manner similar to the method described in (*5*). Briefly, Z-domain was immobilized on microtiter wells at a concentration of 5 μg/ml, blocked, and washed as described. A matrix of mixtures of biotin-IgG-Fc (312 to 0.3 nM) and peptide (215 μM to 0.8 nM) were prepared. These mixtures were incubated with immobilized Z-domain for 1 hour. Plates were then washed and developed as described with avidin–horseradish peroxidase conjugate. Inhibition curves were then computed for each concentration of biotin-IgG-Fc, and the curve of half-maximal inhibition was extrapolated to zero biotin-IgG-Fc concentration to obtain a K_i.

11. The DNA sequence of the peptide was moved to a monovalent phage display format by cassette mutagenesis to give a construct with the STII signal sequence, the peptide KEASCSYWLGELVWCVAGVE, a GGGPGGG linker, and the M13 gene III protein starting at residue 253.
12. A series of second-generation monovalent phage display libraries were constructed based on the sequence KEASCSYWLGELVWCVAGVE, in which five sequential residues were randomized by using NNS codons in each library starting at positions 1, 4, 7,10, 12, and 16, excluding the two cysteines. Each library had a diversity of $\sim 1 \times 10^8$. These libraries were independently screened for binding to IgG-Fc for six rounds and then sequenced.
13. Three additional libraries were constructed by using the degeneracy of the genetic code to recombine the preferred amino acids at each position into one peptide. The DNA sequences for these libraries contained the following mixtures of bases (IUPAC codes): DRG GWA GMA RRC TGC KCT TRS CAC MTG GGC GAG CTG GTC TGG TGC RVC RVM BKC GAS KDW, DRS VWG SVG RRC TGC KCC TRS YRS MTG GGC GAG CTG GTC TGG TGC RNC VVS NBS GWS KDM, and DNS NNS NNS VNS TGC BVG TDS HRS MDS GGC GAG STC KKG WRG TGC RNM NNS NNS NNS NNM. These libraries were also sorted against IgG-Fc for six rounds and then sequenced.
14. Inhibition assays were performed as described (*10*) at pH 7.2 and at pH 6.0. The peptide was found to inhibit fourfold more tightly at the lower pH. Kinetic and steady-state binding to immobilized IgG_1 was also measured directly by BIAcore (Pharmacia), giving $K_{on} = 1.6 \times 10^6\ M^{-1}\ s^{-1}$, $K_{off} = 2.5 \times 10^{-2}\ s^{-1}$, and $K_d = 16$ nM in 25 mM MES (pH 6.0), 0.05% Tween-20.
15. S. R. Fahnestock *et al.*, in *Bacterial Immunoglobulin-Binding Proteins* (Academic Press, New York, 1990), vol. 1, chap. 11; R. Karlsson, L. Jendeberg, B. Nilsson, J. Nilsson, P. Nygren, *J. Immunol. Methods* **183**, 43 (1995).
16. Crystals were grown in 100 mM NaOAc (pH 6.0), 20% polyethylene glycol 4000, and 20% isopropanol by vapor diffusion from 4-μl drops containing 100 μM IgG-Fc, up to 150 μM peptide, and a 50% dilution of reservoir solution. Data were collected to 2.6 Å at the Stanford Synchrotron Radiation Laboratory (SSRL) and were reduced with DENZO [W. Minor and Z. Otwinowski, *Methods Enzymol.* **176**, 307 (1997)]. Phasing was accomplished by molecular replacement with AmoRE [J. Navaza, *Acta Crystallogr.* **A50**, 157 (1994)], with an IgG-Fc subunit derived from Deisenhofer *et al.* (*3*) as a search model. The crystal contained one Fc dimer and two peptide molecules per asymmetric unit. The structure was refined with X-PLOR 3.1 [A. T. Brünger *et al.*, *Science* **245**, 458 (1987)], with noncrystallographic restraints on the Fc dimer over regions >10 Å away from nonequivalent crystal contacts. The final dimeric Fc model consisted of IgG_1 residues 237 to 443 with eight sugars per monomer.
17. Surface area and geometric measurements were made with the Crystallography and NMR System (CNS) [A. T. Brünger *et al.*, *Acta Crystallogr. D.* **54**, 905 (1998)]. A solvent probe radius of 1.4 Å was used, and surface area changes were computed by subtracting complexed from uncomplexed solvent-accessible surface areas. Contact regions were defined as the set of atoms that lie within 5.0 Å of any nonhydrogen atom on the opposing molecule.
18. Protein A domain numbering is from H. Gouda *et al.*, *Biochemistry* **31**, 9665 (1992).
19. The computer program SITEFINDER (WLD) was used to generate 2.5 million patches of contiguous surface atoms having solvent-accessible surface areas of 525 $Å^2$. Patches were randomly distributed across all of the available structures (PDB codes: 1FC1, 1FC2, 1FCC, 1ADQ, and 1DN2) and were of a random globular shape. To ensure even sampling, probabilities were weighted so that each solvent-exposed atom was included in an equal number of surface patches (~10,000 patches per atom) The properties of each site were computed and then compared with those of the consensus binding patch on Fc.
20. L. Young, R. L. Jernigan, D. G. Covell, *Protein Sci.* **3**, 717 (1994); S. Jones and J. M. Thornton, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13 (1996); C. Tsai, S. L. Lin, H. J. Wolfson, R. Nussinov, *Protein Sci.* **6**, 53 (1997); L. L. Conte, C. Chothia, J. Janin, *J. Mol. Biol.* **285**, 2177 (1999).
21. Patches from (*19*) were ranked separately by polarity and solvent-accessible surface fraction. For each atom in the Fc dimer, the average rank of all patches involving the atom was then computed. The average atomic ranks for polarity and accessibility were then combined linearly (⟨accessibility⟩ − ⟨polarity⟩) to give a composite score incorporating both properties.
22. C. Medesan, D. Matesoi, C. Radu, V. Ghetie, E. S. Ward, *J. Immunol.* **158**, 2211 (1997).
23. W. L. DeLano, in preparation.
24. Single alanine mutants of the peptide and Protein A Z-domain were displayed monovalently on gene III, assayed by enzyme-linked immunosorbent assay, and directly compared with unmutated control peptide as described [B. C. Cunningham, D. G. Lowe, B. Li, B. D. Bennett, J. A. Wells, *EMBO J.* **13**, 2508 (1994)]. Assays were performed under conditions where EC_{50}(wt)/EC_{50}(mut) will approximate K_d(wt)/K_d(mut) (EC_{50}, median effective concentration). Additional mutagenesis data are available in L. Jendeberg *et al.*, *J. Mol. Recog.* **8**, 270 (1995).
25. B. K. Kay, A. V. Kurakin, R. Hyde-DeRuysher, *Drug Discovery Today* **8**, 370 (1998).
26. We thank B. C. Cunningham, J. K. Tong, and M. Dennis for assistance in the initial selection experiments against Fc; A. Braisted for training in solid phase peptide synthesis; C. Wiesmann for help with crystallographic refinement; and the SSRL for use of their facility in data collection

19 August 1999; accepted 27 December 2000

Evidence for a High Frequency of Simultaneous Double-Nucleotide Substitutions

Michalis Averof,[1]* Antonis Rokas,[2] Kenneth H. Wolfe,[3] Paul M. Sharp[4]*

Point mutations are generally assumed to involve changes of single nucleotides. Nevertheless, the nature and known mechanisms of mutation do not exclude the possibility that several adjacent nucleotides may change simultaneously in a single mutational event. Two independent approaches are used here to estimate the frequency of simultaneous double-nucleotide substitutions. The first examines switches between TCN and AGY (where N is any nucleotide and Y is a pyrimidine) codons encoding absolutely conserved serine residues in a number of proteins from diverse organisms. The second reveals double-nucleotide substitutions in primate noncoding sequences. These two complementary approaches provide similar high estimates for the rate of doublet substitutions, on the order of 0.1 per site per billion years.

Mutational events can be studied either by direct observation of mutations in the laboratory or by comparing sequences that have been accumulating mutations naturally, during evolution. Studies of the first kind have suggested that some mutations can involve multiple nucleotide changes (*1, 2*), and indeed, mechanisms that affect neighboring nucleotides are known. Examples include template-directed mutations occurring during DNA repair and replication (*1*) or dipyrimidine lesions induced by ultraviolet light (*2, 3*). Some evolutionary comparisons have also suggested that simultaneous double-nucleotide substitutions occur at neighboring sites (*4*), but the significance and generality of these observations have been questioned (*5*). Thus, changes in neighboring nucleotides are usually attributed to coincidence of independent mutations.

We used two independent and complementary approaches based on sequence comparisons to study double-nucleotide substitutions and to obtain estimates of their frequency. The first approach examined changes that have occurred over long evolutionary time scales, between two particular dinucleotides, TC and AG. Serine is unique among amino acids in that it is encoded by two groups of codons, TCN and AGY, which cannot be interconverted by a single-nucleotide mutation. Switches between these groups of codons could occur indirectly, by two separate single-nucleotide mutations

[1]Institute of Molecular Biology and Biotechnology (IMBB)–FORTH, Vassilika Vouton, 711 10 Iraklio, Crete, Greece. [2]Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh, EH9 3JT, UK. [3]Department of Genetics, Trinity College, University of Dublin, Dublin 2, Ireland. [4]Institute of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, UK.

*To whom correspondence should be addressed. E-mail: averof@imbb.forth.gr (M.A.) or paul@evol.nott.ac.uk (P.M.S.)

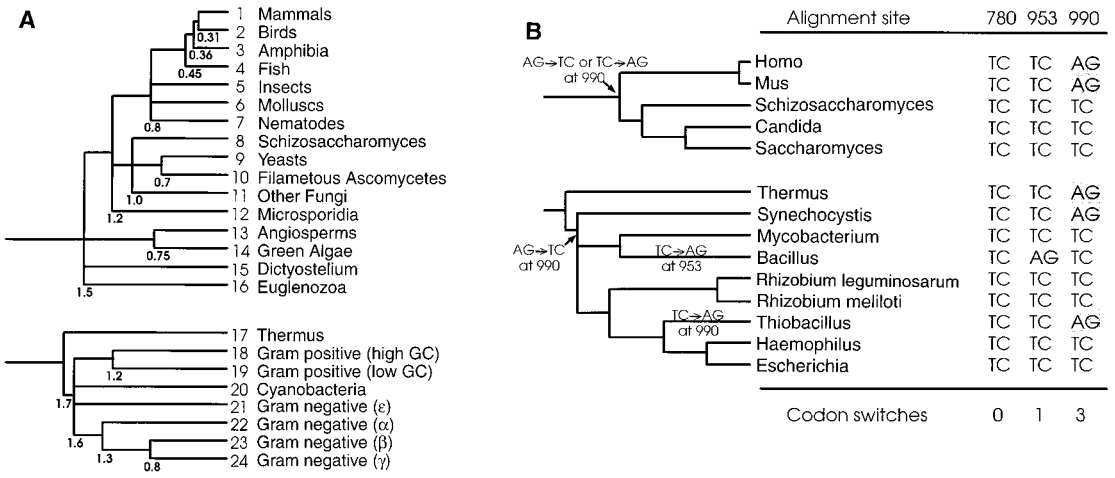

Fig. 1. (**A**) Overview of the phylogeny and divergence times used for the analysis of serine codon switches. The phylogeny is based on a number of recent phylogenetic analyses (*20, 24*), with points of uncertainty shown as unresolved polychotomies. Times of common ancestors are indicated in Gyr before present. (**B**) Determination of serine codon switches. The data set of glutamine fructose-6-phosphate transaminase is shown as an example. There are three sites where serine is absolutely conserved in the protein sequence (alignment sites 780, 953, and 990). At least four codon switches can be observed. The time sampled by this data set (sum of branch lengths) is 3 × 14.71 Gyr.

(TC↔AC↔AG or TC↔TG↔AG), or perhaps directly by simultaneous double-nucleotide mutation (TC↔AG). In the former case, the switch would involve an intermediate step whereby the triplet would encode either threonine (ACN) or cysteine (TGY), residues that are ionically and sterically different from serine (*6*), so such changes are unlikely to be tolerated in critical functional or structural sites of a protein. Nevertheless, TCN↔AGY switches have been observed at sites encoding extremely conserved serine residues, for example in ubiquitin (*7*) and in the active site of serine proteases (*8*). Switches at these sites seem most likely to result from simultaneous double-nucleotide mutations, which in this context are synonymous and most likely selectively neutral.

To investigate the generality and frequency of such switches, we studied 23 data sets of homologous proteins containing serine residues absolutely conserved over a wide range of eukaryotes and/or prokaryotes (Fig. 1A). We analyzed the distribution of TCN and AGY codon types in these conserved serines, inferring the position and frequency of codon switches during evolution (illustrated in Fig. 1B) (*9*). Our analysis reveals a widespread occurrence of codon switches at such sites (Table 1), with an estimated frequency of about 0.1 per site per billion years (94/774 = 0.12 per site per Gyr). This rate appears to be consistent among different phylogenetic lineages and different genes (Fig. 2). Rate estimates from bacteria and eukaryotes are very similar, 0.11 and 0.12 per site per billion years (Gyr), respectively.

Of the 70 switches where the direction of change could be inferred (by parsimony and with reference to outgroups), 60 were in the TC→AG rather than the AG→TC direction. However, independent rate estimates for each direction are very similar, 0.10 and 0.11 per site per Gyr, respectively. The bias therefore reflects a preponderance of TCN-type codons as potential targets, rather than a bias in the direction of mutation [this points to a strong codon bias in the ancestral representation of serines (*10*)].

Most codon switches at such highly conserved serines appear to result from simultaneous double-nucleotide mutations. However,

Table 1. Rates of serine codon switches in 23 data sets of highly conserved proteins. The phylogenetic assemblages (species) represented in each data set are indicated by numbers as specified in Fig. 1A. The inferred number of codon switches and estimated time sampled by each data set (in Gyr) are indicated.

Protein	Species	Switches	Time
Ribosomal protein S7	1,3,4,5,8,9	1	5.51
Ribosomal protein S11	1,3,5,8,9,13,14	2	22.62
Ribosomal protein S12	1,3,5,7,9,16	1	6.36
Ribosomal protein S17	1,2,5,9,10,13,15	0	15.49
Arginosuccinate synthetase	1,9,18,24	3	18.12
Glycine dehydrogenase	1,2,8,9,13	3	28.20
Glutamine fructose-6-phosphate transaminase	1,8,9,17,18,19,20,22,23,24	4	44.13
2-oxaloglutarate dehydrogenase	1,9,23,24	5	32.35
Asparagine synthase	1,9,13	2	38.08
Adenylosuccinate synthase	1,7,8,9,13,15,19,20,22,23,24	2	17.11
dUTP pyrophosphatase	1,9,13,24	0	5.71
Uridine-5-diphosphate glucose-4-epimerase	1,9,18,19,22,23,24	5	32.32
Phosphoenol pyruvate carboxykinase	1,2,5,7,11	7	55.40
Arginosuccinate lyase	1,2,3,8,9,14,18,21,24	9	63.05
1,4-α-glycan branching enzyme	1,9,13,18,19,20,24	6	28.92
Histidine tRNA synthetase	1,4,7,9	2	16.08
Tryptophanyl tRNA synthetase	1,8,9	2	9.38
Ribonucleotide reductase (large subunit)	1,5,7,8,9,12	16	123.80
Fumarate hydratase	1,9,11	3	21.06
Aspartate ammonia lyase	18,19,23,24	3	46.40
DNA topoisomerase 2	1,5,8,9,13,16	5	51.60
Dimethylallyl transferase	1,9,10,13	1	18.21
Ribonucleotide reductase (small subunit)	1,5,6,7,8,9,12,13,15	12	74.10
Total		94	774.00

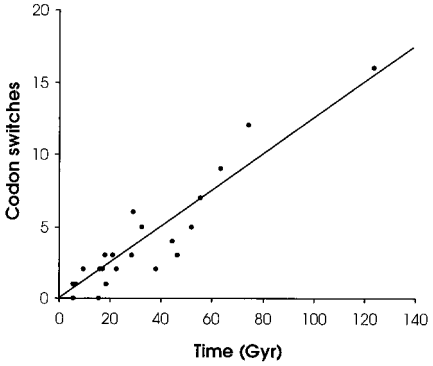

Fig. 2. Rate of observed serine codon switches for 23 proteins. Data is from Table 1. The line has a slope of 0.12 switches per site per Gyr.

Table 2. Analysis of single- and double-nucleotide substitutions in the pseudo eta globin locus on each branch of the catarrhine primate phylogeny (Fig. 3). Positions of substitutions were inferred by parsimony. L, number of aligned nucleotides; ObsS, ObsD, numbers of changes observed as single- or double-nucleotide substitutions, respectively; ExpD, number of doublet substitutions expected by coincidence of two separate single-nucleotide substitutions; RealD, number of excess double changes, inferred to have occurred as simultaneous double-nucleotide substitutions.

Tree branch	L	ObsS	ObsD	ExpD	RealD
Node 1–rhesus monkey	6617	284	19	10.83	8.17
Node 1–gibbon	6996	129	4	2.25	1.75
Node 1–node 2	7187	39	1	0.19	0.81
Node 2–orangutan	6974	92	2	1.17	0.83
Node 2–node 3	7187	46	1	0.28	0.72
Node 3–gorilla	7055	41	0	0.24	−0.24
Node 3–node 4	7187	7	0	0.01	−0.01
Node 4–human	6997	26	0	0.10	−0.10
Node 4–node 5	7187	34	0	0.16	−0.16
Node 5–chimp	7024	23	2	0.05	1.95
Node 5–pigmy chimp	7003	11	1	0.01	0.99
Total		732	30	15.30	14.70

it is conceivable that these switches could occur by two separate single-nucleotide mutations, through intermediates that encode threonine or cysteine. Kimura suggested that slightly deleterious intermediates may sometimes survive to be rescued by rapidly selected compensatory mutations (*11*), but there are a number of observations that argue against this possibility in this case. First, Kimura's model applies to situations where compensatory mutations are relatively frequent (e.g., when many different mutations can have a compensatory effect) or when the selective coefficient against the intermediates is rather low, which seem very unlikely. Second, if deleterious alleles were involved, we would expect these to survive much more frequently in the presence of additional copies of the gene, but we observe very similar rates of codon switches in haploid and diploid genomes, as well as in proteins that belong to multigene families (*12*). Moreover, we have also noticed TCN↔AGY switches among codons encoding highly conserved serines in closely related sequences, with no evidence of a transition through nonserine intermediates (*13*).

Other mechanisms have also been proposed that could explain switches in serine codons through nondeleterious intermediates (*8, 14–16*). For example, a transient substitution of serine by another amino acid could be complemented by the presence of a neighboring serine residue (*16*), an alternative genetic code may have allowed TGN to encode serine (*15*), or the two types of serine codon may reflect independent origins from a different ancestral amino acid (*8*). These explanations may apply in special cases and could contribute to a small proportion of codon switches. However, they are unlikely to account for the widespread distribution of codon switches, as observed in diverse phylogenetic lineages, in different proteins, and in serine residues whose position and identity has been absolutely conserved.

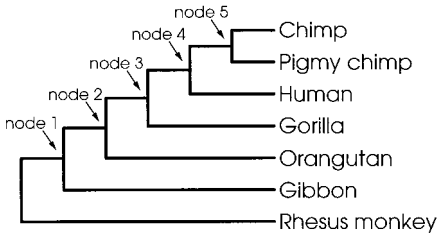

Fig. 3. Phylogeny of the catarrhine primates (*17*) used in the analysis of pseudo eta globin sequences (Table 2). Branch lengths are not to scale.

In our second approach, we examined double-nucleotide substitutions among noncoding sequences of closely related species. In these sequences, substitutions are expected to accumulate in a manner that is unbiased by selection, and so directly reflect mutational processes. We compared a long (about 7 kb) noncoding sequence from the pseudo eta globin locus of seven closely related catarrhine primates (Fig. 3) (*17*) to determine whether mutations in that region involve a significant fraction of clustered nucleotide changes (*18*). Using parsimony analysis, we determined the number of single- and double-nucleotide changes that have occurred during the evolution of these species and found a significant excess of double-nucleotide substitutions relative to what would be expected by coincidence of single-nucleotide changes alone (Table 2). The excess, apparently simultaneous, dinucleotide mutations are estimated to have occurred at a rate of 0.1 per site per Gyr (*19*), on average, at any nucleotide doublet.

These two analyses are complementary: they examine double-nucleotide substitutions in different contexts and over very different time scales. Any concerns that the serine codon switches might have involved compensatory changes via nonserine intermediates are offset by the observation of similarly high levels of doublet changes in closely related noncoding sequences. Equally, although the rates for all dinucleotide changes were estimated from just one particular region of the primate genome, the rates of TC↔AG changes estimated from serine switches apply to a wide range of loci from diverse organisms. Both approaches point to the conclusion that the rate of double-nucleotide substitutions is high compared to expectations based on the coincidence of individual neutral nucleotide substitutions, which typically occur at a rate of around 1 to 10 per site per Gyr (*20, 21*).

We expect that the rates of different doublet mutations will vary considerably depending on a cell's exposure to different mutational mechanisms. For example, we would expect to see a much higher incidence of dipyrimidine lesions in cells that are exposed to ultraviolet light (e.g., exposed unicellular organisms, skin cells) than in cells that are not (e.g., the germ line of large multicellular animals). Such differences might explain why the estimated frequency of specific TC→AG and AG→TC substitutions in serine codons, which may involve dipyrimidines (TC in the coding strand or CT in the noncoding strand, respectively), is higher than would be predicted by the average frequency of double-nucleotide substitution estimated from the eta globin pseudogene. The sequence-specificity of mutational mechanisms could result in different rates of substitution among various doublets in different cell types. These observations may be important in the context of models of molecular evolution and phylogenetic reconstruction, as well as mutational mechanisms of human disease.

References and Notes

1. G. B. Golding and B. W. Glickman, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 8577 (1985).
2. D. M. Hampsey, J. F. Ernst, J. W. Stewart, F. J. Sherman, *J. Mol. Biol.* **201**, 471 (1988).
3. H. Nakazawa *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 360 (1994).
4. K. H. Wolfe and P. M. Sharp, *J. Mol. Evol.* **37**, 441 (1993).
5. D. Mouchiroud, C. Gautier, G. Bernardi, *J. Mol. Evol.* **40**, 107 (1995); J. M. Comeron and M. Kreitman, *Genetics* **150**, 767 (1998); N. G. C. Smith and L. D. Hurst, *Genetics* **153**, 1395 (1999).
6. R. Grantham, *Science* **185**, 862 (1974).
7. P. M. Sharp and W.-H. Li, *J. Mol. Evol.* **25**, 58 (1987).
8. S. Brenner, *Nature* **334**, 528 (1988).
9. We collected data sets of homologous proteins showing wide phylogenetic conservation (representing diverse eukaryotic and eubacterial lineages) and unambiguous relationships (excluding multigene families, cases of horizontal transfer, concerted evolution). For each data set, conserved protein sequences were obtained from the SWISS-PROT database (release 8/98) using BLAST (*22*). Protein sequences were aligned using CLUSTALW (*23*) and were searched for unambiguously aligned sites where serine is absolutely conserved (i.e., present in all available sequences). The corresponding codons were determined from the respective nucleotide sequences, obtained from the GenBank/EMBL (European Molecular Biology Laboratory) database. Changes in serine codon type were determined in the most parsimonious way on the basis of phylogenies (illustrated Fig. 1B), and rates of change were estimated as the number of inferred changes over the time sampled at each site (the sum of all branch lengths). Phylogenetic relationships and times of divergence were based on published data for the respective species (*20, 24*) (Fig. 1A). Trees were also constructed from the protein sequences themselves [using the Neighbor-Joining method (*25*)], and sequences showing an inconsistent phylogenetic placement were eliminated. Because of difficulties in determining

their times of divergence, eubacterial and eukaryotic sequences were treated separately and archaebacterial sequences were excluded. In estimating the rates of codon switches, we tried to be conservative, for example, by overestimating times of divergence in cases of uncertainty. The analysis was also carried out on the basis of alternative published phylogenies and the results were always robust (M. Averof *et al.*, data not shown).

10. Y. Diaz-Lazcoz, A. Henaut, P. Vigier, J. L. Risler, *J. Mol. Biol.* **250**, 123 (1995).
11. M. Kimura, in *Population Genetics and Molecular Evolution*, T. Ohta and K. Aoki, Eds. (Springer-Verlag, Berlin, 1985), pp. 19–39.
12. M. Averof and P. M. Sharp, data not shown.
13. Ubiquitin, one of the most highly conserved proteins known, is encoded by repeats sharing recent common ancestry due to concerted evolution (*7*); serine switches among repeats have been reported [P. M. Sharp and W.-H. Li, *Trends Ecol. Evol.* **2**, 328 (1987); P. M. Sharp, M. Averof, A. T. Lloyd, G. Matassi, J. F. Peden, *Philos. Trans. R. Soc. London Ser. B* **349**, 241 (1995)]. We have also found serine switches in orthologous genes from the two strains of *Helicobacter pylori* for which complete genome sequences have been determined [J.-F. Tomb *et al.*, *Nature* **388**, 539 (1997); R. A. Alm *et al.*, *Nature* **397**, 176 (1999)]. Switches were seen at sites where serine is conserved in a range of other species.
14. D. M. Irwin, *Nature* **336**, 429 (1988).
15. P. S. Goldfarb, *Nature* **336**, 429 (1988).
16. E. V. Koonin and A. E. Gorbalenya, *Nature* **338**, 467 (1989).
17. M. M. Miyamoto, B. F. Koop, J. L. Slightom, M. Goodman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 7627 (1988); W. J. Bailey *et al.*, *Mol. Phylogenet. Evol.* **1**, 97 (1992).
18. Sequences from the pseudo eta globin region of seven catarrhine primates were collected from the EMBL database (accession numbers U01317, K02542, X79724, K02543, M18038, M54985, J03818). The sequences were aligned using CLUSTALW (*23*) and were inspected manually to eliminate ambiguously aligned positions. Single- and double-nucleotide substitutions were then determined by parsimony analysis using PAUP (*26*). The number of double-nucleotide substitutions expected to have occurred by coincidence of single-nucleotide substitutions at adjacent sites was determined by resolution of the following equations: ObsS = RealS + 2(RealD), ObsD = RealD + ExpD, and ExpD = (RealS/L)^{2}L, where L is the total number of aligned nucleotides (sample size), ObsS and ObsD are the numbers of changes observed as single- and double-nucleotide substitutions, respectively, RealS and RealD are the numbers of changes that have occurred as single- and simultaneous double–nucleotide substitutions, respectively, and ExpD is the number of doublet substitutions (observed at adjacent nucleotides) that have occurred by coincidence of two separate single-nucleotide changes. The difference of ObsD and ExpD was evaluated by a chi-squared test. In mammalian genomes, CG dinucleotides are methylated, which renders them susceptible to mutation (by deamination), yielding TG and CA dinucleotides at relatively high frequencies [A. P. Bird, *Nucleic Acids Res.* **8**, 1499 (1980)]. To prevent sequential CN→CG→TG and NG→CG→CA changes (or parallel CG→TG and CG→CA mutations in independent lineages) from being counted as doublet mutations, CN→TG and NG→CA substitutions were excluded from the analysis. We also examined the rates of substitution of each of the four nucleotides separately; the biases in these rates are not sufficient to affect our results.
19. The rate of double-nucleotide substitution was estimated as 2% of the overall rate of single-nucleotide substitution (RealD/ObsS in Table 2). Average rates of neutral nucleotide substitutions are around 5 per site per Gyr (*20, 21*), yielding an estimated doublet rate of 0.1 per site per Gyr. The excess of double-nucleotide substitutions (the difference between ExpD and ObsD) is highly significant, with a chi-squared value of 14.2 (df = 1, $P \ll 0.005$).
20. H. Ochman and A. C. Wilson, *J. Mol. Evol.* **26**, 74 (1987).
21. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997); B. S. Gaut, *Evolutionary Biology*, M. K. Hecht *et al.*, Eds. (Plenum, New York, 1998), pp. 93–120; W. Makalowski and M. S. Boguski, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9407 (1998).
22. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
23. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
24. G. J. Olsen, C. R. Woese, R. Overbeek, *J. Bacteriol.* **176**, 1 (1994); D. F. Feng, G. Cho, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13028 (1997); T. M. Embley, R. P. Hirt, *Curr. Opin. Genet. Dev.* **8**, 624 (1998); J. D. Palmer, R. K. Jansen, H. J. Michaels, M. W. Chase, J. R. Manhart, *Ann. Mo. Bot. Gard.* **75**, 1180 (1988); K. H. Wolfe, M. Gouy, Y. W. Yang, P. M. Sharp, W.-H. Li, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6201 (1989); F. J. Ayala, A. Rzhetsky, F. J. Ayala, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 606 (1998); A. M. A. Aguinaldo *et al.*, *Nature* **387**, 489 (1997); S. Kumar and S. B. Hedges, *Nature* **392**, 917 (1998).
25. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
26. D. Swofford, *PAUP (Phylogenetic Analysis Using Parsimony)*, version 3.0q (Illinois Natural History Survey, Champaign, IL, 1990).
27. This work was initiated in the Genetics Department of Trinity College, University of Dublin, and was continued with the support of C. Delidakis at the IMBB, Crete. It was supported in part by the EPET II programme of the General Secretariat for Research and Technology, Greece, and a UK Biotechnology and Biological Sciences Research Council grant G04905.

12 August 1999; accepted 23 December 1999