Journal of
Molecular Evolution
© Springer-Verlag New York Inc. 1986

Compositional Constraints and Genome Evolution*

Giorgio Bernardi and Giacomo Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Summary. Nucleotide sequences of all genomes are subject to compositional constraints that (1) affect, to about the same extent, both coding and noncoding sequences; (2) influence not only the structure and function of the genome, but also those of transcripts and proteins; (3) are the result of environmental pressures; and (4) largely control the fixation of mutations. These findings indicate (1) that noncoding sequences are associated with biological functions; (2) that the organismal phenotype comprises two components, the classical phenotype, corresponding to the "gene products," and a "genome phenotype," which is defined by the compositional constraints; and (3) that natural selection plays a more important role in genome evolution than do random events.

Key words: Genome composition — Isochores — Neutral theory — Natural selection

Introduction

The evolution of living organisms is caused primarily by mutations that may be eliminated or become fixed in the genome. It is generally agreed that elimination affects deleterious mutations and occurs by negative selection. In contrast, fixation has been visualized as due either to positive Darwinian selection acting on advantageous mutations or to random genetic drift acting on selectively neutral (i.e., selectively equivalent) mutations. It should be immediately pointed out, however, that this central issue in genome evolution is not an all-or-none issue, as indicated by the fact that both advantageous and neutral mutations definitely become fixed in evolution. Indeed, the issue is of a quantitative, not a qualitative, nature, and concerns the predominance of deterministic or stochastic events in genome evolution. This problem has been studied here using a new approach based on our work on the organization of the vertebrate genome [see Salinas et al. (1986) and Zerial et al. (1986a, b) for recent reports], namely by investigating the compositional constraints that affect the genome.

Our initial observations [see Bernardi et al. (1985) for a review] can be summarized as follows: The nuclear genome of warm-blooded vertebrates exhibits a compositional compartmentalization, in that it consists mainly of a mosaic of very long (>300 kilobases) DNA segments, the isochores. These belong to a small number of classes characterized by different GC levels and by fairly homogeneous base compositions (at least in the 3- to 300-kb range), and seem to correspond to the DNA segments present in Giemsa and reverse chromosomal bands. The families of DNA molecules derived from the isochore classes (referred to henceforth as "genome compartments") can be separated and used to study the distribution within the genome of any sequence that can be detected with an appropriate probe. This approach has revealed (1) that the distribution of genes in the genomes of warm-blooded vertebrates is highly biased toward the GC-rich isochores (which are either absent or poorly represented in cold-

Offprint requests to: Giorgio Bernardi

* Presented at the FEBS Symposium on Genome Organization and Evolution, held in Crete, Greece, September 1–5, 1986.

Table 1. Genomes examined and numbers of genes analyzed[a]

Genome		Number of genes
	Prokaryotes	
1A	Lambda (left arm)	23
1B	Lambda (right arm)	39
02	*Agrobacterium tumefaciens*	26
03	*Anacystis nidulans*	1
04	*Bacillus licheniformis*	2
05	*Bacillus megaterium*	1
06	*Bacillus pumilus*	1
07	*Bacillus stearothermophilus*	1
08	*Bacillus subtilis*	7
09	*Erwinia amylovora*	2
10	*Escherichia coli*	43
11	*Haemophilus haemolyticus*	1
12	*Klebsiella pneumoniae*	4
13	*Pseudomonas aeruginosa*	1
14	*Pseudomonas putida*	1
15	*Rhizobium* sp.	3
16	*Salmonella typhimurium*	6
17	*Shigella dysenteriae*	3
18	*Streptomyces fradiae*	1
19	*Thermus thermophilus*	1
20	*Vibrio cholerae*	4
	Viruses	
31	Abelson murine leukemia virus	1
32	Adenovirus type 12	5
33	AKV murine leukemia virus	2
34	Avian sarcoma virus Y73	1
35	Hepatitis B virus	2
36	Herpes simplex virus type 1	2
37	Human adult T-cell leukemia virus	3
38	Human papilloma virus	4
39	Human papovavirus BK	7
40	Mouse hepatitis virus	2
41	Polyoma virus	6
42	Tobacco mosaic virus	6
51	Adenovirus 2 associated virus[b]	
52	Adenovirus type 3[b]	
53	Adenovirus type 25[b]	
54	Adenovirus type 28[b]	
55	ERP[b]	
56	Herpes simplex virus type 2[b]	
57	Pseudorabies[b]	
58	Simian adenovirus SA7[b]	
59	Shope[b]	
60	Frog virus type 3[b]	
	Lower eukaryotes and fungi	
61	*Dictyostelium discoideum*	3
62	*Neurospora crassa*	3
63	*Physarum polycephalum*	2
64	*Saccharomyces cerevisiae*	15
65	*Trypanosoma brucei*	5
	Invertebrates	
71	*Bombyx mori*	2
72	*Chironomus thummi thummi*	1
73	*Drosophila melanogaster*	20
	Vertebrates	
81	*Cyprinus carpio*	1
82	*Lophius americanus*	1
83	*Xenopus laevis*	2
84	Chicken	2
85	Mouse	7
86	Hamster	1
87	Rat	2
88	Dog	1
89	Calf	3
90	Ape	2
91	Human	20
92	Chicken (L2)[c]	4
93	Chicken (H2)[c]	5
94	Mouse (L1)[c]	1
95	Mouse (L2)[c]	5
96	Mouse (H2)[c]	3
97	Rabbit (L2)[c]	1
98	Rabbit (H2)[c]	1
99	Human (L2)[c]	5
100	Human (H1)[c]	2
101	Human (H3)[c]	6

[a] The approximately 340 genes analyzed comprise most of the protein-coding genes with defined ends whose primary structures were available in the GenBank (Bilofsky et al. 1986) as of December 1985. A list of all the genes used in this analysis is available upon request

[b] Viral genome for which only the CpG/GpC ratio was available (Russell 1974)

[c] L1, L2, H1, H2, and H3 are the isochore classes, or genome compartments, in which these genes were localized (Bernardi et al. 1985). These genes are different from those considered in groups 84, 85 and 91

blooded vertebrates) and tends to be conserved within birds and within mammals, and (2) that the GC levels of both coding and noncoding sequences (e.g., introns and families of interspersed repeats), as well as of codon third positions, show a linear dependence on 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 in the compartmentalized genome of warm-blooded vertebrates. [It should be noted that a correlation between GC levels in third positions and GC levels of flanking sequences was independently reported by Ikemura (1985) for several vertebrate genes; moreover, a very recent paper by Aota and Ikemura (1986) has confirmed and extended several of our findings.]

Here we investigate the correlations between the base compositions of the three codon positions of a number of genes and those of the corresponding genomes. Data on about 340 genes from over 50 organisms spanning a very wide phylogenetic range (see Table 1) were used [see Bernardi and Bernardi (1985, 1986) for preliminary reports]. An analysis of these data led us to an assessment of compositional constraints on the genomes, and to an understanding of their causes and effects.

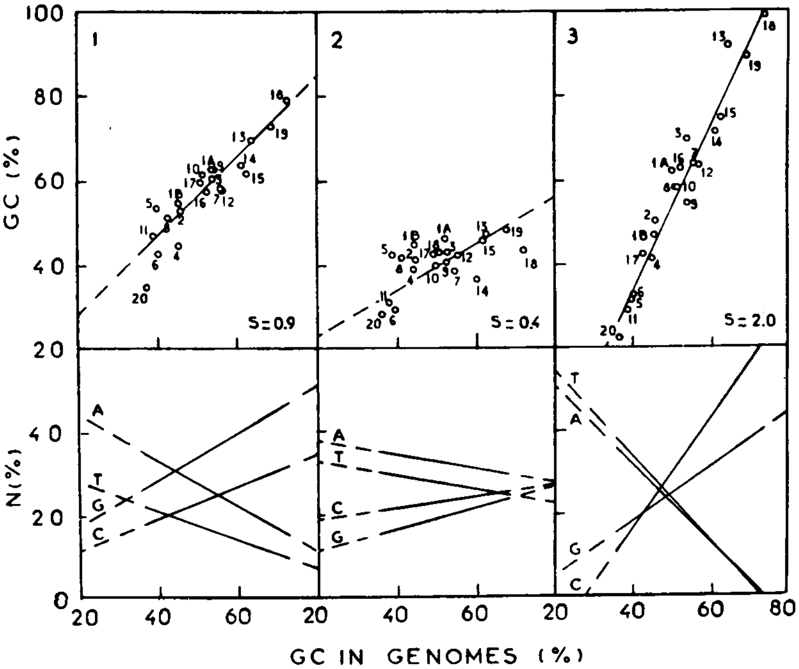

Fig. 1. GC levels (top panels) and nucleotide levels (bottom panels) of the three codon positions (1, 2, and 3) of prokaryotic genes, plotted against GC levels of the corresponding genomes. In the top panels, points belonging to different genes from the same genome were weighted by gene size and averaged. Numbers refer to the genomes listed in Table 1. Lines were drawn using the least-squares method. The slopes (s) are indicated; correlation coefficients were 0.91, 0.58, and 0.97 for positions 1, 2, 3, respectively

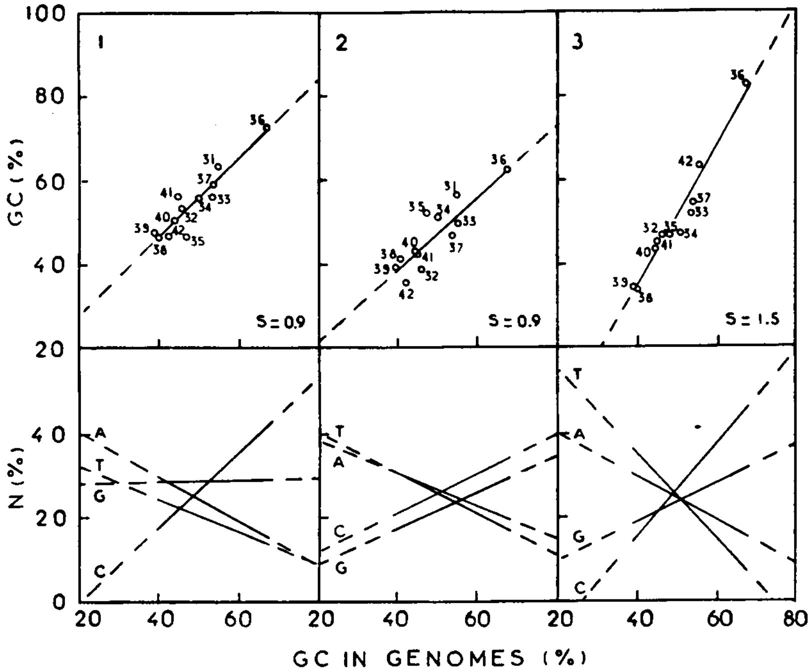

Fig. 2. GC levels (top panels) and nucleotide levels (bottom panels) of the three codon positions (1, 2, and 3) of viral genes, plotted against GC levels of the corresponding genomes. Correlation coefficients of the relationships shown in the top panels were 0.91, 0.88, and 0.95 for positions 1, 2, 3, respectively. Other conventions are as in Fig. 1

Compositional Constraints Affect Both Coding and Noncoding Sequences

If the GC levels of the three codon positions of prokaryotic and viral genes are plotted against the GC levels of the corresponding genomes (or of the genome compartments, in the case of phage lambda), linear relationships are found (Figs. 1 and 2, top panels). For each position (except the second position of viruses), slopes and intercepts are approximately the same in prokaryotes and viruses. Slopes increase from second- to first- to third-position plots; moreover, for any given genome, the second-position GC level is lower than the first-position GC level, whereas the third-position GC level is lower or higher according to whether the genomic GC level is low or high.

Prokaryotic and viral genomes are formed almost exclusively of coding sequences. Eukaryotic genomes comprise, in addition, intra- and intergenic noncoding sequences; these represent more than 90–95% of the genome in mammals. In eukaryotes, GC levels for each codon position can therefore be plotted not only against the GC levels of the corresponding exons, but also against those of the genes and genomes, or genome compartments, in the case of

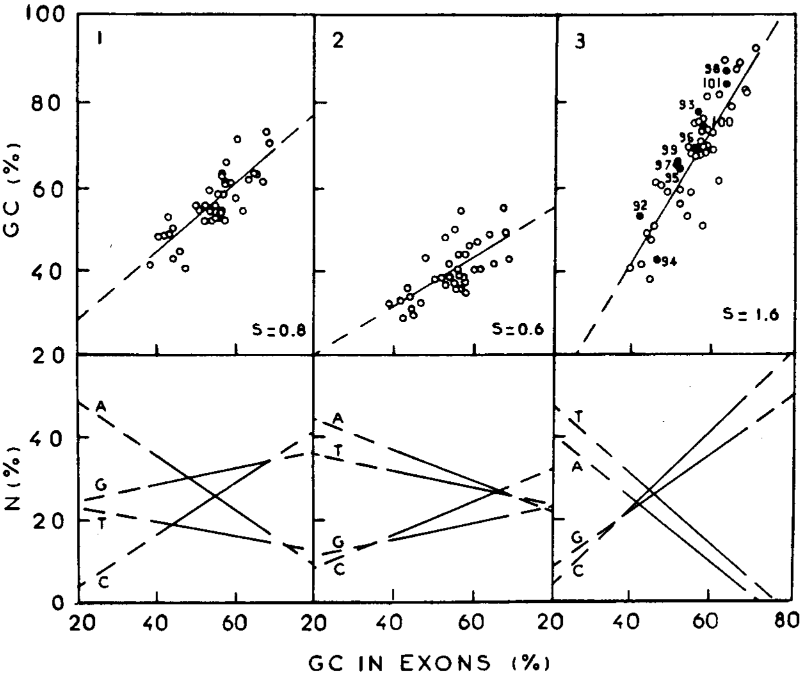

Fig. 3. GC levels (top panels, open circles) and nucleotide levels (bottom panels) of the three codon positions (1, 2, and 3) of individual genes from vertebrates, plotted against GC levels of the corresponding exons. Average values (filled circles) of third-position GC levels of genes belonging to the same compartment of a given genome (see Table 1) are plotted against GC levels of the exons of the genome compartment. Correlation coefficients were 0.81, 0.67, and 0.88, for positions 1, 2, 3, respectively. Other conventions are as in Fig. 1

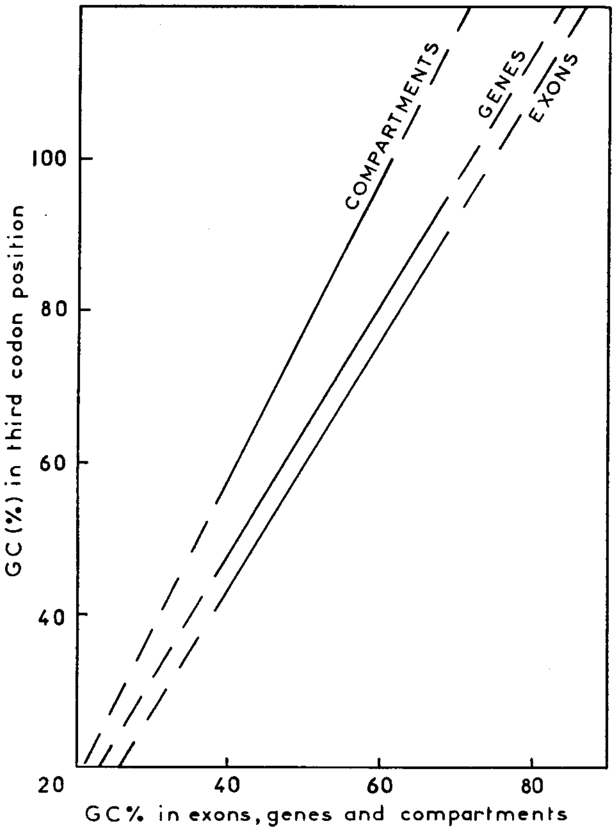

Fig. 4. GC levels of third codon positions of the vertebrate genes of Fig. 3, plotted against GC levels of exons (see Fig. 3), of genes, and of the genome compartments in which the genes are located. The slopes of the exon and the compartment plots are 1.6 and 2.0, respectively

warm-blooded vertebrates. Figure 3 (top panels) presents the "exon plots" obtained with vertebrate genomes; they are essentially identical with those described for prokaryotes and viruses. "Gene plots" are almost coincident with exon plots, whereas "genome plots" show a slightly higher slope as well as a shift to the left, indicating that the GC levels in intergenic sequences are lower than those of exons and that this difference increases with increasing GC level (Fig. 4). The limited results available for the genomes of lower eukaryotes and invertebrates from Table 1 are similar (data not shown).

As far as the levels of individual bases in different codon positions are concerned (Figs. 1–3, bottom panels), G and C levels increase, whereas A and T levels decrease, in all positions as the coding-sequence GC level increases (the only exception is the first-position G level of viral genomes, which is invariant). The G level is always higher than the C level in the first position, whereas the reverse is true in the second and third positions. The A level is higher than the T level in the first and second positions, whereas A and T are represented equally in the third position. This leads to a predominant pyrimidine–purine doublet pattern in the third and first codon positions. The doublet frequency is not constant, however, but increases from 23% (close to the statistical expectation of 25%) to 44% with increasing GC.

An analysis of pseudogenes that have been localized in gene clusters has revealed that their base compositions may be close to those of the active genes located in the same cluster or to those, slightly lower in GC (see above), of the neighboring intergenic sequences. The first situation is exemplified by human pseudo zeta-globin (Proudfoot and Maniatis 1982), pseudo beta-globin (Chang and Slightom 1984), and pseudo I immunoglobulin (from GEA region A; Flanagan and Rabbitts 1982); the second

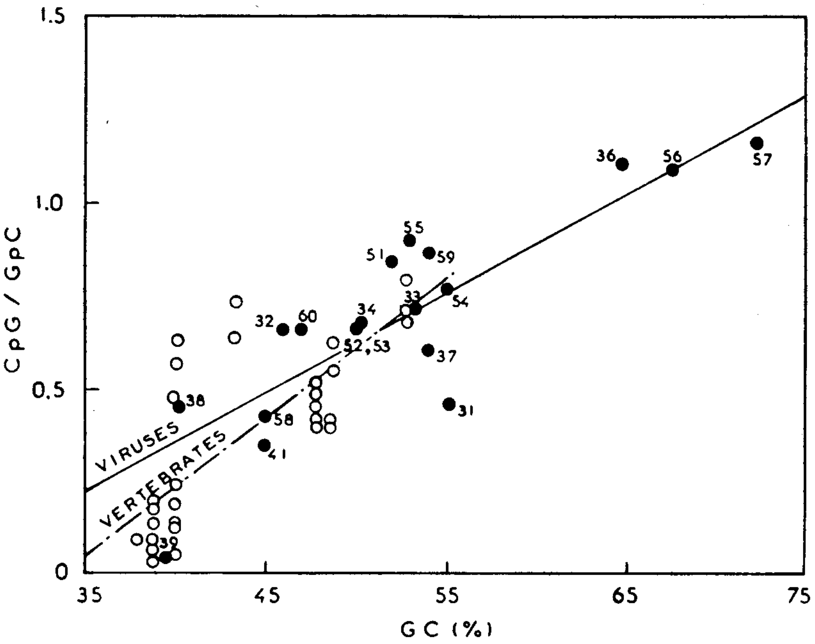

Fig. 5. CpG/GpC ratios for viral genomes, plotted against genomic GC levels (filled circles). Numbers refer to genomes listed in Table 1. Except for frog virus 3 (point 60), all viruses were from warm-blooded vertebrates. The dashed-and-dotted line and the open symbols correspond to a similar plot obtained for vertebrate genes or exons (Bernardi et al. 1985)

one, by human pseudo J1, J2, and J3 (from the mu locus; Flanagan and Rabbitts 1982) and pseudo alpha-I globin (Sawada et al. 1983). These two situations might reflect short or long times of divergence, respectively, of pseudogenes from their parental genes.

A special constraint on the genomes of vertebrates and their viruses is shown by a shortage of the doublet CpG (Swartz et al. 1962), that is, of potential sites of DNA methylation. Such a shortage decreases in degree, however, with increasing genomic GC level in both vertebrates (Bernardi et al. 1985) and their viruses (Fig. 5). Incidentally, these results contradict the suggestions that CpG shortage is associated with constraints exerted by the translation machinery (Subak-Sharpe et al. 1966), and that the bulk of vertebrate DNA derives from, and maintains the CpG shortage of, polypeptide-specifying DNA [Russell et al. 1976; see Bernardi (1985) for a more detailed discussion].

The results of Figs. 1–3 (top panels) indicate that the GC levels of codon positions are precisely related to those of the corresponding genomes, obeying linear relationships (with very high correlation coefficients) that are essentially identical for all genomes (if the differences shown by the second positions of viruses and those existing among exon, gene, and genome plots for eukaryotes are neglected). These relationships indicate the existence of compositional constraints acting on coding sequences (where they affect not only GC levels, but also the levels of individual bases; see Figs. 1–3, bottom panels), as well as on noncoding sequences (as shown by the similarity between exon plots and genome plots of eukaryotes and by the results obtained with pseudogenes). Such compositional constraints appear to predominate over other constraints such as CpG shortage and over "contextual constraints" such as the pyrimidine–purine doublet patterns of contiguous third and first codon positions.

The finding of compositional constraints on the nucleotide sequences of all genomes raises two problems: that of their consequences at the RNA and protein levels, and that of their origin. These points are dealt with in the following two sections.

GC Increases in Coding Sequences Affect mRNA and Protein Stability

All GC changes in second codon positions entail changes in the amino acid compositions of proteins; so also do most first-position changes, and two third-position changes. An analysis of the amino acid replacements that accompany the GC increases in codon positions has revealed that they comprise those (Argos et al. 1979; Zuber 1981) that lead to thermodynamically more stable proteins (see Table 2). Indeed, the amino acids that are most frequently acquired in thermophiles and that most contribute to increased stability (alanine and arginine) increase in frequency with increasing exon GC, whereas those that are correspondingly lost and that diminish stability (serine and lysine) decrease (Fig. 6). In the case of compartmentalized genomes, these changes may take place within the same genome; for instance, in the human genome the (Ala + Arg)/(Ser + Lys) molar ratio varies by a factor of four between proteins encoded by the GC-poorest and the GC-richest genes, respectively (Fig. 6).

In conclusion, the compositional changes that make DNA thermodynamically more stable also in-

Table 2. Amino acid exchanges observed between thermophiles and the accompanying GC-level changes in their codons

Exchange[a]			Codon GC-level change[b]
Mesophiles		Thermophiles	
Gly	→	Ala*	0
Ser	→	Ala*	+
Ser	→	Thr	0
Lys	→	Arg	+
Asp	→	Glu	±
Ser	→	Gly	+
Lys	→	Ala*	+

[a] Observed exchanges are given in order of decreasing frequency; asterisked changes correspond to the largest expected increase in stability (Argos et al. 1979). Only the Lys → Ala exchange requires more than one base change per codon

[b] Increases (+), decreases (−), and no changes (0) in codon GC levels are indicated

crease the thermodynamical stability of the encoded proteins. The same changes obviously also lead to higher GC levels in mRNAs, a factor known (Hasegawa et al. 1979) to increase their base pairing and stability. The limited data available so far suggest that similar changes occur in ribosomal RNAs and tRNAs.

Compositional Constraints Are Due to Environmental Pressures

In the genomes 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 of the formation of the GC-rich compartments of warm-blooded vertebrates (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 (Bernardi et al. 1985). This suggestion has now been tested (G. Bernardi and G. Bernardi, manuscript 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, with the related *Fundulus heteroclitus* used as a reference species, and *Tilapia grahami* from Lake Magadi, Kenya, with three closely related *Tilapia* species used as "controls." In both series of comparisons, the genomes of fishes living at 37°–40°C showed GC-rich components that were absent in the reference species, living at 20°–25°C (Fig. 7). These findings provide precise examples of an environmental factor, temperature, appearing to be responsible for novel compositional constraints on the genome. (The extremely high rates and underlying molecular mechanisms of such changes, and their relevance to the problems of the constancy of mutation rate and gradualism, will be discussed elsewhere.) Moreover, they lead, in conjunction with the points made in the preceding section, to the following initial hypothesis about the formation of GC-rich isochores.

The first step in the formation of GC-rich isochores might have occurred, just as described, in cold-blooded vertebrates exposed to higher temperatures and might have consisted in the preferential fixation of A/T → G/C changes in the genes of functionally important and thermolabile proteins. This process, which was followed by a GC enrichment in the neighboring noncoding sequences, led to a limited formation of GC-rich isochores in cold-blooded vertebrates exposed to higher temperatures (see above) and to a more extended formation of GC-rich isochores in warm-blooded vertebrates. Interestingly, this explanation would account, at least in part, for the higher concentration of genes in GC-rich isochores (Bernardi et al. 1985), and also for the fact that a particular set of genes, the housekeeping genes, appear to be located only in early-replicating reverse bands (Goldman et al. 1984), that is, in GC-rich isochores [see Bernardi et al. (1985) for several examples]. The association of mammalian and avian housekeeping genes with CpG-rich islands (Bird 1986) fits with this hypothesis, since the islands are likely to be located in GC-rich isochores where CpG approaches normal abundance (see Fig. 5). While a number of tissue-specific genes appear to be located in late-replicating Giemsa bands (Goldman et al. 1984; Aota and Ikemura 1986), that is, in GC-poor isochores (Bernardi et al. 1985), others, such as the avian globin genes, were subsequently translocated into GC-rich isochores, where they underwent a GC increase, insuring better protection against DNA breathing and mutability (L. Orgel, personal communication, quoted by Bernardi et al. 1985).

The results presented so far have direct bearing on two important issues in molecular evolution, dealt with in the following two sections: codon usage and the fixation of mutations.

Codon Usage Is Determined Largely by Compositional Constraints

Since nonrandomness of codon usage was discovered, several non-mutually exclusive explanations

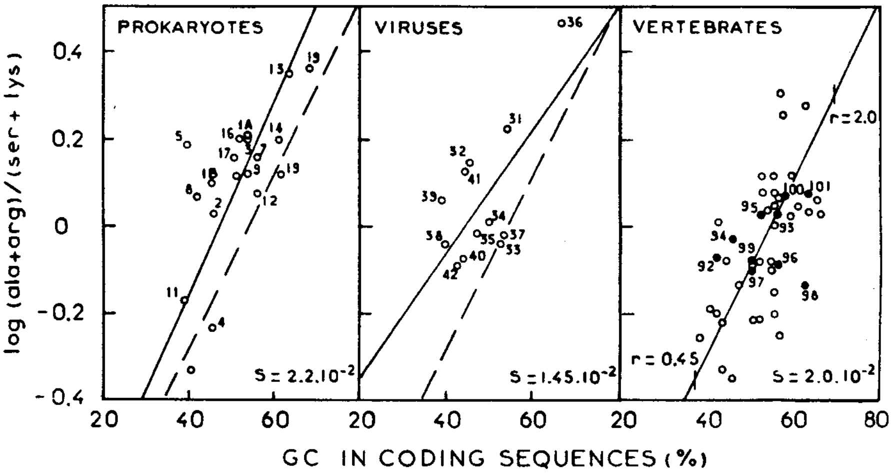

Fig. 6. Plot of (alanine plus arginine)/(serine plus lysine) molar ratio against GC levels of coding sequences from prokaryotes, viruses, and vertebrates (numbers as in Table 1). In the vertebrate plot, open symbols refer to individual genes (from human, chimpanzee, mouse, hamster, rat, calf, and chicken), and closed symbols, to the average values for genes belonging to the same compartment of a given genome (see Table 1); the two r values correspond to the lowest and highest ratios found for human genes. Slopes (s) are indicated; the vertebrate line is shown dashed in the other two diagrams for comparison

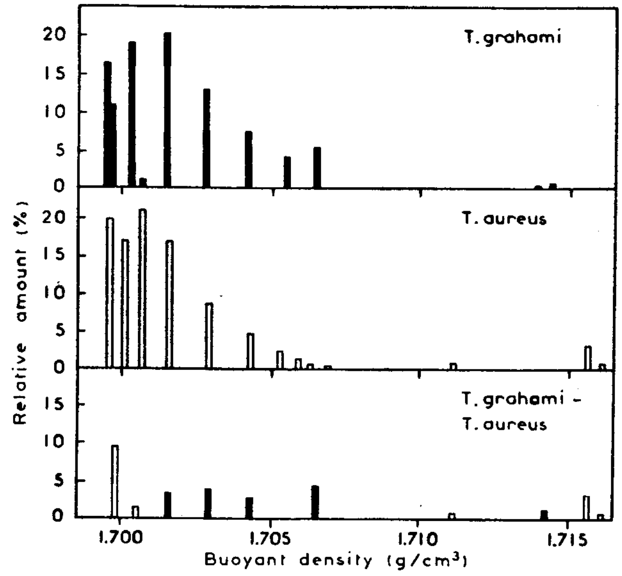

Fig. 7. Histograms showing the relative amounts and buoyant densities in CsCl of DNA fractions obtained by preparative Cs_2SO_4/bis(acetatomercury) dioxane density gradient centrifugation from *Tilapia grahami* and *Tilapia aureus*. The bottom panel shows the difference histogram

have been provided for this phenomenon: (1) the optimization of codon–anticodon interaction energy (Grosjean et al. 1978) and the consequent optimization of translation efficiency in highly expressed genes (Grantham et al. 1981; Bennetzen and Hall 1982); (2) the fulfillment of requirements for mRNA secondary structure and stability (Hasegawa et al. 1979); and (3) adaptation of codons to the actual populations of isoaccepting tRNAs (Post et al. 1979; Ikemura 1985).

These explanations essentially rest on intraspecific differences in usage among synonymous codons. In contrast, our results concern interspecific and intercompartmental differences in usage among synonymous codons, characterized by different GC levels in third positions; the subset of codons represented by these differences corresponds to two-thirds of all synonymous codons. Our results lead to the following conclusions:

1. Interspecific and intercompartmental differences in codon usage depend largely upon the compositional constraints affecting the genome or the genome compartments. This finding provides, to a large extent, a rationale for the "genome strategy of codon usage" (Grantham et al. 1980), which comprises several "compartmental strategies" in compartmentalized genomes. It should be noted that a dependence of codon usage upon genomic GC level had already been noticed in some cases of unusual GC content (Nichols et al. 1981; Kagawa et al. 1984).

2. The proposals that mRNA structure (Hasegawa et al. 1979) and the abundances of synonymous tRNAs (Ikemura 1985) 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 noncoding sequences (Fig. 4), the primary phenomenon is at the DNA level and the effects are at the mRNA or tRNA level. The latter point has already been well demonstrated by the changes in tRNA distribution that occur in the silk gland of *Bombyx mori* in connection with the expression of the fibroin and sericin genes (Chevallier and Garel 1979). As already mentioned, our results do not bear on the intraspecific differences in codon usage that have been shown in such unicellular organisms as *Escherichia coli* and *Saccharomyces cerevisiae* (Ikemura 1985).

3. The results of Figs. 1–3 also account for the "contextual constraints" previously seen in different codon positions (Nussinov 1980, 1981; Lipman and Wilbur 1983; Blaisdell 1985), for the three-base-pair periodicity detected in coding sequences, and for the frequency of pyrimidine–purine doublets in third and first codon positions (Trifonov and Sussman 1980; Shepherd 1981). The dependence of the last frequency on gene GC level casts doubts, however, on the suggestion that the latter phenomenon reflects an archaic code (Shepherd 1981).

Mutations Are Fixed Mainly Through Natural Selection

Intragenomic Changes

Intragenomic changes in all codon positions appear to have been fixed, on the average, under the influence of compositional constraints, in conformity with the base composition of the genome or the genome compartment in which genes are located. For instance, the base changes that occurred over evolutionary time in the genomes of warm-blooded vertebrates since their radiation (about 75 million years in the case of mammals; Britten 1986) preferentially kept low and high GC levels in the 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" (Grantham et al. 1980), one should, therefore, take into consideration a more general "compositional strategy of coding sequences" that concerns also nonsilent changes. This strategy comprises several compartmental strategies in compartmentalized genomes.

Intragenomic changes in noncoding 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 noncoding sequences. Again, this strategy may consist of several compartmental strategies.

The CpG level (and the number of potential methylation sites) in both coding and noncoding sequences of vertebrates appears to be subject to the same compositional constraints as the base changes just discussed; indeed, the degree of CpG shortage differs in different genome compartments in a manner correlated with their GC levels (Fig. 5).

To sum up, intragenomic GC-level changes clearly indicate that most mutations, in both coding and noncoding sequences, are fixed 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 (Kimura 1968, 1983, 1986; King and Jukes 1969) certainly also occurs, but only to an extent such that the general compositional strategy and the relationships of Figs. 1–3 are not blurred.

Intergenomic Changes

Intergenomic GC-level 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 Figs. 1–3 (top panels) are correlated with the different fixation rates that have been detected for different codon positions in a number of genes (Kimura 1983; Li et al. 1985) 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.

A clear directionality is shown by the amino acid substitutions, the silent base changes, the changes in noncoding sequences, and the CpG changes that accompanied the transition from cold-blooded to warm-blooded vertebrates, leading (Bernardi et al. 1985) to the formation of GC-rich genes and GC-rich isochores in the genomes of the latter. These directional changes can be explained only by positive Darwinian selection acting on mutations that confer selective advantages with respect to environmental pressures. These advantages have been identified as far as the transition from cold-blooded to warm-blooded vertebrates is concerned. Silent changes led to an optimization of structure and function at the levels of both DNA and RNA, and nonsilent changes led, 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" (1) "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 (2) "that only a minute fraction of DNA changes are adaptive in nature" (Kimura 1986). Both proposals rest, in fact, on the classical concept that the phenotype of an organism corresponds only to its "gene products"; as a logical consequence, "silent" mutations and changes in noncoding 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 the hypothesis was proposed.

Conclusions

The main finding reported here concerns the demonstration of compositional constraints that affect the genomes of living organisms. These constraints are more general and stronger than other genome constraints, namely (1) fixation-rate constraints (also called "functional constraints," a term that should be abandoned, since all constraints have functional consequences), which limit the fixation of mutations in (in increasing order of effect) third, first, and second codon positions; (2) constraints associated with the CpG (and potential methylation site) levels of vertebrate genomes; and (3) constraints associated with the codon usage of highly expressed genes of unicellular organisms.

Indeed, compositional constraints (1) affect both coding and noncoding sequences from all organisms explored, whereas other constraints are more limited with respect to the range of sequences or organisms affected; (2) are evident in all codon positions independently of differences in fixation rates (Figs. 1–3); (3) lead to the disappearance of the CpG shortage in GC-rich viral genomes and vertebrate genome compartments (Fig. 5); and (4) are not hidden by the preferential use of a subset of codons of highly expressed genes in unicellular organisms.

The existence of compositional constraints indicates that most mutations are fixed not at random, but in relationship to a general compositional strategy of the genome. This strategy appears to be largely the result of natural selection, including positive selection of mutations that are advantageous with respect to environmental pressures. Neutral mutations no doubt also exist, but their fixation occurs at a level low enough that it does not distort the general compositional strategy. These conclusions lead to two general ideas: (1) Genome evolution depends more on natural selection than on random events, and (2) the environment can mold the genome through selection. The latter point has been illustrated here by the effects of temperature on the composition and 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 oligonucleotides (for instance, ultraviolet light affects the level of pyrimidine doublets in bacterial genomes). Indeed, compositional constraints should be visualized as a subset of the "sequence constraints" acting on the genome (Bernardi et al. 1973) and influencing DNA structure (Wada and Suyama 1986).

In eukaryotes, coding and noncoding sequences appear to be under essentially the same compositional constraints, and therefore under the same selection pressures. This finding, first of all, stresses the fundamental unity of the genome, already suggested by the genome strategy of codon usage (Grantham et al. 1980), and contradicts what has been called (Mayr 1976) the "beanbag" view of the genes within the genome. Second, it confirms the idea (Grantham et al. 1980) that the genome is the unit upon which natural selection acts. Third, it does not support the view that noncoding sequences can be equated with functionless "junk DNA" (Ohno 1972). Rather, it suggests that noncoding sequences do play a physiological role, one that may concern the modulation of basic genome functions (see below). This suggestion, although not a new one (Britten and Davidson 1969; Zuckerkandl 1976, 1986; Davidson and Britten 1979) no longer rests on "adaptive stories," which can rightly be criticized (Gould and Lewontin 1979; Doolittle and Sapienza 1980; Orgel and Crick 1980), but rather on the newly demonstrated compositional constraints. Interestingly, identical conclusions have been reached, on the basis of different evidence, for the noncoding sequences of the mitochondrial genome of yeast (Bernardi 1982, 1983; de Zamaroczy and Bernardi 1985, 1986a,b, 1987).

Finally, compositional constraints identify a new component in the organismal phenotype, which may be called the "genome phenotype." Indeed, compositional constraints affect largely the structure and stability of the genome (at its 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 conceivably touch on a number of basic functions, such as replication, recombination, transcription, and translation, that are sensitive to the compositional and structural features just mentioned. This component is in addition to the other, classical component of the phenotype, which is formed by the gene products and is defined by nonsilent mutations in the genes and by mutations in regulatory signals.

Acknowledgments. The senior author thanks the Fogarty International Center for Advanced Study in the Health Sciences, National Institutes of Health, 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 financial help from the French Ministère de la Recherche et la Technologie (Programme Mobilisateur "Essor des Biotechnologies"). This paper is dedicated to the memory of Alberto Monroy and Eduardo Scarano.

References

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

Argos P, Rossmann MG, Grau UM, Zuber A, Franck G, Tratschin JD (1979) Thermal stability and protein structure. Biochemistry 18:5698–5703

Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Bernardi G (1982) The evolutionary origin and the biological role of non-coding sequences in the mitochondrial genome of yeast. In: Attardi G, Borst P, Slonimski PP (eds) Mitochondrial genes. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp 269–278

Bernardi G (1983) Genome instability and the selfish DNA issue. Folia Biol 29:82–92

Bernardi G (1985) The organization of the vertebrate genome and the problem of the CpG shortage. In: Cantoni GL, Razin A (eds) Biochemistry and biology of DNA methylation. Alan R. Liss, New York, pp 3–10

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

Bernardi G, Bernardi G (1986) The human genome and its evolutionary context. Cold Spring Harbor Symp Quant Biol 51, in press

Bernardi G, Ehrlich SD, Thiéry JP (1973) The specificity of deoxyribonucleases and their use in nucleotide sequence studies. Nature 246:36–40

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FL, Rindone WP, Swindell CD, Tung CS (1986) The GenBank genetic sequence data bank. Nucleic Acids Res 14:1–4

Bird A (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209–213

Blaisdell BE (1985) Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eukaryotic nuclear DNA sequences both protein coding and noncoding. J Mol Evol 21:278–288

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398

Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165:349–357

Chang L-YE, Slightom JL (1984) Isolation and nucleotide sequence analysis of the β type globin pseudogene from human, gorilla and chimpanzee. J Mol Biol 180:767–784

Chevallier A, Garel JR (1979) Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in *Bombyx mori* by using two-dimensional polyacrylamide gel electrophoresis. Biochimie 61:245–262

Davidson EH, Britten RJ (1979) Regulation of gene expression: possible role of repetitive sequences. Science 204:1052–1059

de Zamaroczy M, Bernardi G (1985) Sequence organization of the mitochondrial genome of yeast—a review. Gene 37:1–17

de Zamaroczy M, Bernardi G (1986a) The GC clusters of the mitochondrial genome of yeast and their evolutionary origin. Gene 41:1–22

de Zamaroczy M, Bernardi G (1986b) The primary structure of the mitochondrial genome of *Saccharomyces cerevisiae*—a review. Gene 47:155–177

de Zamaroczy M, Bernardi G (1987) The AT spacers and the *varl* gene of the mitochondrial genome of yeast: structure and evolutionary origin. Gene, in press

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603

Flanagan JG, Rabbitts TH (1982) Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing γ, Σ and α genes. Nature 300:709–713

Goldman MA, Holmquist GP, Gray MC, Caston LA, Nag A (1984) Replication timing of genes and middle repetitive sequences. Science 224:686–692

Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptionist program. Proc R Soc Lond [Biol] 205:581–598

Grantham R, Gautier G, Gouy M, Mercier R, Paré A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Grosjean H, Sankoff D, Min Jou W, Fiers W, Cedergren RJ (1978) Bacteriophage MS2 RNA: a correlation between the stability of the codon:anticodon interaction and the choice of code words. J Mol Evol 12:113–119

Hasegawa M, Yasunaga T, Miyata T (1979) Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res 7:2073–2079

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Kagawa Y, Nojima H, Nukima N, Ishizuka M, Nakajima T, Yasuhara T, Tanaka T, Oshima T (1984) High guanine plus cytosine content in the third letter of codons of an extreme thermophile. J Biol Chem 259:2956–2960

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England

Kimura M (1986) DNA and the neutral theory. Philos Trans R Soc Lond [Biol] 312:343–354

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Li WH, Luo CC, Wu CI (1985) Evolution of DNA sequences. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 1–94

Lipman DJ, Wilbur WJ (1983) Contextual constraints on synonymous codon choices. J Mol Biol 163:363–376

Mayr E (1976) Evolution and the diversity of life. Harvard University Press, Cambridge, Massachusetts

Nichols BP, Blumenberg M, Yanofsky C (1981) Comparison of the nucleotide sequence of trpA and sequences immediately beyond the trp operon of *Klebsiella aerogenus, Salmonella typhimurium* and *E. coli.* Nucleic Acids Res 9:1743–1755

Nussinov R (1980) Some rules in the ordering of nucleotides in the DNA. Nucleic Acids Res 8:4545–4562

Nussinov R (1981) The universal dinucleotide asymmetry rules and the aminoacid codon choice. J Mol Evol 17:237–244

Ohno S (1972) An argument for the genetic simplicity of man and other mammals. J Hum Evol 1:651–662

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607

Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP (1979) Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit β in *Escherichia coli.* Proc Natl Acad Sci USA 76:1697–1701

Proudfoot NJ, Gil A, Maniatis T (1982) The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. Cell 31:553–563

Russell GJ (1974) Characterization of deoxyribonucleic acids by doublet frequency analysis. PhD dissertation, University of Glasgow

Russell GJ, Walker PMB, Elton RA, Subak-Sharpe JH (1976) Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 108:1–23

Salinas J, Zerial M, Filipski J, Bernardi G (1986) Gene distribution and nucleotide sequence organization in the mouse genome. Eur J Biochem, in press

Sawada I, Beal MP, Shen CKJ, Chapman B, Wilson AC, Schmid C (1983) Intergenic DNA sequences flanking the pseudo alpha globin genes of human and chimpanzee. Nucl Acids Res 11:8087–8101

Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Subak-Sharpe H, Burk RR, Crawford LV, Morrison JM, Hay J, Keir AM (1966) An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbour base sequences. Cold Spring Harbor Symp Quant Biol 31:737–748

Swartz MN, Trautner TA, Kornberg A (1962) Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbour base sequences in deoxyribonucleic acids. J Biol Chem 237:1961–1967

Trifonov EN, Sussman JL (1980) The pitch of chromatin is reflected in its nucleotide sequence. Proc Natl Acad Sci USA 77:3816–3820

Wada A, Suyama A (1986) Local stability of DNA and RNA secondary structure and its relation to biological function. Prog Biophys Mol Biol 47:113–157

Zerial M, Salinas J, Filipski J, Bernardi G (1986a) Gene distribution and nucleotide sequence organization in the human genome. Eur J Biochem, in press

Zerial M, Salinas J, Filipski J, Bernardi G (1986b) Genomic localization of hepatitis B virus in a human hepatoma cell line. Nucl Acids Res 14:8373–8386

Zuber H (1981) Structure and function of thermophilic enzymes. In: Eggerer H, Huber R (eds) Structural and functional aspects of enzyme catalysis. Springer-Verlag, Berlin, pp 114–127

Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. II. A selectionist model for random fixation in proteins. J Mol Evol 7:269–311

Zuckerkandl E (1986) Polite DNA: functional density and functional compatibility in genomes. J Mol Evol 24:12–27