Nucleic Acids Research, 1993, Vol. 21, No. 2 179–183

# Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure

Paul M.Sharp and Andrew T.Lloyd
Department of Genetics, Trinity College, Dublin 2, Ireland

Received November 25, 1992; Revised and Accepted December 14, 1992

## ABSTRACT

**The recent determination of the complete sequence of chromosome III from the yeast *Saccharomyces cerevisiae* allows, for the first time, the investigation of the long range primary structure of a eukaryotic chromosome. We have found that, against a background G+C level of about 35%, there are two regions (one in each chromosome arm) in which G+C values rise to over 50%. This effect is seen in silent sites within genes, but not in noncoding intergenic sequences. The variation in G+C content is not related to differential selection of synonymous codons, and probably reflects mutational biases. That the intergenic regions do not exhibit the same phenomenon is particularly interesting, and suggests that they are under substantial constraint. The yeast chromosome may be a model of the structure of the human genome, since there is evidence that it is also a mosaic of long regions of different base compositions, reflected in wide variation of G+C content at silent sites among genes. Two possible causes of this regional effect, replication timing, and recombination frequency, are discussed.**

## INTRODUCTION

As vast amounts of DNA sequence data accumulate, it is becoming possible to investigate whether the sequences of genes (and their patterns of evolution) are influenced by their chromosomal location. Perhaps the most striking result has been the demonstration that base composition (G+C content) varies enormously among mammalian genes and appears to be related to their genomic context (1). For example, G+C content at third codon positions (which are generally silent) varies among human genes from less than 30% to over 90% (2), these values are correlated with the G+C content of the introns and flanking sequences of the same genes (3), and also among neighbouring genes (4,5), suggesting that the base composition variation is a local chromosomal effect. This has been interpreted as evidence that the human genome comprises a mosaic of long regions ('isochores') of different base composition (1,6). A similar genomic structure has been suggested for birds (1,6), and certain monocotyledenous plants (7), although the data are as yet less numerous. It is interesting, particularly with respect to elucidating the origins and possible significance of this phenomenon, to ask how widespread it is among other organisms.

Here we investigate base composition variation among genes from the budding yeast *Saccharomyces cerevisiae*. In particular, we have examined the recently determined complete sequence of chromosome III (8); this chromosome is 315kb long, and contains 160–180 presumptive genes (8,9). The results indicate position dependent variation in gene G+C content. Several factors which may be related to, and perhaps cause this variation, namely (i) codon usage, (ii) time of replication, and (iii) frequency of recombination, are investigated. The latter two aspects are probably better understood for this chromosome than for any other, and codon usage has been more extensively analysed in *S.cerevisiae* than in any other eukaryote (2,10–13).

## MATERIALS AND METHODS

From the complete sequence (8) of chromosome III (GenBank/EMBL/DDBJ accession number X59720), all 182 featured open reading frames (ORFs), and the intergenic regions between them, were extracted using the ACNUC sequence retrieval system (14). The featured ORFs are open reading frames of more than 100 codons in length identified in Ref.8. Two transposable elements, Ty2 (8) and Ty5 (15), were excluded because codon usage in transposable element genes differs from that in chromosomal genes (16). Two open reading frames (L26c, R74c; in the terminology of Ref.8) were replaced by others which completely overlap them (LX8c, RX13w; see the GenBank entry), because the latter are longer, and have codon positional nucleotide frequencies more typical of yeast genes (13). One gene identified recently (17) as *RIM1* was added. This yielded a total of 178 genes. There were only 153 intergenic sequences because some ORFs overlap slightly. Introns (only two have been identified on this chromosome) were excluded from the analysis. It is possible that some of these ORFs are not genes, though it is quite unlikely that long open reading frames exist by chance in DNA of this base composition. In addition, in a sequence of this length there may well be some errors (see, for example, Ref.18), but we have found the overall results to be robust against minor changes in the dataset.

To avoid any effects of amino acid composition, G+C content for genes was calculated only from silent third codon positions (i.e., excluding Met, Trp and termination codons). Codon usage

**Table 1.** G+C content at silent sites in *Saccharomyces cerevisiae*.

| | No. of sequences | G+C content[a] Mean ± SD (O/E) | Codon usage bias[b] Mean | CAI | Rho |
|---|---|---|---|---|---|
| All genes[c] | 575 | 0.38 ± 0.07 (2.4) | 0.27 | 0.25 | |
| Chromosome III genes[d] | 178 | 0.40 ± 0.09 (2.8) | 0.16 | −0.05 | |
| Chromosome III noncoding | 153 | 0.34 ± 0.05 (1.7) | 1 | – | |

[a] For genes, G+C content was calculated only for silent third codon positions; O/E is the ratio of the Observed/Expected standard deviations.
[b] Codon usage bias is measured by CAI, the codon adaptation index (19); Rho is the correlation coefficent of G+C content with CAI.
[c] 575 genes from throughout the genome, listed in Ref.13.
[d] 178 open reading frames (presumptive genes) from chromosome III (see Materials and Methods for details).

bias was measured by CAI, the codon adaptation index (19); CAI values can range between 0 and 1, with higher values indicating stronger bias. Values for silent site G+C content and CAI were calculated using the CODONS program (20). The expected standard deviation of G+C content among genes was calculated from the binomial theory, using the harmonic mean length of sequences.

## RESULTS

To analyse possible regional base composition variation in yeast we have examined 'silent' sites, i.e., synonymously variable sites within genes, and noncoding sequences, on the complete sequence (8) of chromosome III. By comparison with a set of 575 gene sequences from throughout the *S.cerevisiae* genome (13), it was established that chromosome III appears to be representative of the genome as a whole (Table 1). The G+C content of the entire chromosome is 39%, while that for the genome has been estimated (21) at 39−40%. The average value for G+C content at silent sites in ORFs on chromosome III is similar to that for all genes, and to the genomic base composition; chromosome III genes exhibit at least as much variation as the genome-wide dataset. The average G+C content in noncoding regions on chromosome III is somewhat lower than the average at silent sites in chromosome III genes; it is also less variable, both when observed standard deviations are compared, and when the ratio of observed/expected standard deviations are considered (Table 1). The G+C content value for each gene was plotted against its position on chromosome III; the more G+C-rich genes did not appear to be randomly distributed (Fig. 1a). When the corresponding data for noncoding sequences were plotted, their lower variability was evident, and there was less sign of any spatial pattern (Fig. 1b). Regional effects became much clearer when moving average values for 15 adjacent sequences (genes, or intergenic regions) were considered (Fig. 2). Genes with a high G+C content were seen to be clustered in two major regions, one on each arm of the chromosome (Fig. 2a). Similar patterns of regionalization of base composition variation were evident for genes coded on each strand of the DNA (Fig. 2b). Noncoding sequences did not show similar patterns: the only apparent (small) peak was near the centromere (Fig. 2a).

To test whether the clustering of the more G+C-rich genes was significantly greater than might occur by chance, we performed simulations in which the order of genes was randomized, and the moving average G+C values for 15 adjacent sequences were recomputed. We then asked whether any values in the simulations were as high as those observed in the analysis in Fig. 2a, where the observed values at the peaks in the left and right chromosome arms were 0.48 and 0.52, respectively. Out of 10,000 simulations in which all genes were shuffled, only 8 yielded a peak value greater than 0.52, but 426 yielded values greater than 0.48. Therefore, separate simulations were performed for genes from the two chromosome arms: 8 out of 10,000 simulations for the left arm genes yielded a peak value greater than 0.48, and 3 out of 10,000 simulations for the right arm yielded a value greater than 0.52. These simulations provide strong evidence that the clustering of G+C-rich genes in the two arms is highly significant ($P < 0.001$).

The chromosome was then considered as if consisting of five regions: two areas of high genic G+C content (the peaks in Fig. 2a, taken to be located at 40−80 kb, and 200−260 kb, respectively), and the three surrounding areas. The mean G+C content values for genes and intergenic regions within these areas are given in Table 2. With respect to mean intergenic G+C content, there was no significant difference among the five regions (analysis of variance: $F_{4,148\ d.f.} = 0.41$, $p = 0.80$); for genes, there was no significant difference among the three regions with low G+C content ($F_{2,107\ d.f.} = 0.64$, $p = 0.53$), or between the two regions with high G+C content (t-test: $t_{66\ d.f.} = 1.4$, $p = 0.16$). Pooling the regions of high and low genic G+C content, respectively, the mean G+C content values for genes differed significantly between the two types of region ($t_{176\ d.f.} = 9.08$, $p < 0.0001$), but the mean values for intergenic sequences did not ($t_{151\ d.f.} = 0.43$, $p = 0.67$). However, these statistical tests must be taken with some caution, because the regions were designated subjectively after considering Fig. 2a.

The relationship between silent site G+C content and codon usage bias (measured by the CAI) of yeast genes was investigated. Among genes from all chromosomes there was a small positive correlation between the strength of codon usage bias and G+C content, but for chromosome III genes there was no significant correlation (Table 1). Furthermore, codon usage bias values showed no relation to chromosome position, either when individual genes were considered (Fig. 3a), or when a moving average was plotted (Fig. 3b).

## DISCUSSION

Mammalian chromosomes appear to be a mosaic of 'isochores', regions several hundred kilobases in length which differ in base composition, but within which base composition is relatively homogeneous (1,6). Recently, it has been suggested that variation in base composition may be similarly dependent on genomic location in a wide range of (if not all) eukaryotes and even

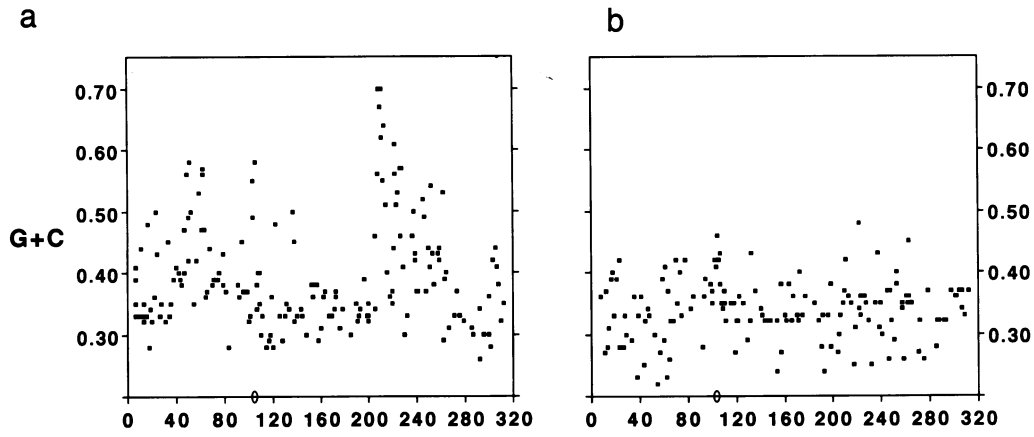


**Figure 1.** G+C content at silent sites along yeast chromosome III. **a**, 178 open reading frames. **b**, 153 intergenic sequences. The ellipse indicates the position of the centromere.

prokaryotes (22,23). However, the latter studies have not considered chromosome position and have focussed on G+C content at the third positions of genes, without taking into account variation in codon usage. This point is critical, because in diverse species (e.g., an insect, *Drosophila melanogaster* (24); a fungus, *Aspergillus nidulans* (25); a slime mould, *Dictyostelium discoideum* (26); and an enteric bacterium, *Serratia marcescens* (27)), analyses have indicated that silent site G+C content variation among genes can be largely accounted for by the frequencies of certain 'optimal' codons; these frequencies are in turn related to the level of gene expression.

Here we have demonstrated that genes from different regions of *S.cerevisiae* chromosome III have different levels of silent site G+C content. This is particularly interesting because it is quite unexpected: an earlier examination of long *S.cerevisiae* sequences revealed their G+C content values 'to be within a narrow range around that of the whole genome' (4), while in the studies cited above (22,23) yeast was the one eukaryote in which those authors did not find much G+C variation among genes. There has been extensive documentation of the fact that yeast genes vary considerably in codon usage bias, depending on their level of expression—can this explain the variation in G+C content? In yeast genes expressed at a high level, codon usage is very biased (10), and 22 translationally 'optimal' codons have been identified (13): in the most strongly biased genes very few other codons are used. However, of these 22 optimal codons, 50% end in C or G, and so there would not necessarily be any correlation between the strength of codon usage bias and silent site G+C content (and none was observed among chromosome III genes). Also, it has been reported that highly expressed transcripts are encoded by genes scattered over the chromosome (9), and we have found that genes with high or low codon usage bias (measured by the CAI) show no particular distribution along the chromosome. Thus, the G+C content variation among chromosome III genes does not appear to be related to gene expression level: we infer that it is due to the genes' location.

There is no reason to expect that chromosome III is atypical among yeast chromosomes. Genes located throughout the genome show levels of G+C content variability similar to those on chromosome III. Interestingly, when multivariate statistical analysis (correspondence analysis) is applied to codon usage in yeast genes, the second most important trend among genes is highly correlated with G+C content at silent sites. [For the 575 genes from throughout the genome (13), the positions of genes on the second axis produced by the correspondence analysis, and the silent site G+C content values for those genes, are correlated with a coefficient of 0.85 ± 0.02.] The primary trend among genes is that already discussed, i.e., variation in the extent of usage of optimal codons, accounting for 34% of the variation among genes; the secondary trend (i.e., G+C content) can explain a further 6.4% of this variation. [Recently, we have reported similar observations from correspondence analysis of codon usage in a related yeast, *Candida albicans* (28), although the number of genes available for analysis was far more limited.] Thus, the G+C content variation is a relatively minor effect by comparison with the major variation in usage of optimal codons; furthermore, without the precise map locations provided by the complete sequence of *S.cerevisiae* chromosome III, it was not apparent that this base composition variation is related to gene location.

**Table 2.** G+C content in regions of yeast chromosome III.

| Region[a] | Genes N[b] | Mean ± SD[c] | Intergenic sequences N[b] | Mean ± SD[c] |
|---|---|---|---|---|
| 0–40 | 25 | 36.9 ± 5.6 | 18 | 32.9 ± 5.3 |
| 40–80 | 26 | 44.2 ± 7.1 | 21 | 32.9 ± 6.1 |
| 80–200 | 62 | 35.7 ± 6.1 | 58 | 34.1 ± 5.0 |
| 200–260 | 42 | 47.0 ± 10.3 | 34 | 33.9 ± 5.1 |
| 260–315 | 23 | 35.0 ± 6.2 | 22 | 34.2 ± 4.1 |
| High G+C regions[d] | 68 | 46.2 ± 9.3 | 55 | 33.5 ± 5.5 |
| Low G+C regions[e] | 110 | 35.8 ± 6.0 | 98 | 33.9 ± 4.9 |

[a] Coordinates of the region are given in kilobases.
[b] Number of sequences.
[c] Mean and standard deviation of G+C content; for genes the value refers to silent third codon positions.
[d] Regions 40–80kb and 200–260kb.
[e] Regions 0–40kb, 80–200kb, and 260–315kb.

An obvious question is whether the G+C content variation in yeast is a similar phenomenon to that observed in the human genome (1). The regions of different G+C content on yeast chromosome III are about 40–120 kb in length (Fig. 2a; Table 2). This is rather shorter than the presumed size of human isochores, although their lengths have not been estimated with any accuracy. Chromosomes in yeast have an average length a

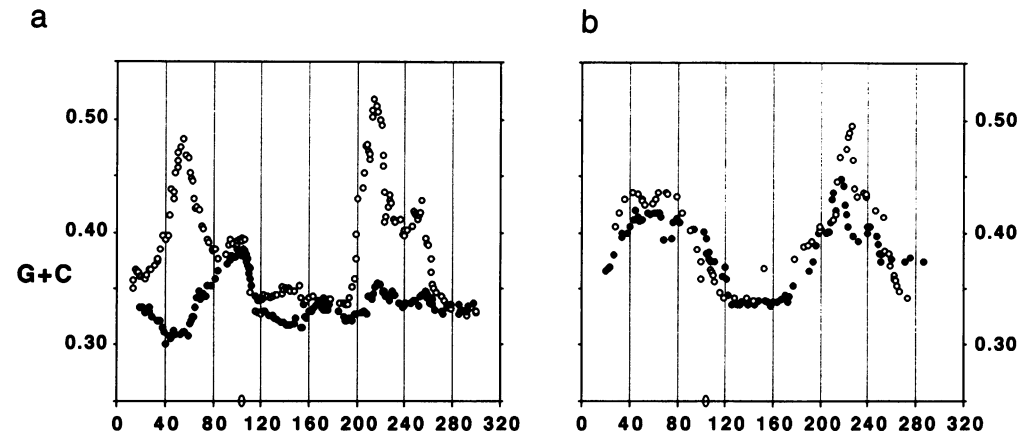


**Figure 2.** Moving average G+C content at silent sites along the yeast chromosome III sequence; each point is the weighted average for 15 adjacent sequences (weighting was by the number of sites in each sequence). **a**, G+C values for coding (open circle) and noncoding (closed circle) sequences. **b**, Coding sequences on the Crick (open circle) and Watson (closed circle) strands. The ellipse indicates the position of the centromere.

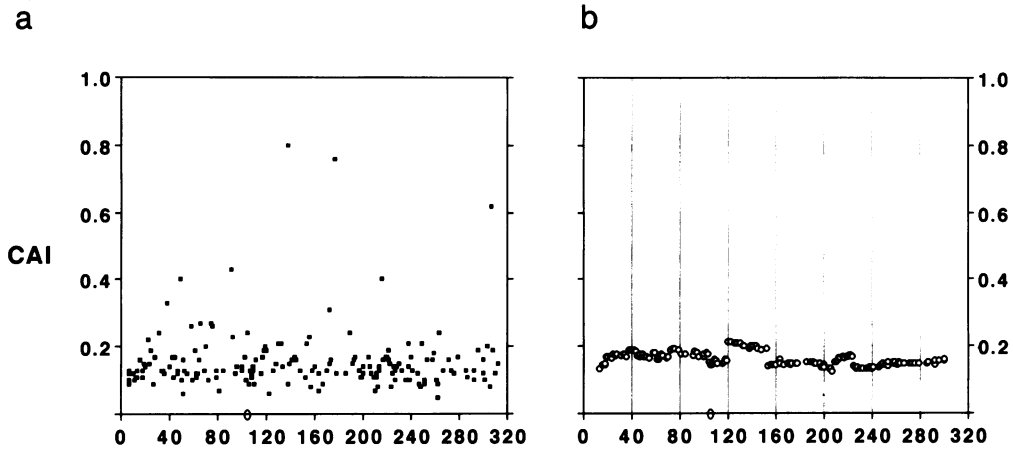


**Figure 3.** Codon usage bias and gene position on yeast chromosome III. **a**, CAI values for individual genes. **b**, Moving average of CAI values for 15 adjacent genes. The ellipse indicates the position of the centromere.

little under 1 megabase, whereas human chromosomes are on average around 100 megabases long. Human genes are also longer than those in yeast (since most human genes contain introns), and the intergenic regions in humans are also longer than those in yeast. Thus, given this difference in scale between the two genomes, if isochores do exist in yeast, they might be expected to be shorter than those on human chromosomes. However, this depends critically on what the causes of this aspect of chromosome structure are, and whether these factors are the same in the two species.

Since silent sites in human genes appear to be under little, if any, selective constraint, their base composition is expected to reflect mutational biases (5,29,30). Consequently, the G+C content variation along the human genome has been interpreted as evidence that different regions of chromosomes are subject to different mutational spectra (5,31−33). Similarly, codon usage (and hence silent site base composition) in yeast genes, other than those which are highly expressed, appears to be largely influenced by mutational biases (11,12).

Why might mutational biases vary among chromosome regions? For mammals, we have speculated that this arises because the regions are replicated at different times (and that the intracellular conditions which influence the spectrum of mutations, such as the relative concentrations of free nucleotide pools, vary during the replication cycle) and that the more G+C-rich regions are those replicated early (5). [It has recently been suggested that there is no correlation between G+C content and time of replication for human genes (34); however, the time of replication varies among tissues, and since the data cited do not refer to germline cells, it may not be relevant.] The time of replication has been measured (35) for several points along the first 200 kb of yeast chromosome III, and the G+C-rich peak in the left arm of the chromosome (Fig. 2a) coincides approximately with an early replicating region; the replication time data do not extend as far as the location of the peak in the right arm.

Alternatively, it has been suggested that base composition variation around the human genome may be correlated with the local frequency of recombination (A.Eyre-Walker, pers.comm.), since recombination events involve DNA repair, and that process is biased to G+C-richness (36). Comparison of distances on the physical and genetic maps of yeast chromosome III has indicated

that recombination frequencies are higher towards the middle of the two chromosome arms (8,9). Thus, on the large scale, the areas of higher G+C content and higher recombination frequency may coincide. On a shorter scale, there does not appear to be a clear correlation between these two factors. Four 'hot spots' of recombination have been identified on chromosome III (reviewed in Ref.37). One hot spot lies between *HIS4* and *BIK1* at about 68 kb, within the G+C-rich region on the left arm, though a little to the right of its peak. A second hot spot lies to the left of (perhaps 10kb away from) *THR4*; *THR4* is in centre of the G+C-rich peak on the right arm. A third hot spot may be close to the small peak to the left of the centromere (Fig. 2a), but the fourth (near 85 kb) does not appear to be in a G+C-rich region. Of course, if G+C-richness and high recombination frequency are correlated, it remains to be determined whether either effect causes the other.

In the human genome, G+C content in noncoding sequences (introns or flanking sequences) is highly correlated with that at silent sites in the neighbouring coding sequences (4). However, the yeast and human genomes differ in this respect: the intergenic regions of yeast chromosome III do not exhibit base composition variation similar to the silent sites in genes. It might have been expected that intergenic sequences are under relatively little constraint, and should be influenced by any regional mutational biases. Thus, the observation that the intergenic sequences within the G+C-rich peaks do not show the same elevated G+C content as the silent sites in genes is particularly interesting. In yeast these intergenic regions are much shorter than in the human genome (the average length on yeast chromosome III is 638bp). Since many regulatory elements are located within these regions, these may constitute a selective constraint on base composition.

In conclusion, it remains to be seen whether the factors determining regional genic base composition variation are the same in the human and yeast genomes. However, whether replication time, recombination frequency, or some other factors are the underlying cause(s), the issue is likely to be more easily resolved for a 'model' organism like yeast.

## ACKNOWLEDGEMENTS

This is a paper from the Irish National Centre for Bioinformatics. We are grateful to André Goffeau, David McConnell, Myra O'Regan and Ken Wolfe for discussion, and particularly to Adam Eyre-Walker for sharing his ideas on the possible link between G+C content and recombination frequencies. This work was supported by grant SC/91/603 from EOLAS (The Irish Science and Technology Agency).

## REFERENCES

1. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) *Science*, **228,** 953- 958.
2. Ikemura, T. (1985) *Mol. Biol. Evol.*, **2,** 13–34.
3. Aota, S.-i. and Ikemura, T. (1986) *Nucleic Acids Res.*, **14,** 6345–6355.
4. Ikemura, T. and Aota, S.-i. (1988) *J. Mol. Biol.*, **203,** 1–13.
5. Wolfe, K.H., Sharp, P.M. and Li, W.-H. (1989) *Nature*, **337,** 283–285.
6. Bernardi, G. (1989) *Annu. Rev. Genet.*, **23,** 637–661.
7. Matassi, G., Montero, L.M., Salinas, J. and Bernardi, G. (1989) *Nucleic Acids Res.*, **17,** 5273–5290.
8. Oliver, S.G. *et al.* (1992) *Nature*, **357,** 38–46.
9. Yoshikawa, A. and Isono, K. (1990) *Yeast*, **6,** 383–401.
10. Bennetzen, J.L. and Hall, B.D. (1982) *J. Biol. Chem.*, **257,** 3026–3031.
11. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) *Nucleic Acids Res.*, **14,** 5125–5143.
12. Bulmer, M. (1990) *Nucleic Acids Res.*, **18,** 2869–2873.
13. Sharp, P.M. and Cowe, E. (1991) *Yeast*, **7,** 657–678.
14. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and di Paola, G. (1985) *CABIOS*, **1,** 167–172.
15. Voytas, D.F. and Boeke, J.D. (1992) *Nature*, **358,** 717.
16. Shields, D.C. and Sharp, P.M. (1989) *J. Mol. Biol.*, **207,** 843–846.
17. Van Dyck, E., Foury, F., Stillman, B. and Brill, S.J. (1992) *EMBO J.*, **11,** 3421–3430.
18. Bork, P., Ouzonis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E. (1992) *Nature*, **358,** 287.
19. Sharp, P.M. and Li, W.-H. (1987) *Nucleic Acids Res.*, **15,** 1281–1295.
20. Lloyd, A.T. and Sharp, P.M. (1992) *J. Hered.*, **83,** 239–240.
21. Mandel, M. (1970) In Sober, H.A. (ed.), Handbook of Biochemistry. CRC, Cleveland, pp. H75-H79.
22. D'Onofrio, G. and Bernardi, G. (1992) *Gene* **110,** 81–88.
23. Sueoka, N. (1992) *J. Mol. Evol.*, **34,** 95–114.
24. Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988) *Mol. Biol. Evol.*, **5,** 704–716.
25. Lloyd, A.T. and Sharp, P.M. (1991) *Mol. Gen. Genet.*, **230,** 288–294.
26. Sharp, P.M. and Devine, K.M. (1989) *Nucleic Acids Res.*, **17,** 5029–5039.
27. Sharp, P.M. (1990) *Mol. Microbiol.*, **4,** 119–122.
28. Lloyd, A.T. and Sharp, P.M. (1992) *Nucleic Acids Res.*, **20,** 5289–5295.
29. Sharp, P.M. (1989) In Hill, W.G. and Mackay, T.F.C. (eds.) Evolution and Animal Breeding. C.A.B. International, Wallingford, pp. 24–32.
30. Eyre-Walker, A.C. (1991) *J. Mol. Evol.* **33,** 442–449.
31. Filipski, J. (1987) *FEBS Letts.*, **217,** 184–186.
32. Sueoka, N. (1988) *Proc. Natl. Acad. Sci. U.S.A.*, **85,** 2653–2657.
33. Filipski, J., Salinas, J. and Rodier, F. (1989) *J. Mol. Biol.* **206,** 563–566.
34. Eyre-Walker, A. (1992) *Nucleic Acids Res.*, **20,** 1497–1501.
35. Reynolds, A.E., McCarroll, R.M., Newlon, C.S. and Fangman, W.L. (1989) *Mol. Cell. Biol.*, **9,** 4488–4494. 36. Brown, T.C. and Jiricny, J. (1988) *Cell*, **54,** 705–711.
37. Zenvirth, D., Arbel, T., Sherman, A., Goldway, M., Klein, S. and Simchen, G. (1992) *EMBO J.*, **11,** 3441–3447.