J Mol Evol (1991) 33:241–250

Journal of Molecular Evolution

Molecular Considerations in the Evolution of Bacterial Genes

Jeffrey G. Lawrence, Daniel L. Hartl, and Howard Ochman

Department of Genetics, Box 8232, Washington University School of Medicine, St. Louis, MO 63110, USA

Summary. Synonymous and nonsynonymous substitution rates at the loci encoding glyceraldehyde-3-phosphate dehydrogenase (*gap*) and outer membrane protein 3A (*ompA*) were examined in 12 species of enteric bacteria. By examining homologous sequences in species of varying degrees of relatedness and of known phylogenetic relationships, we analyzed the patterns of synonymous and nonsynonymous substitutions within and among these genes. Although both loci accumulate synonymous substitutions at reduced rates due to codon usage bias, portions of the *gap* and *ompA* reading frames show significant deviation in synonymous substitution rates not attributable to local codon bias. A paucity of synonymous substitutions in portions of the *ompA* gene may reflect selection for a novel mRNA secondary structure. In addition, these studies allow comparisons of homologous protein-coding sequences (*gap*) in plants, animals, and bacteria, revealing differences in evolutionary constraints on this glycolytic enzyme in these lineages.

Key words: Enteric bacteria — Codon usage bias — Synonymous substitution — Glyceraldehyde-3-phosphate dehydrogenase — Outer membrane protein 3A

Introduction

For the past 30 years, much of the research in molecular evolution has focused upon the comparisons of homologous genes across diverse taxa (Zuckerkandl and Pauling 1962; Wilson et al. 1977; Woese et al. 1990; Li and Graur 1991). In addition to providing a wealth of information on the arrangement, interaction, and regulation of genes, these studies have provided insights into the phylogenetic relationships among organisms and the patterns of change in the molecules themselves.

Nucleotide substitutions within protein-coding regions are generally divided into two classes: nonsynonymous substitutions, whereby changes in the nucleotide sequence alter the amino acid sequence, and synonymous substitutions, which do not alter the amino acid sequence. Nonsynonymous substitution rates vary among genes due to functional constraints. Proteins under little selective pressure evolve at higher rates relative to proteins under intense structural or functional constraint. Similarly, different domains of proteins may experience variable selective pressures. Regions coding for signal peptides evolve rapidly whereas enzyme active sites, ligand binding regions, and other structurally important domains evolve relatively slowly (Fitch 1976; DuBose and Hartl 1990).

Although synonymous substitutions in coding regions do not alter the amino acid sequence and are often considered selectively neutral, or nearly so, many synonymous sites do not evolve as quickly as noncoding DNA and pseudogenes (Li et al. 1985), implying the existence of selective constraints on changes at these sites (Ikemura 1985; Li et al. 1985). In bacteria, several factors may affect the evolution of synonymous sites, including codon usage bias and position on the chromosome. In the case of codon usage bias, highly expressed genes utilize a set of preferred codons (Gouy and Gautier 1982; Ikemura 1985), and selection on codon choice reduces the overall rate of synonymous site evolution among these genes (Sharp and Li 1987a). Sharp et al. (1989) have also demonstrated that the distance of a gene from the bacterial origin of replication is correlated

Offprint requests to: H. Ochman

with its synonymous substitution rate. However, not all of the variation in the rate of synonymous site evolution can be correlated with codon usage bias and map position, as selection on nucleotide sequences may involve other factors (Bossi 1983; Liljenström and von Heijne 1987; Bulmer 1988; Lawrence and Hartl 1991).

Taxa-specific factors may also influence evolutionary rates. Due to variation in population sizes, generation times, spontaneous mutation rates, and overall genetic organization, selective and stochastic processes may act differentially across taxa and influence their long-term rates of evolution. There is evidence that genes evolve more quickly in rodents that in primate lineages, a phenomenon ascribed to the number of cell division cycles (Wu and Li 1985; Li and Tanimura 1987; Li et al. 1990). In addition, Ochman and Wilson (1987a,b) observed that ratios in the accumulation of synonymous and nonsynonymous substitutions varied between bacteria and mammals. Mammalian lineages have, on average, fourfold fewer changes at nonsynonymous sites than bacterial lineages, and this disparity between bacterial and mammalian genes was attributed to several sources: (1) Selection on haploid genomes may be more intense as mutations cannot be maintained through overdominance. (2) Bacterial populations are large and, therefore, less sensitive to the fixation of mildly deleterious alleles by genetic drift. (3) Observed differences could arise from factors not directly related to the biology of the organisms, that is, the class of loci examined in bacteria is not homologous to those analyzed in mammals.

To fully evaluate the evolutionary forces acting within and among genes, we have compared homologous genes among species of varying degrees of relatedness. Comparisons involving multiple taxa allow the detailed investigation of evolutionary events at numerous timepoints and show the patterns of substitutions through time. We have examined nucleotide sequence variation at the *gap* locus, encoding glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and the *ompA* locus, encoding outer membrane protein 3A (OmpA), in 12 species of enteric bacteria. Rates of synonymous and nonsynonymous substitutions at these loci were utilized to investigate variation of selective constraint at synonymous and nonsynonymous sites at each locus. Analysis of the conserved glycolytic enzyme GAPDH within plants, vertebrates, and bacteria also allows comparisons of evolutionary rates and mechanisms in these disparate lineages.

Materials and Methods

Loci. The *gap* locus encodes GAPDH and is located at 39.3 min on the *Escherichia coli* linkage map (Bachmann 1990). The *ompA* locus encodes OmpA and is located at 21.8 min on the linkage map (Bachmann 1990). Portions of *gap* (nucleotides 45–927 of the *E. coli* sequence) and *ompA* (nucleotides 298–1011 of the *E. coli* sequence) from isolates of *Escherichia blattae, E. coli, Escherichia fergusonii, Escherichia hermannii, Escherichia vulneris, Salmonella typhimurium, Citrobacter freundii, Klebsiella pneumonia, Enterobacter aerogenes, Serratia marcescens,* and *Serratia odorifera* were amplified by the polymerase chain reaction (Saiki et al. 1985, 1988) and the nucleotide sequences of both strands of each template were determined. Gene phylogenies based on these data (Lawrence et al. 1991) were constructed by maximum parsimony utilizing the PAUP software package (version 2.4.1, D. Swofford) and tested using the method of Felsenstein (1985).

Sequence Divergence. Each pairwise sequence divergence was calculated and corrected by the method of Perler et al. (1980) using the GCG program package (Devereux et al. 1984). Nucleotide sequences from multiple taxa were then used to define composite sequence divergences for each gene in the following manner:

1) The numbers of potential synonymous and nonsynonymous changes were tabulated for each codon position within each gene; potential changes for codon positions comprising more than one codon class were calculated as a weighted average of potential changes for each class. For example, while the proline-encoding CCC codon allows three potentially synonymous changes at the third position, the histidine-encoding CAC codon allows a single synonymous change. Therefore, codon positions within the *gap* or *ompA* loci that have experienced a CCC to CAC substitution are scored, on average, as allowing two synonymous changes.
2) The numbers of synonymous and nonsynonymous substitutions were tabulated for each codon position in a cladistic manner, utilizing the known phylogenetic relationships among the species. When multiple taxa share a polymorphism, it is scored as a single substitution if the change is identical by descent. Conversely, it is scored as multiple substitutions if the changes are convergent, i.e., identical only in state.
3) Composite divergences were defined as the total number of substitutions within six-codon windows divided by the total number of potential changes for that region. Six-codon windows were selected because they comprised the minimum set of codons containing at least one synonymous substitution in each window of the *ompA* gene. In this manner, a set of composite divergences for overlapping six-codon windows spanning the length of the sequenced region for each gene was generated.

Splines. To define those regions of a gene containing aberrant patterns of nucleotide substitutions, cubic splines were fit to the composite nucleotide divergences for overlapping six-codon windows of each gene utilizing the algorithm of Schluter (1988). This analysis provides an estimator of the data as a series of cubic polynomials, allowing the elucidation of relationships not discernible by conventional linear regression analyses. In addition, splines are of utility to data sets that are autocorrelated (the data are not independent), as is the case with overlapping six-codon windows. The complexity of the estimator is controlled by the smoothing parameter, whose value was chosen as that which maximized the increase in predictability of the spline, as measured by the square of the correlation coefficient, relative to its increase in complexity. Cubic splines were fit to 1000 bootstrapped data sets to establish 95% confidence limits for each composite divergence.

Additional Calculations. Values for the codon adaptation index (CAI)—a measure of codon usage bias—were calculated over

Table 1. Pairwise nucleotide divergences at synonymous (lower diagonal) and nonsynonymous (upper diagonal) sites at the *gap* locus

	Eco	Efe	Sty	Cfr	Ehe	Ev7	Eae	Kpn	Ev6	Ebl	Sma	Sod
Eco	—	0.8	1.2	2.0	3.4	3.6	4.7	3.3	3.1	4.1	5.8	8.1
Efe	19.6	—	0.9	2.3	2.8	3.1	4.1	2.7	2.6	4.4	4.9	7.1
Sty	24.9	27.1	—	2.5	3.0	3.2	4.4	3.0	2.6	4.9	5.4	7.5
Cfr	29.3	28.8	29.2	—	3.1	2.0	3.7	2.5	2.2	3.5	5.0	7.9
Ehe	37.5	30.8	35.3	30.7	—	3.0	3.3	2.8	2.6	5.6	4.5	7.1
Ev7	42.6	36.7	39.9	34.1	27.5	—	3.0	3.0	2.8	5.0	4.8	6.9
Eae	49.0	48.3	52.3	46.3	42.7	36.9	—	2.0	1.9	4.9	4.5	7.0
Kpn	29.2	33.0	32.6	29.2	30.0	28.0	40.3	—	0.2	3.3	4.2	7.1
Ev6	30.6	33.1	31.9	26.3	29.1	27.3	39.6	9.0	—	3.2	4.2	6.8
Ebl	26.8	36.0	39.7	31.1	35.6	30.1	46.0	19.2	22.2	—	4.9	8.1
Sma	59.8	67.1	62.0	56.4	49.2	52.2	54.6	38.3	42.3	44.0	—	6.2
Sod	62.4	57.3	63.8	55.5	47.8	51.5	56.4	46.0	49.3	45.4	42.4	—

Eco—*Escherichia coli* K12, Efe—*E. fergusonii* ATCC 35469, Ev6—*E. vulneris* ATCC 29943, Ev7—*E. vulneris* ATCC 33821, Ebl—*E. blattae* ATCC 29907, Ehe—*E. hermannii* ATCC 33650, Sty—*Salmonella typhimurium* LT2, Cfr—*Citrobacter freundii* 0860, Eae—*Enterobacter aerogenes* E482, Kpn—*Klebsiella pneumonia*, Sma—*Serratia marcescens* ATCC 13880, Sod—*S. odorifera* ATCC 3307. Pairwise divergences calculated and corrected for multiple substitutions by the method of Perler et al. (1980)

Table 2. Pairwise nucleotide divergences at synonymous (lower diagonal) and nonsynonymous (upper diagonal) sites at the *ompA* locus

	Eco	Efe	Sty	Cfr	Ehe	Ev7	Eae	Kpn	Ev6	Ebl	Sma	Sod
Eco	—	0.8	3.8	4.5	5.0	4.9	7.9	7.9	9.8	7.4	16.4	18.7
Efe	14.4	—	3.8	4.5	4.2	7.4	7.8	7.7	9.4	7.8	16.9	19.4
Sty	35.4	34.7	—	2.7	3.2	5.7	6.4	8.0	9.1	7.4	16.6	20.0
Cfr	37.1	35.6	36.5	—	2.8	5.0	6.3	7.7	8.8	8.7	16.4	20.4
Ehe	35.6	37.9	42.0	37.0	—	4.2	5.3	6.1	8.1	8.0	16.2	20.3
Ev7	37.9	40.4	43.4	45.5	37.0	—	5.6	6.7	7.0	9.0	17.5	21.2
Eae	35.8	41.0	41.9	30.9	39.4	45.2	—	6.8	6.4	8.9	16.8	20.0
Kpn	38.6	44.8	43.4	37.2	40.6	47.5	47.8	—	4.7	7.8	13.5	16.8
Ev6	35.2	49.0	49.4	50.0	44.4	47.2	36.9	36.7	—	9.9	14.8	18.6
Ebl	51.4	41.9	60.1	60.3	48.4	55.4	56.2	52.7	51.7	—	15.1	17.7
Sma	49.7	55.5	71.5	59.6	60.8	62.1	51.4	71.0	64.5	69.7	—	7.0
Sod	69.9	70.3	81.5	79.9	78.3	75.6	68.0	84.5	83.8	89.6	41.3	—

Taxa designated as in Table 1. Pairwise divergences calculated and corrected for multiple substitutions by the method of Perler et al. (1980)

the six-codon windows using the method of Sharp and Li (1987b). Folding energy of RNA secondary structures was computed by the method of Zuker and Stiegler (1981), utilizing the GCG program package (Devereux et al. 1984).

Results and Discussion

Sequence Evolution in High Codon Bias Genes

The nucleotide sequences of the *gap* and *ompA* loci, encoding the glycolytic enzyme GAPDH and OmpA, have been determined for 12 species of enteric bacteria. Maximum parsimony analysis of these genes produced congruent and robust phylogenetic trees (Lawrence et al. 1991). The relationships among the taxa are consistent with those inferred from genotypic and phenotypic data (Brenner and Falkow 1971; Cocks and Wilson 1972; Ochman and Wilson 1987b). Because the phylogenies constructed from DNA sequences from these two genes were congruent, they are taken to represent the actual relationships among the taxa (Lawrence et al. 1991). Pairwise divergences for both genes at synonymous and nonsynonymous sites are presented in Tables 1 and 2. When compared to other loci sequenced in enteric bacteria (Ochman and Wilson 1987a; Sharp and Li 1987a), it is apparent that *gap* and *ompA* are evolving slowly at nonsynonymous sites, reflecting selective constraints on the encoded protein, and at synonymous sites, produced in part by selection on codon choice.

Comparisons of a number of homologous genes sequenced in both *E. coli* and *S. typhimurium* reveal that nucleotide divergence at synonymous sites approaches saturation for typical chromosomal loci (Ochman and Wilson 1987a); that is, each site has experienced an average of at least one substitution. Highly expressed genes experience selection at synonymous sites that is reflected in codon usage bias (Gouy and Gautier 1982; Sharp and Li 1987b), and

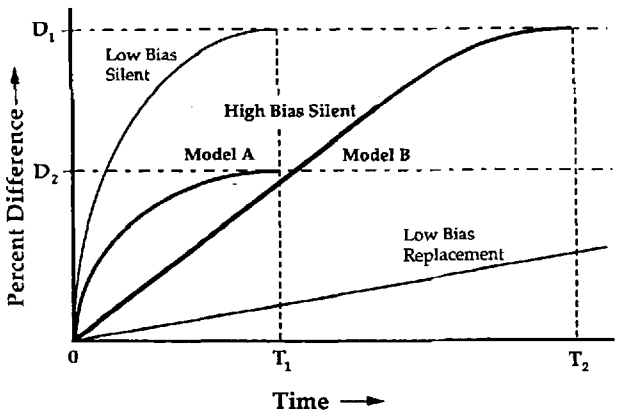

Fig. 1. Effect of codon usage bias on the rate of synonymous site substitution. See text for explanation of models.

these genes have a lower rate of synonymous substitution (Sharp and Li 1987a). Among all bacterial genes for which comparisons are possible, the indices of codon usage bias for *gap* and *ompA* for the 12 enteric taxa were among the highest [CAI (Sharp and Li 1987b) ranging from 0.69 to 0.83 for *gap* and from 0.73 to 0.82 for *ompA* (Lawrence et al. 1991)], and their synonymous substitution rates were among the lowest (Tables 1 and 2). This variability in synonymous substitution rates suggests that each class of bacterial gene may have a different molecular clock. To determine the utility of these molecules for evolutionary studies, and to examine the manner in which genes with high codon usage bias might evolve, we have investigated the accumulation of synonymous substitutions in these genes.

Although bias in codon usage is correlated with synonymous substitution rate, it has been noted that selection for codon choice may not be uniform, or uniformly effective, across coding regions (Liljenström and von Heijne 1987; Bulmer 1988; Lawrence and Hartl 1991). Therefore, synonymous substitution rates could differ within genes, and mutations at different codon positions may be fixed at different rates. In this case, certain sites would quickly accumulate substitutions, and pairwise comparisons of these genes from distantly related taxa would yield the same degree of sequence difference at synonymous sites (model A in Fig. 1). By this model, the extent of sequence divergence in low and high bias genes would reach a maximum in the same period of time. Alternatively, all synonymous sites may be evolving more slowly in highly expressed genes, and local variation in codon usage bias may not be reflected in differences in substitution rates (model B of Fig. 1). Tso et al. (1985) originally proposed that highly expressed genes may harbor two classes of synonymous sites, with one fraction accumulating substitutions at a rate comparable to nonsynonymous sites. To distinguish these models, it is necessary to make pairwise comparisons of genes from species of various degrees of relationship.

Pairwise comparisons of low (*pab*A and *trp*D) and high (*gap* and *ompA*) bias genes among species of enteric bacteria (Table 3) clearly support model B, whereby high-bias genes continue to accumulate synonymous substitutions after low-bias genes have reached saturation. For high-bias genes, more distantly related taxa grow more dissimilar at both synonymous and nonsynonymous sites. (Low-bias genes evolve more rapidly at synonymous sites, and more distantly related taxa do not become substantially more dissimilar.) Assuming that genes maintain similar selection for codon choice in these species, which has been demonstrated in *S. marcescens* (Sharp 1990), these results imply that these genes do not harbor distinct classes of synonymous sites.

Intragenic Variation in Substitution Rates

Although the analysis of synonymous site evolution does not reveal distinct classes of sites, it is not clear whether the rate of evolution at these sites is uniform across the coding region. The nucleotide sequences of the *gap* and *ompA* loci from 12 species of enteric bacteria allow fine-scale investigation of substitution rates at individual regions of these genes. By simultaneously considering the nucleotide sequences from all 12 species, and the phylogenetic relationships between the taxa, one can estimate the number of substitutions that have occurred at each codon over time. Utilizing the evolutionary relationships of the species minimizes problems associated with multiple substitutions inherent to two-taxa comparisons. Composite divergences of six codon windows across the *ompA* gene are presented in Fig. 2B. Not surprisingly, divergences at both synonymous and nonsynonymous sites vary with respect to position within the gene. Nonsynonymous substitutions are often nonrandomly distributed across the gene, as different regions of the encoded protein experience distinct structural and functional constraints (DuBose and Hartl 1990). Although intragenic variation in the distribution of substitutions at nonsynonymous sites can be interpreted as functional or structural constraints on the encoded protein, it is difficult to interpret variation in synonymous substitution rates in the same terms. Although the overall degree of codon usage bias for individual genes is associated with its synonymous substitution rate, rates of local synonymous substitution were not correlated with local codon usage bias ($r^2 = 0.002$; Fig. 2A) and not significantly correlated with local nonsynonymous substitution rates ($r^2 = 0.28$).

Because linear regression does not provide adequate analysis of the data, cubic splines—series of

Table 3. Accumulation of differences at synonymous and nonsynonymous sites among species of enteric bacteria

Species[a]	Time[b]	*trp*D		*pab*A		*gap*		*omp*A	
		S[c]	R	S	R	S	R	S	R
Shigella dysenteriae	25	14.1	0.4	—	—	—	—	—	—
Escherichia fergusonii	70	—	—	—	—	12.7	0.6	8.6	0.7
Salmonella typhimurium	140	44.7	2.3	39.6	7.9	17.2	0.8	20.6	3.4
Klebsiella pneumonia	160	—	—	45.1	11.5	18.2	2.7	25.3	6.5
Serratia marcescens	200	48.1	12.2	43.9	15.2	29.5	4.3	32.2	13.1
Morganella morganii	250	—	—	—	—	39.2	6.0	—	—

[a] Nucleotide sequences obtained from GenBank version 62.0. References: *trp*D: Nichols et al. (1980); *pab*A: Kaplan and Nichols (1983), Kaplan et al. (1985); *gap*: Branlant and Branlant (1985), Lawrence et al. (1991), Ochman (unpublished results); *omp*A: Beck and Bremer (1980), Braun and Cole (1984), Freudl and Cole (1983), Lawrence et al. (1991). GenBank accession numbers are as follows: *trp*D: J01714, J01787, J01811, J01792, *pab*A: K00030, X02603, X02604, X02605; *gap*: X02662, M63368, M63369, M63371, M63369; *omp*A: J01654, M63353, X02006, M63355, X00618

[b] Time since divergence of each species from *Escherichia coli*. Values for *S. dysenteriae*, *S. typhimurium*, and *S. marcescens* were taken from Ochman and Wilson (1987a,b); additional values were interpolated from Fig. 1 in Lawrence et al. (1991)

[c] S: Average percent difference at synonymous sites. R: Average percent difference at nonsynonymous sites. Values were calculated by the method of Perler et al. (1980) without correction for multiple substitutions, and averaged for multiple taxa (e.g., the *S. marcescens* values for the *gap* and *omp*A loci represent averages of the *E. coli*/*S. marcescens*, *S. typhimurium*/*S. marcescens*, and *K. pneumonia*/*S. marcescens* comparisons)

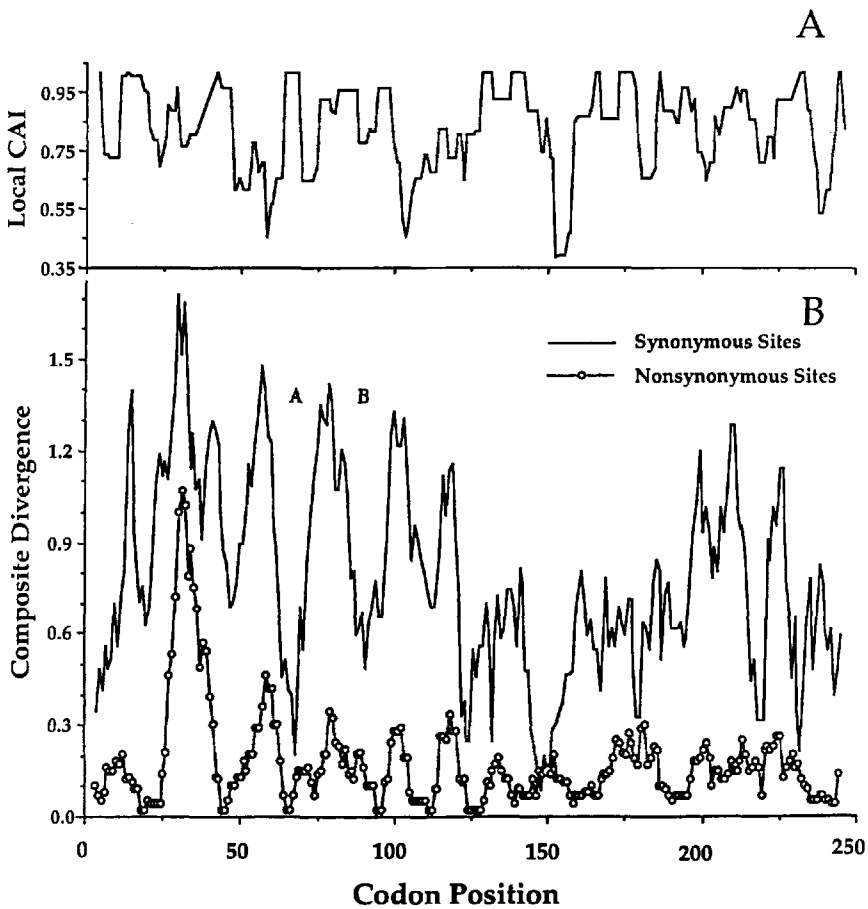

Fig. 2. Molecular evolution at the *omp*A locus. Values for the codon adaptation index (CAI, Sharp and Li 1987b) and composite synonymous and nonsynonymous site divergences were determined for six-codon windows as described in the Materials and Methods section. Codon position 1 corresponds to codon 99 of the *E. coli* sequence (Beck and Bremer 1980). Subsequent numbering follows that designated in Lawrence et al. (1991). 'A' and 'B' denote the shaded regions of Fig. 4.

polynomial estimators fit to portions of the data set—were fit to these composite divergences. Descriptive statistics for a family of splines are presented in Table 4. The complexity of the estimator is controlled by the smoothing parameter, and a value was chosen (−1.0) that maximized the increase in predictability of the spline, as measured by r^2, relative to its increase in complexity, as shown in Fig. 3A. (Splines of greater complexity did not appreciably increase the predictability of the model,

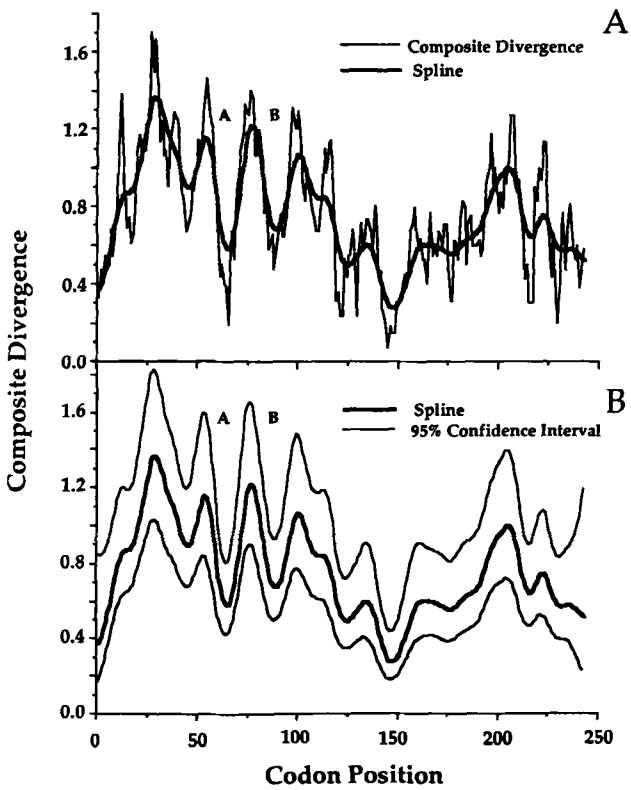

Fig. 3. A Cubic spline of composite synonymous substitution rates calculated for six-codon windows over the *ompA* locus. Codon positions correspond to those designated in Fig. 2. **B** Confidence limits of the spline estimator established from splines fit to 1000 bootstrapped *ompA* composite divergence data sets. 'A' and 'B' denote the shaded regions of Fig. 4.

whereas splines of decreasing complexity lost disproportionate amounts of predictability.)

To test if synonymous substitution rate is uniform across the *ompA* coding region (i.e., if synonymous substitutions are distributed at random across the coding sequence) cubic splines were fit to 1000 bootstrapped data sets to establish 95% confidence intervals (Fig. 3B). It is apparent from these confidence intervals that the synonymous substitution rate along the *ompA* coding region is not uniform and that synonymous substitutions do not occur at random over the length of this molecule. In particular, the rates of synonymous substitution are significantly higher in the regions surrounding codons 30, 80, and 105, and significantly lower rates are detected in the regions surrounding codons 68, 90, and 145. More specifically, any spline estimating the data must pass through maxima and minima at these codon positions. Because such maxima and minima exist, synonymous substitution rate is not uniform across this portion of the *ompA* coding region. These regions are also significantly variant utilizing less complex splines, and it is unlikely that this distribution of synonymous substitutions arose by stochastic factors alone. Moreover, sets of randomly distributed codon positions do not reveal significant variation (data not shown).

Table 4. Descriptive statistics for cubic splines of different smoothing parameters

λ^a	Complexity[b]	r^{2c}	$\Delta r^2/\Delta C^d$
1.0	16.7	46.1	0.899
0.5	18.7	48.1	0.958
0.0	21.1	50.7	1.087
−0.5	23.8	53.9	1.215
−1.0	26.8	57.8	1.280
−1.5	30.2	62.1	1.258
−2.0	34.1	66.6	1.157
−2.5	38.4	71.0	1.010
−3.0	43.4	75.2	0.844

[a] The natural logarithm of the smoothing parameter, describing the complexity of the spline
[b] The complexity of the spline, analogous to polynomial order
[c] Square of the correlation coefficient of the spline, defined as the ratio of the variance of the predicted values to the variance of the data
[d] The ratio of increase in the square of the correlation coefficient to its increase in complexity

Changes at synonymous sites do not alter the encoded protein; therefore, selection at these sites is most likely operating at the level of nucleic acids. For example, autocatalytic genes must maintain particular nucleotide sequences to function (Green et al. 1986) and some RNA secondary structures affect gene expression (Keller and Calvo 1979; Oxender et al. 1979). To determine if features of the *ompA* message include energetically stable secondary structures, overlapping 200-bp regions were examined for stem-loop configurations. One such structure, shown in Fig. 4, was predicted to form along two regions experiencing low substitution rates (regions 'A' and 'B' of Figs. 2 and 3). Its energy of folding, −33.6 kcal/mol, was significantly lower than random sequences of identical nucleotide composition (−27.3 kcal/mol, $P = 0.04$). The homologous structures predicted from the 12 species of enteric bacteria were also thermodynamically stable, with a mean energy of folding of −35.8 kcal/mol. The encoded amino acid sequence for a portion of the structure (Ala–Pro–Val–Val–Ala–Pro–Ala–Pro–Ala–Pro–Ala–Pro) has been described as a "hinge" (Chen et al. 1980), analogous to proline-rich hinges in mammalian immunoglobulins (Edelman et al. 1969; Beale and Feinstein 1976; Liu et al. 1976; Huck et al. 1989).

The *ompA* protein is rich in β-sheet secondary structure (Nakamura et al. 1974; Nakamura and Mizushima 1976), and a β-barrel composed of eight antiparallel β-sheets was proposed to form in the N-terminal portion of the OmpA protein (Vogel and Jähnig 1986). The proline-rich hinge divides OmpA into two regions: the N-terminal β-barrel domain, inserted into the outer membrane, and the C-ter-

minal domain, residing in the periplasmic space. The peptide undergoes at least one conformational change following cleavage of the signal peptide prior to insertion in the outer membrane (Freudl et al. 1986), and an immature tertiary structure, the "molten globule" (Kuwajima 1989; Ptitsyn et al. 1990), has been implicated in facilitating the translocation of proteins across membranes (Bychkova et al. 1988). The mRNA secondary structure predicted from our analysis could induce ribosome pausing, similar to that detected for translation of the *trp* leader peptide (Oxender et al. 1979). This pause would allow a secondary structure of the precursor pro-OmpA protein, notably the transmembrane β-barrel, to fold properly. To test this hypothesis, in vitro mutagenesis could be utilized to create an *ompA* message lacking this secondary structure; however, the function of the OmpA protein is poorly understood and mutants have a variety of phenotypes (Manning et al. 1977; Nikaido et al. 1977; Freudl et al. 1985; Reid and Henning 1987), so it is unclear what phenotype such a mutant may exhibit. Although this particular RNA secondary structure may account for two regions of the DNA sequence displaying low rate of synonymous substitution, we can offer little explanation for the remaining portions of the *ompA* gene showing deviant rates of synonymous substitution. Moreover, it remains unclear how such a structure may affect translation. Although the role of RNA secondary structure in transcription attenuation is well characterized (for a review see Yanofsky 1988), preliminary results suggest that not all energetically stable RNA secondary structures influence rates of translation (Sørensen et al. 1989). However, such structures could affect translation by serving as protein-binding signals (MacDonald 1990), interfering with ribosome loading (Gross et al. 1990; Roy et al. 1990), or interacting directly with the ribosome.

Variation in substitution rates was examined across the *gap* gene in a similar manner. Splinal analysis of local composite divergences at synonymous and nonsynonymous sites revealed two regions, those around codon positions 175 and 245, which show significant deviation. The region surrounding codon position 175 immediately precedes the S-loop of GAPDH (Biesecker et al. 1977) and is characterized by an unusually low rate of synonymous substitution, comprising 10 codons with neither synonymous nor nonsynonymous change. Analysis of GAPDH-encoding sequences also reveals conservation of the amino acid sequences in this region among distantly related taxa (Conway et al. 1987). Although too small to be implicated in mRNA secondary structure, it is possible that this region harbors a DNA sequence important for the structure, function, or expression of this gene.

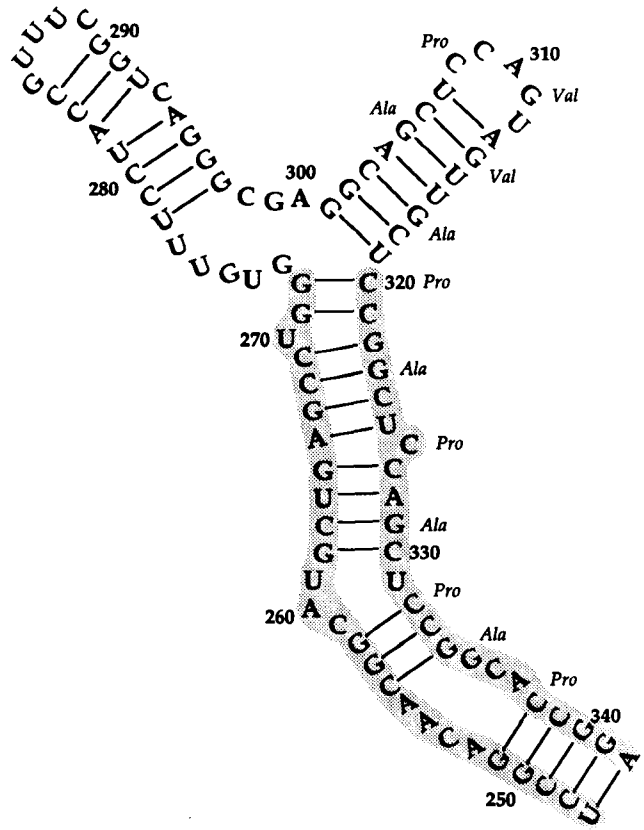

Fig. 4. Potential secondary structure of a region of the *ompA* mRNA. Nucleotides are numbered according to the *E. coli* DNA sequence (Beck and Bremer 1980); corresponding amino acid translations are included for the "hinge region." Shaded bases correspond to regions 'A' and 'B' of Fig. 2.

The region surrounding *gap* codon position 245 is characterized by significantly higher rates of both synonymous and nonsynonymous substitutions. The high rate of synonymous substitution in this region results from the numerous nonsynonymous substitutions in that region, followed by fixation of mutations to preferred codons. This region (Fig. 5) has experienced numerous nonconservative substitutions during the divergence of the enteric bacteria (10 of 23 amino acids conserved) relative to the flanking regions (25 of 25 and 19 of 20 amino acids conserved). The ability to tolerate numerous nonconservative changes, including substitutions involving proline and various charged residues, implies that this region is not of critical structural or functional importance. Not surprisingly, the corresponding region is also highly variable among GAPDH-encoding genes of other organisms (Conway et al. 1987).

Patterns and Rates of Sequence Evolution Across Taxa

GAPDH catalyzes the oxidative phosphorylation of glyceraldehyde-3-phosphate to 1,3-diphosphoglyc-

Table 5. Comparison of the numbers and rates of nucleotide substitutions per site for GAPDH from mammals, plants, and bacteria

	Divergence times[a] ($\times 10^6$ years)	K_a[b]	Rate ($\times 10^{-9}$)	K_s	Rate ($\times 10^{-9}$)	K_s/K_a
Mammals (*Homo–Rattus*)	60–90	0.03	0.16–0.25	0.41	2.3–3.4	13.7
Plants (*Hordeum–Zea*)	50–90	0.03	0.16–0.29	0.56	3.1–5.6	18.7
Bacteria (*Escherichia–Salmonella*)	120–160	0.01	0.03–0.04	0.30	0.9–1.3	30.0

[a] Estimates of divergence times after Ochman and Wilson (1987a) and Wolfe et al. (1989)

[b] K_a and K_s are the numbers of nonsynonymous and synonymous substitutions, respectively, as calculated by the method of Li et al. (1985). Rates are the numbers of substitutions per site per year and based on the range of divergence times shown

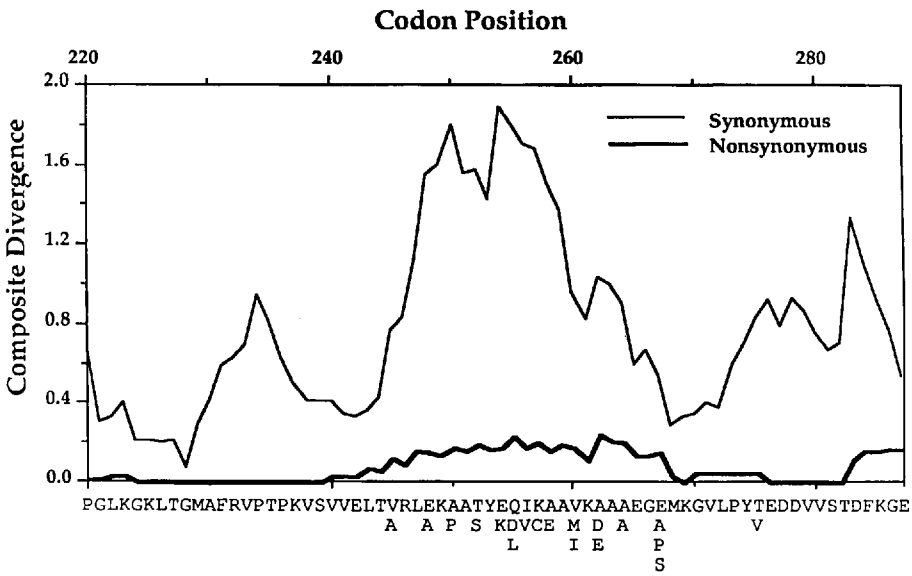

Fig. 5. Molecular evolution of a highly variable region in the bacterial *gap* locus. Values for composite synonymous and nonsynonymous site divergence were determined for six-codon windows as described in the Materials and Methods section. The amino acid sequence for the *E. coli* GAPDH is shown as well as amino acid substitutions among 12 species of enteric bacteria (Lawrence et al. 1991). Codon position 220 corresponds to codon 235 of the *E. coli* sequence (Branlant and Branlant 1985).

erate in glycolysis and is conserved among all lineages. Analysis of this gene allows comparisons of the rates and patterns of evolution of a homologous protein-coding region within bacteria, plants, and mammals. To provide consistency with the results of other investigators (Li et al. 1985; Martin et al. 1989), we employed the method outlined by Li et al. (1985) to examine the extent of sequence divergence of GAPDH from *E. coli* and *S. typhimurium*. Substitution rates at synonymous and nonsynonymous sites for GAPDH in mammals, plants, and bacteria, along with the estimated divergence times, are presented in Table 5. This table includes only those comparisons among genes where the extent of sequence divergence is certainly not in saturation and is limited to relatively closely related taxa. Given the range of divergence times shown in Table 5, there is overlap in the nonsynonymous and synonymous substitution rates of GAPDH in plants and mammals (Martin et al. 1989). These eukaryotic GAPDHs are evolving much more rapidly—about two to three times faster at synonymous sites and approximately five times faster at nonsynonymous sites—than the corresponding gene from *E. coli* and *S. typhimurium*. Because GAPDH is constitutively expressed at high levels, this lower synonymous substitution rate is presumably the result of constraints imposed by codon bias. Most genes in bacteria are not homologous to those analyzed from higher eukaryotes, and any incongruencies may be attributed to the data set rather than to systematic differences between the organisms. Although we have provided only a single example, these data suggest that population or genome-specific parameters contribute to the relatively reduced rate of evolution at nonsynonymous sites in the bacterial *gap* comparison.

For *gap*, the ratio of synonymous to nonsynonymous changes in the *E. coli* to *S. typhimurium* comparison is about twice that calculated for mammals or plants (Table 5). For both groups, the ratio for the GAPDH-encoding locus is much higher than that calculated for other genes within the taxa; K_s/K_a averages 5.0 in mammals and 20.0 in bacteria. There is a tendency for genes with low nonsynonymous substitution rates to have low substitution rates at synonymous sites (Li et al. 1985), and substitution rates for GAPDH at both synonymous and nonsynonymous sites are lower than those in most other genes within these same taxa (Ochman and Wilson 1987b; Sharp and Li 1987a; Martin et al.

1989). In enteric bacteria, the average divergence at synonymous sites is over three times that calculated for the *gap* locus. The ratio of synonymous to nonsynonymous substitutions in more distant comparisons of enteric bacteria approaches that observed in mammalian and plant GAPDH-encoding loci. Nonsynonymous substitution rates for bacterial *gap* genes increase almost fivefold in the *E. coli* to *S. marcescens* comparison using the estimated divergence times presented in Table 3. The decline in the ratio of synonymous to nonsynonymous substitutions can be attributed to multiple substitutions at synonymous sites as well as to an increase in compensatory nonsynonymous substitutions among distantly related taxa.

In conclusion, by utilizing nucleotide sequences of homologous genes among closely related taxa of known phylogenetic relationships, factors influencing the rate of evolution of bacterial genes can be investigated in greater scope and detail than is available in two-taxa comparisons. When nonsynonymous and synonymous substitutions are tabulated cladistically, that is when evolutionary events can be distinguished as being identical by descent or identical due to convergence, variation in evolutionary rates can be detected within protein-coding regions. Such analysis has revealed a novel mRNA secondary structure implicated in the proper expression of an outer membrane protein. In addition, this method is applicable to analyses focused upon nucleotide sequences of structural or functional importance.

Acknowledgments. We thank J. Carulli, J. Cheverud, C. Kurland, P. Green, D. Krane, and A. Templeton for helpful discussions, C.-I. Wu for aid with divergence calculations, M. Sturmoski for technical assistance, and P. Sharp for many helpful comments on the manuscript. This work was supported by grants GM 40322 (D.L.H.) and GM 40995 (H.O.) from the National Institutes of Health.

References

Bachmann BJ (1990) Linkage map of *Escherichia coli* K-12, ed 8. Microbiol Rev 54:130–197

Beale D, Feinstein A (1976) Structure and function of the constant regions of immunoglobins. Quart Rev Biophys 9:135–180

Beck E, Bremer E (1980) Nucleotide sequence of the gene *ompA* coding the outer membrane protein II* of *Escherichia coli* K-12. Nucleic Acids Res 8:3011–3024

Biesecker G, Harris JI, Thierry JC, Walker JE, Wonacott AJ (1977) Sequence and structure of D-glyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus*. Nature 266:328–333

Bossi L (1983) Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. J Mol Biol 164:73–87

Branlant G, Branlant C (1985) Nucleotide sequence of the *Escherichia coli gap* gene: different evolutionary behavior of the NAD^+-binding domain and of the catalytic domain of the D-glyceraldehyde-3-phosphate dehydrogenase. Eur J Biochem 150:61–66

Braun G, Cole ST (1984) DNA sequence analysis of the *Serratia marcescens ompA* gene: implication for the organisation of an enterobacterial outer membrane protein. Mol Gen Genet 195:321–328

Brenner DJ, Falkow S (1971) Molecular relationships among members of the Enterobacteriaceae. Adv Genet 16:81–118

Bulmer M (1988) Codon usage and intragenic position. J Theor Biol 133:67–71

Bychkova VE, Pain RH, Ptitsyn OB (1988) The 'molten globule' state is involved in the translocation of proteins across membranes? FEBS Lett 238:231–234

Chen R, Schmidmayr W, Kramer C, Chen-Schemeisser U, Henning U (1980) Primary structure of outer membrane protein II (*ompA* protein) of *Escherichia coli* K12. Proc Natl Acad Sci USA 77:4592–4596

Cocks GT, Wilson AC (1972) Enzyme evolution in the Enterobacteriaceae. J Bacteriol 110:793–802

Conway T, Sewell GW, Ingram LO (1987) Glyceraldehyde-3-phosphate dehydrogenase gene from *Zymomonas mobilis*: cloning, sequencing, and identification of promoter region. J Bacteriol 169:5653–5662

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

DuBose RF, Hartl DL (1990) The molecular evolution of alkaline phosphatase: correlating variation among enteric bacteria to experimental manipulations of the protein. Mol Biol Evol 7:547–577

Edelman GM, Cunningham BA, Gall WE, Gottlieb PD, Rutishauser U, Waxdal MJ (1969) The covalent structure of an entire γG immunoglobin molecule. Proc Natl Acad Sci USA 63:78–85

Felsenstein J (1985) Confidence limits on phylogenies with a molecular clock. Syst Zool 34:152–161

Fitch WM (1976) The molecular evolution of cytochrome c in eukaryotes. J Mol Evol 8:13–40

Freudl R, Cole ST (1983) Cloning and molecular characterization of the *ompA* gene from *Salmonella typhimurium*. Eur J Biochem 134:497–502

Freudl R, Braun G, Hindennach I, Henning U (1985) Lethal mutations in the structural gene of an outer membrane protein (OmpA) of *Escherichia coli* K-12. Mol Gen Genet 201:76–81

Freudl R, Schwarz H, Stierhof Y-D, Gamon K, Hindenach I, Henning U (1986) An outer membrane protein (OmpA) of *Escherichia coli* K-12 undergoes a conformational change during export. J Biol Chem 261:11355–11361

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Green PJ, Pines O, Inouye M (1986) The role of antisense RNA in gene regulation. Annu Rev Biochem 55:569–597

Gross G, Mielke C, Hollatz I, Blöcker H, Frank R (1990) RNA primary sequence or secondary structure in the translational initiation region controls expression of two variant interferon-β genes from *Escherichia coli*. J Biol Chem 265:17627–17636

Huck S, Lefrane G, Lefrane M-P (1989) A human immunoglobin *IGHG3* allele (*Gmbo, b1, c3, c5, u*) with an *IGHG4* converted region and three hinge exons. Immunogenetics 30:250–257

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–33

Kaplan JB, Nichols BP (1983) Nucleotide sequence of *Escherichia coli pabA* and its evolutionary relationship to the *trp*(G)D. J Mol Biol 168:451–468

Kaplan JB, Merkel WK, Nichols BP (1985) Evolution of the

glutamine amidotransferase genes: nucleotide sequences of the *pab*A genes from *Salmonella typhimurium, Klebsiella aerogenes,* and *Serratia marcescens.* J Mol Biol 183:327–340

Keller EB, Calvo JM (1979) Alternative secondary structures of leader operons and the regulation of the *trp, phe, thr,* and *leu* operons. Proc Natl Acad Sci USA 76:6186–6190

Kuwajima K (1989) The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins 6:87–103

Lawrence JG, Hartl DL (1991) Unusual codon usage bias occurring within insertion sequences in *Escherichia coli.* Genetica (in press)

Lawrence JG, Ochman H, Hartl DL (1991) Molecular and evolutionary relationships among enteric bacteria. J Gen Microbiol (in press)

Li W-H, Graur D (1991) Molecular evolution. Sinauer Associate, Sunderland MA

Li W-H, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–96

Li W-H, Wu C-I, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li W-H, Gouy M, Sharp PM, O'hUigin C, Yang Y-W (1990) Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. Proc Natl Acad Sci USA 87:6703–6707

Liljenström H, von Heijne G (1987) Translation rate modification by preferential codon usage: intragenic position effects. J Theor Biol 124:43–55

Liu Y-SV, Low TLK, Infante A, Putnam FW (1976) Complete covalent structure of a human IgA1 immunoglobin. Science 193:1017–1019

MacDonald PM (1990) *bicoid* mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. Development 110:161–171

Manning PA, Pugsley AP, Reeves P (1977) Defective growth functions in mutants of *Escherichia coli* K12 lacking a major outer membrane protein. J Mol Biol 116:285–300

Martin W, Gierl A, Saedler H (1989) Molecular evidence for pre-Cretaceous angiosperm origins. Nature 339:46–48

Nakamura K, Mizushima S (1976) Effects of heating in dodecyl sulfate solution on the conformation and electrophoretic mobility of isolated major outer membrane proteins from *Escherichia coli* K-12. J Biochem (Tokyo) 80:1411–1422

Nakamura K, Ostrovsky DN, Miyazawa T, Mizushima S (1974) Infrared spectra of outer and cytoplasmic membranes of *Escherichia coli.* Biochim Biophys Acta 332:329–335

Nichols BP, Miozzari GF, VanCleemput M, Bennett GN, Yanofsky C (1980) Nucleotide sequences of the trpG region of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium,* and *Serratia marcescens.* J Mol Biol 142:503–517

Nikaido H, Song SA, Shaltiel L, Nurminen M (1977) Outer membrane of *Salmonella.* XIV. Reduced transmembrane diffusion rates in porin deficient mutants. Biochem Biophys Res Commun 76:324–330

Ochman H, Wilson AC (1987a) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74–86

Ochman H, Wilson AC (1987b) Evolutionary history of enteric bacteria. In: Niedhardt FD (ed) *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. American Society of Microbiology, Washington DC, pp 1649–1654

Oxender DL, Zurawski G, Yanofsky C (1979) Attenuation in the *Escherichia coli* tryptophan operon: role of RNA secondary structure involving the tryptophan codon region. Proc Natl Acad Sci USA 76:5524–5528

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J (1980) The evolution of genes: the chicken preproinsulin gene. Cell 20:555–566

Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnik E, Razgulyaev OI (1990) Evidence for the molten globule state as a general intermediate in protein folding. FEBS Lett 262:20–24

Reid G, Henning U (1987) A unique amino acid substitution in the outer membrane protein OmpA causes conjugation deficiency in *Escherichia coli* K-12. FEBS Lett 223:387–390

Roy P, Rondeau SB, Vézina C, Boileau G (1990) Effect of mRNA secondary structure on their efficiency of translation initiation by eukaryotic ribosomes. FEBS Lett 191:647–651

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim NA (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491

Schluter, D (1988) Estimating the form of natural selection on a quantitative trait. Evolution 42:849–861

Sharp PM (1990) Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. Mol Microbiol 4:119–122

Sharp PM, Li W-H (1987a) Rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sharp PM, Li W-H (1987b) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Shields DC, Wolfe KH, Li W-H (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. Science 246:808–810

Sørensen MA, Kurland CG, Pedersen S (1989) Codon usage determines translation rate in *Escherichia coli.* J Mol Biol 207:365–377

Tso JY, Sun X-H, Kao T-H, Reese KS, Wu R (1985) Isolation and characterization of rat and human glyceraldehyde-3-phosphate dehydrogenase cDNAs: genomic complexity and molecular evolution of the gene. Nucleic Acids Res 13:2485–2502

Vogel H, Jähnig F (1986) Models for the structure of outer membrane proteins of *Escherichia coli* derived from Raman spectroscopy and prediction models. J Mol Biol 190:191–199

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Annu Rev Biochem 46:573–639

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Wolfe KH, Gouy M, Yang Y-W, Sharp PM, Li W-H (1989) Date of monocot–dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86:6201–6205

Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745

Yanofsky C (1988) Transcription attenuation. J Biol Chem 263: 609–612

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New York, pp 189–225

Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148

Received January 21, 1991/Accepted April 18, 1991