Volume 221, number 1, 3-10 FEB 05006 August 1987

Co-evolution of base composition and codon usage in *Xenopus laevis* and human globin genes with long-range DNA organization of their genome

J. Moreau and K. Scherrer

Institut Jacques Monod, 2, Place Jussieu, 75251 Paris Cedex 05, France

Received 21 May 1987

Eucaryotic DNA is punctuated by many A+T-rich segments that we named A+T-rich linkers. Two types of these A+T-rich linkers can be distinguished: (i) isolated A+T-rich linkers, and (ii) A+T-rich linkers crowded in clusters. We have analysed the distribution of A+T-rich linker across the α- and β-globin gene domain in *Xenopus laevis* and human genomes using isodenaturation and electron microscopy. Comparison of our data with those previously obtained for the avian globin genes leads us to conclude that genes can be harboured indifferently in either domain. A correlation is established between the presence of A+T-rich linker inside introns and flanking regions and the A+T content of the coding sequence. For the coding sequence, a high A+T content is strongly correlated with high A+T content in the codon's third position and weakly in the first position.

Globin gene, A + T-rich sequence; Electron microscopy; Codon usage

1. INTRODUCTION

A characteristic feature of the eukaryotic genome is its high adenine and thymine content (55–60% A + T). We have shown previously by partial denaturation of DNA and visualisation in the electron microscope that the distribution of A + T-rich sequences in the eukaryotic genome is not random but follows a particular pattern [1,2]. Two types of A + T-rich sequences can be described on the basis of their arrangement in DNA: either in long clusters ('A + T clusters') or in single elements isolated inside long stretches of undenatured DNA ('A + T linkers'). The short A + T-rich elements contain from 62 to more than 75% A + T, and their mean length is 800 ± 300 bp. The frequency of A + T-rich clusters and linkers varies in different parts of the genome. Genes may be localized either in A + T-rich domains where A + T clusters are frequent, or in more G + C-rich regions. For example, some avian proto-oncogenes, such as c-*mil* or c-*erb* B, are in A + T clusters [3], whereas the human proto-oncogenes c-*myc* and c-*sis [3] or the avian* α- and β-globin genes [2,4] are in rather G + C-rich domains. In the former arrangement, the A + T-rich sequences are distributed in long stretches framing the genes and are also found within intervening sequences; in the second, A + T linkers are most often located near the 5′- and 3′-ends of a given gene domain [2].

Little is known about the function of A + T-rich linkers in the genome, except for some genomic domains where correlations seem to exist with the organizational domain of gene families [2] and the position of matrix attachment sites [5,6]. To approach the possible role of A + T-rich sequences in the organization and function of the genome, we have systematically studied their arrangement in the vicinity of the globin genes in various

Correspondence address: J. Moreau, Institut J. Monod, 2, place Jussieu, 75251 Paris Cedex 05, France

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/87/$3.50

organisms. To complete the previously published data concerning the chicken [2] and duck genes [4], the globin gene domains of *Xenopus laevis* and human were studied.

The present results indicate differences between these different organisms with regard to the organization of the A + T-rich elements. Thus, the duck and chicken α- and α-globin genes, the human α-globin gene and one globin *X. laevis* domain are localized in domains of DNA relatively stable to denaturation, whereas the human β-globin genes, and the region containing the *X. laevis* α,β_1-globin genes are localized in A + T-rich domains. The results are discussed with regard to the relation between the average base content of long stretches of genomic DNA and the base composition of the gene and coding sequences.

2. MATERIALS AND METHODS

Partial denaturation was carried out as described [2,6]. Briefly, all cloned DNAs were partially denatured using the isodenaturation technique in 0.1 M Tris-HCl (pH 8.5)/10 mM EDTA, 80% (unless otherwise stated) formamide, at 24.5°C (±0.1°C). Cytochrome *c* (30 μg/ml) was added and the solution spread on a hypophase containing the same buffer diluted 1:10 with a formamide concentration 30% below that of the hyperphase. Molecules were analysed by visualisation in an electron microscope (Philips 410) and photographic negatives were measured by a digitizer connected to an HP 9820 calculator. Denaturation maps are the compilation of all results (each clone has been denatured in two or three different experiments). Histograms represent the analysis of at least 40–50 denatured molecules.

3. RESULTS

New data concerning the human species and *X. laevis* reported here will be discussed together with previously published data about the arrangement of A + T-rich DNA in the avian globin gene domains.

3.1. *The human globin genes*

In the human genome, the α- and β-globin genes are separated into two domains; one contains (in 5′ to 3′ order) the $\psi_{\alpha 1}$, α_2 and α_1 genes, and the second the ϵ, γ_G, γ_A, pseudo ψ_β, δ and β genes (fig.1). The DNA of each domain was available in several recombinant clones which were analysed by partial denaturation as described in section 2.

3.1.1. The α-globin gene domain

A subclone of λHαG2 containing the pseudo α_1, α_2 and α_1 genes was obtained by *Bam*H1 digestion of the λHαG2 clone [7]. This 13 kbp DNA fragment is very resistant to partial denaturation. In 90% formamide, no denaturation was observed (fig.1A). This result suggests that within the fragment studied, the α-globin gene domain does not contain A + T-rich sequences, i.e. no DNA segments longer than 50 bp and exceeding a 60% A + T content. The average distance between A + T-rich elements in eukaryotic DNA is of the order of 10–40 kbp [1]. It appears that in this case the domain framed by A + T-rich elements exceeds that of the immediate gene cluster.

3.1.2. The human β-globin gene domain

The human β-globin genes extend over 45 kbp of DNA, from the ϵ to the β gene. 85 kbp of DNA, including all β genes, were studied in a total of 6 recombinants.

Denaturation analysis was carried out on the DNA of three recombinant cosmid clones, HG3, HG15 and HG28 [8], each containing an average of 35 kbp of genomic DNA inserted in the pJB8 vector. The total domain analysed extended from 40 kbp upstream of the ϵ gene to 30 kbp downstream of the β gene.

In contrast to the stability of the human α-globin domain, all cosmid clones were found to be easily denatured in 80% formamide (fig.1C): 100% of the observed molecules showed denaturation extending over 50–60% of their length. Only the segment corresponding to the pJB8 vector DNA conserved full stability. A control carried out under the same conditions with pJB8 confirmed the absence of denaturation in the vector (not shown). Due to this extensive denaturation we did not attempt to localise each individual A + T-rich segment with precision; as examples, the electron micrographs of the denatured HG15 and HG28 cosmid DNA are shown in fig.1C.

Two λ recombinant clones, however, λHϵG1 and λHγG4 were analysed in detail. These (fig.1B) con-

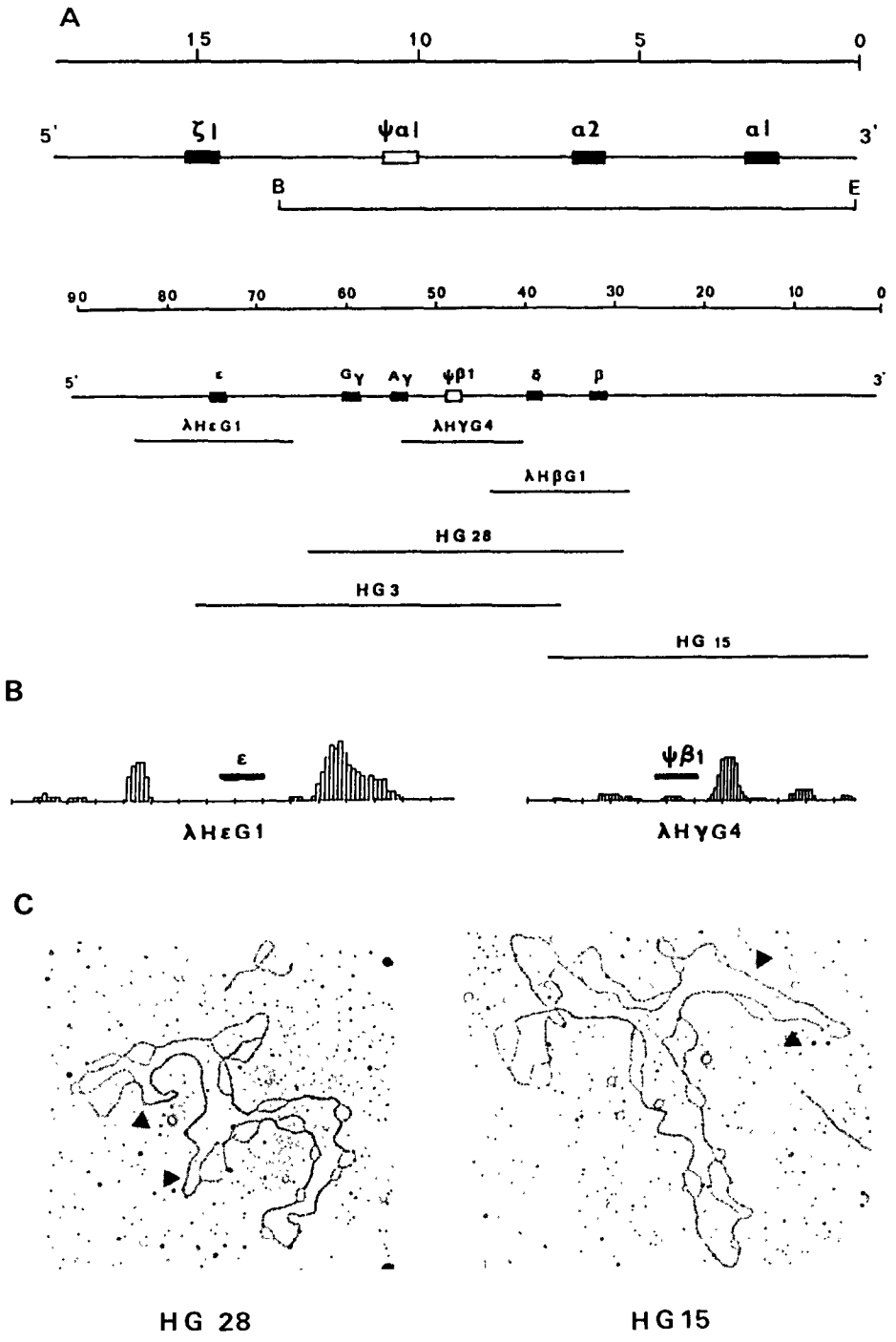

Fig.1. Denaturation map of the human α- and β-type globin gene cluster. (A) α- and β-type globin gene map. (B) Denaturation maps obtained in 75% formamide for the recombinants $\lambda H\epsilon G1$ and $\lambda H\gamma G4$. The abscissa gives the position of a given segment on the map and the ordinate the relative frequency of denaturation observed in that segment. (C) Electron micrographs showing the denaturation of two cosmid recombinant DNAs exposed to 80% formamide. Arrow shows the limit of the vectors.

tain the ϵ gene and the pseudo β gene, respectively [9]. In 80% formamide, the DNA inserts in these clones show the same behaviour as in the corresponding cosmid clones. In 75% formamide, the $\lambda H\epsilon G1$ clone displays three denatured regions, one localised 3.5 kbp upstream of the ϵ gene and two others, 3 and 4.5 kbp downstream. Similarly, in clone $\lambda H\gamma G4$ more than 40% of the molecules display a denaturation bubble in 75% formamide 2 kbp downstream of the pseudo β_1 gene. The $\lambda H\beta G1$ clone proved to be very unstable and thus a precise A + T map could not be obtained.

These results indicate a high A + T content of the human β-globin gene domain. This result is in agreement with sequence analysis [10] of a 16.5 kpb fragment including the δ and β genes (this fragment contains 62.6% A + T).

3.2. *The* X. laevis *globin gene domains*

We next studied a set of 11 overlapping recombinant DNA clones carrying the *X. laevis* globin genes, which are organised in three genomic domains [11] (fig.2).

3.2.1. The larval β_2-globin gene domain

The larval β_{2a} gene was analysed in a genomic 14 kb fragment cloned (.XG13) in a λ vector. The denaturation map obtained in 80% formamide shows only two characteristic denaturations (fig.2B), one approx. 500 bp downstream of the gene (present in 70% of the molecules) and another (present in 30% of the molecules) 4 kbp downstream. In addition, a very weak denaturation, which could be localised in the second intron of the gene, appears inside the gene in less than 10% of the molecules.

3.2.2. The genomic domain containing the α-globin gene cluster

The genomic domain including three α-globin genes (larval α_{2a}, larval α_{2b} and adult α_2), was analysed in three overlapping clones, XG10, XG175 and XG143. This region covers a total of 40 kb.

The thermodynamic behaviour of these three fragments is similar to that of the larval β_{2a} gene. In 80% formamide (fig.2A), a few denatured regions appeared. Four were present in 50–70% of the molecules, and were localised 7 kbp downstream of the larval α_{2a} gene, immediately before the larval α_{2b} gene; two other denaturations frame the adult α_2 gene. The respective distance between each denaturation (or cluster of denaturations) is 14 and 11 kbp. Within the genomic region a few denaturations are present in less than 40% of the molecules. One is localised just on the 3′-side of the larval α_{2b} gene which, like the adult α_2 gene, is framed by A + T-rich sequences.

3.2.3. The domain containing α_1- and β_1-globin genes

Nine overlapping recombinants, covering a DNA domain of about 70 kbp and containing the five genes, larval α_{1a} and α_{1b}, adult α_1, adult β_1 and larval β_{1a}, were analysed (XG66, -140, -82, -69, -38, -121, -4, -35 and -32). The denaturation map in 80% formamide (fig.2C) was much more complex than the other two domains analysed. Indeed, a great number of A + T-rich sequences were present. A large part of the DNA was denatured in 80–100% of the molecules analysed. Among the globin genes, the adult α_1 and β_{1a} genes are directly framed by A + T-rich sequences. In three other genes, the larval α_{1a}, larval α_{1b} and β_{1a}, one denatured region appeared within the gene. The precision of our denaturation map (± 50 bp at best) does not permit the identification of the intron containing the A + T-rich sequence. The instability is not homogeneously distributed througout the domain; the domain including the larval α_{1a} and up to 15 kbp of DNA downstream from the adult β_1 gene is the least stable to denaturation (in 75% formamide, the same denaturations were again observed). The second region extending downstream of this area is relatively stable although it is still less stable than the two other globin α_2 and β_2 domains of *X. laevis*.

3.3. *Relation between the average base composition of domain and its coding sequence*

Denaturation maps show a large disparity in average base composition within the genomic domains containing the globin genes. The DNA surrounding the globin genes can be divided into two classes according to thermodynamic behaviour: (i) quite stable regions, including the chicken α- and β-globin genes, the human α-globin genes and the *X. laevis* α_2-globin gene domain; (ii) unstable

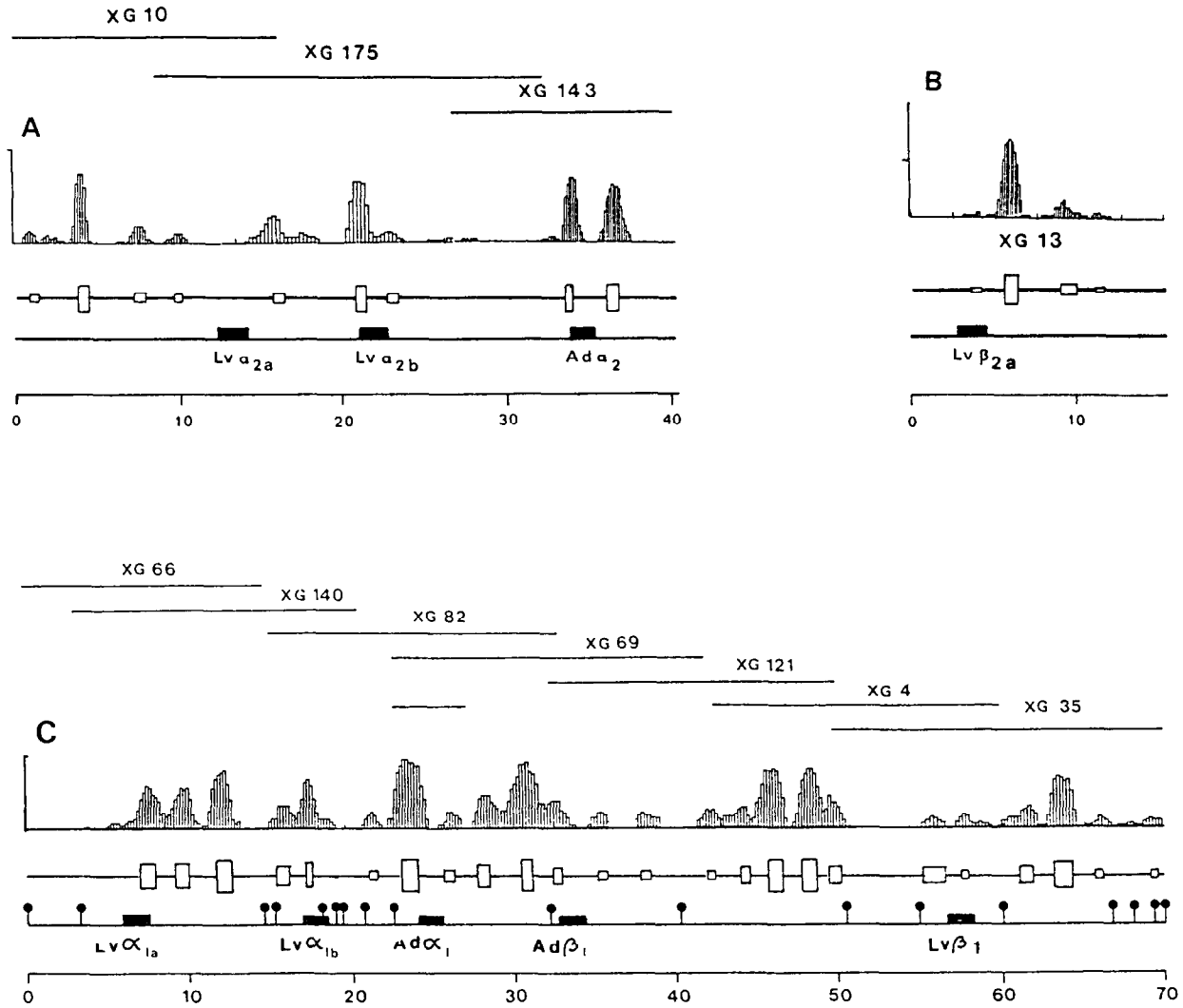

Fig.2. Denaturation map of the *X. laevis* globin gene clusters. For each domain are shown: on the upper line the *X. laevis* DNA insert in the set of overlapping clones; below, the histogram of denaturation obtained in 80% formamide. For (A,B) maximal height in the histogram of peaks corresponds to denatured regions present in 50% of the molecules, and for (C) the peak maxima correspond to denaturations present in 100% of the molecules studied. The third line shows the location of AT-rich sequences by boxes after internal alignment and position correction. On the last line, the position of globin genes is indicated by black boxes.

regions, such as the human β-globin gene domain and the *X. laevis* α_1- and β_1-globin gene domains.

These melting patterns apply to the extragenic sequences as well as to the intervening sequences. Thus, the introns of larval α_{1a}-, larval α_{1b}- and larval β_{1a}-globin genes of *X. laevis*, like the introns of human δ- and β-globin genes [10], are very A + T-rich. This observation can be compared with the correlation between the A + T content of introns and that of surrounding sequences in mosaic genes [3].

First, we analysed exon and intron sequences of 93 genes to define their relative A + T content. For the exons the results showed a single peak with a Gaussian distribution (fig.3) of modal value 45.5% A + T content. For the introns the distribution was

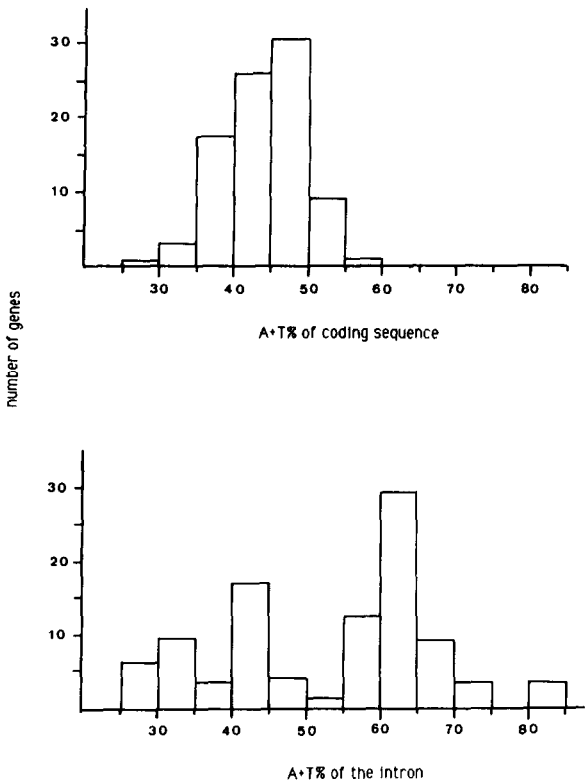

Fig.3. Histogram of the AT content of introns and exons in 93 genes of different organisms (globin family and genes for which the complete sequence, intron and exon, are known).

clearly bimodal, with one peak containing more than 55% A + T (modal value 62.5% A + T) and the other less than 50% A + T (modal value 40% A + T). We arbitrarily considered as an 'A + T-rich' coding sequence containing more than 45% A + T and introns containing more than 55% A + T.

Next, we determined whether the base composition of the coding sequences was influenced by the A + T content of the domain. Fig.4 shows the relation established between the base composition of exons and introns of the α- and β-globin genes, including all species for which there are sequence data. α and β genes are distributed in 2 classes, separated by borders that characterize the base distribution in the coding and intervening sequences. The α and β genes belong to distinct areas. The β-globin genes are located in the A + T-rich, and the α genes in the G + C-rich domain.

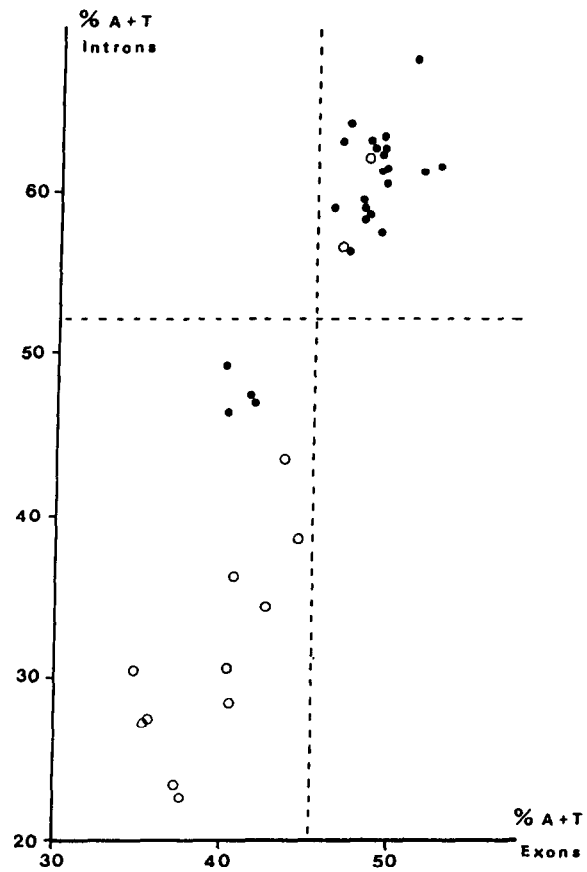

Fig.4. Plot relating the AT content of introns to that of exons: (○) α-globin genes; (●) β-globin genes. The vertical broken line indicates the arbitrary border between AT-rich and GC-rich exons. The horizontal broken line indicates the border between AT-rich and GC-rich introns.

Two exceptions are the avian β-globin genes and the avian π-globin genes which are G + C-rich. In both cases, however, the relation between A + T content of intron and exon persists.

According to the analysis shown, these data can be related to larger stretches of DNA framing the genes. Thus, not only does the base content of the introns seem to be dependent on that of the genomic environment, but also the base content of the gene sequence itself seems to be related to that of the domain in which it is located. Bernardi et al. [12] and Aota and Ikemura [13] have recently shown that the base composition of coding sequences corresponds to that of the flanking genomic region. This relation is principally at-

Table 1

A + T base composition of the three codon positions and of introns in globin genes for which the complete intron sequences are available

Species	Genes	% A + T			
		1	2	3	Introns
X. laevis	β-globin L	43.53	63.26	49.66	68.50
	β_{1a}-globin	42.75	60.0	48.96	66.31
Duck	α_A-globin	35.21	64.08	23.94	36.22
	α_D-globin	37.75	65.95	20.56	30.70
	π-globin	46.10	56.73	42.55	62.00
	β-globin	39.72	61.64	15.07	49.06
Chicken	α_A-globin	40.14	61.07	17.60	28.25
	π-globin	42.25	60.56	19.72	56.82
	β-globin	40.41	61.64	18.49	49.06
	ϵ-globin	45.58	61.90	17.68	46.62
	ϱ-globin	45.30	64.53	11.67	46.07
Rabbit	α-globin	41.96	60.14	13.98	30.00
	β-globin	36.73	65.98	34.69	63.00
Human	α-globin	36.00	57.88	10.58	27.50
	β-globin	32.65	62.58	31.97	62.50
	δ-globin	36.70	63.94	36.05	63.08
	ϵ-globin	41.89	63.51	33.78	64.08

tributable to the third nucleotide of codons. The same observation can be established for globin genes (table 1). This phenomenon is illustrated by the β-globin gene: in mammals the gene is located in an A + T-rich region and the third base in the codon is at least 31.9% A + T; in birds the gene is in a G + C-rich region and the third base is less than 18.5% A + T.

4. DISCUSSION

All eukaryotic genomes are heterogeneous in their base distribution. Analysis by denaturation and electron microscopy has allowed the observation of long-range organization in mosaic domains of the eukaryotic genome as distinguished by the abundance and distribution pattern of A + T-rich linkers or clusters. The precise localization of the A + T-rich elements relative to specific genes allows the analysis to be taken beyond that relating the base composition of genes and codon usage of domains in general. In our previous studies we have shown that genes may be found in both A + T-rich and G + C-rich domains [1,6,14]. In the case of mosaic genes it was shown that the thermodynamic behaviour and hence A + T content of intervening sequences correspond to that of the genomic domain in which the gene is located [3].

In agreement with Bernardi et al. [12] and Aota and Ikemura [13], our analysis of globin genes shows a parallel evolution of the coding sequences and their flanking genomic regions: this is principally attributable to the third codon position and to a lesser degree to the first position. This relation seems to be largely independent of the properties of a given coding sequence, but rather related to its placement in the genome.

We can extend this relation to the entire range of eukaryotic genomes, as well as the mitochondrial and prokaryotic genomes: the human [15] and *Dictyostelium discoideum* [16] genomes contain 56 and 78% A + T, respectively, and the third coding nucleotide 34.04 and 77.44% A + T; mitochondrial DNAs of human [17], mouse [18] and the dipterous *Drosophila yakuba* [19], contain 55.6, 63.2 and 75.8% A + T, respectively, and the third nucleotide in codons is 52.25, 69.73 and 93.8% A + T. In the case of bacteria, the overall base composition of genes has been shown to conform to codon usage [20]. Such values correlate well with data on the globin genes (table 1).

The pronounced base homogeneity throughout coding and non-coding sequences forces one to consider the mechanisms of drift in A + T content during evolution. Accepted processes affecting base composition include mutation and selection. However, these processes by themselves do not explain a spreading of the modification in base composition to domains such as the *X. laevis* globin genes. An additional mechanism may be proposed for which DNA repair efficiency during DNA replication varies according to nucleotide composition.

In conclusion, the base composition of genes, within the limitation of coding function, does not seem to depend on individual genes, but rather on properties of large genomic domains. The example of chromosome banding [21] shows a possible correlation between A + T-rich domains [13] and functional properties such as replication [22].

ACKNOWLEDGEMENTS

The authors are grateful to Drs R. Weber and J.

Stalder for their gift of recombinant DNA containing the globin genes of *X. laevis*, and to Dr R.A. Flavell for the gift of a clone containing human globin genes. We thank J.A. Lepesant and P. Brooks for helpful discussions concerning the preparation of the manuscript. Sequence data were treated using computer facilities at CITI2 in Paris on a PDP8 computer with the help of the French Ministère de la Recherche et de l'Enseignement Supérieur (Programme Mobilisateur 'Essor des Biotechnologies'). This investigation was supported by the French CNRS, the INSERM and the Association pour la Recherche sur le Cancer.

REFERENCES

[1] Moreau, J., Matyash-Smirniaguina, L. and Scherrer, K. (1981) Proc. Natl. Acad. Sci. USA 78, 1341–1345.
[2] Moreau, J., Marcaud, L., Maschat, F., Kejzlarova-Lepesant, J., Lepesant, J.-A. and Scherrer, K. (1982) Nature 295, 260–262.
[3] Moreau, J. (1986) Thesis, Université Paris VII.
[4] Kretsovali, A., Marcaud, L., Moreau, J. and Scherrer, K. (1986) Mol. Gen. Genet. 203, 193–201.
[5] Mircovitch, J., Mirault, M.-E. and Laemmli, U.K. (1984) Cell 39, 223–232.
[6] Moreau, J., Kejzlarova-Lepesant, J., Brock, H., Lepesant, J.-A. and Scherrer, K. (1985) Mol. Gen. Genet. 199, 357–364.
[7] Lauer, J., Shen, C.-K.J. and Maniatis, T. (1980) Cell 20, 119–130.
[8] Grosveld, F.G., Dahl, H.H., De Boer, E. and Flavell, R.A. (1981) Gene 13, 227–237.
[9] Fritsch, E.F., Lawn, R.M. and Maniatis, T. (1980) Cell 19, 959–972.
[10] Poncz, M., Schwartz, E., Ballantine, M. and Surrey, S. (1983) J. Biol. Chem. 258, 11599–11609.
[11] Hosbach, H.A., Wyler, T. and Weber, R. (1983) Cell 32, 45–53.
[12] Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salina, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science 228, 953–958.
[13] Aota, S.-I. and Ikemura, T. (1986) Nucleic Acids Res. 14, 6345–6356.
[14] Scherrer, K. and Moreau, J. (1985) Proc. 16th FEBS Congress, VNU Science Press, pp. 105–122.
[15] Chen, H.R. and Barker, W.C. (1985) Trends Genet. 1, 221–223.
[16] Kimmel, A.R. and Firtel, R.A. (1983) Nucleic Acids Res. 11, 541–552.
[17] Anderson, S., Bankier, A.T., Barrell, B.G., De Bruijn, M.H.-L., Coulson, A.R., Drouin, J., Eperon, C.I., Nierlich, D.P., Roe, B.A., Sanger, F., Schrier, P.M., Smith, A.J.M., Staden, R. and Young, I.G. (1981) Nature 290, 457–465.
[18] Bibb, M.J., Van Etten, R.A., Wrigt, C.T., Walberg, M.W. and Glayton, D.A. (1981) Cell 26, 167–180.
[19] Clary, D.O. and Wolstenholme (1985) J. Mol. Evol. 22, 252–271.
[20] Bibb, M.J., Findlay, P.R. and Johnson, M.W. (1984) Gene 30, 157–166.
[21] Comings, D.E. and Drets, M.E. (1976) Chromosoma 56, 199–211.
[22] Holmquist, G., Gray, M., Porter, T. and Jordan, J. (1982) Cell 31, 121–129.