Theoretical Population Biology **61**, 435–447 (2002)
doi:10.1006/tpbi.2002.1588

TPB

Microevolutionary Genomics of Bacteria

I. King Jordan, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building. 38A, 8600 Rockville Pike, Bethesda, Maryland 20894
E-mail: koonin@ncbi.nlm.nih.gov

Received January 22, 2002

The availability of multiple complete genome sequences from the same species can facilitate attempts to systematically address basic questions in genome evolution. We refer to such efforts as "microevolutionary genomics". We report the results of comparative analyses of complete intraspecific genome (and proteome) sequences from four bacterial species—*Chlamydophila pneumoniae, Escherichia coli, Helicobacter pylori* and *Neisseria meningitidis*. Comparisons of average synonymous (K_s) and nonsynonymous (K_a) substitution rates were used to assess the influence of various biological factors on the rate of protein evolution. For example, *E. coli* experiences the most intense purifying selection of the species analyzed, and this may be due to the relatively larger population size of this species. In addition, essential genes were shown to be more evolutionarily conserved than nonessential genes in *E. coli* and duplicated genes have higher rates of evolution than unique genes for all species studied except *C. pneumoniae*. Different functional categories of genes were shown to evolve at significantly different rates emphasizing the role of category-specific functional constraints in determining evolutionary rates. Finally, functionally characterized genes tend to be conserved between strains, while uncharacterized genes are over-represented among the unique, strain-specific genes. This suggests the possibility that nonessential genes are responsible for driving the evolutionary diversification between strains. © 2002 Elsevier Science (USA)

INTRODUCTION

Molecular evolutionary studies have long relied on the availability of genomic sequence data (Li, 1997; Hughes, 1999). An explosion of such data hailed the beginnings of the genomics era and has served to stimulate research in genome evolution. The first complete genome sequence of a cellular life form was published in 1995 (Fleischmann *et al*., 1995) and the nascent field of comparative genomics began in earnest shortly thereafter. By the end of 1996, five complete prokaryotic genome sequences were available for comparison (Fleischmann *et al*., 1995; Fraser *et al*., 1995; Bult *et al*., 1996; Himmelreich *et al*., 1996; Kaneko *et al*., 1996). These included representatives of three distinct taxonomic groups of bacteria and one archaeal species. A substantial portion of the entire scope of prokaryotic phylogenetic diversity was covered by these genomes (Fig. 1a). This broad coverage facilitated the ability of investigators to address a number of fundamental questions about deep evolutionary relationships among the proteins encoded in prokaryotic genomes and, by inference, among the genomes themselves (Kolsto, 1997; Koonin and Galperin, 1997; Koonin *et al*., 1997; Tatusov *et al*., 1997; Trevors, 1997; Huynen and Bork, 1998). For example, it was shown that while protein sequences are, in general, highly conserved across genomes, there are surprisingly few universally conserved gene families and very little conservation of genome organization (Mushegian and Koonin, 1996; Koonin and Galperin, 1997; Koonin *et al*., 1997; Watanabe *et al*., 1997). Early comparative genomics studies also uncovered evidence of numerous horizontal gene transfer events (reviewed in Doolittle, 1999; Koonin *et al*., 2001).

The recent proliferation of completely sequenced prokaryotic genomes has resulted in considerable increase in the arborescence of the phylogenetic tree of

0040-5809/02 $35.00

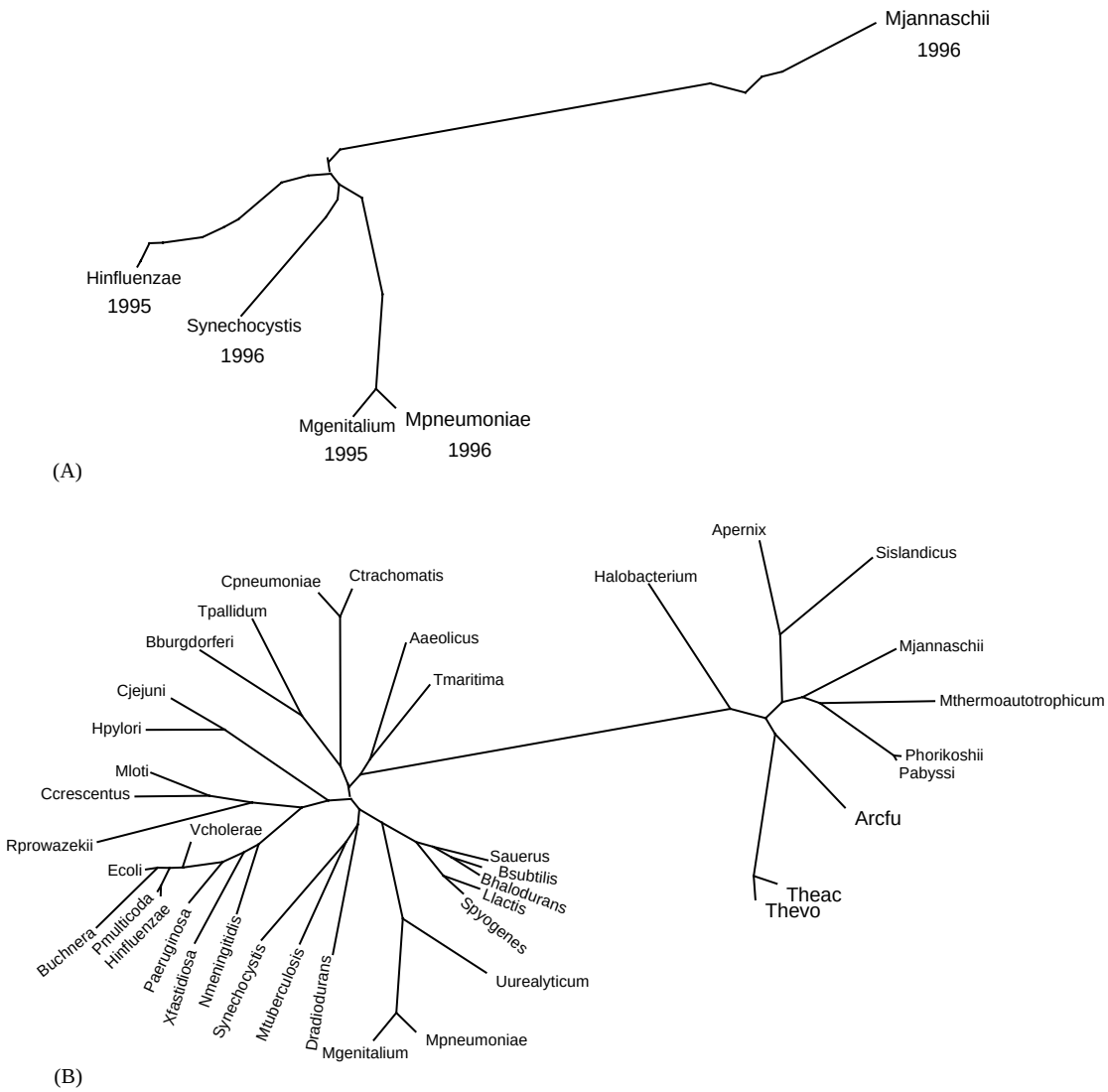

FIG. 1. Phylogeny of complete prokaryotic genomes based on a concatenated alignment of ribosomal proteins (Wolf *et al.*, 2001). (A) Phylogenetic relationship of species with a complete genome sequence by 1996. (B) Phylogenetic relationship of species with a complete genome sequence by 2001.

completely sequenced genomes (Fig. 1b). There are now numerous instances of closely related species that are represented in the set of publicly available completed genomes. There are even a number of cases where more than one complete bacterial genome has been sequenced for a single species. By 2001, there were at least four bacterial species—*Chlamydophila pneumoniae*, *Escherichia coli*, *Helicobacter pylori* and *Neisseria meningitidis*—with more than one completely sequenced genome (Blattner *et al.*, 1997; Tomb *et al.*, 1997; Alm *et al.*, 1999; Kalman *et al.*, 1999; Parkhill *et al.*, 2000; Read *et al.*, 2000; Shirai *et al.*, 2000; Tettelin *et al.*, 2000; Perna *et al.*, 2001). Comparative sequence analysis of these closely related genomes has the potential to enable the deduction of qualitatively unique inferences about the mechanisms and processes that characterize bacterial molecular evolution. The simultaneous comparison of multiple orthologs from complete intraspecific genome sequences can also be beneficial in the sense that it provides a high degree of power and resolution for attempts to address questions relating to genome evolution.

We have coined the phrase "microevolutionary genomics" to describe this class of endeavors. Microevolution is a notoriously vague term that can generally be considered to describe evolutionary phenomena that

occur within species (Futuyma, 1986). The molecular changes between strains that we observe here are consistent with this definition. However, microevolutionary studies also tend to focus on polymorphism, which is defined as variation within a population. Divergence between two or three strains of the same bacterial species, as observed here, may well be fixed and if so would not correspond to polymorphism. Our use of the phrase microevolutionary genomics should be considered in this context.

It may be useful to define several molecular evolutionary concepts that will show up repeatedly in this work. Purifying or negative selection is a type of natural selection where deleterious variants are removed because of the lower fitness of the individuals that carry them (Hughes, 1999). Purifying selection has the effect of reducing the rate of protein evolution. The synonymous substitution rate (K_s) measures the rate of change for synonymous substitutions, which are those that do not change the encoded amino acid sequence (they typically occur in the third positions of codons). These changes can be considered to be unaffected by natural selection (but see Results and Discussion). The nonsynonymous substitution rate (K_a) measures the rate of change for nonsynonymous substitutions that do change the encoded amino acid sequence (typically, substitutions in the second and first positions of codons). These changes do tend to be affected by natural selection. Homologous genes or proteins are those that can be demonstrated to share a common ancestor. Orthologs are homologs that have diverged due to speciation (accordingly, orthologous genes from two species derive from the same ancestral gene in the last common ancestor of the species in question) and paralogs are homologs that have diverged due to gene duplication (Fitch, 1970, 2000).

In this study, we sought to exploit the existence of multiple complete genome sequences from single bacterial species to ask a series of explicit questions about the nature of bacterial genome evolution and, in particular, about the factors that influence the rate of protein evolution. We asked: Does the intensity of purifying selection vary among bacterial species? What is the nature of the functional distribution for genes conserved between bacterial strains of the same species (orthologs)? Do different functional classes of genes evolve at different rates? Do duplicated genes evolve at the same rate as unique genes? Do essential genes evolve at the same rate as nonessential genes? Orthologs shared between completely sequenced strains of *C. pneumoniae*, *E. coli*, *H. pylori* and *N. meningitidis* were identified and their sequences were analyzed with respect to these questions.

RESULTS AND DISCUSSION

Genomic Selection Levels

The relative level of purifying selection is thought to be the most important factor governing the rate of evolution for protein-coding genes (Kimura, 1983; Li, 1997). Proteins that are under severe functional constraints are subject to high levels of purifying selection and evolve relatively slowly. Conversely, proteins that are less functionally constrained experience lower levels of purifying selection and therefore evolve more rapidly. The level of purifying selection for a gene can be assessed by comparing K_s and K_a (Hughes, 1999). K_s can be considered to reflect the rate of neutral change since synonymous substitutions are generally not (strongly) affected by natural selection. The ratio—K_a/K_s—reflects the rate of amino acid substitution normalized by the amount of neutral change and is used to evaluate the level of purifying selection acting on a gene. This only works for closely related genes where K_s is not saturated due to too many cases of multiple change at a site. In addition, synonymous changes might in fact be subject to some purifying selection due, for example, to codon bias (Shields *et al.*, 1988). However, on average, the level of purifying selection that acts on synonymous changes is much lower than that for nonsynonymous changes. So although the formal assumption of no selection on synonymous changes that underlies the use of K_a/K_s is often violated, in practice this ratio serves as a robust approximation of the level of purifying selection.

The K_a/K_s ratio is most often used to compare the strength of purifying selection acting on different genes. Here we employ this analytical tool to compare the strength of purifying selection at the level of whole genomes. To do this, all orthologous genes shared between strains of the same species were identified and their sequences aligned. Then K_a and K_s were determined for each orthologous pair (Fig. 2). The average values of K_a and K_s for all analyzed genomes can be compared to give some indication as to the level of purifying selection for each species (Table I). *E. coli* has the lowest average K_a/K_s ratio indicating that purifying selection is most stringent for this species. This is consistent with a previously published

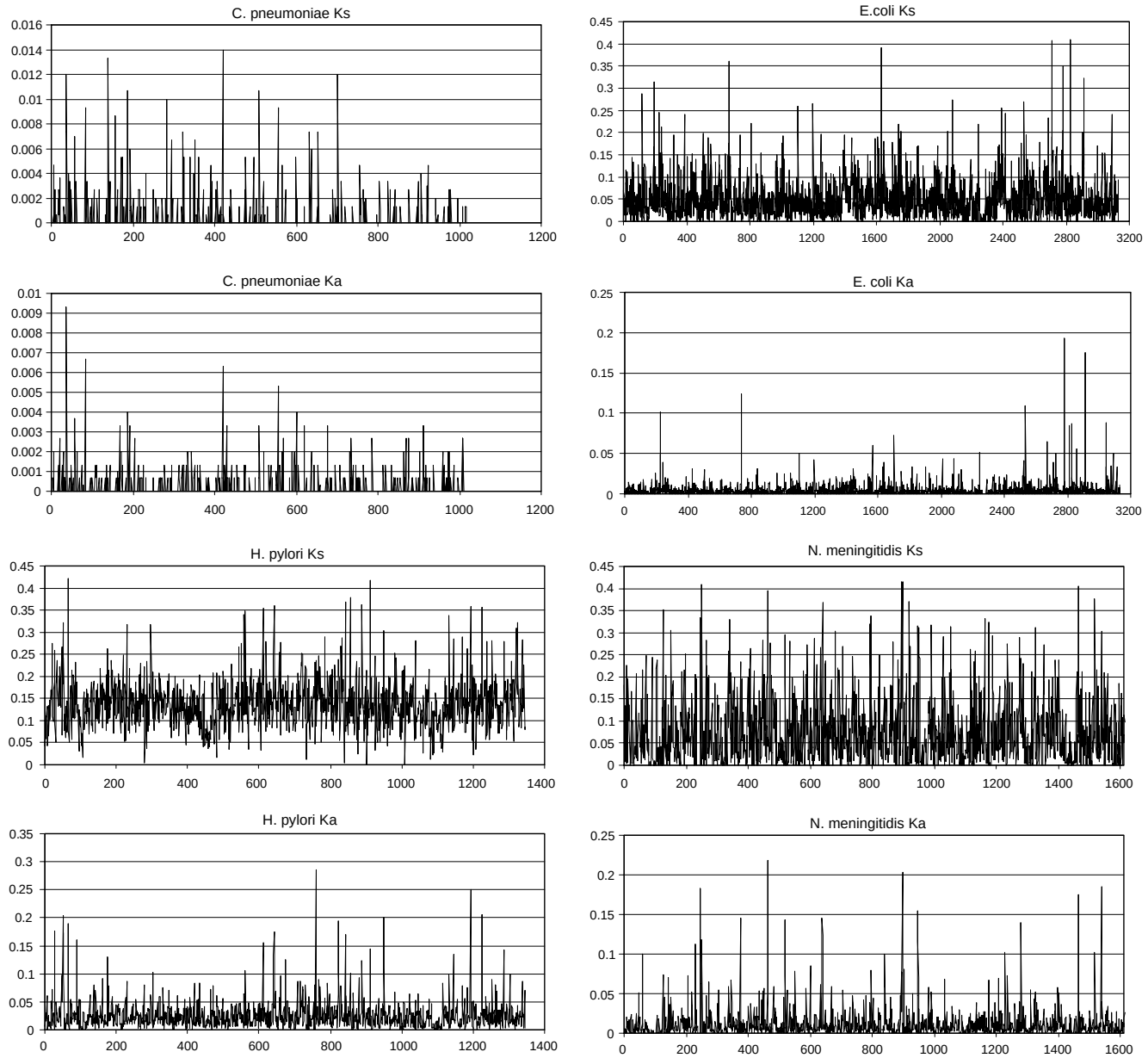

FIG. 2. Levels of K_s and K_a for orthologs shared between strains of *C. pneumoniae*, *E. coli*, *H. pylori* and *N. meningitidis*. The order of values (*x*-axis) corresponds to the order of the orthologous genes in the originally sequenced genome.

interspecific comparison of prokaryotic nucleotide substitution rates (Ochman *et al.*, 1999). *E. coli* is followed by *N. meningitidis*, *H. pylori* and, finally, *C. pneumoniae*. The average values of K_a/K_s for *N. meningitidis* and *H. pylori* are very similar relative to any other pairwise comparison. The average K_a/K_s for *C. pneumoniae* is much higher than the corresponding value for any of the other species. This is probably an artifact that results from the fact that the *C. pneumoniae* genomes are all very closely related (Fig. 1 and Table I). The lack of sequence divergence among the *C. pneumoniae* genomes results in a lack of accuracy of the K_a and K_s calculation due to the very small number of nucleotide changes that occur in each orthologous gene pair. That the unusually high K_a/K_s value of *C. pneumoniae* is an artifact, is supported by the coefficient of variation for K_a and K_s (Table I). This value indicates the normalized amount of variation seen

TABLE I

Average Sequence Variation between Intraspecific Orthologs

	C. pneumoniae	*E. coli*	*H. pylori*	*N. meningitidis*
No. of orthologs	1018	3106	1327	1607
aa % id ± SE[a]	99.93 ± 0.4	99.15 ± 1.4	95.35 ± 4.2	97.96 ± 2.9
nt % id ± SE[b]	99.96 ± 0.3	98.40 ± 1.4	94.68 ± 2.5	97.40 ± 2.6
K_s[c]	4.63×10^{-4}	4.85×10^{-2}	1.38×10^{-1}	7.22×10^{-2}
K_s—SE[d]	1.48×10^{-6}	1.35×10^{-5}	4.05×10^{-5}	4.52×10^{-5}
K_s—coeef[e]	3.27	0.86	0.39	1.01
K_a[c]	2.63×10^{-4}	3.93×10^{-3}	2.60×10^{-2}	1.14×10^{-2}
K_a—SE[d]	7.34×10^{-7}	1.26×10^{-6}	2.03×10^{-5}	1.19×10^{-5}
K_a—coeff var[e]	2.84	2.15	1.04	1.69
K_a/K_s	0.568	0.081	0.188	0.158

[a] Average amino acid percent identity with standard error for all orthologs shared between strains.
[b] Average nucleotide percent identity with standard error for all orthologs shared between strains.
[c] Average synonymous (K_s) or nonsynonymous (K_a) substitution rate for all orthologs shared between strains.
[d] Standard error of the synonymous (K_s) or nonsynonymous (K_a) substitution rate for all orthologs shared between strains.
[e] Coefficient of variation is the average of the synonymous (K_s) or nonsynonymous (K_a) substitution rate divided by the standard deviation of the synonymous (K_s) or nonsynonymous (K_a) substitution rate for all orthologs shared between strains.

for K_a and K_s. If nonsynonymous changes are more subject to purifying selection, K_a will vary more than K_s because of a broad distribution of the strength of selection constraints (Grishin *et al.*, 2000), and this will be reflected in substantially higher coefficients of variation for K_a than for K_s. This is seen for three of the four species analyzed here. *C. pneumoniae* is the only species with a higher coefficient of variation for K_s than K_a.

It is tempting to speculate as to the biological reason for the differences in the level of purifying selection between *E. coli* and *N. meningitidis–H. pylori*. *E. coli* is a free-living organism, whereas *N. meningitidis* and *H. pylori* are parasites. It can be expected that a free-living organism like *E. coli* will have higher population numbers than species with a parasitic life style. The efficiency of natural selection is determined, in large part, by the population size (Ohta, 1992). Specifically, purifying selection is more efficient at removing deleterious variants in large populations. The relatively low value of K_a/K_s for *E. coli* is consistent with an enhanced purifying selection caused by the large population size. However, a study of nucleotide substitution rates between prokaryotic genomes revealed a number of cases of free-living species with a lower level of purifying selection than *E. coli* (Ochman *et al.*, 1999). More controlled comparisons of the same type with additional species will show whether or not the pattern of a stronger purifying selection for free-living organisms holds up.

Functional Distribution of Orthologs

Orthologs are homologous genes shared between different species (or strains of the same species) that have diverged as a result of the diversification event(s) that delimits the species (strains). With closely related genomes, such as those analyzed here, one can expect that the vast majority of genes in any genome will have an ortholog in the corresponding genome of the other strain(s) of the same species. Interestingly, this turns out not to be the case. For three of the four species analyzed here, a substantial fraction of the genes in any genome are strain specific. The presence of strain-specific genes may be attributed to a number of different factors. They might arise because of lineage-specific expansion of paralogous gene families via gene duplication that occurred after the strains split (Jordan *et al.*, 2001b). Alternatively, lineage-specific loss (Aravind *et al.*, 2000; Braun *et al.*, 2000) of a gene in one strain will result in a unique gene (no ortholog) in the other strain. Perhaps, even more interesting is the possibility that numerous strain-specific genes were introduced via horizontal transfer. In fact, there is compelling evidence that is consistent with this last scenario. For example, in both *E. coli* and *H. pylori*, it appears that the strain-specific insertion of phage genomes has resulted in the introduction of numerous unique genes (Alm *et al.*, 1999; Perna *et al.*, 2001).

Comparison of complete genome sequences from the same species was used to assess whether certain

functional classes of genes are more prone to be retained between strains, whereas other classes are more likely to be unique. This was done by comparing the observed functional distribution of orthologous proteins shared between strains with the expected distribution based on examination of the complete predicted proteomes (Table II). Proteins were assigned functions on the basis of their designation in the clusters of orthologous groups of proteins (COG) database (Tatusov *et al*., 1997, 2000, 2001). The COG database assigns proteins to 18 specific functional categories using a modification of a previously designed functional classification scheme (Riley, 1993). These 18 categories are subsumed into four broad functional classes. While the accuracy of protein placement into the specific categories is prone to error and uncertainty on some occasions, the broad classification scheme of COGs is more robust and was employed here. Expected values for the number of proteins in the four functional classes were determined by evaluating the proportion of proteins in each class for the whole genome. Then the distribution of functional classes among proteins with shared orthologs in multiple strains was determined. Expected and observed values were compared with chi-square tests for the four species analyzed here (Table II).

The *C. pneumoniae* strains only diverged very recently and as such there are very few strain-specific genes among these genomes. Consistent with this fact, there was no statistically significant difference between the observed and expected functional distributions of *C. pneumoniae* orthologs. All three of the other species did show significant differences between the observed numbers of orthologs in each functional class versus the expected values based on the whole genome. This indicates that genes that tend to be conserved between genomes may have distinct functional characteristics relative to those that make up the strain-specific sets. For *E. coli*, *H. pylori* and *N. meningitidis*, there was an under-representation of poorly characterized proteins among the orthologs shared between strains. In each of these species, all three of the other functional classes were over-represented among orthologs. This probably indicates, not surprisingly, that the genes responsible for encoding critical, house-keeping functions are most conserved between strains. Meanwhile, poorly characterized genes with less critical cellular functions may be responsible for many of the strain-specific differences in

TABLE II

Chi-square Test of the Observed versus Expected Distribution of Orthologs in Different Functional Classes

Species	Functional class[a]	Observed[b]	Expected[c]
C. pneumoniae	Information storage and processing	184	180.0
	Cellular processes	143	139.3
	Metabolism	208	203.2
	Poorly characterized	483	495.5
	$\chi^2 = 0.61$	$p = 0.89$	
E. coli	Information storage and processing	448	423.8
	Cellular processes	611	526.1
	Metabolism	1003	872.5
	Poorly characterized	1072	1311.6
	$\chi^2 = 78.38$	$p = 6.81 \times 10^{-17}$	
H. pylori	Information storage and processing	213	207.0
	Cellular processes	280	249.4
	Metabolism	315	276.3
	Poorly characterized	537	612.3
	$\chi^2 = 18.61$	$p = 3.29 \times 10^{-4}$	
N. meningitidis	Information storage and processing	281	252.4
	Cellular processes	306	263.7
	Metabolism	426	352.9
	Poorly characterized	615	759
	$\chi^2 = 52.44$	$p = 2.41 \times 10^{-11}$	

[a] Functional class designations based on the COGs database (Tatusov *et al*., 1997, 2000, 2001).
[b] Observed number of orthologs in each functional class.
[c] Expected number of orthologs in each functional class calculated as the fraction of genes in the genome (species) for a functional class multiplied by the total number of orthologs (shared between strains) of that functional class.

the biology of these organisms. This raises an interesting paradox whereby the genes that are most interesting from a functional biologist's standpoint (because they have more direct importance for cellular functions) may in fact be the least interesting from an evolutionary biologist's perspective. In other words, uncharacterized genes that encode proteins with no ancient conserved regions might actually represent the protein-coding component of the genome most responsible for driving the diversification between evolutionary lineages.

An interesting corollary to these results was obtained through a recent comparative genomic analysis of orthologous proteins from closely related (but not intraspecific) complete genome sequences from the Chlamydiaceae family (Jordan *et al.*, 2001a). It was shown that the relative rate of evolution among orthologs shared between species was extraordinarily constant. This is consistent with a conservative mode of evolution for orthologs dominated by purifying selection. However, approximately 1% of the orthologous proteins analyzed did show evidence of lineage-specific enhanced rates of protein evolution consistent with functional diversification between orthologs. The proteins that showed this anomalous pattern of evolution have similar structural and functional characteristics. They are known or predicted to function at the periphery of the cell and may mediate the interaction of the cells with their environment. More specifically, the majority of these proteins were predicted membrane proteins and, possibly, receptors. This is consistent with the notion that bacterial and archaeal proteomes consist of a stable conserved core of proteins and a variable shell (Makarova *et al.*, 1999). The stable core encodes mostly informational proteins (e.g., those involved in translation), whereas the variable shell includes membrane-associated proteins (Rivera *et al.*, 1998). In addition, the orthologous proteins with enhanced relative rates of evolution also had very narrow phyletic distributions; homologs for 6 out of 7 could only be detected among the Chlamydiacea. Thus, even in the case of orthologs shared between species, it is the lineage-specific component that appears to be most prone to evolutionary diversification.

Functional Class and Evolutionary Rate

Given that different broad functional classes of genes have different evolutionary trajectories in terms of the retention of orthologs between strains, we wished to assess whether functional class also influences the rate of evolution. Levels of K_s and K_a were determined for individual orthologs shared between strains (Fig. 2). Orthologs were then grouped into 18 specific functional categories based on their classification in the COGs database. For each species, analysis of variance (ANOVA) was used to test whether genes belonging to different functional categories have significantly different rates of evolution between strains (Table III).

C. pneumoniae was the only species that did not show any evidence of significantly different rates of evolution for orthologs from different categories. As described above, this is probably an artifact of the relative lack of resolution provided by the extremely recently diverged *C.pneumoniae* genomes. For the three remaining species, the levels of K_a varied significantly between the different functional categories. This is not too surprising when the influence of protein functional constraints on evolutionary rate is considered. Apparently, different functional classes of genes have markedly different levels of functional constraint. Therefore, purifying selection will act with more or less intensity in different functional groups resulting in the significantly different rates of evolution observed. However, this does not result from uniform rates of evolution for all or even most orthologs within functional categories. In fact, there is a great amount of variation in evolutionary rates within functional groups as well as between them (data not shown). Instead, the difference between groups seen here reflects the high level of resolution afforded by the simultaneous analysis of thousands of genes. Such differences would most likely not have been detected without the use of a genomic-scale approach.

Data for the three genomes that showed differences in the rates of evolution between functional groups were pooled, and the strength of purifying selection acting on orthologs of each functional group was assessed by comparing K_a and K_s/K_a (Fig. 3). Determination of

TABLE III

Analysis of Variation (ANOVA) Results for Average K_s and K_a from Different Functional Categories

	K_s[a]	K_a[a]
C. pneumoniae	0.81	0.44
E. coli	3.14×10^{-17}	1.60×10^{-13}
H. pylori	3.14×10^{-2}	4.49×10^{-28}
N. meningitidis	7.59×10^{-3}	1.01×10^{-7}

[a] Level of statistical significance for the ANOVA test of the null hypothesis of equal levels of average synonymous (K_s) or nonsynonymous (K_a) substitution rate among the different functional categories. This is the probability that the differences observed (data not shown) are due to chance.

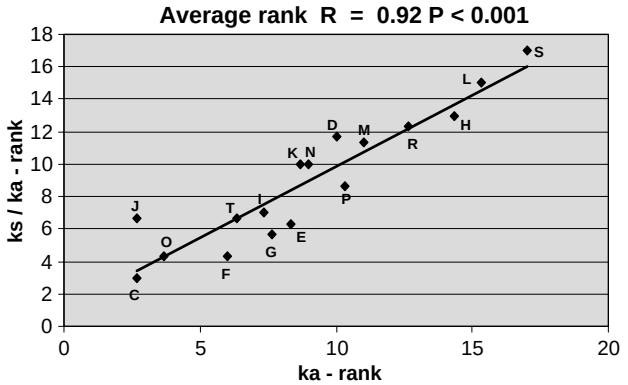

FIG. 3. Rank correlation of average K_a and K_s/K_a for orthologs from different COG functional categories. Metabolism: C—energy production and conversion, G—carbohydrate transport and metabolism, E—amino acid transport and metabolism, F—nucleotide transport and metabolism, H—coenzyme metabolism, I—lipid metabolism. Information storage and processing: J—translation, ribosomal structure and biogenesis, K—transcription, L—DNA replication, recombination and repair. Cellular processes: D—cell division and chromosome partitioning, O—posttranslational modification, protein turnover, chaperones, M—cell envelope biogenesis, outer membrane, N—cell motility and secretion, P—inorganic ion transport and metabolism, T—signal transduction mechanisms. Poorly characterized: R—general function prediction only, S—function unknown.

the Spearman rank correlation coefficient for these two measures indicates that they are strongly correlated. Among the most evolutionarily conserved functional categories are energy production and conversion as well as translation. These classes are consistently among the predicted highly expressed genes (Karlin and Mrazek, 2000). This suggests that there may be a correlation between the level of expression and the level of conservation for prokaryotic genes as has been demonstrated for mammalian genes (Duret and Mouchiroud, 2000). The least conserved groups are the unknown function and, more unexpectedly, DNA replication, recombination and repair.

Another surprising finding that resulted from this analysis was that levels of K_s also varied significantly between functional groups (Table III). This most likely indicates that synonymous substitutions are also subject to purifying selection based on functional constraint, although some effect of mutational biases cannot be ruled out.

Gene Duplication and the Rate of Evolution

Gene duplication is an important mechanism that can lead to the emergence of genetic novelty (Ohno, 1970; Li, 1997; Hughes, 1999). Studies of recently duplicated paralogous genes indicate that while paralogs do evolve under considerable selective constraint (Hughes and Hughes, 1993), there is an elevated rate of evolution between paralogs immediately following duplication (Lynch and Conery, 2000; Kondrashov *et al.*, 2002). This is likely to be due to changes in functional constraint after duplication.

We sought to assess whether there was evidence for different functional constraints between duplicated and nonduplicated genes using completely sequenced intraspecific bacterial genomes. Previous studies that have addressed similar questions relied on the comparison of evolutionary rates between paralogs and those between orthologs with a comparable level of divergence (Kondrashov *et al.*, 2002). The availability of closely related completely sequenced genomes allows for a different approach that relies strictly on the comparisons of the same type of genes—orthologs. Because orthologs reflect divergence since the diversification event that defines the lineages under analysis, this approach results in the comparison of genes that have been changing for precisely the same amount of evolutionary time. This seems to provide some conceptual advantage over studies that compare evolutionary rates between paralogs that diverged by gene duplications that occurred at different times and/or the rates of evolution between paralogs to those between orthologs.

For each species, individual predicted proteomes were self-compared and clustered using the BLASTCLUST program (ftp://ncbi.nlm.nih.gov/blast/documents/README.bcl). BLASTCLUST was run using a range of percent identity cut-offs (50–60–70–80–90) for clustering proteins. For each BLASTCLUST run, this procedure resulted in two classes of genes for any genome: unique and duplicated. Evolutionary rates (K_a and K_s) were then compared for the sets of orthologs that were defined as either unique or duplicated in the different BLASTCLUST iterations. Consistent with the model of relaxed functional constraints following gene duplication, the set of orthologs that were defined as belonging to duplicated groups showed clear and consistent evidence of higher average evolutionary rates (data not shown). However, there was a very wide range of observed evolutionary rates between orthologs within both the unique and duplicated sets. This wide range of variation resulted in a lack of statistical significance for the differences between the average rates of evolution for unique versus duplicated genes.

Increasing the percent identity at which BLASTCLUST was run had the effect of progressively decreasing the number of orthologous genes that were

included in the duplicated set. It is likely that the iterations run with higher percent identities are enriched for genes that have duplicated relatively recently (although the time of duplication cannot be unequivocally ascertained by percent identity between paralogous proteins because of the differences in evolutionary rates among different classes of proteins). If relaxation of functional constraint is most pronounced immediately after gene duplication (Lynch and Conery, 2000), then recently duplicated genes, identified by running BLASTCLUST with a high percent identity cut-off, should have the relatively highest levels of K_a (greater K_a/K_s values). So, as the percent identity cut-offs rise for BLASTCLUST, there should be a progressively greater difference in the average K_a between the duplicated and unique gene sets. This prediction was evaluated for each genome by taking the difference between the average duplicated K_a and the average unique K_a at each BLASTCLUST percent identity cut-off (50–60–70–80–90). These differences were ranked and compared to the ranks of the percent identity values used for the different BLASTCLUST runs (Fig. 4). For *C. pneumoniae*, there was a very slight and nonsignificant negative correlation between these two ranks. This could, again, result from the lack of resolution afforded by the extremely closely related *C. pneumoniae* genomes. For all other species, there are strong and statistically significant positive correlations between these two ranks. While the values of the Spearman rank correlation coefficient (R) were very high, the p values associated with these correlations were marginal simply because of the low number (5) of observations. This trend is consistent with an evolutionary pattern where there was an initial relaxation of functional constraint following gene duplication.

Evolutionary Rate of Essential versus Nonessential Genes

Molecular genetic 'knock-out' techniques have long been used to define genes as being either essential or nonessential for viability. Essential genes are predicted

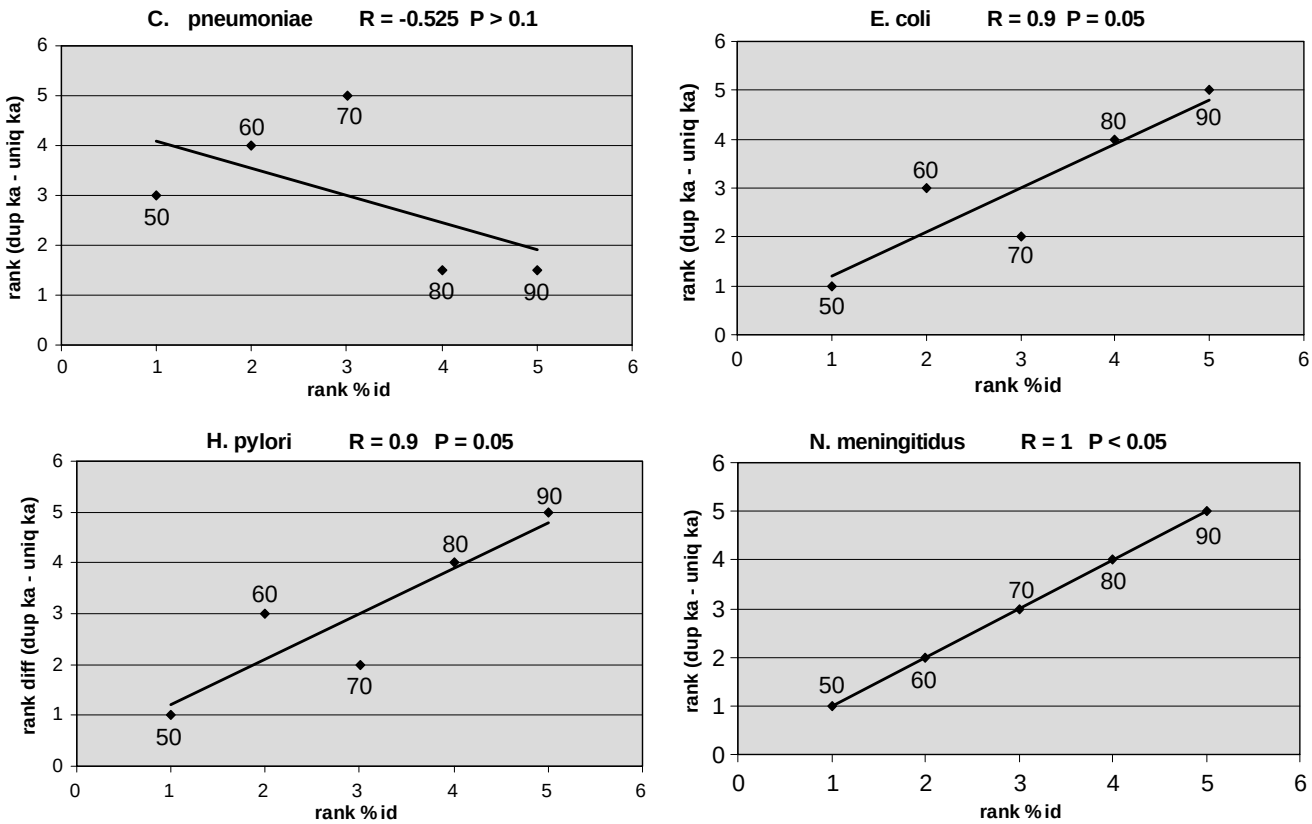

FIG. 4. Rank correlation of BLASTCLUST percent identity cut-off values and the difference between the average K_a of duplicated versus unique orthologs shared between strains.

to be more evolutionarily conserved than nonessential genes that are more dispensable or redundant (Wilson *et al*., 1977). The reasoning behind this prediction is that purifying selection should be more intense for essential genes. Recent tests of this prediction have yielded equivocal results (Hurst and Smith, 1999; Hirsh and Fraser, 2001). While these works do suggest some relationship between a gene's fitness class and its rate of evolution, they did not find any significant difference between the rate of evolution for essential versus nonessential genes. Analysis of complete genome sequences, such as conducted here, has the potential to address this question with an unprecedented level of resolution.

E. coli genes are identified as essential or nonessential (or unknown) in the Profiling of the *E .coli* Genome (PEC) database (http://www.shigen.nig.ac.jp/ecoli/pec/). These characterizations are based mainly on experimental evidence. Genes that have conditional lethal mutants are considered to be essential. Genes that have null mutations in a strain that is still able to grow are considered to be nonessential. In a few cases, general functional considerations are used by PEC to classify genes as essential or nonessential. For example, ribosomal structural genes are classified as essential, while genes involved in chemotaxis are considered nonessential.

Levels of K_a and K_s were calculated for *E. coli* orthologs classified as essential or nonessential (Table IV). The ranges of K_a and K_s for essential genes are fully included in the ranges for nonessential genes. However, average K_a and K_s were significantly lower for essential genes than for nonessential genes. The lower levels of K_a for essential genes indicate an average reduction in the rate of essential protein evolution consistent with the knock-out rate prediction. Interestingly, K_s is also lower for essential genes. This is most likely due to a correlation between K_s and K_a (Fig. 5), which also has been observed in a number of other cases (Wolfe and Sharp, 1993; Ohta and Ina, 1995; Makalowski and Boguski, 1998). This correlation reflects, to some extent, the action of purifying selection on synonymous sites due perhaps to codon bias. However, the higher value of K_s/K_a for essential genes indicates that K_a is decreased to a greater extent than K_s among essential genes. The greater influence of purifying selection on nonsynonymous sites is also indicated by the greater coefficients of variation seen for K_a (Table IV).

The decrease in the intraspecific rates of evolution for essential genes reflects the action of purifying selection over relatively short (micro) evolutionary time spans. Another measure of gene conservation is the taxonomic or phyletic distribution of homologous genes across different species. This measures the action of purifying selection over much greater periods of time. A phyletic distribution parameter was determined for each orthologous *E. coli* protein using the COGs database. Each COG is made up of representatives of anywhere from 3 to 26 different taxonomic groups of organisms. This value of this number for the COG that any gene maps to is taken as its phyletic distribution parameter and indicates the extent to which homologs of the gene are conserved during macroevolution. Consistent with what was found with the microevolutionary analysis, essential *E. coli* genes have a significantly higher phyletic

TABLE IV

Average Sequence Variation and Phyletic Distribution for Essential versus Nonessential *E. coli* Orthologs

	Essential ($n = 205$)	Nonessential ($n = 1813$)
$K_s \pm$ SE[a]	$2.70 \times 10^{-2} \pm 1.4 \times 10^{-4}$	$5.10 \times 10^{-2} \pm 2.3 \times 10^{-5}$
K_s—range[b]	0.000–0.219	0.000–0.409
K_s—coeff var[c]	1.12	0.82
$K_a \pm$ SE[a]	$1.11 \times 10^{-3} \pm 9.8 \times 10^{-6}$	$3.60 \times 10^{-3} \pm 4.4 \times 10^{-6}$
K_a—range[b]	0.000–0.019	0.000–0.175
K_a—coeff var[c]	1.83	2.22
K_s/K_a	24.37	14.21
Phyletic dist.[d]	20.46 ± 5.5	13.41 ± 7.7

[a] Average synonymous (K_s) or nonsynonymous (K_a) substitution rate for all orthologs shared between strains.
[b] Low and high values observed for synonymous (K_s) or nonsynonymous (K_a) substitution rate between orthologous pairs.
[c] Coefficient of variation is the average of the synonymous (K_s) or nonsynonymous (K_a) substitution rate divided by the standard deviation of the synonymous (K_s) or nonsynonymous (K_a) substitution rate for all orthologs shared between strains.
[d] Average phyletic distribution parameter for all orthologs shared between strains.

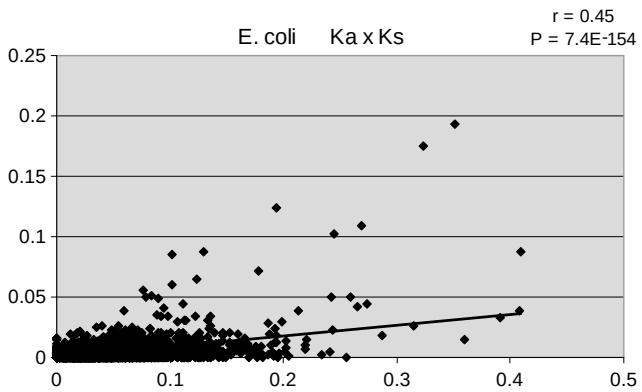

FIG. 5. Correlation between K_a (*x*-axis) and K_s (*y*-axis) for orthologs shared between strains of *E. coli*.

distribution parameter, and are thus more conserved, than nonessential genes.

A dense sampling of orthologous genes allowed us to unequivocally demonstrate that essential *E. coli* genes are more conserved than nonessential genes (Jordan *et al.*, in press). Previous studies on much smaller numbers of eukaryotic genes were unable to detect such a difference (Hurst and Smith, 1999; Hirsh and Fraser, 2001). It is not clear whether the findings here reflect a molecular evolutionary difference between eukaryotes and bacteria or merely result from the increased resolution afforded by a genomic-scale approach.

Conclusion

The ever-increasing availability of multiple complete genome sequences from very closely related species, and even strains of the same species, represents a potential boon for molecular evolutionists. The use of such data here allowed for the systematic analysis of several fundamental questions about bacterial genome evolution. It is our hope that, in addition to the empirical results reported here, this study may be of some use as a heuristic for future attempts at microevolutionary genomics.

MATERIALS AND METHODS

Ortholog Identification

Complete bacterial genome sequences (gene and protein) were downloaded from the National Center for Biotechnology (NCBI) ftp server (ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria). All-against-all BLAST (Altschul *et al.*, 1990, 1997) searches between the predicted protein sequences from intraspecific complete genomes were performed to identify orthologs shared between strains of the same species. The SEALS suite of programs (Walker and Koonin, 1997) was used to run BLAST in batch mode and to post-process the results of all BLAST searches. Proteins that were reciprocal best hits in BLAST searches with an lscore cut-off > 0.7 were taken as orthologs. Sets of orthologous genes were refined by examining the distribution of K_s (determined as described below). The distribution of K_s for orthologs of each species is approximately normal with a sparsely populated right tail of relatively high K_s values (data not shown). Genes with anomalously high K_s were excluded from further analysis. This step resulted in the elimination of small fraction (0–1.3%) of the total set of orthologous genes identified for each species.

Determination of Sequence Variation

The amino acid sequences of orthologus protein pairs were aligned using ClustalW (Higgins *et al.*, 1996) with default options. ClustalW was run under the SEALS environment to allow for the execution of numerous simultaneous sequence alignments. Nucleotide sequences were aligned to correspond to the amino acid sequence alignments (maintain reading frame) using the SEALS package. Levels of amino acid and nucleotide percent sequence identity for orthologous pairs were determined using the program ORTHOLOG_ID (IKJ, unpublished, available on request). Levels of K_a and K_s were determined using the PBLTEST program (IBR, unpublished, available on request). This program is a modification of the Li93 program and calculates substitution rates based on the Pamilo–Bianchi–Li method (Li, 1993; Pamilo and Bianchi, 1993).

REFERENCES

Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., *et al.* 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*, *Nature* **397**, 176–180.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool, *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and

PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res*. **25**, 3389–3402.

Aravind, L., Watanabe, H., Lipman, D. J., and Koonin, E. V. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes, *Proc. Natl. Acad. Sci. USA* **97**, 11,319–11,324.

Blattner, F. R., Plunkett 3rd, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al*. 1997. The complete genome sequence of *Escherichia coli* K-12, *Science* **277**, 1453–1474.

Braun, E. L., Halpern, A. L., Nelson, M. A., and Natvig, D. O. 2000. Large-scale comparison of fungal sequence information: Mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*, *Genome Res*. **10**, 416–430.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al*. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*, *Science* **273**, 1058–1073.

Doolittle, W. F. 1999. Lateral genomics, *Trends Cell Biol*. **9**, M5–8.

Duret, L., and Mouchiroud, D. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate, *Mol. Biol. Evol*. **17**, 68–74.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins, *Syst. Zool*. **19**, 99–113.

Fitch, W. M. 2000. Homology a personal view on some of the problems, *Trends Genet*. **16**, 227–231.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., *et al*. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science* **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., *et al*. 1995. The minimal gene complement of *Mycoplasma genitalium*, *Science* **270**, 397–403.

Futuyma, D. J. 1986. "Evolutionary Biology," Sinaauer Associates, Sunderland, MA.

Grishin, N. V., Wolf, Y. I., and Koonin, E. V. 2000. From complete genomes to measures of substitution rate variability within and between proteins, *Genome Res*. **10**, 991–1000.

Higgins, D. G., Thompson, J. D., and Gibson, T. J. 1996. Using CLUSTAL for multiple sequence alignments, *Methods Enzymol*. **266**, 383–402.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., and Herrmann, R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*, *Nucleic Acids Res*. **24**, 4420–4449.

Hirsh, A. E., and Fraser, H. B. 2001. Protein dispensability and rate of evolution, *Nature* **411**, 1046–1049.

Hughes, A. L. 1999. "Adaptive Evolution of Genes and Genomes," Oxford Univ. Press, Oxford, UK.

Hughes, M. K., and Hughes, A. L. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*, *Mol. Biol. Evol*. **10**, 1360–1369.

Hurst, L. D., and Smith, N. G. 1999. Do essential genes evolve slowly?, *Curr. Biol*. **9**, 747–750.

Huynen, M. A., and Bork, P. 1998. Measuring genome evolution, *Proc Natl. Acad. Sci. USA* **95**, 5849–5856.

Jordan, I. K., Kondrashov, F. A., Rogozin, I. B., Tatusov, R. L., Wolf, Y. I., and Koonin, E. V. 2001a. Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins, *Genome Biol*. **2**, 53.51–53.59.

Jordan, I. K., Makarova, K. S., Spouge, J. L., Wolf, Y. I., and Koonin, E. V. 2001b. Lineage-specific gene expansions in bacterial and archaeal genomes, *Genome Res*. **11**, 555–565.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. 2002. Essential genes are more evolutionarily conserved than non-essential genes in bacteria, *Genome Res*., in press.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R. W., Olinger, L., Grimwood, J., Davis, R. W., and Stephens, R. S. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*, *Nat. Genet*. **21**, 385–389.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., *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, *DNA Res*. **3**, 109–136.

Karlin, S., and Mrazek, J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes, *J. Bacteriol*. **182**, 5238–5250.

Kimura, M. 1983. "The Neutral Theory of Molecular Evolution," Cambridge Univ. Press, New York, NY.

Kolsto, A. B. 1997. Dynamic bacterial genome organization, *Mol. Microbiol*. **24**, 241–248.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. 2002. Selection in the evolution of gene duplications, *Genome Biol*. **3**, 8.1–8.9.

Koonin, E. V., and Galperin, M. Y. 1997. Prokaryotic genomes: The emerging paradigm of genome-based microbiology, *Curr. Opin. Genet. Dev*. **7**, 757–763.

Koonin, E. V., Makarova, K. S., and Aravind, L. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification, *Annu. Rev. Microbiol*. **55**, 709–742.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y., and Walker, D. R. 1997. Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea, *Mol. Microbiol*. **25**, 619–637.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution, *J. Mol. Evol*. **36**, 96–99.

Li, W. H. 1997. "Molecular Evolution," Sinauer Associates, Sunderland, MA.

Lynch, M., and Conery, J. S. 2000. The evolutionary fate and consequences of duplicate genes, *Science* **290**, 1151–1155.

Makalowski, W., and Boguski, M. S. 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences, *Proc. Natl. Acad. Sci. USA* **95**, 9407–9412.

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I., and Koonin, E. V. 1999. Comparative genomics of the Archaea (Euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell, *Genome Res*. **9**, 608–628.

Mushegian, A. R., and Koonin, E. V. 1996. Gene order is not conserved in bacterial evolution, *Trends Genet*. **12**, 289–290.

Ochman, H., Elwyn, S., and Moran, N.A. 1999. Calibrating bacterial evolution, *Proc. Natl. Acad. Sci. USA* **96**, 12,638–12,643.

Ohno, S. 1970. "Evolution by Gene Duplication," Springer-Verlag, New York, NY.

Ohta, T. 1992. The nearly neutral theory of molecular evolution, *Annu. Rev. Ecol. Syst*. **23**, 263–286.

Ohta, T., and Ina, Y. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences, *J. Mol. Evol.* **41**, 717–720.

Pamilo, P., and Bianchi, N. O. 1993. Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes, *Mol. Biol. Evol.* **10**, 271–281.

Parkhill, J., Achtman, M., James, K. D., Bentley, S. D., Churcher, C., Klee, S. R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., *et al.* 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491, *Nature* **404**, 502–506.

Perna, N. T., Plunkett 3rd, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., *et al.* 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7, *Nature* **409**, 529–533.

Read, T. D., Brunham, R. C., Shen, C., Gill, S. R., Heidelberg, J. F., White, O., Hickey, E. K., Peterson, J., Utterback, T., Berry, K., *et al.* 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39, *Nucleic Acids Res.* **28**, 1397–1406.

Riley, M. 1993. Functions of the gene products of *Escherichia coli*, *Microbiol. Rev.* **57**, 862–952.

Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. 1998. Genomic evidence for two functionally distinct gene classes, *Proc. Natl. Acad. Sci. USA* **95**, 6239–6244.

Shields, D. C., Sharp, P. M., Higgins, D. G., and Wright, F. 1988. "Silent" sites in Drosophila genes are not neutral: Evidence of selection among synonymous codons, *Mol. Biol. Evol.* **5**, 704–716.

Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba, T., Ishii, K., Hattori, M., Kuhara, S., *et al.* 2000. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA, *Nucleic Acids Res.* **28**, 2311–2314.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Res.* **28**, 33–36.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. 1997. A genomic perspective on protein families, *Science* **278**, 631–637.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.* **29**, 22–28.

Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, A. C., Nelson, K. E., Eisen, J. A., Ketchum, K. A., Hood, D. W., Peden, J. F., Dodson, R. J., *et al.* 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58, *Science* **287**, 1809–1815.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., *et al.* 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*, *Nature* **388**, 539–547.

Trevors, J. T. 1997. Evolution of bacterial genomes, *Antonie Van Leeuwenhoek* **71**, 265–270.

Walker, D. R., and Koonin, E. V. 1997. SEALS: A system for easy analysis of lots of sequences, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 333–339.

Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. 1997. Genome plasticity as a paradigm of eubacteria evolution, *J. Mol. Evol.* **44**, S57–S64.

Wilson, A. C., Carlson, S. S., and White, T. J. 1977. Biochemical evolution, *Annu. Rev. Biochem.* **46**, 573–639.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L., and Koonin, E. V. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades, *BMC Evol. Biol.* **1**, 8.

Wolfe, K. H., and Sharp, P. M. 1993. Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat, *J. Mol. Evol.* **37**, 441–456.