Variation in G+C-content and codon choice: differences among synonymous codon groups in vertebrate genes

A.Marín*, J.Bertranpetit[1], J.L.Oliver[2] and J.R.Medina

Departimento de Genética y Biotecnia, Facultad de Biología, Universidad de Sevilla, Aptdo 1095, E-41080 Seville, [1]Departimento de Biología Animal, Facultad de Biología, Universidad de Barcelona, Avda Diagonal 645, E-08028 Barcelona and [2]Unidad de Genética, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, Spain

Received March 17, 1989; Revised June 1, 1989, Accepted June 26, 1989

ABSTRACT

The relationship between G+C-content and codon usage in genes of human, mus, rat, bovine and chicken nuclear genomes was investigated. Correlation and lineal regression analyses were carried out on plots that related the frequency of each codon within each synonymous codon group to the G+C-content of the coding sequence as a whole. Under GC pressure, in most of the quartet codon groups there is a preferential choice of the C-ending codon, except in leucine and valine codon groups where the choice of the G-ending codon is preferred. Among duets, the choice of codons specifying phenylalanine and glutamate shows the strongest dependence on G+C-content. The relationship found between G+C-content and codon usage in these genomes correlate with taxonomic distance.

INTRODUCTION

Early works showed that the base ratio A+T/G+C of overall DNA differs among different species (1), and that different DNA segments isolated from organisms of a given species have the same base ratio (2,3).

However, in vertebrates there are large differences in base composition between different DNA regions in the same genome. It has been shown recently (4), that the nuclear genome of warm-blooded vertebrates exhibits a compositional compartmentalization, being a mosaic of very long DNA sequences which are relatively homogeneous in their G+C-content (either G+C-rich or A+T-rich). These segments are named 'isochores', and it has been proposed that genes that map in chromosomal R-bands belong to G+C-rich isochores, and those in G-bands to A+T-rich isochores (4−8).

According to the degeneracy of the genetic code one might expect that differences in G+C-content among genes could be accounted for by changes in G+C-content at the third, and to a lesser extent, at the first codon position, while the second codon position is constrained by the choice of amino acid. The G+C level found at each codon position in vertebrate genes shows a positive linear relationship to the G+C-content of the corresponding coding sequences; the slopes of the regression lines increase from second, to first, to third positions. Furthermore, a high correlation is found between the G+C-content at the third codon position of exons and the neighboring introns and flanking sequences. These relationships indicate the existence of compositional constraints operating on both coding and noncoding sequences (5−7,9,10).

The variation in G+C-content through the genome is accompanied by changes in codon usage (4−7). In vertebrate genes, the G+C-content at the third codon position is distributed over a wide range (7), and subsequently, the codon-choice patterns of various genes in the same organism may differ considerably.

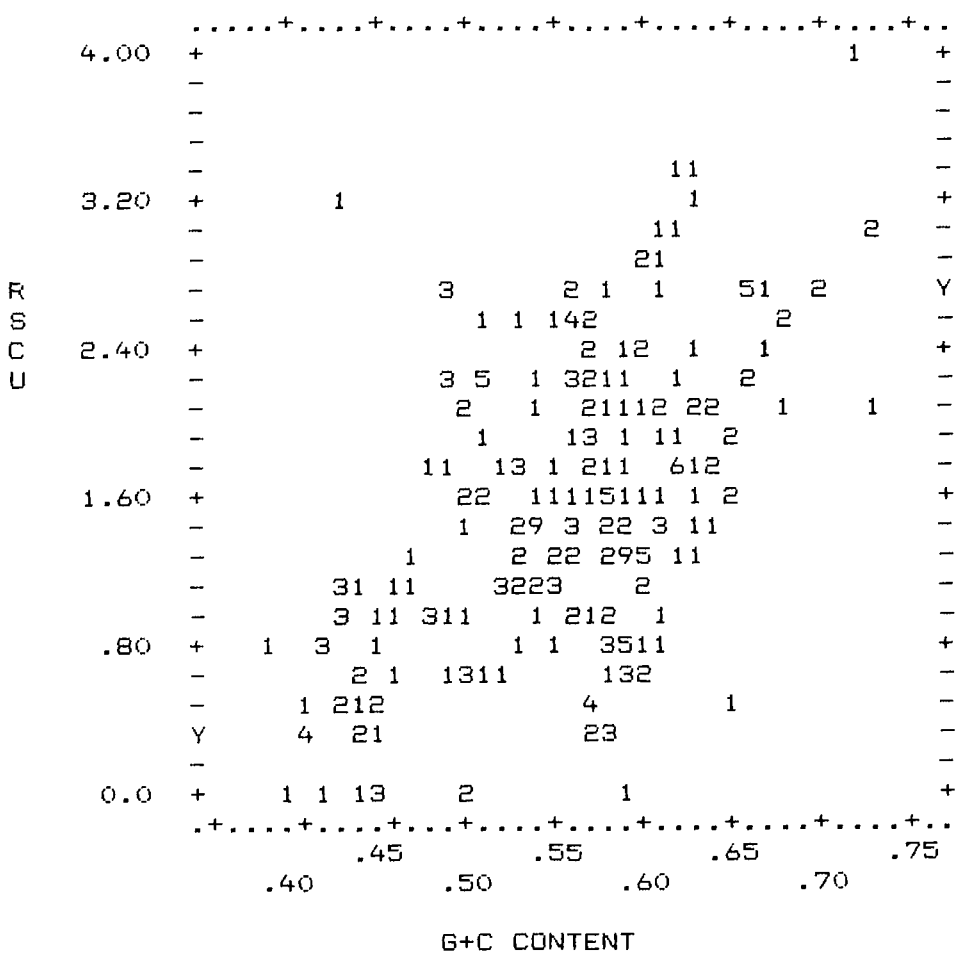

Fig. 1 RSCU values of the CGC codon (Arg4) plotted against the G+C content (G+C/G+C+A+U) of the coding sequence in human genes. The regression line may be drawn connecting the Y marks on the plot frame. No. of genes = 277, correlation coefficient = .511, slope = 5.76.

In this paper we deal with the effect of differences in the G+C-content of genes on codon usage. Our approach was to investigate at the level of synonymous codon groups, examining the choice of codons within each codon group in relation to the G+C-content of the coding sequence.

DATA AND METHODS

Our starting material is the compilation by Ikemura and coworkers (11) from the GenBank Genetic Sequence Data Bank (release 50.0) of codon appearances in the nuclear genomes of *Homo sapiens* (HUM), *Mus musculus* (MUS), *Rattus norvegicus* (RAT), *Bos taurus* (BOV) and *Gallus gallus* (CHK).

In this work we have tentatively divided the amino acids encoded by six codons (arginine, leucine, and serine) into four-fold degenerate (Arg4, Leu4 and Ser4) and two-fold degenerate (Arg2, Leu2 and Ser2) codon groups, which contain codons differing only in the third nucleotide. Methionine and tryptophan which have only one codon each, are not relevant in this study. Termination codons have not been considered as they appear only once in each gene.

To allow codon choice to become apparent, in our analyses we included only those genes for which at least five codons of the pertinent codon-group were present, because of this the number of genes differs in each analysis. The number of five codons was chosen after carrying out trials setting 1, 3, 5, 7 and 10 codons as lower limits, the slopes obtained in these conditions were in most of cases within the interval of 1 standard error of those computed for the limit of five codons.

To express the nature of the relationships between codon usage and G+C-content we proceeded as follows. In the first place, we computed the relative synonymous-codon usage value (RSCU) of each codon as defined by Sharp et al. (12). The RSCU value for a codon is the observed frequency of that codon divided by the frequency expected if usage of synonymous codons was uniform. This draws attention to the particular choice of codons irrespective of the amino acid composition of the gene product, and enables comparison between data sets of different sizes.

In the second place, for each genome and codon we made a diagram, in which each gene is represented by a point whose coordinates are the RSCU value of the codon under analysis (Y axis), and the G+C-content of the coding sequence (X axis). Thus we obtained a scatter of points relating the choice of each codon to the G+C-content of the coding sequence. Then, we computed the correlation coefficient, and a regression line ($y = a + bx$) was drawn in each graph, as an example the plot for the CGC codon (Arg4) in human genes is given in Fig. 1. For these tasks we used BMDP6D and BMDP1R programs (13).

Finally, we made a numerical evaluation of the extent of genomic differences by computing a 'distance' between each pair of genomes. The distance measure used is simply the sum of differences between the values of the slopes of the regression lines for two genomes. This distance is a version of the 'Manhattan metric' often used by numerical taxonomists (14).

RESULTS AND DISCUSSION

The G+C-content of a coding sequence can be increased by the choice of codons ending in G or C. In the case of four-fold degenerate codon groups, both G- and C-ending codons are available. In the case of two-fold degenerate codon groups there are two modalities: the pyrimidine-restricted codon groups (where the third codon position is U or C), and the purine-restricted codon groups (where the third position is A or G). The codon group of isoleucine is unique in that the choice can be made among A-, U- and C-ending codons.

Codon choice in four-fold degenerate codon groups

In Table 1 the results of correlation and regression analyses for four-fold degenerate and isoleucine codon groups are given; as indicated in the method section, these were carried out after plotting the RSCU values of codons ending in each of the four bases A, T, C and G in each codon group against the G+C-content of the coding sequence as a whole.

As the correlation coefficients indicate, in most quartet codon groups, the usage of both G- and C-ending codons increases as the G+C-content of the coding sequence does, while

Table 1. Quartet codon groups. Correlation coefficients (R) and slopes (s) when plotting the RSCU values of codons ending in bases A, T, C and G against the G+C content of the coding sequence as a whole. N = number of genes. The first two bases of the codon are given in brackets.

		Ending base of the codon							
		A		U		C		G	
	N	R	s	R	s	R	s	R	s
Arg4 (CG-)									
HUM	277	−.532	−4.68	−.425	−3 52	.511	5.76	249	2.44
MUS	125	−.420	−5.16	−.349	−3.80	.591	8.16	.064*	0 80*
RAT	166	−.103*	−1 28*	−.381	−3 96	.428	6.56	−.130*	−1.36*
BOV	42	−.615	−3.92	−.545	−2.96	.272*	3.00*	.323	3.88
CHK	49	−.315	−2.36	−.668	−8.48	.623	7.72	.331	3.12
Leu4 (CU-)									
HUM	396	−.557	−2.08	−.631	−3.80	.076*	0.44*	.631	5.44
MUS	177	−.377	−1.80	−.542	−3.56	−.026*	0.20*	.551	5.56
RAT	224	−.367	−2.28	−.446	−3.76	.074*	0.56*	.512	5.52
BOV	68	− 642	−2.40	−.457	−3.24	.276	1.52	.587	4.12
CHK	68	− 411	−1.40	−.759	−4.64	.290	2.20	.436	3 84
Ser4 (UC-)									
HUM	374	−.534	−4.04	−.445	−4 04	599	5.96	.365	2 12
MUS	173	−.439	−4.36	−.202	−2 52	397	4.96	.296	1.96
RAT	158	−.465	−5.16	−.119*	−1.44*	.339	4.32	.347	2.28
BOV	59	−.553	−4.00	−.465	−3.28	.470	4.04	668	3 24
CHK	56	−.460	−3.76	− 570.	−7.24	.458	6.84	437	4.16
Thr (AC-)									
HUM	371	−.536	−4.00	−.547	−4.20	.578	5.72	.397	2.48
MUS	172	−.394	−3.60	−.505	−4.56	.407	4.96	.452	3.24
RAT	203	−.249	−2.92	−.291	−3.64	.354	5.12	.220	1.44
BOV	59	−.424	−3.00	−.483	−3.08	.401	3.24	.446	2.84
CHK	65	−.784	−8.60	−.593	−5.56	.665	9.00	.538	5.16
Pro (CC-)									
HUM	370	−.450	−4.00	−.488	−4.16	.466	4.84	.570	3 32
MUS	161	−.314	−3.08	−.285	−2.56	.285	3.28	.350	2.32
RAT	199	−.436	−4.52	− 275	−2.60	.501	6.96	.022*	0.20*
BOV	59	− 492	−3.44	−.386	−3.24	.514	4.48	.379	2.20
CHK	59	−.786	−7.56	−.688	−6.52	.755	10.72	.602	3.36
Ala (GC-)									
HUM	395	−.526	−3.68	− 588	−4.32	.606	5.60	.453	2.40
MUS	176	−.430	−3.88	− 122*	−1.08*	.180	1.96	.423	3.00
RAT	216	−.359	−3.24	−.385	−3.40	.341	4.16	.402	2.48
BOV	65	−.555	−4.20	−.562	−3.48	.615	5.32	.560	2.36
CHK	68	−.728	−5.56	−.526	−5.12	.646	7.40	.480	3.32
Gly (GG-)									
HUM	381	−.499	−4.52	−.406	−2.32	.506	4.64	.326	2.20
MUS	166	−.591	−5 76	−.211	−1 64	.583	6.12	.164	1 28
RAT	217	−.457	−4.56	−.204	−1 60	.496	6.24	−.013*	0.12*
BOV	62	−.525	−4.00	−.307	−1.48	.364	2.64	.496	2.84
CHK	69	−.640	−5.32	−.511	−4.40	.634	6.72	.374	3.00
Val (GU-)									
HUM	379	−.543	−3.12	−.621	−4.28	.209	1.52	.570	5.84
MUS	171	−.328	−1.84	−.632	−5.36	−.033*	0.28*	.640	7 56
RAT	202	−.272	−1.92	−.569	−4.80	.067*	0.76*	.389	5.96
BOV	67	−.601	−2.32	−.374	−2 16	.404	2.68	.253	1.84
CHK	68	−.652	−3.16	−.694	−4.84	.290	2.48	.546	5.52

(continued)

Table 1 (*continued*)

		Ending base of the codon							
		A		U		C		G	
	N	R	s	R	s	R	s	R	s
Ile (AU-)									
HUM	323	−.445	−2.04	−.571	−4.59	.682	6.63		
MUS	165	− 547	−3.48	−.504	−4.80	.718	8.28		
RAT	195	−.360	−2.43	−.359	−5.49	.523	7 92		
BOV	53	−.511	−2.37	−.411	−2.85	.556	5.22		
CHK	31	−.651	−2.67	−.732	−4.95	.847	7.62		

* Not significant at 0.05 level

the opposite occurs for A- and U-ending codons. However, in some codons the expected behavior is not found, and the choice of these codons seems not to be dependent on the genomic G+C-content. The codons which show non statistically significant correlation coefficients are: CGG (in MUS and RAT genomes); CUC (in HUM, MUS and RAT); UCU (in RAT); CCG (in RAT); GCU (in MUS); GUC (in MUS and RAT); and GGG (in RAT). Some of them also show rather low correlation coefficients in the other genomes. This observation undoubtedly requires further investigation in order to elucidate the constraints involved in the usage of these codons. As a common feature, these codons are G+C-rich, with the exception of UCU, and it is remarkable that all they are quite frequently used in all the five genomes, being their averaged RSCU values slightly greater than 1, (that means that they are used more than 25 % of the times to code for the corresponding amino acid). Thus, it is not the avoidance of these codons that makes them insensitive to changes in the G+C content of genes.

We used linear regression analysis to quantify the association shown by correlation. The slopes of the regression lines drawn in each plot (Table 1) indicate the increase (or decrease) in the relative frequency of the codon under analysis when the G+C-content of the coding sequence increases. The higher the slope, the greater the preference (or avoidance) for that codon when the G+C-content of the coding sequence increases.

Results in Table 1 show that in most of the codon groups with four-fold degeneracy there is a clear bias favoring the choice of C-ending codons under GC pressure, except in the cases of leucine (quartet) and valine, where G-ending codons are used more frequently as the G+C-content increases (the exception is the bovine genome where the C-ending codon of valine is preferred). This general trend of C prevailing over G in the third codon position was already noted by previous authors (5), by investigating the correlations between base compositions of the three codon positions and those of the corresponding exons.

The increase in usage of G- or C-ending codons is concomitant with a decrease in the usage of A- and U-ending codons (negative correlation coefficients and slopes). Codon groups of leucine (quartet), valine and isoleucine show higher slopes in the case of the U-ending codon (the exception, again, is the bovine genome). The glycine codon group show higher slopes for the A-ending codon in all genomes. In the remaining codon groups, there is a slight tendency towards higher slopes in the A-ending codon, although the standard errors of slopes (not shown) are often overlapping.

Codon choice in two-fold degenerate codon groups

In the case of duets there is an excess of pyrimidine-restricted groups (which increase C-ending codons under GC pressure) over purine-restricted groups (which increase G-ending codon use under GC pressure). In Table 2 the correlation coefficients and slopes

Table 2. Duet codon groups. Correlation coefficients (R) and slopes (s) when plotting the RSCU value of the C-ending codon (pyrimidine restricted duets) or the G-ending codon (purine restricted duets) against the G+C content of the coding sequence as a whole. N = number of genes. The first two bases of the codon are given in brackets

	Pyrimidine-restricted duets (C-ending codon)								
	N	R	s	N	R	s	N	R	s
	Ser2 (AG-)			Asn (AA-)			His (CA-)		
HUM	324	.571	3 26	326	.638	4.10	261	.664	4.36
MUS	126	.423	2.82	162	.566	3.68	128	.306	2.20
RAT	166	.363	2.96	172	.368	3.16	98	.434	3.64
BOV	51	.696	4.18	45	767	4.22	40	.638	4.78
CHK	41	.592	3.52	54	.829	4.66	43	706	4.82
	Asp (GA-)			Tyr (UA-)			Cys (UG-)		
HUM	366	.661	3.48	292	.627	3.58	276	609	3.68
MUS	167	.661	4.26	133	497	3.38	120	.483	3.50
RAT	202	468	3.26	140	.302	3.06	111	.497	3.18
BOV	59	.671	3.46	47	.633	3.34	47	.617	2.98
CHK	55	.729	4.74	42	.646	4.78	29	.531	3.92
	Phe (UU-)								
HUM	346	.710	4.18						
MUS	165	.671	4.26						
RAT	177	.611	4.38						
BOV	62	.787	4 30						
CHK	53	.655	4.44						
	Purine-restricted duets (G-ending codon)								
	N	R	s	N	R	s	N	R	s
	Arg2 (AG-)			Leu2 (UU-)			Lys (AA-)		
HUM	239	.605	3.36	161	.648	4.80	372	688	3.58
MUS	116	.254	1.66	75	.261	1.92	172	.537	3 18
RAT	150	.297	2.40	75	.330	1.96	213	.397	2.32
BOV	38	.743	5.06	19	.529	2.40	61	.523	2.68
CHK	31	.778	6.90	22	.630	4.16	68	.797	3.68
	Gln (CA-)			Glu (GA-)					
HUM	367	.605	2.88	391	.791	4.44			
MUS	150	.513	3.46	172	.747	5.48			
RAT	198	.541	3 42	219	.727	4.50			
BOV	55	.449	1.90	63	822	4.62			
CHK	53	.646	3.22	63	.745	4.48			

of the regression lines for duet codon groups are given. Results in Table 2 correspond to the C-ending codon in the case of pyrimidine-restricted groups, and to the G-ending codon in the case of purine-restricted groups. The figures for the synonymous U- and A-ending codons are the same with a negative sign.

In duet codon groups the highest slope values are found in the phenylalanine codon group among the pyrimidine-restricted, and in the glutamate codon group among the purine-restricted codon groups. It is worth noting that in the leucine (duet) codon group, the slope in human genes is more than twice that of rodents; in most of cases, the chicken genes show the highest slopes.

Codon choice at first position

Some degeneracy at the first codon position creates a mechanism that allows the overall G+C composition of the coding sequence to be reflected in the codon usage at that position.

Table 3. Correlation coefficients (R) and slopes (s) in plots of the ratio Arg4/Arg2, Leu4/Leu2 and Ser4/Ser2 against the G+C content of genes.

	RARG			RLEU			RSER		
	N	R	s	N	R	s	N	R	s
HUM	367	.391	0.21	377	.522	0.57	401	−.085*	−0.01*
MUS	159	.325	0.15	164	531	0.75	177	−.070*	0.02*
RAT	199	.358	0.23	203	.467	0.46	219	.055*	0.02*
BOV	61	.322	0.17	60	.508	0.41	66	−.004*	0 00*
CHK	55	.601	0.25	53	514	0 62	65	198*	0.04*

* Not significant at 0.05 level

Arginine and leucine provide a choice at first position between C (Arg4) or A (Arg2), and C (Leu4) or U (Leu2), respectively. In the case of serine the choice at the first codon position is between A (Ser4) or U (Ser2), and, therefore, no effect of the G+C content is expected on that choice.

To examine the effect of G+C content on the choices at first codon position, we have plotted the proportion of codons Arg4/Arg2 (RARG), Leu4/Leu2 (RLEU), and Ser4/Ser2 (RSER) in each gene against the G+C content of the coding sequence as a whole. Correlation coefficients and linear regression analyses are shown in Table 3.

These results confirm the relationship between base composition and codon usage. Correlation coefficients and slopes of regression lines are higher in plots for RLEU than in plots for RARG in all five genomes, this indicates that leucine codons are more affected by G+C content than are arginine codons. Correlation coefficients and slopes in plots of RSER were all nonsignificant.

It is interesting to note that the two sets of serine codons cannot be converted into each other by single nucleotide mutations. This fact has been beautifully exploited by Brenner (15) to make a reconstruction of the pathway of molecular evolution of the active-site sequences of enzymes that have analogous essential serine residues; thus, serine residues coded for by the duet and the quartet codon groups are preserved through evolution due to mutational distance and to insensitivity to genomic changes in the G+C content.

Comparison of genomic differences

In the previous sections we made a statistical characterization of the compositional constraint (G+C-content) that could affect codon choice in vertebrate genes. To examine the degree of relatedness between these genomes, we have computed a matrix of distances by using a 'Manhattan metric' with the slopes of regression lines obtained for each codon (Table 4). It should be mentioned that the value of these distances has no biological meaning and only their relative magnitudes are important. It can be seen that there is a certain correlation between the computed genomic distance and taxonomic distance between these

Table 4. Distances between genomes. The distance between a pair of genomes in the sum of the differences between the values of the slopes obtained in each plot.

HUM	—				
MUS	47.7	—			
RAT	48.6	42.4	—		
BOV	47.2	67.9	63.5	—	
CHK	66.2	93.5	91.5	91.4	—
	HUM	MUS	RAT	BOV	CHK

species, thus mus and rat are the more closely related species, and each mammal is more distant to chicken than to any other mammal. The basis of these relationships is unknown, but could reflect similarities and differences in the availabilities of tRNA species or in the isochore organization among these genomes. For example, it has been reported the existence of differences in the heavy components of human and mouse genomes (16) concerning the presence in the human genome of a G+C-rich fraction containing a large number of genes and specific repetitive sequences which is not represented in the mouse genome.

Concluding remarks

Many factors have been considered to influence the non-random usage of synonymous codons. Most of them are related to translational efficiency through the stability of the codon-anticodon complex (17), or in terms of tRNA availability (18) and the level of gene expression (12, 19−23). Other hypotheses have considered nucleic acid secondary structure (24, 25) and contextual constraints (26−28). Evolutionary aspects of codon usage have also been considered(29−31).

The role of the organism's G+C-content in relation to codon choice has been noted in a number of papers (32, 33 and references therein), being referred to as a rule by Ikemura and Ozeki (34) in explaining non-random patterns of codon choice. It is worth mentioning that in enterobacterial genes it has been shown that the bias in codon choice due to genomic G+C content is greater in modestly expressed genes than in highly expressed genes (35).

The importance of the G+C-content in molding codon choice has been emphasized by Bernardi and coworkers (5, 6) who claimed that codon usage is largely determined by compositional constraints concerning both G+C content and the content of individual bases. They suggested that there exists a compositional strategy of the genome, providing a rationale for the 'genome hypothesis' of Grantham and coworkers (36, 37). Recently, it has been proposed that differences in base composition in vertebrate genomes are caused by different mutational bias of DNA polymerases in germline cells (38), or by variation in mutation patterns along the replication timing of different chromosomal regions in the germline (39).

Our work provides further support for and substantiation of the relationship between G+C-content and codon choice, giving a picture of how it works on individual codons in vertebrate genes. One possible bias in our approach is that genes included in the study may be submitted to different functional constraints which can be superimposed on the compositional constraint analysed here. Perhaps the most important finding is that there are different linear relationships between the G+C-content of coding sequences and the choice of particular codons, and that these relationships, which are a reflection of the coding strategy of each genome, seem to be correlated with taxonomic distance. This strategy seems to have been well conserved ever since the radiation of mammals (about 80 million years) and is somewhat different in birds (divergence time from mammals 270 million years).

ACKNOWLEDGEMENTS

We are most grateful to T. Ikemura for send us an early copy of the table of codon usage (11) employed in this study, to B. Cubero for her valuable discussion, and to E. Martínez-Force for help in preparing the manuscript. This work was partially supported by the DGICYT of the Spanish Government (PB87−0881) to JLO.

*To whom correspondence should be addressed

REFERENCES

1. Chargaff,E. (1955) The Nucleic Acids. Chargaff,E. and Davidson,J.M. (eds) Vol I. Academic Press, New York.
2. Rolfe,R. and Meselson,M. (1959) Proc. Natl. Acad. Sci. USA 45, 1039–1044.
3. Sueoka,N. (1962) Proc. Natl. Acad. Sci. USA 48, 582–592.
4. Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) Science 228, 953–958.
5. Bernardi,G. and Bernardi,G. (1986a) J. Mol. Evol. 24, 1–11.
6. Bernardi,G. and Bernardi,G. (1986b) Cold Spring Harbor Symp. Quant. Biol. 51, 479–487.
7. Aota,S. and Ikemura,T. (1986) Nucl. Acids Res. 14, 6345–6355 & 8702.
8. Ikemura,T. and Aota,S. (1988) J. Mol. Biol. 203, 1–13.
9. Bernardi,G. and Bernardi,G. (1985) J. Mol Evol. 22, 363–365.
10. Ikemura,T. (1985) Mol. Biol. Evol. 2, 13–34.
11. Aota,S., Gojobori,T., Ishibashi,F., Maruyama,T. and Ikemura,T. (1988) Nucl. Acids Res. 16, r315–r402.
12. Sharp, P. M., Tuohy, T. M. F. and Mosurski, R. (1986) Nucl. Acids Res. 14, 5125–5143.
13. Dixon, W. J. and Brown, M. B. (1979) BMDP-79 Biomedical computer programmes P Series. University of California Press, Berkeley.
14. Sneath,P.H.A. and Sokal,R.R. (1973) Numerical Taxonomy. Freeman, San Francisco.
15. Brenner,S. (1988) Nature 334, 528–530.
16. Zerial, M., Salinas, J., Filipski, J. and Bernardi, G. (1986) Eur. J. Biochem. 160, 479–485.
17. Grosjean,H. and Fiers,W. (1982) Gene 18, 199–209.
18. Ikemura,T. (1982) J. Mol. Biol. 158, 573–597.
19. Bennetzen, J. L. and Hall, B. D. (1982) J. Biol. Chem. 257, 3026–3031.
20. Gouy, M. and Gautier, C. (1982) Nucl. Acids Res. 10, 7055–7074.
21. Sharp, P. M. and Li, W.-H. (1986) Nucl. Acids Res. 14, 7737–7749.
22. Sharp, P. M. and Li, W.-H. (1987) Nucl. Acids Res. 15, 1281–1295.
23. Thomas, L. K., Dix, D. B. and Thompson, R. C. (1988) Proc. Natl. Acad. Sci. USA 85, 4242–4246.
24. Fitch,W.M. (1976) Science 194, 1173–1174.
25. Hasegawa,M., Yasunaga,T. and Miyata,T. (1979) Nucl. Acids Res. 7, 2073–2079.
26. Nussinov,R. (1980) Nucl. Acids Res. 8, 4545–4562.
27. Lipman,D.J. and Wilbur,W.J. (1983) J. Mol. Biol. 163, 363–376.
28. Ohno, S. (1988) Proc. Natl. Acad. Sci. USA 85, 4378–4382.
29. Sharp, P. M. and Li, W.-H. (1986) J. Mol. Evol. 24, 28–38.
30. Bulmer, M. (1987) Nature 325, 728–730.
31. Li, W.-H. (1988) J. Mol. Evol. 24, 337–345.
32. Osawa,S., Ohama,T., Yamao,F., Muto,A., Jukes,T.H., Ozeki,H. and Umesono,K. (1988) Proc. Natl. Acad. Sci. USA 85, 1124–1128.
33. West,S.E.H. and Iglewski,B.H. (1988) Nucl. Acids Res. 16, 9323–9335.
34. Ikemura,T. and Ozeki,H. (1983) Cold Spring Harbor Symp. Quant. Biol. 47, 1087–1097.
35. Nichols, B. P., Blumenberg, M. and Yanofsky, C. (1981) Nucl. Acids Res. 9, 1743–1755.
36. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) Nucl. Acids Res. 8, r49–r62.
37. Grantham,R., Gautier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981) Nucl. Acids Res. 9, r43–r74.
38. Filipski, J. (1987) FEBs Lett. 217, 184–186.
39. Wolfe, K. H., Sharp, P. M. and Li,W.-H. (1989) Nature 337, 283–285.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.