Randomness and Natural Selection in Genome Evolution

GIORGIO BERNARDI and GIACOMO BERNARDI
Laboratoire de Génétique Moléculaire, Institut Jacques-Monod, 2, Place Jussieu, 75005 Paris, France.

1. Introduction

The evolution of living organisms is caused primarily by mutations that may subsequently be eliminated or become fixed in the genome. While it is generally agreed that elimination affects deleterious mutations and occurs by negative selection, fixation has been visualized as due either (i) to positive Darwinian selection acting on advantageous mutations or (ii) to random genetic drift acting on selectively neutral (i.e. selectively equivalent) mutations. Since both advantageous and neutral mutations definitely can be fixed in evolution, the issue is of a quantitative and not of a qualitative nature and concerns the predominance of deterministic or stochastic events in genome evolution. That the issue is a difficult one is proved by the fact that this debate has gone on, in the form just stated, for almost twenty years.

We will summarize here a new approach to the problem (see [1—3] for recent reports). The question we have asked basically concerns the degree of freedom in the fixation of mutations. According to the neutral theory [4—7], the fixation of mutations is free of constraints except for 'functional constraints', like the requirement of given aminoacids in certain positions of polypeptide chains. As a consequence, fixation of mutations in third codon positions and in non-coding sequences was considered to be essentially random. Our starting observation was that this certainly was not the case in the genome of warm-blooded vertebrates [8].

2. The Compositional Compartmentalization of Genomes

The genomes of warm-blooded vertebrates are highly compartmentalized, in that they mainly consist of a mosaic of very long ($\gg 300$ kilobases) DNA segments, the *isochores*, which (i) belong to a small number of classes characterized by different GC* levels and by fairly homogeneous base compositions (at least in the 3—300 kbase range); and (ii) seem to correspond to the DNA segments present in Giemsa and Reverse chromosomal bands. Since the families of DNA molecules derived from the isochore classes (or genome compartments) can be separated, it is possible to study the genome distribution of any sequence that can be detected with an appropriate probe. This approach revealed (i) that the distribution of

* GC = mole % of deoxyguanosine + deoxycytidine.

Jean Maruani (ed.), Molecules in Physics, Chemistry, and Biology, Vol. IV, 3—12.

genes is highly biased towards the GC-rich isochores (which are either absent or poorly represented in cold-blooded vertebrates) and tends to be conserved within birds and within mammals; and (ii) that the GC levels of both coding and non-coding sequences (e.g. introns, families of interspersed repeats), as well as of codon third positions, show a linear dependence upon the GC levels of the genome compartments harboring the sequences. While the strongly biased gene distribution largely came as a surprise, the GC relationships indicated the existence of compositional constraints. (It should be noted that a correlation between GC levels in third positions and GC of flanking sequences was independently reported by Ikemura [9] for several vertebrate genes; moreover, a very recent paper [10] has confirmed and extended some of the points made above).

3. Compositional Constraints Affect both Coding and Non-Coding Sequences

The compositional constraints first detected in the genomes of warm-blooded vertebrates have been investigated in over 60 genomes of prokaryotes, viruses and vertebrates. The results of Figure 1 indicate that the GC levels of each codon position of the genes from a given genome (or genome compartment) are linearly related to those of the corresponding coding sequences; slopes and intercepts for each position are very close for all genomes explored (except for a higher slope in the case of second positions of viruses). In the case of vertebrates, plots against the GC levels of genome (or genome compartments) are very close to those against GC levels of coding sequences, in spite of the fact that non-coding sequences represent over 90% of the genome. These relationships indicate the existence of compositional constraints acting on coding sequences (where they also affect the levels of individual bases; not shown), as well as non-coding sequences.

These findings raise two problems, that of their consequences at the RNA and protein levels, and that of their origin.

4. GC Increases in Coding Sequences Affect mRNA and Protein Stability

All GC changes in second codon positions entail changes in the amino acid composition of proteins; so also do most first position changes, and two third position changes. An analysis of the amino acid replacements which accompany the GC increases in codon positions has revealed that they comprise those [11] that lead to thermodynamically more stable proteins (see Table I). Indeed, the amino acids (alanine and arginine) that are most frequently acquired in thermophiles and that most contribute to an increased stability *increase*, whereas those (serine and lysine) that are correspondingly lost and that diminish stability *decrease* with increasing exon GC (not shown). In the case of compartmentalized genomes, these changes may take place within the same genome.

In conclusion, the compositional changes that make DNA thermodynamically

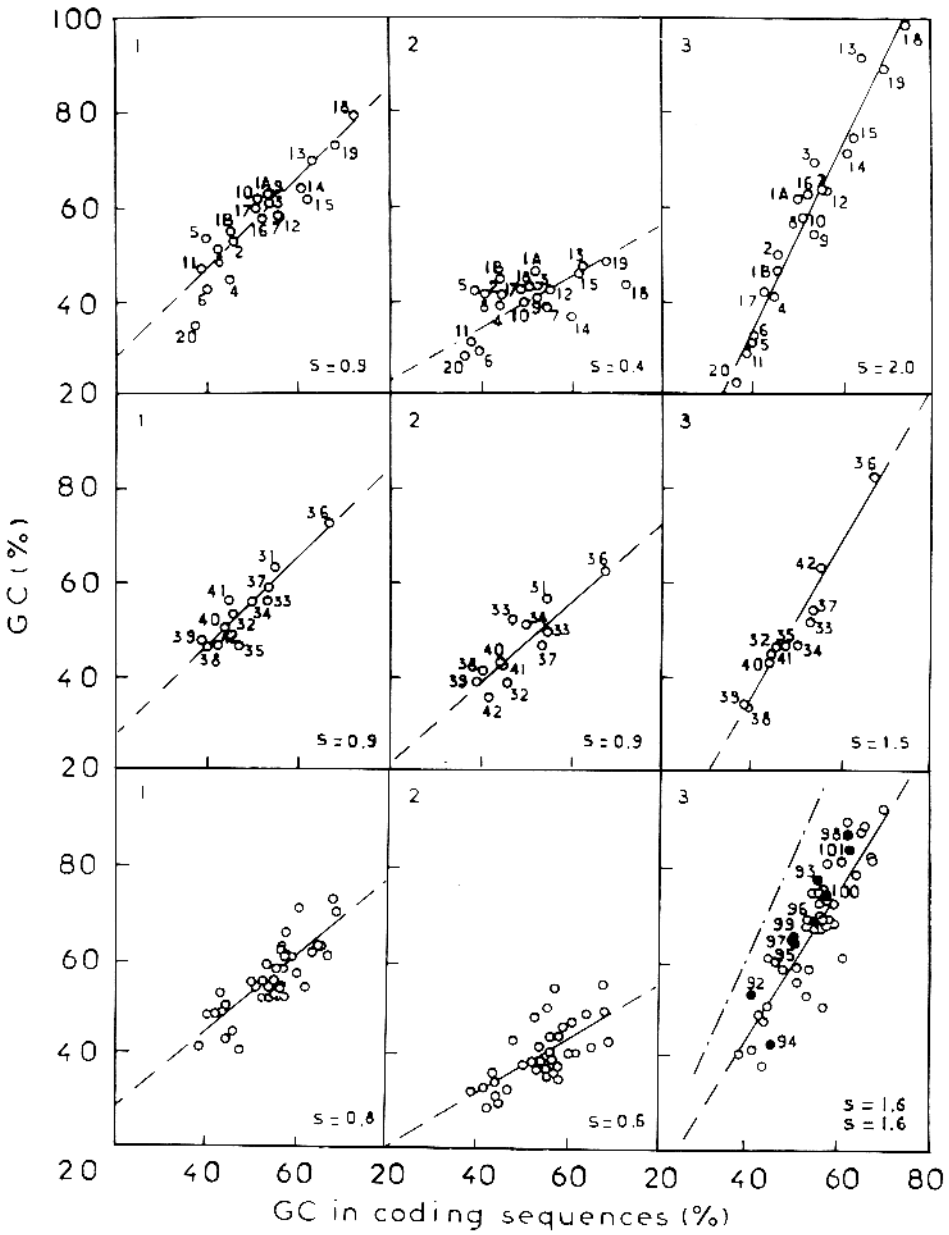

Fig. 1. (Top and middle frames) GC levels of the three codon positions (1, 2, 3) of prokaryotic (top frames) and viral (middle frames) genes are plotted against the GC levels of the corresponding genomes (the list of genomes is given in [1]). The scatter of the points belonging to different genes from the same genome was small and average values, weighted for gene size, were therefore used. Numbers refer to the genomes listed in Table I. Lines were drawn using the least-squares method; the slopes are indicated; correlation coefficients were 0.91, 0.58, and 0.97 for prokaryotic genomes, and 0.91, 0.88 and 0.95 for viral genomes, respectively. (Bottom frames) (○) GC levels of codon positions of individual genes from vertebrates are plotted against the GC levels of the corresponding exons; (●) average values for third position GC of genes belonging to the same compartment of a given genome are plotted against the GC levels of coding sequences of the genome compartment. Correlation coefficients were 0.81, 0.67 and 0.88, respectively. (Dash-and-point line) the third position plot against GC of genome compartments.

Table I. Amino acid exchanges observed in thermophiles and the accompanying GC changes in their codons.[a]

Exchanges			Codon GC
Mesophiles		Thermophiles	
*Gly	→	Ala	0
*Ser	→	Ala	+
Ser	→	Thr	0
Lys	→	Arg	+
Asp	→	Glu	±
Ser	→	Gly	+
*Lys	→	Ala	+

[a] The observed exchanges are given in order of decreasing frequency, asterisked changes corresponding to the largest expected increase in stability [11]. Only the Lys → Ala exchange requires more than one base change per codon. The right column indicates increases (+), decreases (−), and no changes (0) in codon GC levels.

more stable, also increase the thermodynamical stability of the encoded proteins. The same changes obviously also lead to higher GC levels in mRNAs, a factor known [12] to increase their base pairing and stability.

5. Compositional Constraints are Due to Environmental Pressures

In the genome of warm-blooded vertebrates, different compositional constraints are associated with different genome compartments. One way to understand the origin of compositional constraints is, therefore, to investigate the causes for the formation of the GC-rich compartments of warm-blooded vertebrates; (as already mentioned, these compartments do not exist or are poorly represented in cold-blooded vertebrates). We know that the formation of GC-rich isochores is due to regional increases in GC and we attributed such increases to the requirements of chromosome structure and function at the temperatures prevailing in warm-blooded vertebrates [8]. This suggestion has now been tested (Giacomo Bernardi and Giorgio Bernardi, paper in preparation) by comparing the genomes of fishes trapped by geological events in hot springs, streams or lakes, with those of closely related species living in colder environments. The species analyzed so far comprise several *Cyprinodontidae* from the Death Valley basin, California, using the related *Fundulus heteroclitus* as a reference species, and *Tilapia grahami* from Lake Magadi, Kenya, using three closely related *Tilapia* species as 'controls'. In both series of comparisons, the genomes of fishes living at 37°—40° showed GC-rich components that were absent in the reference species living at 20°—25° (Figure 2).

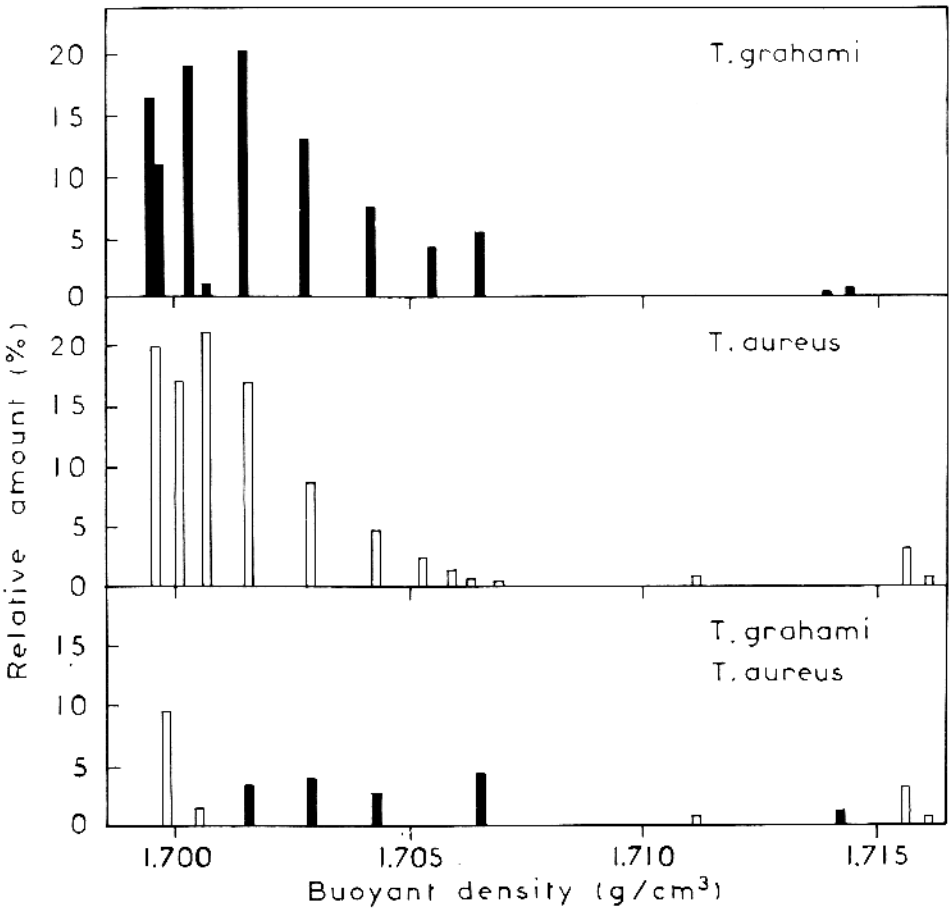

Fig. 2. Histograms showing the relative amounts and buoyant densities in CsCl of DNA fractions obtained by preparative Cs_2SO_4/bis(acetatomercury)dioxane density gradient centrifugation from *Tilapia grahami* and *Tilapia aureus*. The bottom panel shows the difference histogram.

These findings provide precise examples in which an environmental factor, temperature, appears to be responsible for novel compositional constraints in the genome. (The extremely high rates and the underlying molecular mechanisms of such changes and their relevance for the problems of the constancy of mutation rate and gradualism will be discussed elsewhere).

The results presented so far have a direct bearing on two important issues in molecular evolution, namely codon usage and the fixation of mutations.

6. Codon Usage is Largely Determined by Compositional Constraints

Since non-randomness of codon usage was first discovered, several, not mutually exclusive, explanations were provided for this phenomenon. These comprise: (i)

the optimization of codon-anticodon interaction energy [13] and the consequent optimization of translation efficiency in highly expressed genes [14, 15]; (ii) the fulfillment of requirements for mRNA secondary structure and stability [12]; (iii) an adaptation of codons to the actual populations of isoaccepting tRNAs [9, 16].

These explanations essentially rested on intraspecific differences in the usage of all synonymous codons. In contrast, our results concern interspecific and intercompartmental differences in the usage of synonymous codons characterized by different GC levels in third positions; this sub-set of codons corresponds to 2/3 of all synonymous codons. Our results lead to the following conclusions.

(i) Interspecific and intercompartmental differences in codon usage largely depend upon the compositional constraints affecting the genome, or the genome compartments. This provides, to a large extent, a rationale for the 'genome strategy of codon usage' [17] which comprises several 'compartmental strategies' in compartmentalized genomes. It should be noted that a dependence of codon usage upon genome GC was already noticed in some cases of unusual GC content [18, 19].

(ii) The proposals that mRNA structure [12] or the abundance of synonymous tRNAs [9] are the causes and not the effects of codon usage should be reversed. Since third codon positions are under essentially the same compositional constraints as non-coding sequences (Figure 1), the primary phenomenon is at the DNA level and the effects are at the mRNA or tRNA levels. The latter point was already well demonstrated by the changes in tRNA distributions that occur in the silk gland of *B. mori* in connection with the expression of the fibroin and sericin genes [20]. As already mentioned, our results do not bear on the intraspecific differences in codon usage which have been shown in unicellular organisms, like *E. coli* and *S. cerevisiae* [9].

7. Mutations are Mainly Fixed through Natural Selection

The isochore pattern of mammals is very similar, in that roughly the relative amounts of different isochore classes are quite close in different species. Unless one attributes this situation to convergent evolution, this should be seen as the result of the common descent of mammals from a warm-blooded tetrapod endowed with a genome similar to those of present day mammals. Since this common ancestor dates back to over 100 million years ago, one has to reach the conclusion that the base changes which occurred since, preferentially kept low and high GC levels in different codon positions of genes located in GC-poor and GC-rich compartments, respectively. In other words, on the average, changes were conservative as far as base composition is concerned. By analogy with the 'genome strategy of codon usage' [17] one should, therefore, take into consideration a more general 'compositional strategy of coding sequences', which also concerns non-silent changes. This strategy comprises several compartmental strategies in compartmentalized genomes. (b) Changes in non-coding sequences of eukaryotes

conform to the same general rules as changes in coding sequences. In eukaryotes, the 'compositional strategy of coding sequences' is therefore part of a 'general compositional strategy', that also affects non-coding sequences. Again this strategy may consist of several compartmental strategies. (c) The CpG level (and the level of potential methylation sites) in both coding and non-coding sequences of vertebrates also appears to be subject to the same compositional constraints as the base changes just discussed; indeed, the CpG shortage is different in different genome compartments and is correlated with their GC levels.

To sum up, intragenomic GC changes clearly indicate that most mutations are fixed, in both coding and non-coding sequences, not at random, but under the influence of compositional constraints, in compliance with a 'general compositional strategy' involving, in all likelihood, both negative and positive selection. Random fixation of neutral mutations [4—7] certainly also occurs, but only to an extent such that the 'general compositional strategy' and the relationships of Figure 1 are not blurred.

As far as intergenomic changes are concerned, two points should be made. (a) GC changes in the three codon positions do not proceed in parallel: second position changes lag behind first position changes which, in turn, lag behind third position changes. Such decreasing extents of change appear to be correlated with the corresponding increasing impacts on amino acid composition of proteins. In other words, the different slopes of Figure 1 are correlated with the different fixation rates that have been detected in different codon positions of a number of genes [4] and indicate the existence of constraints other than the compositional ones. The higher slope of second position GC of viral genomes is likely to reflect lower amino acid constraints in viral proteins. (b) A clear directionality is shown by the amino acid substitutions, the silent base changes, the changes in non-coding sequences and the CpG changes which accompanied the transition from cold-blooded to warm-blooded vertebrates, to lead [8] to the formation of GC-rich genes and GC-rich isochores in the genomes of the latter. These directional changes can only be explained by a positive Darwinian selection acting on mutations that confer selective advantages in relationship with environmental pressures. These advantages have been identified as far as the transition from cold-blooded to warm-blooded vertebrates is concerned. Indeed, silent changes led to an optimization of structure and function at the level of both DNA and RNA, non-silent changes leading, in addition, to an optimization of structure and function at the protein level.

Obviously, our conclusions reverse the proposals of the 'neutral-mutation-random-drift hypothesis' (i) 'that the great majority of evolutionary changes at the molecular level are caused not by Darwinian selection acting on advantageous mutations, but by random fixation of selectively neutral or nearly neutral mutants' and (ii) 'that only a minute fraction of DNA changes are adaptive in nature' [6]. Both proposals rest, in fact, on the classical concept that the phenotype of an organism only corresponds to its 'gene products'; as a logical consequence, 'silent'

mutations and changes in non-coding sequences were visualized as having no evolutionary impact. Moreover, random fixation of mutations was perfectly compatible with the limited sequence data analyzed at the time it was proposed.

8. Conclusions

In conclusion, compositional constraints indicate that most mutations are not fixed at random, but in relationship to a 'general compositional strategy' of the genome. This appears to be largely the result of natural selection including positive selection of mutations, that are advantageous as far as environmental pressures are concerned. Neutral mutations no doubt also exist, but their fixation occurs at a level low enough so as not to distort the 'general compositional strategy'. These conclusions lead to two general ideas: (i) that genome evolution depends more on natural slection than on random events; and (ii) that the environment can mould the genome through selection. The latter point has been illustrated here by the effects of temperature on the composition and on the compartmentalization of the vertebrate genome; other environmental factors certainly also play a role and may affect not base composition, but the frequencies of di- and oligo-nucleotides (for instance, ultraviolet light affects the level of pyrimidine doublets in bacterial genomes). Indeed, compositional constraints should in fact be visualized as a sub-set of the 'sequence constraints' acting on the genome [21] and influencing DNA structure [22].

In eukaryotes, both coding and non-coding sequences appear to be under essentially the same compositional constraints, and therefore under the same selection pressures. This finding stresses, first of all, the fundamental unity of the genome, already suggested by the genome strategy of codon usage [17], and contradicts what has been called [23] the 'bean bag' view of the genes within the genome. Second, it confirms the idea [17] that the genome is the unit upon which natural selection acts. Third, it does not support the view that non-coding sequences can be equated with functionless 'junk DNA' [24]. In contrast, it rather suggests that non-coding sequences do play a physiological role, which may have to do with the modulation of basic genome functions. This suggestion, although not a new one [25—28] does not rest anymore on 'adaptive stories', which can be rightly criticized [29—31], but on the newly demonstrated compositional constraints. Interestingly, the same conclusions have been reached on the basis of different evidence for the non-coding sequences of the mitochondrial genome of yeast [32—37].

Finally, compositional constraints identify a new component in the organismal phenotype, which may be called the 'genome phenotype'. Indeed, compositional constraints largely affect the structure and stability of the genome (at its different DNA, chromatin, and chromosome levels), of the transcripts and even of proteins (as exemplified by the stability changes accompanying GC increases in the genome), as well as codon usage. At the same time, they also conceivably touch on

a number of basic functions, such as replication, recombination, transcription and translation, that are sensitive to the compositional/structural features just mentioned. This component adds on the other classical component of the phenotype, which is formed by the 'gene products', and is defined by non-silent mutations in the genes and by mutations in regulatory signals.

Acknowledgements

The senior author thanks the Fogarty International Center for Advanced Study in the Health Sciences, National Institutes of Health, Bethesda, 20205, U.S.A. for a scholarship during which this work was initiated. Sequence data treatments were performed using computer facilities at CITI2 in Paris on a PDP8 computer, with the help of the French Ministère de la Recherche et la Technologie (Programme Mobilisateur 'Essor des Biotechnologies').

References

1. G. Bernardi and G. Bernardi: *J. Mol. Evol.* **24**, 1—11 (1986).
2. J. Salinas, M. Zerial, J. Filipski, and G. Bernardi: *Eur. J. Biochem.* **160**, 469—478 (1986).
3. M. Zerial, J. Salinas, J. Filipski, and G. Bernardi: *Eur. J. Biochem.* **160**, 479—485 (1986).
4. M. Kimura: *Nature* **217**, 624—626 (1968).
5. M. Kimura: *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, England (1983).
6. M. Kimura: *Phil. Trans. R. Soc. Lond. (Biol.)* **312**, 343—354 (1986).
7. J. L. King and T. H. Jukes: *Science* **164**, 788—798 (1969).
8. G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier: *Science* **228**, 953—958 (1985).
9. T. Ikemura: *Mol. Biol. Evol.* **2**, 13—34 (1985).
10. S. I. Aota and T. Ikemura: *Nuclei Acids Res.* **14**, 6345—6355 (1986).
11. P. Argos, M. G. Rossmann, U. M. Grau, A. Zuber, G. Franck, and J. D. Tratschin: *Biochemistry* **18**, 5698—5703 (1979).
12. M. Hasegawa, T. Yasumaga, and T. Miyata: *Nucleic Acids Res.* **7**, 2073—2079 (1979).
13. H. Grosjean, D. Sankoff, W. Min Jou, W. Fiers, and R. J. Cedergren: *J. Mol. Evol.* **12**, 113—119 (1978).
14. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier: *Nucleic Acids Res.* **9**, r43—r74 (1981).
15. J. L. Bennetzen and B. D. Hall: *J. Biol. Chem.* **257**, 3026—3031 (1982).
16. L. E. Post, G. D. Strycharz, M. Nomura, H. Lewis, and P. P. Dennis: *Proc. Natl. Acad. Sci. USA* **76**, 1697 (1979).
17. R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Paré: *Nucleic Acids Res.* **8**, r49—r62 (1980).
18. B. P. Nichols, M. Blumenberg, and C. Yanofsky: *Nucleic Acids Res.* **9**, 1743—1755 (1981).
19. Y. Kagawa, H. Nojima, N. Nukima, M. Ishizuka, T. Nakajima, T. Yasuhara, T. Tanaka, and T. Oshima: *J. Biol. Chem.* **259**, 2956—2960 (1984).
20. A. Chevallier and D. R. Garel: *Biochimie* **61**, 245—262 (1979).
21. G. Bernardi, S. D. Ehrlich, and J. P. Thiéry: *Nature* **246**, 36—40 (1973).
22. A. Wada and A. Suyama: *Mol. Biol.* **47**, 113—157 (1986).
23. E. Mayr: *Evolution and the Diversity of Life*. Harvard University Press, Cambridge, Massachusetts (1976).
24. S. Ohno: *J. Hum. Evol.* **1**, 651—662 (1972).

25. R. J. Britten and E. H. Davidson: *Science* **165**, 349—357 (1969).
26. E. Zuckerkandl: *J. Mol. Evol.* **7**, 269—311 (1976).
27. E. Zuckerkandl: *J. Mol. Evol.* **24**, 12—27 (1986).
28. E. H. Davidson and R. J. Britten: *Science* **204**, 1052—1059 (1979).
29. S. J. Gould and R. C. Lewontin: *Proc. R. Soc. Lond. (Biol.)* **205**, 581—598 (1979).
30. W. F. Doolittle and C. Sapienza: *Nature* **284**, 601—603 (1980).
31. L. E. Orgel and F. H. C. Crick: *Nature* **284**, 604—607 (1980).
32. G. Bernardi: in: G. Attardi, P. Borst, and P. P. Slonimski (eds.): *Mitochondrial Genes*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 269—278 (1982).
33. G. Bernardi: *Folia Biol.* **29**, 82—92 (1983).
34. M. de Zamaroczy and G. Bernardi: *Gene* **37**, 1—17 (1985).
35. M. de Zamaroczy and G. Bernardi: *Gene* **41**, 1—22 (1986a).
36. M. de Zamaroczy and G. Bernardi: *Gene* **47**, 155—177 (1986b).
37. M. de Zamaroczy and G. Bernardi: *Gene* **54**, 1—22 (1987).