonzalez JM, DiRuggiero J, Robb FT: **Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences.** *Genetics* 1999, **152**:1299-1305.

41. Fukuda Y, Washio T, Tomita M: **Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*.** *Nucleic Acids Res* 1999, **27**:1847-1853.

42. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D,
• Chillingworth T, Davies RM, Feltwell T, Holroyd S *et al.*: **The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences.** *Nature* 2000, **403**:665-668.
In addition to presenting the complete genome sequence of the second species in the ε-Proteobacteria, the authors show how genome sequences can be used to identify hypervariable loci. Most remarkably, they find extensive polymorphisms in the library used for sequencing showing that the mutation rate in this species, especially at microsatellite loci, is very high.

43. Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH:
•• **Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases.** *Nucleic Acids Res* 1999, **27**:1642-1649.
A detailed comparative analyses of proteome composition and codon usage between two related genomes (*T. pallidum* and *B. burgdorferi).* The authors find that the primary factor influencing codon usage is that the direction a gene is transcribed relative to the direction of replication. This suggests that unusual nucleotide composition cannot be used as a reliable indicator of horizontal gene exchange and that gene finding programs might be improved if they incorporated information on direction of replication.

44. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D:
•• **Detecting protein function and protein–protein interactions from genome sequences.** *Science* 1999, **285**:751-753.
This paper, along with [45••], documents how distribution patterns of protein domains can be used to help predict protein functions.

45. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein**
•• **interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.
See annotation for [44••].

46. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO:
•• **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.
An simple and elegant approach to use the distribution patterns of proteins in different species to help predict functions.

47. Eisen JA, Hanawalt PC: **A phylogenomic study of DNA repair genes, proteins, and processes.** *Mutat Res* 1999, **435**:171-213.