J Mol Evol (1999) 49:325–329

JOURNAL OF MOLECULAR EVOLUTION

Compositional Correlations in the Chicken Genome

Héctor Musto,[1,2] Héctor Romero,[1,2] Alejandro Zavala,[1] Giorgio Bernardi[3]

[1] Laboratorio de Organización y Evolución del Genoma, Sección Bioquímica, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
[2] Departamento de Genética, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay
[3] Stazione Zoologica Anton Dohrn, Laboratorio di Evoluzione Molecolare, Villa Comunale, 80121 Napoli, Italy

Received: 9 December 1998 / Accepted: 18 April 1999

Abstract. This paper analyses the compositional correlations that hold in the chicken genome. Significant linear correlations were found among the regions studied—coding sequences (and their first, second, and third codon positions), flanking regions (5′ and 3′), and introns—as is the case in the human genome. We found that these compositional correlations are not limited to global GC levels but even extend to individual bases. Furthermore, an analysis of 1037 coding sequences has confirmed a correlation among GC_3, GC_2, and GC_1. The implications of these results are discussed.

Key words: *Gallus gallus* — Isochores — Genome organization — Genome evolution

Introduction

The genomes of vertebrates are mosaics of isochores, which are long, compositionally homogeneous DNA segments (in the range between 3 and 300 kb) and belong to a small number of families characterized by different GC levels covering a 30–60% range (GC is the percentage of guanine + cytosine in a given genome, sequence, or codon position) (Bernardi 1995). Among the different classes of vertebrates, different compositional patterns are recognized. Briefly, cold-blooded vertebrates are characterized by low intermolecular compositional heterogeneities and CsCl band asymmetries and generally display only GC-poor isochores (Hudson et al. 1980, Bernardi and Bernardi 1990a, 1990b). On the other hand, the genomes of warm-blooded vertebrates are by far more heterogeneous and reach very high values in GC levels, displaying besides two GC-poor isochore families (L1 and L2, which are similar in buoyant density with the isochores representing the vast majority of the genomes of cold-blooded species), some GC-rich, and very GC-rich isochores, the H1, H2 and H3 families of isochores. It is important to note, however, that the compositional patterns of all warm-blooded vertebrates are not identical.

Within mammals, a general pattern, characterized by large amounts of very GC-rich (>50%) isochores was found in eight out of nine mammalian orders studied (Sabeur et al. 1993), and hence is thought to be the most common within this class. The isochore families defining this pattern are L1, L2, H1, H2 and H3. The human genome is the most studied representative of this general pattern (Bernardi 1995, Clay et al. 1996). Several special patterns were described in some infraorders or families of the orders Rodents, Chiropters and Insectivores, as well as in the species Pangolin. Although different among them, they all have in common the lack of the GC-richest isochores present in the general pattern, and the isochore families making up these genomes are L1, L2, H1 and H2 (Salinas et al. 1986, Mouchiroud et al. 1988, Sabeur et al. 1993, Cacciò et al. 1994). The best characterized species displaying these patterns are mouse and rat. For a review of the evolutionatry history of these patterns (see Bernardi 1995). In sharp contrast with the differences just described within mammals, the available

Correspondence to: G. Bernardi; *e-mail:* bernardi@alpha.szn.it

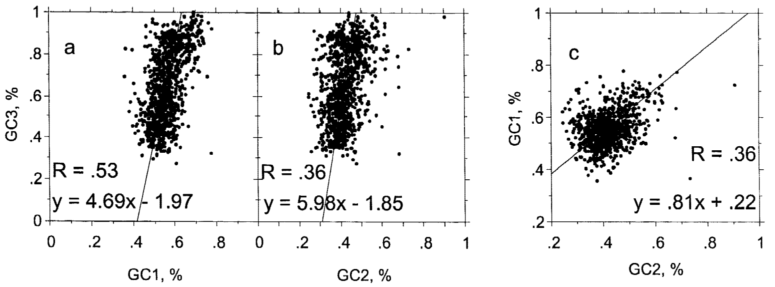

Fig. 1. Plots of GC levels of third versus first **(a)** and second **(b)** and first versus second **(c)** codon positions of 1037 chicken genes. The correlation coefficients and equations of the orthogonal regression lines of each plot are given.

data suggest that, within avian orders, the compositional patterns are almost-identical and are characterized by a compositional distribution of the DNA molecules very similar to that of mammals, but displaying larger amounts of very GC-rich isochores (Cortadas et al. 1979; Olofsson and Bernardi 1983; Kadi et al. 1993).

In the case of the human genome, compositional correlations exist between coding (translated) sequences (and their codon positions) and the isochores or flanking sequences in which each gene is located, as well as between exons and the corresponding introns. These findings are not trivial, since translated sequences in the human genome constitute less than 5% of the whole genome, while the rest is noncoding DNA, and hence imply that compositional constraints operate in the same direction (increasing or decreasing GC) on exons, on introns, and on the isochores surrounding each gene. It is important to note that these correlations are very well established since they have been detected both experimentally by compositional mapping and by the analysis of DNA sequence data (see, e.g., Bernardi et al. 1985; Aota and Ikemura 1986; Clay et al. 1996; reviewed by Bernardi 1995).

From the point of view of the evolution of the genomes of warm-blooded vertebrates, it is clear that verifying whether these compositional correlations hold even within the avian genome is an important goal, since the compositional transitions from GC-poor isochores of cold-blooded vertebrates leading to the GC-rich isochores present only in warm-blooded vertebrates occurred twice independently during evolution. Indeed, it is well established that mammals arose about 220 Mya from therapsids and birds about 150 Mya from dinosaurs. The aim of this paper is to analyze this problem taking the chicken genome as a model.

Materials and Methods

Sequences were from GenBank (Release 109.0, October 1998). Only complete genes (i.e., including initiation and stop codons) were analyzed. To study flanking regions and introns, genomic sequences were preferred over cDNAs. A total of 1037 nonredundant genes was analyzed. To eliminate putative regulatory regions, flanking sequences (5′ and 3′) were defined as the DNA segments over 500 bp upstream or downstream from initiation or stop codons, respectively. The mean size of flanking regions was 3765 bp (minimum, 1310 bp; maximum, 8030 bp).

Results and Discussion

Compositional Correlations Among Codon Positions

Figures 1a and b display scatterplots of the GC levels of third codon positions (GC_3) versus the GC levels of first (GC_1) and second (GC_2) codon positions, respectively, for a collection of 1037 coding sequences. Although the variation of GC_3 is much larger than those for the other codon positions, covering a range of 72.4%, the plots also show strong and significant correlations. Indeed, the *R* values are 0.53 and 0.36, respectively, and the levels of significance are $p < 0.0001$ in both cases. Figure 1c shows the plot of GC_1 versus GC_2. The *R* value is 0.36, and again, the correlation is highly significant ($p < 0.0001$). The slopes of the orthogonal regression lines are similar to those found for human genes (D'Onofrio et al. 1999), the main difference being that the slope is slightly higher for chicken genes in the GC_3 vs GC_1 plot (4.69 vs 3.91). Similar results have been reported previously by Sueoka (1992).

These results lead to two conclusions. First, the correlations among codon positions are likely to be detected within any compositionally heterogeneous genome. Second, compositional constraints in the chicken genome, as in the mammalian genome, are working in the same direction over the three codon positions, although not with the same amplitude, being higher on third than on first and, finally, on second codon positions.

Compositional Correlations of Coding Sequences with Flanking Regions and Introns

Figure 2a shows a plot of GC levels of coding sequences versus the GC levels of the flanking regions (5′ + 3′). To eliminate the influence of regulatory regions we analyzed only from base 500 upstream or downstream from initiation or stop codons, respectively. Although this approach reduced the sample studies, the correlation is highly significant ($p < 0.0001$), and hence it can be concluded that there is a match between the GC levels of exons and the GC levels of the flanking regions. It is important to note that in the human genome, the slope of the orthogonal regression line of the equivalent plot is almost the same (Fig. 1b of Clay et al. 1996), even if, in the latter case, the number and length of sequences analyzed are higher. This indicates that the relation that holds between coding sequences and flanking regions is

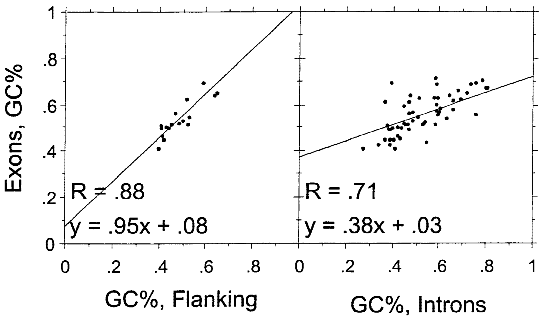

Fig. 2. Plots of GC levels of coding sequences versus those of the corresponding flanking sequences **(a)** and introns **(b)**. The correlation coefficients and equations are shown as in Fig. 1.

constant in the compositionally heterogeneous genomes of vertebrates. Figure 2b displays the plot of GC levels of coding sequences versus the GC levels of the corresponding introns. Even considering the relative scarcity of sequences available, it is again clear that the compositions of both translated and nontranslated regions are strongly correlated. Very similar results have been reported within the human genome (Bernardi 1995; Clay et al. 1996). However, in this case the slope is lowest in the human genome, which is probably caused by the fact that introns within the chicken genome reach higher GC levels, while GC levels of exons reach approximately the same GC levels. Finally, the GC levels of first and second codon positions (GC_1 and GC_2) are both significantly correlated with the GC levels of the corresponding introns (not shown): indeed, the *R* value of the former correlation is 0.53 ($p < 0.0001$), while that for the latter is $R = 0.33$ ($p < 0.02$).

These compositional correlations are not limited to global GC levels but even extend to individual bases. For example, Fig. 3 shows plots of the frequencies of each base in coding regions versus the same base in flanking sequences. In all cases, there are positive and highly significant correlations ($p < 0.001$). An identical situation is found when the frequency of each base is compared between exons and introns (not shown).

Compositional Correlations of GC_3 with Flanking Regions and Introns

Figure 4a displays a plot of GC_3 levels versus the GC levels of the flanking regions (as defined previously). The value of the slope of the equation of the orthogonal regression line is very close to 2.31, which is the value obtained in the human genome (Clay et al. 1996). This correlation is important because it can be used for defining the localization and distribution of genes within the different isochores, as was done for the human genome (Bernardi et al. 1985; Bernardi 1995; Clay et al. 1996). As is the case in mammals, chicken genes exhibit different concentrations in the different isochores (Caccio et al., 1994; McQueen et al., 1996, 1998). Although it can be argured that the flanking regions analyzed here are rather short (mean, 3900 bp), it is important to note that when genes with flanking regions longer than 7000 bp were analyzed, the results were the same (not shown). Furthermore, in these longer sequences, the GC levels ranged from 36 to 53%, which is similar to the GC levels of the isochores in the chicken genome (Olofsson and Bernardi 1983).

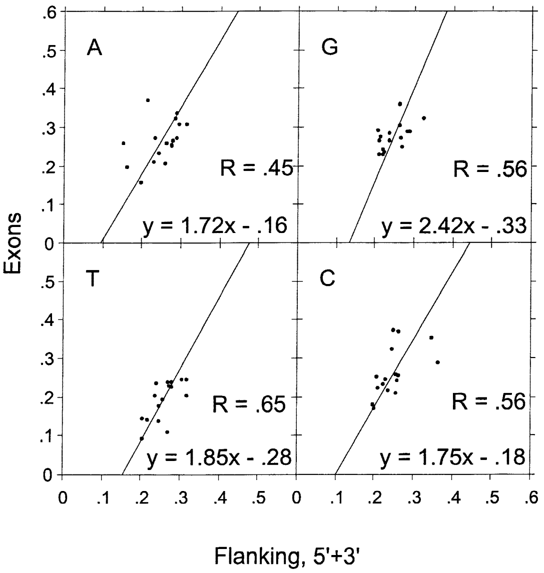

Fig. 3. Plots of the frequencies of each base in flanking regions versus those in coding sequences. The correlation coefficients and equations are shown as in Fig. 1.

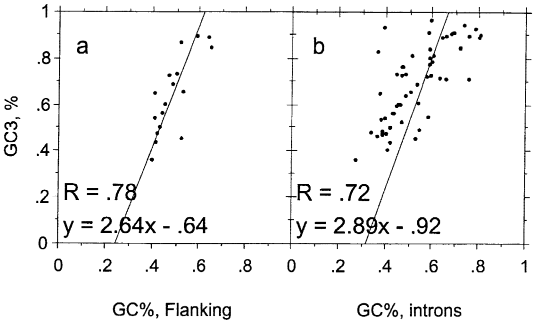

Fig. 4. Plots of GC levels of third codon positions against the GC levels of the corresponding flanking regions **(a)** and introns **(b)**. The correlation coefficients and equations are shown as in Fig. 1.

Figure 4b shows the correlations of GC levels of third codon positions versus the GC levels of the corresponding introns. As expected, there is a very significant correlation which demonstrates that the compositional constraints acting on GC_3 from different genes work in the same direction and with the same amplitude on the corresponding introns.

Compositional Correlations of Noncoding Regions

Figure 5 displays the compositional correlations between the GC levels of noncoding DNA, i.e., introns vs flanking regions. In Fig. 5a the analysis was done on the complete flanking regions, and a strong correlation was

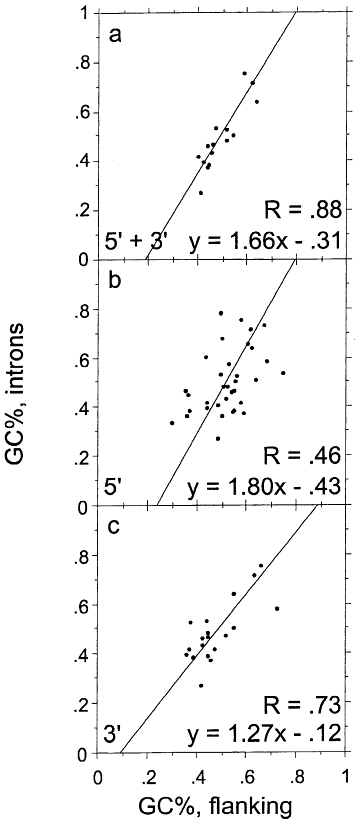

Fig. 5. Plots of GC levels of introns against the corresponding GC levels of flanking (5′ + 3′) regions **(a)**, 5′ alone **(b)**, and 3′ alone **(c)**. The correlation coefficients and equations are shown as in Fig. 1.

found (0.88; $p < 0.0001$). This correlation is not the result of a bias of either 5′ or 3′ regions (see Figs. 5b and c, respectively), since the GC levels of separate sequences are correlated with the GC levels of introns. In the human genome the slopes of the orthogonal regression lines of the same correlations are not as similar as for other plots reported here (1.22 and 1.18 for introns, vs 5′ and 3′, respectively). Two (not mutually exclusive) explanations might account for this. First, as mentioned previously, the GC levels of introns in the chicken genome reach higher values than in the human genome, and second, the data set that we analyzed is rather small given the scarcity of long genomic sequences. As expected, the GC levels of 5′ and 3′ flanking regions are significantly correlated between them ($R = 0.58$, $p < 0.02$).

Finally, Figure 6 shows that the compositional correlations just mentioned extend to the individual bases frequencies (as in the case in the correlation between exons and flanking regions; Fig. 3).

Conclusions

We analyzed the compositional correlations that hold in the chicken genome between coding sequences (and their different codon positions), and noncoding DNA (as a whole or 5′ flanking, 3′ flanking, and introns). This should help in understanding how compositional constraints operate in the avian genome, which is characterized by a higher compositional heterogeneity compared to mammalian genomes, as shown by the comparison of the compositional patterns of birds and mammals, at both the DNA and the GC_3 levels (Bernardi 1995).

We report that the same correlations that hold in the human genome hold in the chicken genome and, remarkably, with similar equations.

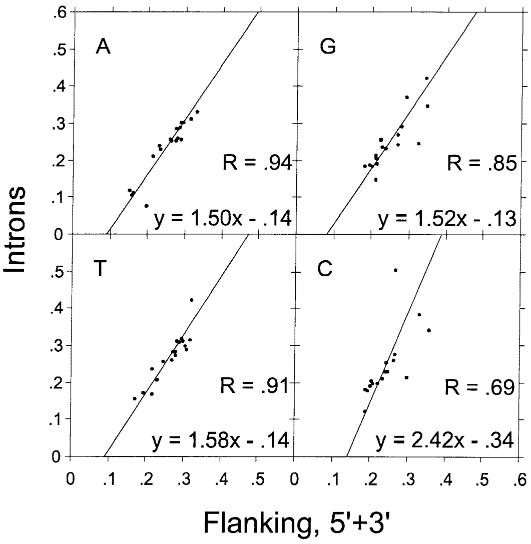

Fig. 6. Plots of the frequencies of each base in introns versus flanking regions. The correlation coefficients and equations are shown as in Fig. 1.

References

Aota S, 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

Bernardi G (1995) The human genome: Organization and evolutionary history. Annu Rev Genet 29:445–476

Bernardi G, Bernardi G (1990a) Compositional patterns in the nuclear genomes of cold-blooded vertebrates. J Mol Evol 31:265–281

Bernardi G, Bernardi G (1990b) Compositional transitions in the nuclear genomes of cold-blooded vertebrates. J Mol Evol 31:282–293

Bernardi G, Olofsson B, Filipski J, et al. (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Cacciò S, Perani P, Saccone S, Kadi F, Bernardi G (1994) Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates. J Mol Evol 39:331–339

Clay O, Cacciò S, Zoubak S, Mouchiroud D, Bernardi G (1996) Human coding and noncoding DNA: Compositional correlations. Mol Phylogenet Evol 5:2–12

Cortadas J, Olofsson B, Meunier-Rotival M, Macaya G, Bernardi G, (1979) Eur J Biochem

D'Onofrio G, Jabbari K, Musto H, Bernardi G (1999) The correlations of protein hydropathy with the composition of coding sequences. Gene (in press)

Hudson AP, Cuny G, Cortadas H, Haschemeyer A, Bernardi G (1980) An analysis of fish genomes by density gradient centrifugation. Eur J Biochem 112:203–210

Kadi F, Mouchiroud D, Sabeur G, Bernardi G (1993) The compositional patterns of the avian genomes and their evolutionary implications. J Mol Evol 37:544–541

McQueen HA, Fantes J, Cross SH, Clark VH, Archibald AL, Bird AP (1996) CpG islands of chicken are concentrated on microchromosomes. Nature Genet 12:321–324

McQueen HA, Siriaco G, Bird AP (1998) Chicken microchromosomes are hyperacetylated, early replicating, and gene rich. Genome Res 8:621–630

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in humans and murids. J Mol Evol 27:311–320

Olofsson B, Bernardi G (1983) Organization of nucleotide sequences in the chicken genome. Eur J Biochem 130:241–245

Sabeur G, Macaya G, Kadi F, Bernardi G (1993) The isochore patterns of mammalian genomes and their phylogenetic implications. J Mol Evol 37:93–108

Salinas J, Zerial M, Filipski J, Bernardi G (1986) Gene distribution and nucleotide sequence organization in the mouse genome. Eur J Biochem 160:469–478

Sueoka N (1992) Directional mutation pressure, selective constraints and genetic equilibria. J Mol Evol 34:95–114