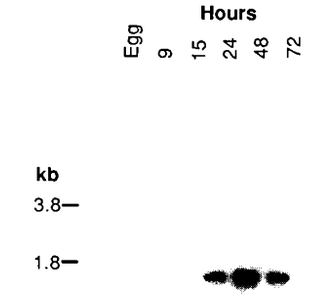

Fig. 3. Accumulation of SpEGF2 mRNA during development. Two micrograms of total RNAs from eggs and embryos at the indicated times of development were analyzed by blot hybridization as described (*15*) with a probe corresponding to the sequence of amino acid 303 to the 3′ end of the mRNA. Exposure was for 1 day with two intensifying screens. Developmental times of 9, 15, 24, 48, and 72 hours correspond to late cleavage, early blastula, mesenchyme blastula, late gastrula, and pluteus stages, respectively. Size calibration was provided by positions of the ribosomal RNAs.

The exact developmental role of the SpEGF2 peptide (or peptides) is unknown. The EGF-like peptides present in the blastocoelic fluid of *A. crassispina* cause exogastrulation of embryos of several different sea urchin species, including a species of *Strongylocentrotus*, when added to the outside of embryos (*11*). Their immediate morphological effect on blastulae, thinning of the vegetal plate, implies that precursor cells of endoderm and secondary mesenchyme can respond to these peptides and suggests a role in early events of gastrulation.

REFERENCES AND NOTES

1. K. A. Wharton, K. M. Johansen, T. Xu, S. Artavanis-Tsakonas, *Cell* **43**, 567 (1985).
2. H. Vässin *et al.*, *EMBO J.* **6**, 3431 (1987).
3. I. Greenwald, *Cell* **43**, 583 (1985).
4. J. Yochem and I. Greenwald, *ibid.* **58**, 553 (1989); J. Austin and J. Kimble, *ibid.*, p. 565.
5. D. J. G. Rees *et al.*, *EMBO J.* **7**, 2053 (1988).
6. H. Marquardt, M. W. Hunkapiller, L. E. Hood, G. J. Todaro, *Science* **223**, 1079 (1984).
7. A. W. Burgess, *Br. Med. Bull.* **45**, 401 (1989).
8. P. Y. Chou and G. D. Fasman, *Adv. Enzymol.* **47**, 45 (1978); J. Garnier, D J. Osguthorpe, B. Robson, *J. Mol. Biol.* **120**, 97 (1978).
9. K. H. Mayo, P. Schaudies, C. R. Savage, A. De Marco, R. Kapstein, *Biochem. J.* **239**, 13 (1986); G. T. Montelione, K. Wüthrich, E. C. Nice, A. W. Burgess, H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 5226 (1987); K. Makino *et al.*, *ibid.*, p. 7841; R. M. Cooke *et al.*, *Nature* **327**, 339 (1987).
10. M. L. Birnstiel *et al.*, *Cell* **41**, 349 (1985).
11. T. Suyemitsu *et al.*, *Cell Diff. Dev.* **26**, 53 (1989).
12. A. B. Smith, *Mol. Biol. Evol.* **5**, 345 (1988).
13. D. A. Hursh, M. E. Andrews, R. A. Raff, *Science* **237**, 1487 (1987).
14. Q. Yang, L. M. Angerer, R. C. Angerer, unpublished observations.
15. ———, *Dev. Biol.* **135**, 53 (1989).
16. D. D. Sabatini, G. Kreibich, T. Morimoto, M. Adesnik, *J. Cell Biol.* **92**, 1 (1982).
17. W. C. Barker, L. T. Hunt, D. G. George, *Protein Seq. Data Anal.* **1**, 363 (1988).
18. R. C. Angerer *et al.*, in *Society for Experimental Biology Seminar Series*, N. Harris and D. Wilkinson, Eds. (Cambridge Univ. Press, Cambridge, in press).
19. L. M. Angerer *et al.*, *Genes Dev.* **2**, 239 (1988).
20. H.-M. Jantzen *et al.*, *Cell* **49**, 29 (1987).
21. C. Moskaluk and D. Bastia, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 1215 (1987).
22. J. Pustell and F. C. Kafatos, *Nucleic Acids Res.* **10**, 51 (1982); *ibid.* **12**, 643 (1984).
23. A. Gray, T. J. Dull, A. Ullrich, *Nature* **303**, 722 (1983); J. Scott *et al.*, *Science* **221**, 236 (1983).
24. R. Derynck, A. B. Roberts, M. E. Winkler, E. Y. Chen, D. V. Goeddel, *Cell* **38**, 287 (1984).
25. K. Kurachi and E. W. Davie, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6461 (1982).
26. L. M. Angerer, K. H. Cox, R. C. Angerer, *Methods Enzymol.* **152**, 649 (1987).

27. Supported by NIH grant GM25553 to R.C.A. and L.M.A. R.C.A. is the recipient of a Career Development Award from USPHS (HD602). We thank E. Davidson for the pluteus cDNA library; S. Reynolds for providing RNA for determination of the transcription start site by primer extension; New England Nuclear for ^{35}S-labeled uridine 5′-triphosphate; and M. Gorovsky, P. Hinkle, V. Stathopoulos, and Y. Xiong for helpful comments on the manuscript. The SpEGF2 cDNA sequence has been submitted to GenBank and assigned the accession number m29004.

31 May 1989; accepted 15 September 1989

Chromosomal Location and Evolutionary Rate Variation in Enterobacterial Genes

PAUL M. SHARP, DENIS C. SHIELDS, KENNETH H. WOLFE, WEN-HSIUNG LI

The basal rate of DNA sequence evolution in enterobacteria, as seen in the extent of divergence between *Escherichia coli* and *Salmonella typhimurium*, varies greatly among genes, even when only "silent" sites are considered. The degree of divergence is clearly related to the level of gene expression, reflecting constraints on synonymous codon choice. However, where this constraint is weak, among genes not expressed at high levels, divergence is also related to the chromosomal location of the gene; it appears that genes furthest away from *oriC*, the origin of replication, have a mutation rate approximately two times that of genes near *oriC*.

EVOLUTIONARY RATES AT THE DNA sequence level are determined by the nucleotide mutation rate and the subsequent effects of natural selection; generally, only the latter is considered to vary among genes in a genome. If "silent" codon positions (that is, those which can be changed without affecting the amino acid sequence of the gene product) are effectively neutral as suggested (*1*), the nucleotide substitution rate at silent sites should reflect only the mutation rate (*2*). Thus, synonymous nucleotide substitutions should accumulate at similar rates in different genes and potentially constitute a useful molecular clock (*3*). However, silent substitution rates have been found to vary among genes in animals (*4, 5*) and bacteria (*6, 7*); the question is whether this reflects variation in mutation rates, selective constraints, or both.

In *Escherichia coli* it is clear that synonymous codons are not selectively equivalent: codon usage is highly biased, the "preferred" codons are those recognized most accurately and/or efficiently by the most abundant tRNA species, and the degree of bias is greater in genes expressed at high levels (*8*). Both tRNA levels (*9*) and patterns of codon usage (*7, 9*) are conserved between *E. coli* and *Salmonella typhimurium*, and so it would be expected that genes retaining high codon usage bias in both species must have accumulated fewer substitutions. This is indeed the observation: comparison of 67 pairs of homologous gene sequences (*10*) from the two species (Fig. 1A) shows that the extent of divergence at silent sites (expressed as the corrected number of base substitutions per synonymous site, K_S) is significantly negatively correlated with the degree of bias in codon usage (Table 1, model 1). (Codon usage bias is measured by the codon adaptation index, CAI; see legend to Fig. 1.) Thus, selection among synonymous codons, which is most prevalent in highly expressed genes, forms an evolutionary constraint (7). However, this cannot explain all of the variation in degree of divergence; in particular there are some genes with low codon bias but surprisingly low divergence (such as those indicated by open circles in Fig. 1A).

We noticed that these anomalous (that is, having low codon bias and low divergence) genes are located near *oriC*, the origin of replication of the bacterial chromosome at

P. M. Sharp, D. C. Shields, K. H. Wolfe, Department of Genetics, Trinity College, Dublin 2, Ireland.
W.-H. Li, Center for Demographic and Population Genetics, University of Texas, Houston, TX 77030.

84 min. [Gene order in *E. coli* and *S. typhimurium* is highly conserved (*11, 12*)]. To investigate this further, we plotted the degree of divergence at silent sites in a gene as a function of its distance from *oriC* (Fig. 1B). If genes with highly biased codon usage are (for the moment) disregarded, it is clear that the divergence increases significantly with distance from *oriC* (Fig. 1B); this effect is only slightly less convincing when the highly biased genes are included (Table 1, model 2). Thus, genes near the terminus of replication (for example, the *che* and *trp* operons) are typically twice as divergent (between *E. coli* and *S. typhimurium*) as genes located near the origin (for example, *dnaA* and *metB*), though they have similar levels of codon bias.

When considered simultaneously (assuming a linear additive relation), codon bias and map position are predictive of the degree of divergence; the combined model (Table 1, model 3) is a better predictor than either model 1 or model 2. Note that there is no statistically significant linear relation between codon bias and distance from the origin (Table 1, model 6), reflecting the fact that the highly expressed genes are not predominantly clustered near *oriC*. Thus, there is no significant indication of any differences in selective constraint on synonymous codon choice that are dependent on map position. Rather, it appears that the variation in rate of sequence divergence seen in Fig. 1B most likely reflects differences in the rate of mutation; we infer that the nucleotide mutation rate increases with distance from the origin of replication.

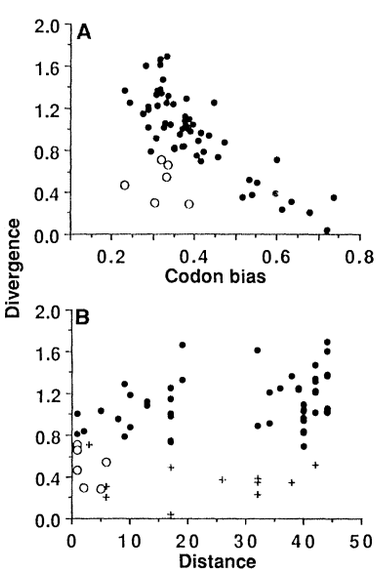

Fig. 1. Relation between codon usage bias, map position, and silent substitution rate among 67 enterobacterial genes. Each point is a pair of homologous genes compared between *E. coli* and *S. typhimurium*. (**A**) Divergence at silent sites (K_S) as a function of codon usage bias (CAI). K_S is the number of nucleotide substitutions per silent site, after correction for multiple hits (*22*). CAI is the codon adaptation index; we estimate the optimality of a gene as a function of the optimality of its constituent codons by means of the CAI (*23*); genes with a high frequency of optimal codons have a high CAI value. The mean value is used for each pair of genes. Open circles indicate six genes with low CAI and K_S values. Linear regression of K_S on CAI is given in Table 1, model 1. (**B**) Divergence at silent sites (K_S, see above) as a function of map position, given as the distance from *oriC* (in either direction) in minutes. Map positions of the *E. coli* genes are used (*11*). Open circles indicate the six genes highlighted in (A). Crosses indicate genes with high codon bias (CAI > 0.50). Regressions of K_S on distance are given in Table 1, models 2 and 4.

The source of such a variation in mutation rate is open to conjecture. It is known that sequences nearer *oriC* are present in higher copy number on average, an effect that is exaggerated under conditions of rapid growth (*13*); could this reduce the mutation rate? Post-replication repair ("recombination repair") in *E. coli* is used to deal with ultraviolet (UV) light–induced thymine dimers and may also repair double-stranded breaks (*14*). This mechanism involves the correction of a lesion in one DNA duplex by means of a homologous sequence derived from a sister duplex (*15*), and so cannot occur at a locus in a single-copy state. It is possible that recombination repair is more prevalent for sequences nearer *oriC*, simply because they are at a higher copy number.

If the effect is due to copy number, we would expect the relation of divergence to distance from *oriC* to be nonlinear, since average copy number should increase exponentially with proximity to the origin of replication. Then, it is interesting that models based on log distance from *oriC* appear to be somewhat better than the linear models (contrast model 2 with model 4, and model 3 with model 5).

Horizontal transfer of genetic material has the potential to disrupt molecular clocks. Thus, an alternative explanation of our observation is that interstrain recombination more often involves sequences nearer the origin of replication. However, when we consider the twofold difference in degree of divergence, this hypothesis would require regular exchange to have taken place among very distantly related strains during the emergence of *E. coli* and *S. typhimurium*. Whereas evidence for recombination among *E. coli* strains has been increasing recently (*16*), the strains involved do not appear to be sufficiently divergent to explain the current observations.

This is not the first suggestion of intragenomic variation in mutation patterns, but it is novel in (i) the nature of the variation, that is, a gradient of increasing mutation rate (*17*), and (ii) the proposed explanation, that is, recombination repair. For example, in mammals, there is evidence that mutation patterns and rates differ among regions of the genome (*18, 19*), but the variation is thought to be discontinuous since regions have discrete base compositions; the mechanism could be either replication under different cellular conditions (*18*) or repair by different polymerase activities (*19*).

Silent molecular clocks vary among genes in mammals and *Drosophila* as well as in bacteria. The mechanisms responsible may be different in each case: for example, selection on codon choice in *Drosophila* constrains the silent rate to varied extents in different genes (*5*), but in addition there are unexplained local differences in nucleotide substitution rate that may reflect the primary mutation rate (*20*). Since there is also evidence that silent substitution rates vary consistently among taxonomic lineages (*4, 5, 21*), a general conclusion can be drawn that silent molecular clocks are not as straightforward as first proposed; nevertheless, the current work suggests that some of the variation among genes can be explained.

Table 1. Regression analyses of relation of silent site divergence to codon usage bias and map position. Divergence at silent sites measured by K_S, codon usage bias measured by CAI, and map position (map) expressed as distance from *oriC*; see Fig. 1 for details. The regression equation for each model is given: a is a constant, and b is the coefficient of the predictor variable or variables, as derived from the regression analysis; t value tests the significance of the regression coefficient: $^{**}P < 0.001$; $^{***}P < 0.0001$. Variation reduction due to fitted model is in r^2 column.

Model	n	b	$b \pm$ SE	t value	r^2
1) $K_S = a + b \times$ CAI	67		−2.32 ± 0.31	−7.53***	47
2) $K_S = a + b \times$ map	67		0.010 ± 0.003	3.59**	17
$K_S = a + b \times$ map*	56		0.010 ± 0.002	4.55***	28
3) $K_S = a + b_1 \times$ CAI + $b_2 \times$ map	67	b_1	−2.20 ± 0.28	−7.95***	58
	67	b_2	0.008 ± 0.002	4.18***	
4) $K_S = a + b \times$ ln map*	56		0.147 ± 0.029	5.03***	32
5) $K_S = a + b_1 \times$ CAI + $b_2 \times$ ln map	67	b_1	−2.32 ± 0.26	−8.79***	61
	67	b_2	0.129 ± 0.026	4.88***	
6) CAI = $a + b \times$ map	67		−0.0007 ± 0.0009	−0.84	1

*Genes with low codon bias (CAI < 0.50) only.

REFERENCES AND NOTES

1. J. L. King and T. H. Jukes, *Science* **164**, 788 (1969); M. Kimura, *Nature* **267**, 275 (1977).
2. M. Kimura, *Nature* **217**, 624 (1968).
3. T. Miyata, T. Yasunaga, T. Nishida, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7328 (1980); A. C. Wilson, H. Ochman, E. M. Prager, *Trends Genet.* **3**, 241 (1987).
4. W.-H. Li *et al.*, *J. Mol. Evol.* **25**, 330 (1987).
5. P. M. Sharp and W.-H. Li, *ibid.* **28**, 398 (1989).
6. H. Ochman and A. C. Wilson, *ibid.* **26**, 74 (1987).
7. P. M. Sharp and W.-H. Li, *Mol. Biol. Evol.* **4**, 222 (1987).
8. M. Gouy and C. Gautier, *Nucleic Acids Res.* **10**, 7055 (1982); T. Ikemura, *Mol. Biol. Evol.* **2**, 13 (1985).
9. T. Ikemura, in *Population Genetics and Molecular Evolution*, T. Ohta and K. Aoki, Eds. (Japan Scientific Societies Press, Tokyo, 1985), pp. 385–406.
10. The genes are (in increasing order of silent site divergence): *rpsU, hupA, ptsH, metJ, pyrE, rpoB, ompA, ptsI, hupB, crr, ilvM, rpoD, prsA, metB, dnaA, hisG, glnA, ilvY, araA, carA, ompR, envZ, cyaA, hisC, hisA*, orf (5′ to *pyrE*), *crp, cysK, hisP, hisH, dnaB, hisF, araB, araD, ilvA, cheZ, cheR, trpB, metF, hisD, trpE, hisB, trpD, pyrI, hisIE, pyrB, dnaG, pabA, cheB, hisM, cheW, pyrC, mglB, araC, aroA, glgC, cheA*, orf (5′ to *metC*), *tar, trpC, sulA, pabB, cheY, cysB, metC, cysZ, trpA*. Data were taken from the GenBank/EMBL DNA sequence data library wherever possible; all sequence data and sources are available on request.
11. B. J. Bachmann, *Microbiol. Rev.* **47**, 180 (1983).
12. M. Riley and S. Krawiec, in Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 967–981; K. E. Sanderson and J. R. Roth, *Microbiol. Rev.* **52**, 485 (1988).
13. M. B. Schmid and J. R. Roth, *J. Bacteriol.* **169**, 2872 (1987).
14. S. R. Kushner, in Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 1044–1053.
15. W. D. Rupp and P. Howard-Flanders, *J. Mol. Biol.* **31**, 291 (1968); A. K. Ganesan, *ibid.* **87**, 103 (1974); S. C. West, E. Cassuto, P. Howard-Flanders, *Nature* **294**, 659 (1981).
16. R. F. DuBose, D. E. Dykhuizen, D. L. Hartl, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 7036 (1988); R. Milkman and A. Stoltzfus, *Genetics* **120**, 359 (1988).
17. Direct investigation of mutation rates at different map positions in *E. coli* revealed one potential hotspot at 58 to 60 min [J. van Brunt and G. Edlin, *J. Mol. Evol.* **5**, 279 (1975)]. In the related enterobacterium *Serratia marcescens*, there is G+C content heterogeneity among genes, which may be due to mutation pattern variation [M. Nomura, F. Sor, M. Yamagishi, M. Lawson, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 658 (1987)], though it could be explained more simply by the differential action of natural selection on codon choice in the face of a constant mutation pressure to G+C–richness (P. M. Sharp, *Mol Microbiol.*, in press).
18. K. H. Wolfe, P. M. Sharp, W.-H. Li, *Nature* **337**, 283 (1989).
19. J. Filipski, *J. Theor. Biol.* **134**, 159 (1988).
20. C. H. Martin and E. M. Meyerowitz, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8654 (1986).
21. C.-I. Wu and W.-H. Li, *ibid.* **82**, 1741 (1985); R. Kikuno, H. Hayashida, T. Miyata, *Proc. Japan Acad. Ser. B* **61**, 153 (1985); R. J. Britten, *Science* **231**, 1393 (1986).
22. W.-H. Li, C.-I. Wu, C.-C. Luo, *Mol. Biol. Evol.* **2**, 150 (1985).
23. P. M. Sharp and W.-H. Li, *Nucleic Acids Res.* **15**, 1281 (1987).
24. This paper is dedicated to the memory of Alan Robertson, who died on 25 April 1989. Supported in part by grants from the European Community (to P.M.S.) and NIH (to W.-H.L.). We thank D. J. McConnell for his comments on the manuscript.

21 June 1989; accepted 22 September 1989

Genomic Sequencing and Methylation Analysis by Ligation Mediated PCR

GERD P. PFEIFER, SABINE D. STEIGERWALD, PAUL R. MUELLER, BARBARA WOLD, ARTHUR D. RIGGS*

Genomic sequencing permits studies of in vivo DNA methylation and protein-DNA interactions, but its use has been limited because of the complexity of the mammalian genome. A newly developed genomic sequencing procedure in which a ligation mediated polymerase chain reaction (PCR) is used generates high quality, reproducible sequence ladders starting with only 1 microgram of uncloned mammalian DNA per reaction. Different sequence ladders can be created simultaneously by inclusion of multiple primers and visualized separately by rehybridization. Relatively little radioactivity is needed for hybridization and exposure times are short. Methylation patterns in genomic DNA are readily detectable; for example, 17 CpG dinucleotides in the 5′ region of human X-linked PGK-1 (phosphoglycerate kinase 1) were found to be methylated on an inactive human X chromosome, but unmethylated on an active X chromosome.

METHYLATION OF CpG DINUCLEOTIDES in critical regions of many vertebrate genes may be part of a gene silencing mechanism involved in cell differentiation, X chromosome inactivation, and genomic imprinting (*1, 2*). Methylation-sensitive restriction endonucleases are commonly used to determine in vivo methylation patterns, but this limits the analysis to a small subset of all CpG dinucleotides. Another method for methylation analysis is genomic sequencing (*3*), a method that retains information normally lost during cloning, such as the location of 5-methylcytosines (*3*) and DNA-protein interactions (*4*). Genomic sequencing has, however, been difficult, requiring large amounts of radioactivity and long autoradiographic exposures (*5*). Primer extension has been used to simplify genomic sequencing, but these procedures still require the special preparation of primers labeled to extremely high specific activity and up to 50 μg of DNA per sequencing lane (*6*).

We now describe a genomic sequencing method in which we use a ligation mediated polymerase chain reaction (PCR) procedure [see figure 1 in (*7*)]. Briefly, step 1 of our genomic sequencing procedure is base-specific chemical cleavage of DNA samples (*8*) at either G, G+A, T+C, or C (*9*), generating 5′ phosphorylated molecules. Step 2 is gene-specific primer extension of an oligonucleotide (primer 1) by a DNA polymerase to give molecules that have a blunt end on the side opposite the primer (*10*). Step 3 is the ligation of an unphosphorylated linker to the blunt ends (*11*). Step 4 is the exponential amplification of the linker-ligated fragments with the use of the longer oligonucleotide of the linker (as a linker-primer) and a second gene-specific primer (primer 2) in a PCR reaction (*12*). After undergoing 15 to 18 amplification cycles, the DNA fragments are separated on a sequencing gel, transferred by electroblotting to nylon membranes (*13*), and hybridized with a single-stranded gene-specific probe (*14*). This procedure works well for all bases, sensitivity is improved, and the background is minimized by the transfer and hybridization steps. Moreover, several different sequences can be analyzed in a single experiment by rehybridization of the membrane.

The human X-linked phosphoglycerate kinase (PGK-1) gene is a housekeeping gene that is subject to X inactivation. The 5′ region is a CpG-rich island (*15*), but, unlike most autosomal CpG islands that are characteristically unmethylated, the Hpa II sites in the region shown (Fig. 1) are methylated on the inactive X chromosome (*16, 17*).

In an experiment with HeLa cell DNA, two different primer sets (Fig. 1, D and E) were included simultaneously in the primer extension and amplification reactions. The sequence defined by primer set D was visualized first (Fig. 2A) by hybridization with an Eco RI–Dde I hybridization probe. After stripping of the first probe from the membrane and rehybridization with an Xma III–

G. P. Pfeifer, S. D. Steigerwald, A. D. Riggs, Molecular Biology Section, Beckman Research Institute of the City of Hope, Duarte, CA 91010.
P. R. Mueller and B. Wold, California Institute of Technology, Division of Biology, Pasadena, CA 91125.

*To whom correspondence should be addressed.