Molecular Evolution

Letter to the Editor

Codon Usage and Genome Composition

Giacomo Bernardi and Giorgio Bernardi

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

Summary. The GC levels of codon third positions from 49 genomes covering a wide phylogenetic range are linearly correlated with the GC levels of the corresponding genomes. Three different relationships have been found: one for prokaryotes and viruses, one for lower eukaryotes, and one for vertebrates. All points not fitting the first relationship can be brought into quasi coincidence with it when plotted against GC levels of coding sequences.

Key words. Codon usage — Genome composition — Coding sequences

Investigations from our laboratory (Bernardi et al. 1985) have shown that the nuclear genome of warm-blooded vertebrates exhibits a compositional compartmentalization in that it consists of a mosaic of very long (≫200 kb) DNA segments, the *isochores,* which are fairly homogeneous in base composition and belong to a small number of major classes distinguished by different GC levels. The families of DNA molecules derived from such classes can be separated and used to study the genome distribution of any sequence that can be probed. This approach has revealed that the base composition and the CpG/GpC ratio of both coding and noncoding sequences depend on the GC levels of the isochores harboring the sequences. These compositional constraints appear also to determine to a large extent the GC levels of codon third positions. Here we have investigated whether the GC levels of codon third positions are *generally* related to those of the corresponding genomes.

Our approach has consisted of plotting the GC levels of codon third positions of 302 genes from 49 genomes covering a wide phylogenetic range (Table 1) against the GC levels of these genomes (or their compositional compartments in the case of warm-blooded vertebrates and phage lambda); for genes belonging to the same genome, average values were used. The results obtained (Fig. 1) indicate that a common linear relationship exists for all prokaryotes and viruses; lower eukaryotes fall on a line showing a slightly different slope; invertebrates seem to fall on the same line, except for *D. melanogaster.* In the case of vertebrates, points (mainly from Bernardi et al. 1985) fall on a line shifted to the left relative to the prokaryotic–viral line.

These different relationships can be understood if one takes into consideration the following:

(a) The first relationship is followed by viral and bacterial genomes, namely by genomes that are almost only formed by coding sequences; eukaryotic genomes in which noncoding (intergenic and intervening) sequences are scarce or close in GC to coding sequences follow relationships that are not very divergent from the first one. These cases are apparently common in lower eukaryotes and invertebrates.

(b) The first relationship is not followed when noncoding sequences are abundant and different in GC from coding sequences. This situation was found in vertebrates and in *D. melanogaster.* In the first case, points appear to fit a common relationship. In contrast, the second case should be considered as an exception because another phylogenetically close

Offprint requests to: Giacomo Bernardi

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

Genomes	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.	1
16 Rhizobium japonicum	2
17 Salmonella typhimurium	6
18 Shigella dysenteriae	3
19 Streptomyces fradiae	1
20 Thermus thermophilus	1
21 Vibrio cholerae	4
Viruses	
22 Abelson Murine leukemia virus	1
23 Adenovirus type 12	5
24 AKV Murine leukemia virus	2
25 Avian Sarcoma virus Y73	1
26 Hepatitis B virus	2
27 Herpes simplex virus type 1	2
28 Human adult T-Cell leukemia virus	3
29 Human papilloma virus	4
30 Human papovavirus BK	7
31 Mouse hepatitis virus	2
32 Polyoma virus	6
33 Tobacco mosaic virus	6
Lower eukaryotes and fungi	
34 Dictyostelium discoideum	3
35 Neurospora crassa	3
36 Physarum polycephalum	2
37 Saccharomyces cerevisiae	15
38 Trypanosoma brucei	5
Invertebrates	
39 Bombyx mori	2
40 Chironomus thummi thummi	1
41 Drosophila melanogaster	20
42 Strongylocentrotus purpuratus	2
Vertebrates	
43 Cyprinus carpio	1
44 Lophius americanus	1
45 Xenopus laevis	2
46 Chicken (L2)	4
47 (H2)	5
48 Mouse (L1)	1
49 (L2)	5
50 (H2)	3
51 Rabbit (L2)	1
52 (H2)	1
53 Man (L2)	5
54 (H1)	2
55 (H3)	6

[a] These comprised almost all the genes whose primary structure was available in the GenBank, as of April 1985. In the case of warm-blooded vertebrates, the table lists the isochores (L1, L2, H1, H2, H3) in which genes are located (Bernardi et al. 1985)

organism (*C. thummi thummi*) does not show the same large difference in GC between noncoding and coding sequences.

(c) It is of great interest to note that if all the points not fitting the prokaryotic–viral relationship are plotted againt GC levels of coding sequences, they are brought into quasi-coincidence with the prokaryotic viral line (Fig. 1).

We conclude that there is a general linear relationship between the GC levels of codon third positions (which correspond, to a large extent, to codon usage) and those of coding sequences. We also con-

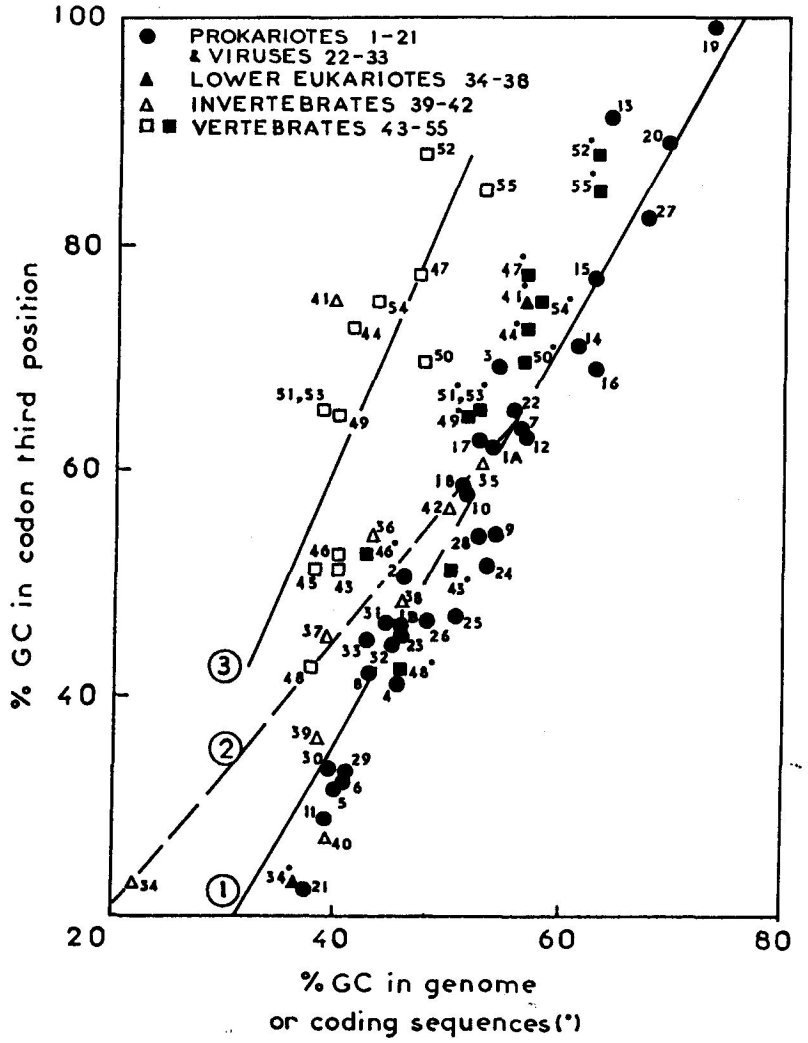

Fig. 1. Plot of GC levels of codon third positions against the GC levels (1) of the genomes of prokaryotes, viruses, lower eukaryotes, and invertebrates, and (2) of the genomes or the isochores of vertebrates. Numbers correspond to the different species listed in Table 1. The relationships shown by prokaryotic and viral genomes, by lower eukaryotes, and by vertebrates are indicated by lines 1, 2 and 3, respectively. Values not fitting relationship 1 have been replotted (filled symbols with asterisked numbers) against GC levels of coding sequences

clude that there are several linear relationships between codon usage and genome composition.

These results have a number of general implications which mainly concern codon strategy, genome organization, and the neutral theory of evolution. These will be discussed elsewhere.

References

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

Received October 1, 1985