DNA sequence evolution: the sounds of silence

PAUL M. SHARP[1], MICHALIS AVEROF[2], ANDREW T. LLOYD[3], GIORGIO MATASSI[1] AND JOHN F. PEDEN[1]

[1] *Department of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, U.K.*
[2] *Wellcome/CRC Institute, Tennis Court Road, Cambridge CB2 1QR, U.K.*
[3] *Department of Genetics, Trinity College, Dublin 2, Ireland*

SUMMARY

Silent sites (positions that can undergo synonymous substitutions) in protein-coding genes can illuminate two evolutionary processes. First, despite being silent, they may be subject to natural selection. Among eukaryotes this is exemplified by yeast, where synonymous codon usage patterns are shaped by selection for particular codons that are more efficiently and/or accurately translated by the most abundant tRNAs; codon usage across the genome, and the abundance of different tRNA species, are highly co-adapted. Second, in the absence of selection, silent sites reveal underlying mutational patterns. Codon usage varies enormously among human genes, and yet silent sites do not appear to be influenced by natural selection, suggesting that mutation patterns vary among regions of the genome. At first, the yeast and human genomes were thought to reflect a dichotomy between unicellular and multicellular organisms. However, it now appears that natural selection shapes codon usage in some multicellular species (e.g. *Drosophila* and *Caenorhabditis*), and that regional variations in mutation biases occur in yeast. Silent sites (in serine codons) also provide evidence for mutational events changing adjacent nucleotides simultaneously.

1. INTRODUCTION

The elucidation of the genetic code revealed that many sites within genes are potentially 'silent', in that they can undergo 'synonymous' nucleotide substitutions which do not change the polypeptide sequence encoded. If changes at such silent sites are truly neutral, we may study them to infer mutation patterns. However, it is possible to think of ways in which even synonymous changes can affect fitness (Clarke 1970). If evidence can be found that silent changes can be 'heard' in evolution, this would be an indication of perhaps the most subtle, and yet pervasive, form of natural selection. We have used two complementary approaches to address these questions: analyses of patterns of synonymous codon usage, and of relative rates of evolution at silent sites.

There are now a large number of species in which a sufficient number of genes have been sequenced to examine codon usage patterns. Such analyses have revealed that alternative synonymous codons are not used in equal frequencies, and that patterns of codon usage vary both among species, and among genes from the same genome. The question that arises is whether these differences among genes (and among species) reflect variation in underlying mutation patterns or the effects of natural selection (Sharp *et al.* 1993). In a small subset of these species there is a sufficient density of information that it is also possible to ask whether gene location influences codon usage (Sharp & Matassi 1994).

To investigate evolutionary rates at silent sites it is necessary to compare quite closely related species: synonymous substitutions occur rapidly (on an evolutionary timescale), and soon reach saturation. As yet, for many of the species in which codon usage has been examined, there are not much sequence data from a sufficiently close relative. In those cases that have been looked at, the extent of divergence at silent sites varies among genes, again begging the question whether these differences reflect variation in underlying mutation rates or in levels of selective constraint.

2. TWO PARADIGMS: *SACCHAROMYCES CEREVISIAE* AND *HOMO SAPIENS*

The budding yeast *Saccharomyces cerevisiae* is the eukaryote in which synonymous codon usage has been most extensively studied. Codon usage varies considerably among yeast genes (Bennetzen & Hall 1982; Ikemura 1982; Sharp *et al.* 1986). Multivariate statistical analyses reveal that there is a single major trend among genes, and that the position of a gene along this trend (i.e. the pattern of codon usage in that gene) is related to its expression level (Sharp & Cowe 1991). At one end of this trend lie highly expressed genes, such as those encoding ribosomal proteins and glycolytic enzymes, with very biased codon usage; at the other extreme are lowly expressed genes, with much more uniform codon usage. The one or (in some cases) two codons for each amino acid which are heavily used in highly expressed genes are those best recognized by the most abundant tRNAs (Ikemura 1982). Thus, the yeast cell shows a high degree of co-adaptation of its codon usage and tRNAs, and this species stands as a paradigm of how natural selection can influence silent sites in eukaryotic genes.

Phil. Trans. R. Soc. Lond. B (1995) **349**, 241–247
Printed in Great Britain

It is not yet possible to examine the relative rates of silent substitution in various yeast genes. The most closely related species from which a significant number of gene sequences are known is *Kluyveromyces lactis*. Codon usage patterns in *K.lactis* appear to be essentially similar to *S.cerevisiae* (Lloyd & Sharp 1993; Freire-Picos *et al.* 1994), but these two species are sufficiently divergent that silent sites have been almost saturated with substitutions. An appropriately closely related species is *Saccharomyces douglasii* (Adjiri *et al.* 1994), but as yet, very few genes from that species have been studied.

Codon usage also varies considerably among human genes. However, in contrast to yeast, the patterns of codon usage in different human genes have not been related (directly) to any aspects of their expression. In fact, different genes expressed in similar amounts in the same tissues (for example, those encoding alpha and beta globin; Sharp *et al.* 1993) have quite different codon usage. Nevertheless, multivariate statistical analyses reveal that there is a single major trend in codon usage among human genes. In this case, genes at one end of the trend are G+C-rich at silent sites, while genes at the other end are relatively A+T-rich, and there is a consistent trend in G+C content between these two extremes which affects all sets of synonymous codons (Marin *et al.* 1989). Because the G+C content at silent sites in a gene is correlated with the G+C content in the 5′ and 3′ flanking regions and introns of the gene (Aota & Ikemura 1986) this appears to be a regional effect. In fact, this is consistent with the 'isochore' hypothesis (Bernardi *et al.* 1985; Bernardi 1993): the results of fractionation of mammalian genomic DNA followed by isopycnic ultracentrifugation have been taken to indicate that G+C content is relatively homogeneous along long (perhaps 300–1000 kb) chromosomal region ('isochores'), but varies among regions.

The question then arises as to why different chromosomal regions should have different G+C content. Bernardi has consistently argued (see, for example, Bernardi 1993) that this must be adaptive, but admits that the mechanism is 'elusive'. From a consideration of population genetics, it is clear that natural selection cannot control G+C content at the individual nucleotide level across the entire human genome. Others have suggested that isochores may result from mutation patterns that vary around the mammalian genome (reviewed in Holmquist & Filipski 1994), and codon usage patterns are consistent with this hypothesis (Eyre-Walker 1991). Indeed, if natural selection has shaped the base composition of these large chromosomal regions, it would most likely act through influencing mutation rates; again the reason is unknown, and whether such selection could be effective is unclear.

Mammals are the eukaryotes in which rates of synonymous substitution have been most extensively studied. From the earliest reports, it was clear that synonymous substitution rates vary much less among genes than do nonsynonymous rates (Miyata *et al.* 1980; Kimura 1981). Nevertheless, while this observation is still valid, synonymous substitution rates vary much more than would be expected by chance alone. For example, among 363 genes compared between mouse and rat, the standard deviation of synonymous substitution rates was about twice that expected (Wolfe & Sharp 1993). Furthermore, this variation appears to be systematic, in that genes retain their relative rates in different mammalian lineages (Bulmer *et al.* 1991). If synonymous substitutions in mammalian genes are largely effectively neutral, we may infer that mutation rates vary systematically among genes.

3. CODON SELECTION IN METAZOA

Does the difference between yeast, where silent sites are shaped by natural selection, and mammals, where they seem to be determined by mutation patterns, reflect a dichotomy between unicellular and multicellular organisms? (Ikemura 1985). More than 20 years ago it was shown that amino acid usage and tRNA abundance are highly coadapted in cells in the silk gland of *Bombyx mori* (Garel 1974). The posterior silk gland produces fibroin (which is rich in glycine and alanine), while the middle silk gland produces sericin (rich in serine); in each case these proteins are produced in very large amounts. At the end of the larval stage of development, the tRNA populations in each type of cell change appropriately. However, this is an extreme situation, and appears to be an adaptation of tRNA pools to the increased requirement for certain amino acids, rather than any selection of codon usage.

Another insect, *Drosophila melanogaster*, is the invertebrate species in which codon usage has been most extensively studied. As with the species discussed above, codon usage varies considerably among genes, and the obvious question is whether the *Drosophila* genome follows the yeast or human paradigm (or is different again). Multivariate analyses again reveal a single major trend among *Drosophila* genes, with genes at one extreme exhibiting strong preference for a restricted subset of codons (Shields *et al.* 1988). Assessing gene expression level in a multicellular organism is rather more difficult than in yeast, but in *Drosophila* there does appear to be a positive correlation between the strength of codon usage bias in a gene and its expression level. When sequences are compared among *Drosophila* species, the rate of synonymous substitution varies quite considerably among genes (Sharp & Li 1989; Moriyama & Hartl 1993). In genes with high codon usage bias, the extent of silent site divergence is much lower, consistent with constraint imposed by codon selection. Thus, it has been deduced that natural selection shapes codon usage in *D.melanogaster* (Shields *et al.* 1988; Moriyama & Hartl 1993; Sharp & Lloyd 1993*b*).

It has been suggested that the *Drosophila* genes with the strongest codon usage bias are not (necessarily) those expressed at the highest levels, but rather those encoding proteins whose primary structure is the most important; that is, selection is for accuracy, rather than efficiency, of translation (Akashi 1994). More conserved amino acids, and those in functionally more

important protein domains, tend to be more often encoded by optimal codons. The data presented are quite striking, but there are a sufficient number of counter-examples to suggest that both accuracy and efficiency are important in this species (Sharp & Matassi 1994).

Recently we have investigated codon usage patterns in a member of another invertebrate phylum, namely the nematode *Caenorhabditis elegans* (Stenico *et al.* 1994). Codon usage variation among *C.elegans* genes appears to be related to gene expression level, in a manner very similar to *Drosophila*. A small number of genes, including representatives with different levels of codon usage bias, can be compared between *C.elegans* and *C.briggsae* (Kennedy *et al.* 1993; Stenico *et al.* 1994). As in *Drosophila*, genes with higher codon usage bias exhibit lower silent site divergence; in this case, genes with low codon usage bias are almost saturated with synonymous changes.

To understand why natural selection can influence codon usage in some metazoa but not others, we need to consider the population genetic aspects of codon selection. The selective differences between alternative synonymous codons are expected to be very small. Natural selection is expected to overcome the randomizing effects of genetic drift, and thus influence codon usage, only if the long-term evolutionary effective population size (N_e) is larger than the reciprocal of the selection coefficient (Wright 1931; Li 1987). We suggested that N_e in *Drosophila melanogaster* has probably been just larger than this critical value (Shields *et al.* 1988). This appears to have been borne out by a recent analysis (Kliman & Hey 1993). Natural selection should be less effective in shaping codon usage in genes located in regions of the genome with reduced recombination, for two reasons. First, selection has great difficulty choosing, simultaneously, the best variants at multiple sites if those sites are tightly linked (Hill & Robertson 1966). Second, any effect of natural selection in determining the fate of alternative synonymous alleles may be swamped by selection at closely linked sites at which alternative alleles (e.g. nonsynonymous changes) confer larger fitness differences (Maynard Smith & Haigh 1974). The lower the local recombination rate, the more 'closely linked' sites there are likely to be. In *D.melanogaster*, regions of the chromosome near the centromeres and telomeres have lower recombination rates, and genes located in these regions were found to have lower codon usage bias than other genes (Kliman & Hey 1993).

4. MUTATION PATTERN VARIATION AROUND THE MAMMALIAN GENOME

If mutation patterns vary among chromosomal regions, this should be reflected in variation in both codon usage and the rate of synonymous substitution. As noted above, there is considerable variation in codon usage among mammalian genes: values of the G+C content at synonymously variable third positions of codons ($GC3_s$) for human genes range from about 30% to about 90% (Ikemura 1985). If this is indeed

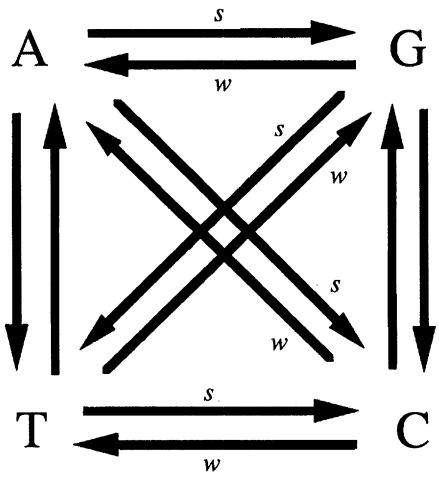

Figure 1. Mutation rates and patterns. If the ratio of sum of the rates marked *w* to the sum of the rates marked *s* varies among genes, then G+C-content at neutral sites will vary (note that different rates denoted by the same symbol are not presumed to be equal). If the sum of all 12 rates varies among genes, then the neutral substitution rate will vary.

caused by mutation patterns, we can infer that the balance between the sum of the four mutation rates from A or T to C or G and the sum of the four rates from C or G to A or T (figure 1) varies among genes. If these individual rates vary, then it would perhaps be surprising if the sum of all 12 rates did not also vary among genes.

If both mutation rates and patterns vary among genes, they might be correlated. Initially it seemed that this was the case, with mammalian genes of intermediate G+C content having faster substitution rates (Wolfe *et al.* 1989*a*), but as more data have accumulated the observed relation has become progressively weaker (Bulmer *et al.* 1991; Wolfe & Sharp 1993). Nevertheless, if mutation rates vary around the genome, we might expect closely linked genes to have both similar G+C contents and similar substitution rates. Some studies have provided evidence for this (Ikemura & Aota 1988; Wolfe *et al.* 1989*a*; Ikemura *et al.* 1990). However, in each case only a small number of genes were examined, and these often included members of duplicate gene families; the latter may have similar evolutionary patterns caused by common ancestry.

To examine this question in more detail, we have investigated codon usage ($GC3_s$), and rate of synonymous substitution (K_s) in more than 50 groups of genes which are tightly linked (located at the same genetic map position) in both the human and mouse (or rat) genomes (G. Matassi, C. Gautier & P. M. Sharp, in preparation). If evolutionary processes are influenced by genomic location, we would expect genes within these groups to exhibit rates which are more similar to each other than to genes from other locations. We have calculated correlation statistics and then asked in how many of 10000 simulated datasets (in which genes were randomly shuffled into groups of similar sizes to the observed dataset) a similarly large correlation was found. For both $GC3_s$ and K_s, less than 0.1% of the simulation datasets exhibited as high a

correlation. This was true even when duplicated genes were excluded. Thus, silent sites in genes located close to one another evolve in a relatively similar fashion, suggesting that their mutation patterns are indeed influenced by chromosome location.

5. MUTATION PATTERN VARIATION IN OTHER GENOMES

Saccharomyces cerevisiae is the first organism for which complete chromosome sequences have been determined. It has therefore been possible to ask whether the evolution of a yeast gene is influenced by its genomic position. Following the completion of the sequence of yeast chromosome III (Oliver *et al.* 1992), we found that gene location does indeed seem to influence silent sites (Sharp & Lloyd 1993*a*). This chromosome is about 315 kb in length, with the centromere 113 kb from the left end. Genes located in regions of about 40–60 kb in the middle of each chromosome arm were found to be more G+C-rich at silent sites; genes in these regions have an average $GC3_s$ of 46 %, compared with 36 % in the surrounding regions. The complete sequences of three other yeast chromosomes have subsequently been published (Dujon *et al.* 1993; Johnston *et al.* 1994; Feldmann *et al.* 1994). These chromosomes have similar regions wherein genes have elevated G+C content; each of these chromosomes is longer than chromosome III, and they have a proportionately larger number of G+C-rich regions (see figure 2).

Whether this pattern of regional G+C differences in the yeast genome reflects the same evolutionary process(es) as the isochores found in the mammalian genome remains an open question. The only factor, so far, correlated with base composition in yeast is gene density: the G+C-rich regions are also more gene rich (Dujon *et al.* 1993; Johnston *et al.* 1994; Feldmann *et al.* 1994; Sharp & Matassi 1994). This also appears to be true in the mammalian genome (reviewed in Sharp & Matassi 1994).

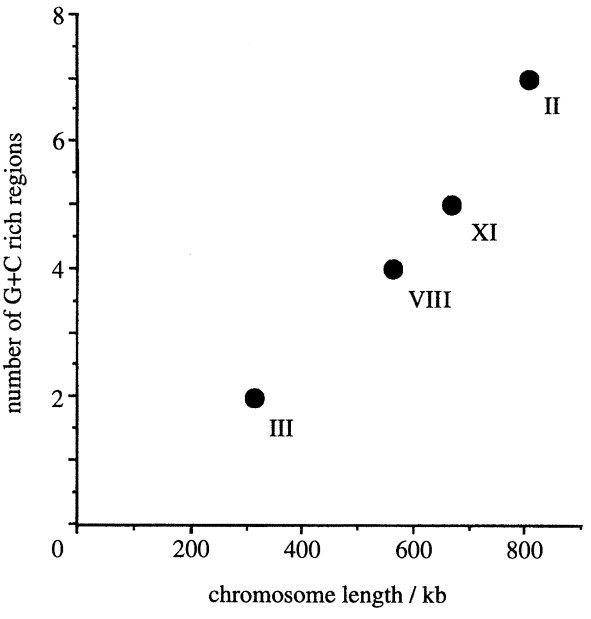

Figure 2. G+C-rich regions in the *Saccharomyces cerevisiae* genome. The number of G+C-rich regions is plotted as a function of chromosome length, for each of the four chromosomes with complete sequences published. Data drawn from: chromosome II (Feldmann *et al.* 1994), chromosome III (Oliver *et al.* 1992; Sharp & Lloyd 1993); chromosome VIII (Johnston *et al.* 1994); chromosome XI (Dujon *et al.* 1994).

The existence of genomic regions of different G+C content in organisms as diverse as mammals and yeast prompts the question of how widespread this phenomenon is. Genes in *Drosophila melanogaster* vary greatly in their G+C content at silent sites, but this can be largely explained by codon selection; in this species many of the optimal codons end in C (Shields *et al.* 1988; Sharp & Lloyd 1993*b*). However, it has recently been reported that, among lowly biased genes, there is a correlation between G+C content at silent sites and in introns (Kliman & Hey 1994), suggesting that there may be regional effects.

Investigation of whether there are G+C regions in the *Caenorhabditis elegans* genome is also complicated by the fact that many optimal codons are C-ending (Stenico *et al.* 1994). A 2200 kb region of *Caenorhabditis elegans* chromosome III has been published (Wilson *et al.* 1994), but our analyses have (as yet) revealed no significant evidence for G+C regionalization in that genome.

6. DOUBLET MUTATIONS

An unusual type of silent site in serine codons may reveal a previously unheralded category of mutational events. The structure of the genetic code is such that serine is unique in being encoded by two sets of triplets (TCN and AGY; N is any base, Y is a pyrimidine) that cannot be interconverted by one nucleotide mutation. The likelihood of a single mutational event simultaneously changing both nucleotides in the first two positions of a serine codon has been considered to be very low, and so a switch between one type of serine codon and the other would require going through intermediate triplets that encode other amino acids (TGY for cysteine or ACY for threonine). At sites where serine is essential to protein function, codon usage is thus thought to be locked into one set of triplets or the other (for example, see Marin *et al.* 1989). For example, it has been suggested that the presence of one or the other type of codon for the serine at the active site of serine proteases must reflect a fundamental dichotomy in this family of genes (Brenner 1988).

However, while examining the evolution of ubiquitin genes we found evidence of several instances of switching between TCN and AGY codons (Sharp & Li 1987*a*, *b*). Ubiquitin genes are unusual in a number of ways. First, they encode an extraordinarily highly conserved protein. Second, some ubiquitin genes consist of tandem repeats of the 76 codon unit required to encode this protein. Within these polyubiquitin genes, there are numerous nucleotide sequence variations between the repeats but (with a very small number of exceptions) these differences are all silent. Interestingly, some of these silent differences involve the first two positions of serine codons. For example, a

	codon 19	20	57	65
unit 1	TCG	TCG	TCC	TCT
unit 2	TCG	TCC	AGC	TCT
unit 3	TCT	TCC	TCC	AGC
unit 4	TCT	TCG	TCT	AGC

Figure 3. Serine codons at four sites in the four repeat units of a *Neurospora crassa* polyubiquitin gene (Taccioli *et al.* 1989).

tetraubiquitin gene from *Neurospora crassa* (Taccioli *et al.* 1989) encodes four identical protein sequences, each with four serine residues: at two of these sites there are serine codon switches among the repeats (see figure 3). Residue 65 appears to be conserved as serine among fungi, plants, animals and even a range of protists; the only exception of which we are aware is *Giardia lamblia* (a representative of probably the earliest diverging eukaryotic lineage; Hashimoto *et al.* 1994), where serine is replaced by alanine. Residue 57 is also conserved as serine among fungi and animals, although it is replaced by alanine in plants and some protists and by glutamic acid in *Giardia*. The various repeats in polyubiquitin genes undergo concerted evolution (Sharp & Li 1987*a*, *b*; Tan *et al.* 1993), and so it is remarkable to find both types of serine codon at homologous positions in different repeat units within the same gene. Serine codon switches are also seen within polyubiquitin genes in *S.cerevisiae* and *Dictyostelium discoideum*.

Prompted by this observation we have examined codon usage at extremely highly conserved serine residues in a number of other highly conserved proteins: switches between the two types of serine codon are surprisingly common (M. Averof & P. M. Sharp, in preparation). By placing the proteins on phylogenetic trees we have estimated the minimum number of switches required to explain the current distribution of serine codon types. By summing the total branch lengths in these trees (in terms of time), we have estimated the rate to be approximately 0.1 switches per site per 10^9 years. Rates of (single nucleotide) synonymous substitution have been estimated to be of the order of 1–20 substitutions per site per 10^9 years across a range of eukaryotes (Wolfe *et al.* 1989*b*). Thus, the rate of serine codon switches is very much higher than would be expected from the coincidental occurrence of independent nucleotide substitutions at adjacent sites. This may indicate that the switches are the result of doublet mutations of the form $T_pC \leftrightarrow A_pG$, or of $C_pT \leftrightarrow G_pA$ on the complementary DNA strand.

A possible explanation that does not invoke doublet mutation involves two successive, compensatory, nucleotide substitutions. In an allele with a deleterious mutation to either a threonine or cysteine codon, a second mutation reverting to a serine codon would be advantageous. This 'reversion' could be a precise back mutation, but could also be a forward mutation to the other type of serine codon. Compensatory substitutions of this kind are expected to be fixed quickly (Kimura 1985), so that the intermediate stage of a non-serine codon may be fleeting, and not observed. However, this raises the question of how the deleterious non-serine codon was initially able to persist in the population.

Another line of evidence suggests that doublet mutations are indeed more common than might have been expected. A large region of noncoding DNA from the beta-globin cluster has been sequenced in a number of higher primates (Bailey *et al.* 1992). Within this region, substitutions at adjacent sites are far more common than expected (K. H. Wolfe & P. M. Sharp, unpublished observations). In the absence of any indication that these sites have any functional importance, we infer that the excess doublet substitutions have resulted from doublet mutations. In the beta globin region these do not involve only the types of mutations seen in serine codons, and of course we would have had no reason to believe that only mutations of the form $T_pC \leftrightarrow A_pG$ (or $C_pT \leftrightarrow G_pA$) occur at high rates. Rather, we conclude that this unique category of silent sites in serine codons are simply highlighting a more general phenomenon.

7. CONCLUSIONS AND PERSPECTIVES

Study of silent sites and silent substitutions has produced a number of interesting insights into both subtle natural selection and mutational processes. Burgeoning gene sequence databases will allow the generality of these phenomena to be investigated. Perhaps the most exciting new dimension in molecular evolutionary studies in the last few years has been the growing realization that genomic location can influence the way in which a gene evolves (Sharp & Matassi 1994). The various rapidly advancing genome projects will enable this to be examined in more detail, and may well provide insights as to how this effect is brought about. It is worth emphasizing that, in studying the fundamental processes of gene evolution, and indeed in evaluating the functional importance of sites in genes (and thus in interpreting the results of genome sequencing), the comparative analysis of closely related species can be invaluable. We have reviewed the results of some such studies above, but for many of the model organisms that are the subject of genome projects there has been (as yet) no concerted effort to determine homologous sequences from a closely related species.

REFERENCES

Adjiri, A., Chanet, R., Mezard, C. & Fabre, F. 1994 Sequence comparison of the *ARG4* chromosomal regions from the two related yeasts, *Saccharomyces cerevisiae* and *Saccharomyces douglasii*. *Yeast* **10**, 309–317.

Akashi, H. 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935.

Aota, S-i. & Ikemura, T. 1986 Diversity in G + C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.* **14**, 6345–6355.

Bailey, W. J., Hayasaka, K., Skinner, C. G. *et al.* 1992

Reexamination of the African hominoid trichotomy with additional sequences from the primate β-globin gene cluster. *Molec. Phylogenet. Evol.* **1**, 97–135.

Bennetzen, J. L. & Hall, B. D. 1982 Codon selection in yeast. *J. Biol. Chem.* **257**, 3026–3031.

Bernardi, G. 1993 The vertebrate genome: isochores and evolution. *Molec. Biol. Evol.* **10**, 186–204.

Bernardi, G., Olofsson, B., Filipski, J. *et al.* 1985 The mosaic genome of warm-blooded vertebrates. *Science, Wash.* **228**, 953–958.

Brenner, S. 1988 The molecular evolution of genes and proteins: a tale of two serines. *Nature, Lond.* **334**, 528–530.

Bulmer, M., Wolfe, K. H. & Sharp, P. M. 1991 Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationships of mammalian orders. *Proc. natn. Acad. Sci. U.S.A.* **88**, 5974–5978.

Clarke, B. C. 1970 Darwinian evolution of proteins. *Science, Wash.* **168**, 1009–1011.

Dujon, B., Alexandraki, D., Andre, B. *et al.* 1994 Complete DNA sequence of yeast chromosome XI. *Nature, Lond.* **369**, 371–378.

Eyre-Walker, A. C. 1991 An analysis of codon usage in mammals: selection or mutation bias? *J. molec. Evol.* **33**, 442–449.

Feldmann, H., Aigle, M., Aljinovic, G. *et al.* 1994 Complete DNA sequence of yeast chromosome II. *EMBO J.* **13**, 5795–5809.

Freire-Picos, M. A., Gonzalez-Siso, M. I., Rodriguez-Belmonte, E., Rodriguez-Torres, A. M., Ramil, E. & Cerdan, M. E. 1994 Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes. *Gene* **139**, 43–49.

Garel, J. P. 1974 Functional adaptation of tRNA population. *J. theor. Biol.* **43**, 211–225.

Hashimoto, T., Nakamura, Y., Nakamura, F., Shirakura, T., Adachi, J., Goto, N., Okamoto, K.-i. & Hasegawa, M. 1994 Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. *Molec. Biol. Evol.* **11**, 65–71.

Hill, W. G. & Robertson, A. 1966 The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294.

Holmquist, G. P. & Filipski, J. 1994 Organization of mutations along the genome, a prime determinant of genome evolution. *Trends Ecol. Evol.* **9**, 65–69.

Ikemura, T. 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. *J. molec. Biol.* **158**, 573–597.

Ikemura, T. 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Molec. Biol. Evol.* **2**, 13–34.

Ikemura, T. & Aota, S.-i. 1988 Global variation in G+C content along vertebrate genome DNA. *J. molec. Biol.* **203**, 1–13.

Ikemura, T., Wada, K.-n. & Aota, S.-i. 1990 Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. *Genomics* **8**, 207–216.

Johnston, M., Andrews, S., Brinkman, R. *et al.* 1994 Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science, Wash.* **265**, 2077–2082.

Kennedy, B. P., Aamodt, E. J., Allen, F. L., Chung, M. A., Heschl, M. F. P. & McGhee, J. D. 1993 The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J. molec. Biol.* **229**, 890–908.

Kimura, M. 1981 Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. natn. Acad. Sci. U.S.A.* **78**, 454–458.

Kimura, M. 1985 Diffusion models in population genetics with special reference to fixation time of molecular mutants under mutational pressure. In *Population genetics and molecular evolution*, (ed. T. Ohta & K. Aoki), pp. 19–39. Berlin: Springer-Verlag.

Kliman, R. M. & Hey, J. 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Molec. Biol. Evol.* **10**, 1239–1258.

Kliman, R. M. & Hey, J. 1994 The effects of mutation and natural selection on codon bias in the genes of Drosophila. *Genetics* **137**, 1049–1056.

Li, W.-H. 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. molec. Evol.* **24**, 337–345.

Lloyd, A. T. & Sharp, P. M. 1993 Synonymous codon usage in *Kluyveromyces lactis*. *Yeast* **9**, 1219–1228.

Marin, A., Bertranpetit, J., Oliver, J. L. & Medina, J. R. 1989 Variation in G+C-content and codon choice: differences among synonymous codon groups in vertebrate genes. *Nucleic Acids Res.* **17**, 6181–6189.

Maynard Smith, J. & Haigh, J. 1974 The hitch-hiking effect of a favorable gene. *Genet. Res.* **23**, 23–35.

Miyata, T., Yasunaga, T. & Mishida, T. 1980 Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc. natn. Acad. Sci. U.S.A.* **77**, 7328–7332.

Moriyama, E. N. & Hartl, D. L. 1993 Codon usage bias and base composition of nuclear genes in Drosophila. *Genetics* **134**, 847–858.

Oliver, S. G., van der Aart, Q. J. M., Agostoni-Carbone, M. L. *et al.* 1992 The complete DNA sequence of yeast chromosome III. *Nature, Lond.* **357**, 38–46.

Sharp, P. M. & Cowe, E. 1991 Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7**, 657–678.

Sharp, P. M. & Li, W.-H. 1987*a* Ubiquitin genes as a paradigm of concerted evolution of tandem repeats. *J. molec. Evol.* **25**, 58–64.

Sharp, P. M. & Li, W.-H. 1987*b* Molecular evolution of ubiquitin genes. *Trends Ecol. Evol.* **2**, 328–332.

Sharp, P. M. & Li, W.-H. 1989 On the rate of DNA sequence evolution in *Drosophila*. *J. molec. Evol.* **28**, 398–402.

Sharp, P. M. & Lloyd, A. T. 1993*a* Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. *Nucleic Acids Res.* **21**, 179–183.

Sharp, P. M. & Lloyd, A. T. 1993*b* Codon usage. In *An atlas of* Drosophila *Genes*, (ed. G. P. Maroni), pp. 378–397. New York: Oxford University Press.

Sharp, P. M. & Matassi, G. 1994 Codon usage and genome evolution. *Curr. Opin. Genet. Devel.* **4**, 851–860.

Sharp, P. M., Stenico, M., Peden, J. F. & Lloyd, A. T. 1993 Codon usage: mutational bias, translational selection, or both? *Biochem. Soc. Trans.* **21**, 835–841.

Sharp, P. M., Tuohy, T. M. F. & Mosurski, K. R. 1986 Codon usage in yeast: Cluster analysis clearly differentiates between highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125–5143.

Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. 1988 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among alternative synonymous codons. *Molec. Biol. Evol.* **5**, 704–716.

Stenico, M., Lloyd, A. T. & Sharp, P. M. 1994 Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.* **22**, 2437–2446.

Taccioli, G. E., Grotewold, E., Aisemberg, G. O. & Judewicz, N. D. 1989 Ubiquitin expression in *Neurospora crassa*: cloning and sequencing of a polyubiquitin gene. *Nucleic Acids Res.* **17**, 6153–6165.

Tan, Y., Bishoff, S. T. & Riley, M. A. 1993 Ubiquitins revisited: further examples of within- and between-locus concerted evolution. *Molec. Phylogenet. Evol.* **2**, 351–360.

Wilson. R., Ainscough, R., Anderson, K. *et al.* 1994 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature, Lond.* **368**, 32–38.

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

Wolfe, K. H., Sharp, P. M. & Li, W.-H. 1989*a* Mutation rates differ among regions of the mammalian genome. *Nature, Lond.* **337**, 283–285.

Wolfe, K. H., Sharp, P. M. & Li, W.-H. 1989*b* Rates of synonymous substitution in plant nuclear genes. *J. molec. Evol.* **29**, 208–211.

Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97–159.