FEBS Letters 352 (1994) 7–10

FEBS 14511

Distribution profiles of GC content around the translation initiation site in different species

Masahiko Mizuno, Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

Received 21 July 1994

Abstract We have analyzed the distribution of guanine–cytosine (GC) content around the translation initiation site in genomic DNA sequences of different species. A set of sequences belonging to one species is aligned at the translation initiation site, and the average GC content is calculated for 100 base windows over a range of 500 bases each for upstream and downstream region. Consistent with previous observations that coding regions are more GC-rich than non-coding regions, we observe a jump in the GC content at the translation initiation site, except for vertebrate sequences. It was also found that the overall shape of the GC content profile is similar within each organism group even though the average GC contents can be very different. Furthermore, by examining different profiles for different species, we have found a negative correlation between the average GC content and the jump size at the translation initiation site. Apparently, more AT-rich genomes, which tend to lack macroscopic mosaic structures, exhibit more marked differences in the GC content at the microscopic level.

Key words: GC content; Base composition; Translation initiation site; Species specificity; Genome organization

1. Introduction

The GC content of a genome differs in various species. There are also variations in the GC content within each genome from the level delineating coding and non-coding sequences to a higher level of chromosomal bands. For example, human chromosomes have Giemsa and Reverse chromosomal bands, which seem to correspond to AT- and GC-rich regions, respectively, in the mega-base range [1]. The nuclear genome of warm-blooded vertebrates exhibits a compositional compartmentalization, in that it consists of a mosaic of very long (>300 kb) DNA segments, the isochores. They belong to a small number of classes characterized by different GC levels and by fairly homogeneous base compositions [2,3]. The genomes of several angiosperms are characterized by a compositional compartmentalization and an isochore organization. There are striking differences in the compositional distributions of coding and intron sequences between the dicots (pea, sunflower and tobacco) and monocots (maize, rice and wheat [4]. A bacterial genome such as *E. coli* is not presumed to have a macroscopic mosaic structure in GC content [5], however, its protein-coding regions tend to be richer in GC content than non-coding ones, as in other genomes [6].

There have been interesting observations of GC content variations at the sequence level; notably, the positive correlation between the GC content at the codon third position in genes of vertebrates and the GC content of the genome portion surrounding each gene [7,8]. We focussed our analysis on the differences in different species, specifically in the region surrounding the translation initiation site over the range up to 1000 bases long. This region is important to gene expression and may reflect different control mechanisms in different species.

*Corresponding author. Fax: (81) (774) 32-8235.
E-mail: kanehisa@kuicr.kyoto-u.ac.jp

2. Materials and methods

2.1. Selection of data

We selected from the GenBank database (releases 72 and 73) those genomic sequence entries that contained complete protein coding sequences. Also incuded were the coding sequences of 500 bases or longer without a stop codon. We excluded mRNA entries, pseudogenes, and genes encoded by mitochondria, chloroplasts, kinetoplasts, and plasmids.

Data sets were selected for the following organism groups and species: mammals (human, mouse, rat, bovine, and rabbit), other vertebrates (chicken and *Xenopus*), invertebrates (*Drosophila, Caenorhabditis*, and *Plasmodium*), plants (*Arabidopsis, Zea,* rice, and tomato), fungi (*Saccharomyces* and *Dictyostelium*), and bacteria (*Escherichia, Bacillus, Salmonella*, and *Pseudomonas*). In each organism group, most of the species were selected according to the abundance of GenBank entries.

In order to examine the possible statistical bias of these data sets, homologous sequences were removed and given different threshold levels according to amino acid sequence homology. For each pair of DNA sequences, a pairwise alignment was made for the translated amino acid sequences using the Dayhoff PAM250 matrix. The DNA sequences coding for less than 20 amino acids were not considered here. The degree of homology was defined by the number of exact matches divided by the length of the shorter sequence. If any pair of sequences exceeded the threshold homology level, only the longer sequence was retained in the data set. We chose five threshold values: 20%, 25%, 50%, 75% and 100% (no threshold). Table 1 shows the numbers of sequences in the data sets utilized in the present analysis.

2.2. Averaging method

A set of DNA sequences were aligned at the translation initiation site and the distribution of GC content was calculated both upstream and downstream of this site. A window length of 100 bases was shifted by 50 bases at a time along the nucleotide sequence from the −500 base position to the +500 base position, where the first nucleotide of the initiator codon was designated as +1, and the preceding base position by −1. The average GC content in each of the 100 base windows was calculated accordingly. When a window contained sequences of less than 100 bases they were not included in the averaging. If a 3′ untranslated region existed between the base positions +1 and +500, the sequence was truncated at the stop codon, i.e. our GC content profiles did not include 3′ untranslated regions. The average GC content in a 100 base window was plotted at the base position in the center of the window.

SSDI 0014-5793(94)00898-1

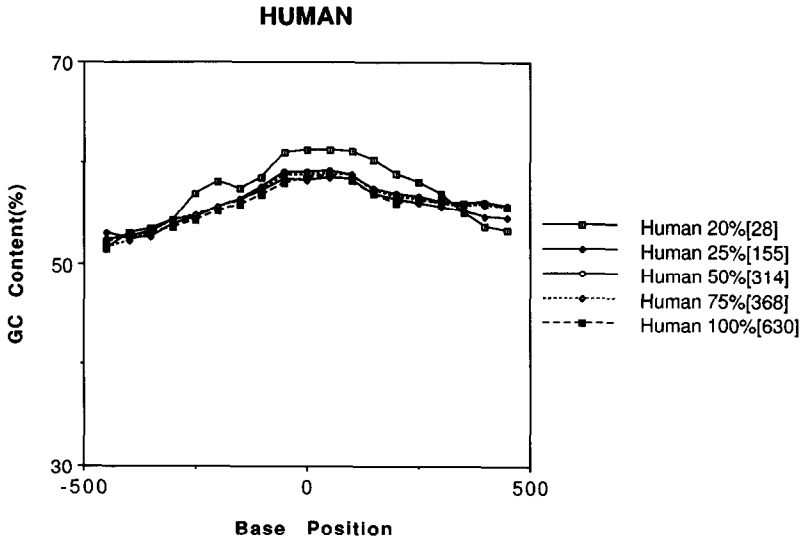

Fig. 1. The GC content profiles around the translation initiation site in *Homo sapiens* sequences with different threshold values for sequence homology in the data set. The average GC content in a 100 base window is plotted against the base position, where the center corresponds to the translation initiation site. The threshold is represented by the percentage of amino acid sequence homology, and the value in brackets is the number of sequences in the data set.

3. Results

The distribution profiles of the GC content around the translation initiation site in human sequences are shown in Fig. 1 with different threshold values for amino acid sequence homology. Except in the case of 20% homology threshold (no pair of sequences share more than 20% homology), where the data set size may have been too small, the profiles are virtually identical. This was also true for all other data sets shown in Table 1, except the profiles for chicken and *Xenopus* which depended somewhat on the choice of threshold values. Again, the data sets for these species may have been too small. Since the choice of the homology threshold did not affect the GC profile results much, as reported here, we present only the results with no homology threshold.

Fig. 2 shows the GC content profiles around the translation initiation site in the six organism groups. Fig. 2a and b show the profiles for the mammalian and other vertebrate groups, respectively, which are more-or-less flat, although there are breaks observed around the translation initiation site. In contrast, there are marked increases in the GC content around the translation initiation site in lower organisms as shown in Fig. 2c–f. Furthermore, although the average GC contents are fairly different in different species, the patterns of relative changes from the upstream to the downstream regions are similar in the same organism group. For example, the four profiles for the

Table 1
The number of sequences in the data sets when different homology thresholds were used, together with the average and standard deviation of the GC content for the 100% data set

Organism group	Genus (species)	Homology thresholds					Average	
		20%	25%	50%	75%	100%	GC content	S.D.
Mammals	*Homo sapiens*	28	155	314	368	630	56.0	9.7
	Mus	20	84	192	229	362	52.6	8.5
	Rattus	22	86	147	164	202	53.0	8.2
	Bos	19	27	32	36	56	52.4	9.7
	Oryctolagus	7	15	20	22	37	54.5	9.5
Other vertebrates	*Gallus gallus*	28	44	62	69	121	57.3	10.4
	Xenopus	12	19	24	27	42	43.2	5.8
Invertebrates	*Drosophila*	67	148	204	219	324	47.8	6.2
	Caenorhabditis	28	37	52	64	74	40.2	5.1
	Plasmodium	27	42	55	72	131	29.9	8.7
Plants	*Arabidopsis thaliana*	33	52	59	64	81	40.5	4.6
	Zea	25	32	35	38	64	55.3	10.1
	Lycopersicon	17	24	33	35	50	36.3	5.9
	Oryza	15	27	29	34	47	51.0	8.5
Fungi	*Saccharomyces*	155	598	795	830	1064	39.2	3.4
	Dictyostelium	18	34	39	40	53	24.3	7.5
Bacteria	*Escherichia*	79	452	724	746	992	48.9	5.3
	Bacillus	80	200	284	325	456	40.4	6.3
	Salmonella	58	116	129	131	158	50.2	5.4
	Pseudomonas	45	91	102	104	115	59.7	6.7

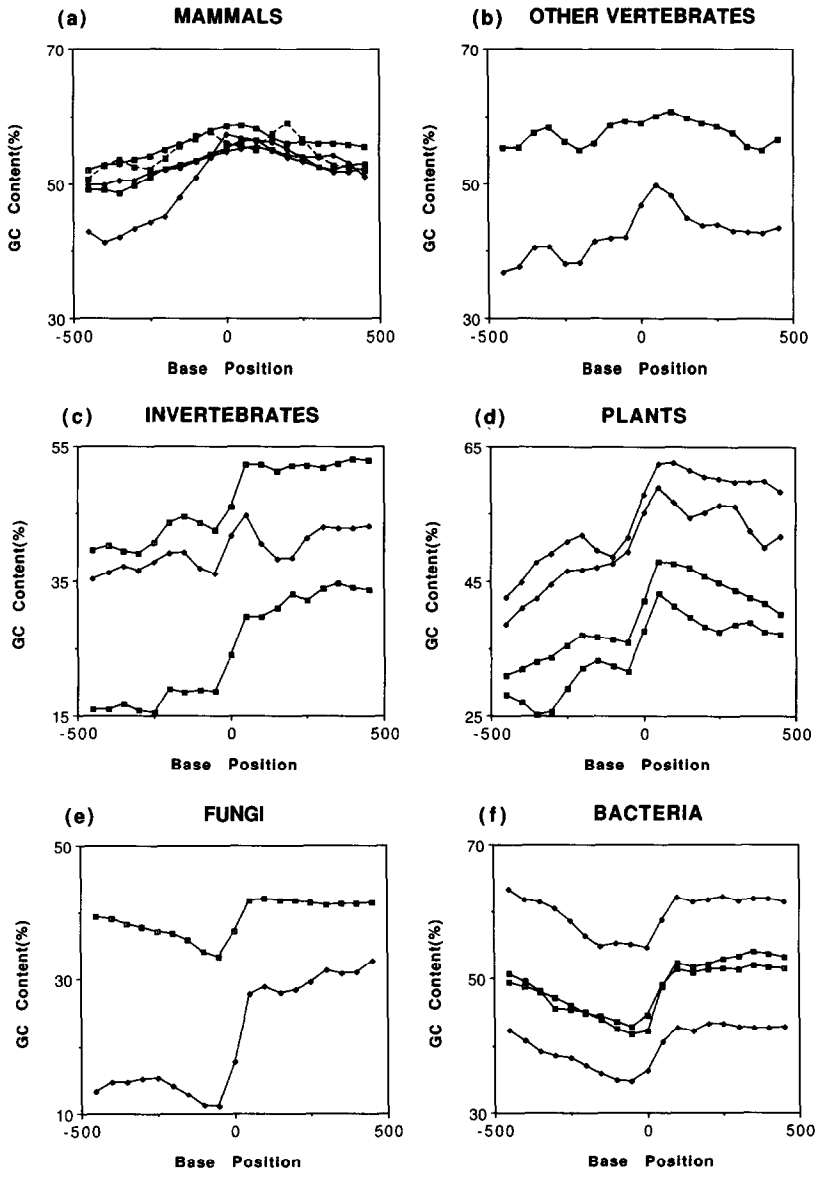

Fig. 2. The GC content profiles around the translation initiation site in different species. Each genus or species is classified into one of the six organism groups: (a) mammals, (b) other vertebrates, (c) invertebrates, (d) plants, (e) fungi, and (f) bacteria. The actual data sets are: (a) *Homo sapiens, Oryctolagus, Mus, Rattus*, and *Bos*; (b) *Gallus gallus* and *Xenopus*; (c) *Drosophila, Caenorhabditis*, and *Plasmodium*; (d) *Zea, Oryza, Arabidopsis thaliana*, and *Lycopersicon*; (e) *Saccharomyces* and *Dictyostelium*; (f) *Pseudomonas, Salmonella, Escherichia*, and *Bacillus*; according to the order from top at the left end of each curve. The data sets used here correspond to the 100% homology threshold (no threshold) in Table 1.

plants in Fig. 2d or the bacteria in Fig. 2f may be superimposed by vertical movement.

For invertebrates, plants, and fungi in Fig. 2c–e, a sharp increase in the GC content appears at the 100 base window from −50 to +50 (position 0 in Fig. 2) which corresponds to the middle point in the sigmoidal curve. However, for bacteria the curve is shifted somewhat towards the downstream region in Fig. 2f; the window from +50 to +150 seems to be at the center of the sigmoidal curve. Thus we analyzed bacterial sequences with a smaller window size. By moving a 10-base window with 5 bases overlapping, it was found that AT-rich regions existed between +1 and about +40 bases.

In the five mammals and four plants studied there is apparently a gradual decrease in the GC content in the downstream region between +1 and +500 (Fig. 2a and d). We suspected this was due to the presence of introns. In fact, when we excluded introns in the data sets the GC content profile became more-or-less flat. In the four bacteria and *Saccharomyces* there is also a gradual increase in the GC content towards the more upstream region (Fig. 2e and f. It is possible that this reflects the existence of other coding regions preceding the ones being considered. The intergenic length of these species is shorter than that of other species in Fig. 2. However, we could not confirm the actual existence of such genes, for no information was available in the GenBank database.

The different GC content profiles in different species shown in Fig. 2 were then examined by two parameters: the average GC content over the 1000 base range and the jump size at the translation initiation site. The characteristics of each profile represented by the two parameters is shown in Fig. 3, which

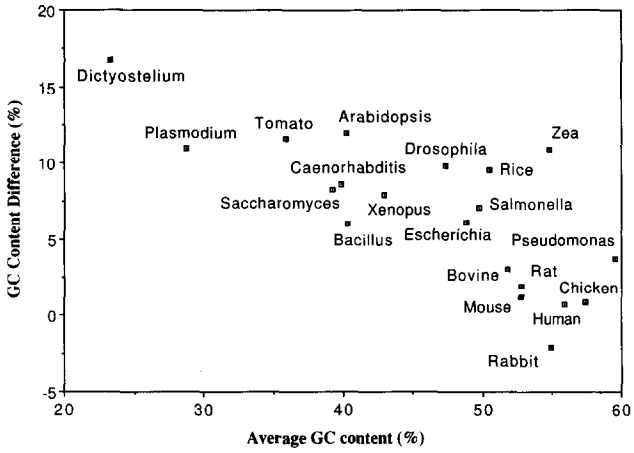

Fig. 3. The correlation between the average GC content and the difference in the GC content around the translation initiation site. The average GC content was calculated by averaging −500 to +500 base positions and the differences by the two 100 base windows immediately upstream and downstream of the translation initiation site. The correlation coefficient was −0.77.

indicates the negative correlation between these parameters: the correlation coefficient was −0.77. When the genome becomes less GC rich (more AT rich), the jump of the GC content at the translation initiation site increases. Namely, the difference in the GC content between the coding region and the preceding non-coding region becomes more distinguished as the genome becomes more AT rich.

4. Discussion

Our results confirm, first of all, previous observations that coding regions are generally GC rich. Furthermore, our results provide a more detailed picture of the local GC profile change between non-coding and coding regions, which is apparently different for different species. We observed more marked changes in lower organisms than in mammals and other vertebrates (Fig. 2). In contrast, Ikemura and Aota [5] observed global variations of GC content in vertebrate genomes, but not in *E. coli* and *S. cerevisiae* genomes. Thus, although the vertebrate genome may be a mosaic of GC rich and AT rich regions at the macroscopic level, the local variation is much smaller at the microscopic level. Conversely, although *E. coli* and *S. cerevisiae* genomes may be homogeneous at the macroscopic level, they show mosaic structures of GC-rich coding regions and AT-rich non-coding regions at the microscopic level. These differences may reflect different control mechanisms for gene expression; possibly, as the genome becomes larger the control mechanism may also involve longer-range interactions.

If such macroscopic variations are inherent in vertebrate sequences then there should be more variations of the GC content in the data set. The last two columns of Table 1 show the average and the standard deviation of the GC content in each of the 20 data sets. The GC contents of individual sequences can vary widely, but the variation is, in fact, larger in mammals than bacteria and yeast, for example. The variation is also larger in monocots (*Zea* and *Oryza*) than dicots (*Arabidopsis* and *Lycopersicon*), which is consistent with the isochore organization. When the data in Table 1 were plotted, we could see a weak positive correlation (correlation coefficient of 0.44) between the average GC content and the standard deviation. Similarly, there was a weak negative correlation (correlation coefficient of −0.44) between the GC standard deviation (macroscopic variation) and the GC jump size at the translation initiation site (microscopic variation).

Next, we divided human, *Zea*, and *Arabidopsis* sequences into two groups, relatively AT-rich and GC-rich ones according to the average value for each species, and calculated the GC content profiles for both groups. The overall shapes of the GC content profiles of the two groups were similar to those shown in Fig. 2 for the respective species. It appears that one gene can be shifted to AT rich or GC rich while retaining the relative GC content profile characteristic of the species. It is the relative GC content profile that is conserved rather than the absolute values of the GC content.

In GenBank, homologous sequences may be registered more frequently in the same organism group than in different groups. Thereby, the GC content profiles might be similar in the same organism group. However, when *Escherichia* sequences without any sequence homologies to *Bacillus* sequences were selected, they demonstrated almost the same GC content profile as in Fig. 2. Thus, the GC content profile seems to reflect a common characteristic of the same organism group irrespective of sequence homology. Again, this suggests a common mechanism of gene expression.

In conclusion, the relative shape of the GC content profile surrounding the translation initiation site is roughly invariant in the same organism group even though the average GC contents can be very different in different species and individual sequences. The local GC variation we observe here becomes smaller as the genome size becomes larger and as the global GC variation at a macroscopic level becomes more significant. The local GC variation is also dependent on the average GC content; as the genome becomes more AT rich, the difference in the GC content becomes more significant for coding and non-coding regions.

Acknowledgements: We are grateful for helpful comments from K. Nakai, A. Ogiwara, and Y. Akiyama. This work was supported by a Grant-in-Aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, the Institute for Chemical Resarch, Kyoto University.

References

[1] Comings, D.E. (1978) Annu. Rev. Genet. 12, 25–46.
[2] Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science 228, 953–958.
[3] Bernardi, G. and Bernardi, G. (1986) J. Mol. Evol. 24, 1–11.
[4] Salinas, J., Matassi, G., Montero, L.M. and Bernardi, G. (1988) Nucleic Acids Res. 16, 4269–4285.
[5] Ikemura, T. and Aota, S. (1988) J. Mol. Biol. 203, 1–13.
[6] Kanehisa, M.I. and Goad, W.B. (1982) Nucleic Acids Res. 10, 265–278.
[7] Aota, S. and Ikemura, T. (1986) Nucleic Acids Res. 14, 6345–6355 and 8702.
[8] Ikemura, T. and Wada, K. (1991) Nucleic Acids Res. 19, 4333–4339.