Journal of Biochemistry and Molecular Biology, Vol. 39, No. 3, May 2006, pp. 240-246

JOURNAL OF Biochemistry and Molecular Biology

Compositional Correlations in Canine Genome Reflects Similarity with Human Genes

Faustin Joy[1], Surajit Basak[2,*], Sanjib Kumar Gupta[2], Pranab Jyoti Das[1], Shankar Kumar Ghosh[1] and Tapash Chandra Ghosh[2]

[1]Department of Animal Genetics and Breeding, West Bengal University of Animal & Fishery Sciences, 37 & 68, Kshudiram Bose Sarani, Kolkata-700037, India

[2]Bioinformatics Centre, Bose Institute, P1/12 CIT Scheme VII M, Kolkata-700054, India

Received 28 May 2005, Accepted 17 Jan 2006

The base compositional correlations that hold among various coding and noncoding regions of the canine genome have been analysed. The distribution pattern of genes, on the basis of GC_3 composition, shows a wide range similar to that observed in human. However the occurrence of maximum number of genes was observed in the range of 65-75% of GC_3 composition. The correlation between the coding DNA sequences of canine with the different non-coding regions (introns and flanking regions) is found to be significant and in many cases the degree of correlation show similarity to human genome. We found that these correlations are not limited to the GC content alone, but is holding at the level of the frequency of individual bases as well. The present study suggests that canines ideally belong to the predicted 'general mammalian pattern' of genome composition along with human beings.

Keywords: Canine genome, Compositional correlations, Exon, Flanking region, GC content, Gene distribution, Intron

Introduction

Once the human genome project was over, scientists all over the world felt that comparing the human genome sequence with those of other organisms would help to find regions of similarity and dissimilarity, providing valuable clues about the structure and function of human genes. The decision to sequence the canine genome highlights the dogs evolutionary and physiological position between the mouse and human and its importance as a model for the study of mammalian genetics and human hereditary diseases (Greer *et al*., 2003). Dog genome project was initiated in June, 2003 at Whitehead Center/MIT Center for Genome Research in Cambridge, Massachusetts, and in July, 2004 the first draft of fully sequenced canine genome was presented to the research community worldwide.

Importance of canine genome lies in the fact that many diseases that have been occurring in human beings were also reported from canines by natural occurrence. Canines were domesticated some 10,000 to 15,000 years back and since then they have evolved into more than 400 distinct breeds with wide variety of morphological and behavioral genetic differences suitable for genetic exploration (Ostrander *et al*., 2000). Many new breeds are evolved from a few founders and have been inbred for desired characteristics. This has led to a species with enormous phenotypic diversity, but with significant homogenization of the gene pool within breeds (Kirkness *et al*., 2003). The combination of genetic homogeneity and phenotypic diversity also provide an opportunity to understand, the genetic basis of many complex developmental processes in mammals (Chase *et al*., 2002). Dogs enjoy a medical surveillance and clinical literature second only to humans, succumbing to 360 genetic diseases that have human counterparts (Patterson, 2002). In addition, the well-preserved lineages and medical records of purebred dog ascribe it a very important model role, in the studies of human genetic diseases.

Studies conducted in various sequenced genomes established different compositional patterns in different classes of animals. The gene concentration pattern present in human genome is common to all vertebrate genomes including canines. The distribution of genes in the human genome is strikingly non-uniform (Bernardi *et al.*, 1985; Bernardi, 1995). It has been reported that highest gene concentration is associated with GC richest isochore families of the genome. Isochores are long (>300 kb) DNA segments, which are compositionally

*To whom correspondence should be addressed.
Tel: 91-33-2355 6626; Fax: 91-33-2335 3886
E-mail:surajit@bic.boseinst.ernet.in

homogenous and belong to small families characterized by different GC levels covering a 30-60% range (GC is the molar fraction of guanine + cytosine) (Bernardi *et al.,* 1985; Bernardi, 1995). If the isochore distribution in chromosomes is known, it will provide information regarding the distribution of genes in the chromosomes. Similarly, the correlations between the GC levels in the third codon positions (GC_3) and intergenic sequence GC levels can be used to assess the distribution of genes in the genome (Mouchiroud *et al.,* 1991; Bernardi 1995).

To study the conditions or abnormalities that can lead to genetic disease, it is essential to have a sound knowledge of the genetic makeup in the normal animal in healthy condition. Here is a modest attempt to study the compositional pattern of a representative sample of canine genome and the correlations that hold between the various regions of gene like exons, introns and 5' and 3' flanking regions. The degree of compositional correlations at the level of individual bases are also analysed. Our studies suggest a strong link in the compositional built up among various regions of the canine genome. Moreover canines show a significant similarity with human genome more than those reported from any other mammals.

Materials and Methods

Four canine chromosomes (three autosomes and X chromosome) have been downloaded from ftp.ncbi.nlm.nih.gov/genbank/genomes. Our own program developed in C was used to retrieve the coding sequences from the complete genome and has been successfully used earlier in retrieving the coding sequences (Gupta *et al*., 2004; Banerjee *et al*., 2005). Those sequences with internal stop codons and not containing proper start and stop codons have been removed. Finally 2210 genes were selected for data analysis. 5' and 3' flanking regions were selected from 500 base pairs upstream from start codon and 500 base pairs downstream from stop codon respectively. Only those cds with more than 500 base pairs flanking region have been selected for our study to remove the influence of promoter and other regulatory regions of the gene. The adjacent coding sequences with overlapping flanking regions were removed and eventually the maximum length of flanking region becomes 6000 base pairs. Base compositions were calculated by using the program CodonW1.3 (available at www.molbiol.ox.ac.uk/cu). The statistical toolkit SPSS 10.0 was used to calculate the correlation coefficients in this study.

Results and Discussion

The compositional distribution of coding sequences. The distribution of GC_3 (third codon position GC levels) values in the canine genes was studied by plotting them on a histogram (Fig. 1). The wide variation in the GC_3 values is in agreement with the findings in human and pig genome indicating a general pattern among the higher mammals (Federico *et al*., 2004). The long range of the GC_3 values also suggest the presence of compositional heterogeneity in canine genome similar to those reported in other mammalian models (Bernardi, 2004). Indeed, the distribution of GC_3 in canines cover a wide range from 26% to 98% (spreading over 72%) having peak at around 65-75% GC_3 level and it falls gradually on either sides. The gene richness in the high GC_3 range reiterates the finding that gene density is higher in the GC rich isochores of mammalian genomes (Mouchiroud *et al*., 1991).

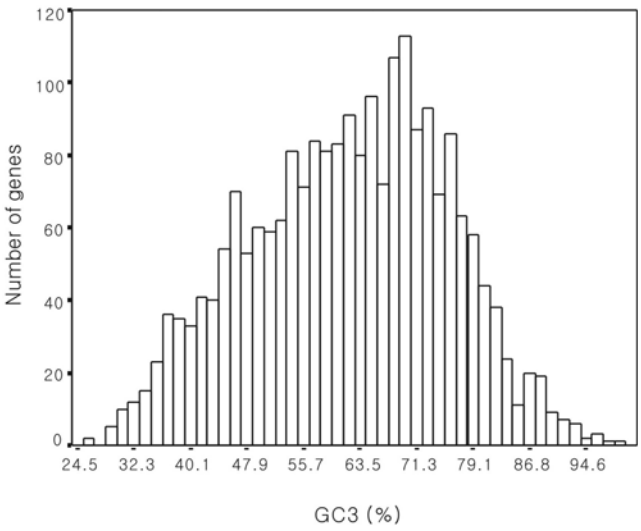

Fig. 1. The compositional distribution of the canine genome. Histogram of GC_3 value (GC value of 3rd codon position) was obtained using a set of 2210 coding sequences.

The compositional correlations between different codon positions. Scatterplots of the GC levels of third codon position (GC_3) against the GC level of first (GC_1) and second (GC_2) codon position, for a data set of 2210 canine coding sequence are shown in Fig. 2(a) and 2(b) respectively. It is noted that the GC_3 versus GC_1 plot is showing a correlation coefficient: $r = 0.62$ ($p < 0.01$). Plot of GC_3 against GC_2 is showing a comparatively weaker, but significant correlation ($r = 0.47$, $p < .01$). When GC_1 is plotted against GC_2 (Fig. 2(c)), the correlation coefficient obtained is 0.62, which is again highly significant at $p < 0.01$.

The above findings indicate that the forces that are shaping the compositional patterns of the canine genome are the same for all codon positions and acting on the three codon positions in a similar way. The constraints operating at the silent sites are also acting on the first and second codon position, but with reduced amplitude.

The compositional correlations between the coding sequences and their flanking regions and introns. We analyzed the compositional correlations between the GC levels of exons and that of corresponding introns and flanking regions (both 3' and 5'). Fig. 3(a) depicts the scatter plot of GC level of exons against the GC level of introns. The correlation coefficient

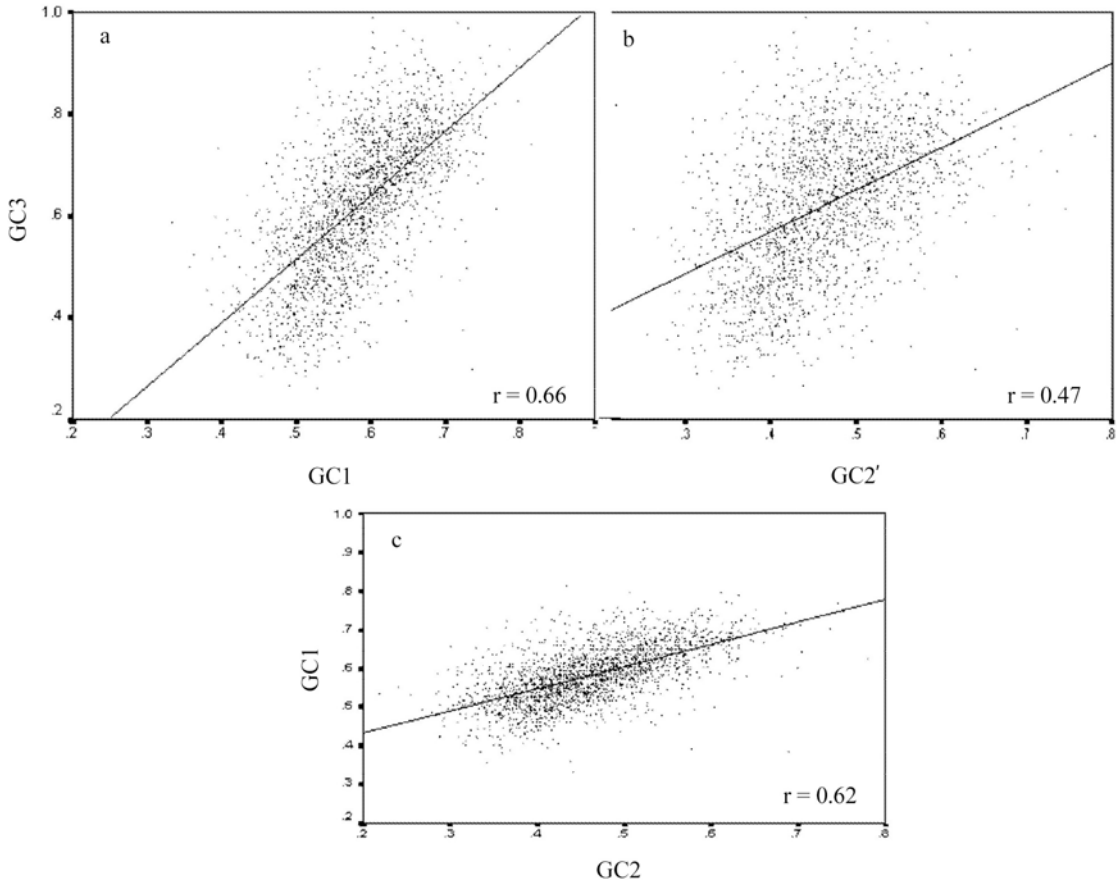

Fig. 2. Scatter plots of GC_3 versus GC_1 (panel a), GC_3 versus GC_2 (panel b) and GC_1 versus GC_2 (panel c).

obtained here is very high ($r = 0.77$, $p < 0.01$). This implies that both the translated and untranslated regions of the genes are matching in their compositional makeup despite the difference of these two regions in producing functional gene product. This supports the notion that introns do not contrast sharply with coding regions with respect to their variation at base compositional levels. Interestingly the correlation coefficient value obtained for the compositional correlation between intron and exon GC contents in human beings is 0.78 (Clay *et al.*, 1996). This similarity in correlation coefficients again indicates the homogenous compositional distribution among various regions of the genes in higher mammals (Bernardi, 2004).

Fig. 3(b) displays the scatter plot of GC levels of exons against the GC levels of flanking regions (5' + 3'). While selecting the flaking regions we have avoided the first 500 bp upstream from the start codon and downstream from the stop codon. This was done to avoid the influence of promoter and other regulatory regions of the gene, which are known to be associated with a definite function. The study showed a significant and strong correlation ($r = 0.68$, $p < 0.01$). These results demonstrate the strong cross-links in GC composition between exons and the corresponding flanking regions. In short, the protein coding regions of the genome as well as those non-protein coding regions is having significant compositional correlations. This highlight the fact that the forces acting on the genome, whether it be selection pressure or mutational forces, are acting independent of the functional significance of the regions.

We analysed whether those correlations are relevant at the level of individual bases. For this purpose, we plotted the frequencies of individual bases in exons against those of flanking regions. The scatter plot (Fig. 4) obtained proves that the correlations at the level of individual bases among exons and flanking regions are also highly significant ($p < 0.01$). On analysis of the individual base correlations among exons and introns, similar results were obtained (data not shown).

The compositional correlation of introns and different codon positions of coding sequences. Fig. 5 shows the correlations of GC content of introns towards that of the first (Fig. 5a), second (Fig. 5b) and third (Fig. 5c) codon positions in exons. According to the expectations, the correlation coefficient of intron GC level with GC_3 levels of exons reaches the highest ($r = 0.84$). This correlation coefficient is highly significant and points out that the constraints acting on the third codon position of exons are also acting in same direction and amplitude in the corresponding introns. The degree of

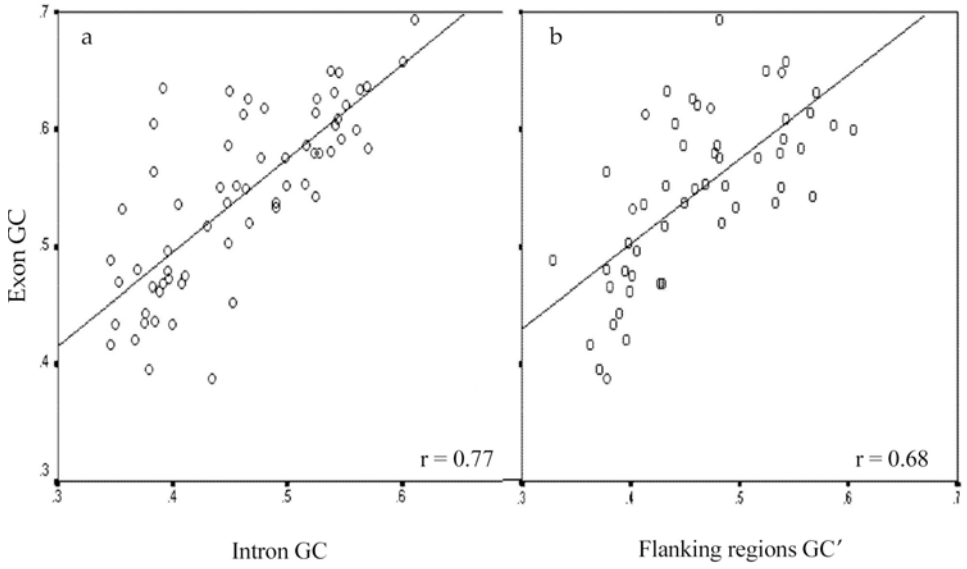

Fig. 3. Scatter plot between the GC levels of coding DNA sequences and the GC levels of corresponding introns (panel a). Scatter plot between the GC levels of coding DNA and the GC levels of flanking (5' + 3') regions (panel b).

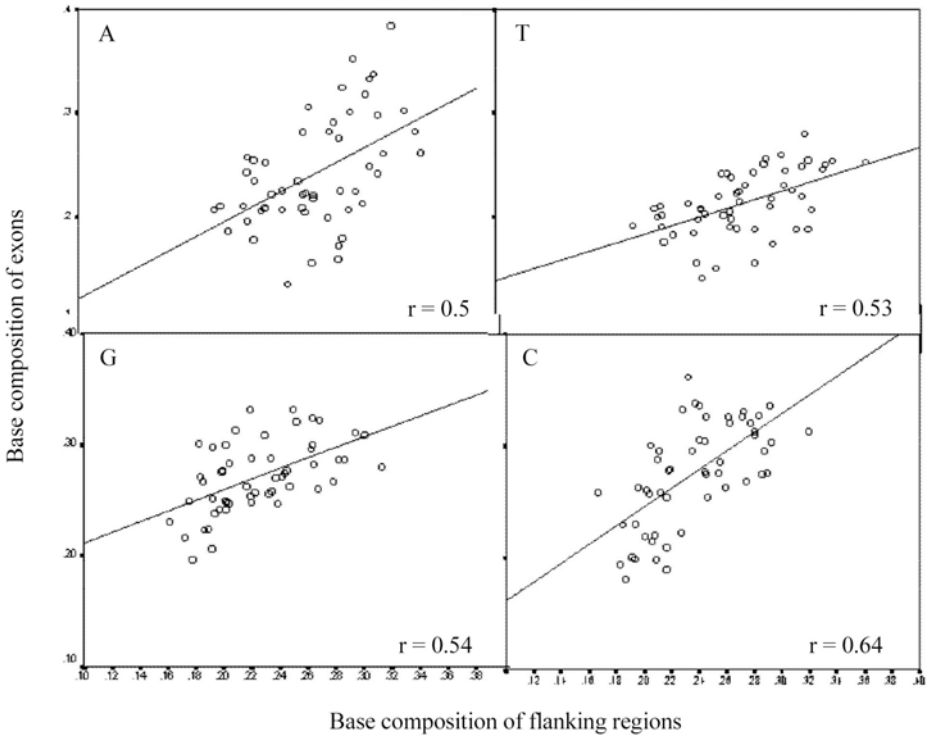

Fig. 4. Frequencies of each base in the coding sequences plotted against the same base in flanking regions (5' + 3').

correlation of introns with GC_1 and GC_2 of exons are comparatively less but highly significant. Surprisingly, the r values obtained i.e., 0.54 and 0.33 respectively are identical with those in avian genome (Musto *et al.*, 1999). These findings point to the fact that a uniform pattern of compositional distribution observed in various gene regions in canine genome can extend to different classes of vertebrates.

The compositional correlations of GC_3 with flanking regions. The correlations between the GC_3 level of coding regions and GC level of flanking regions assume importance from the fact that it can be used to determine the distribution of genes in the entire genome (Bernardi *et al.*, 1985; Bernardi, 1995; Clay *et al.*, 1998). The GC_3 levels of exons are plotted against the GC level of flanking regions in Fig. 6a. The

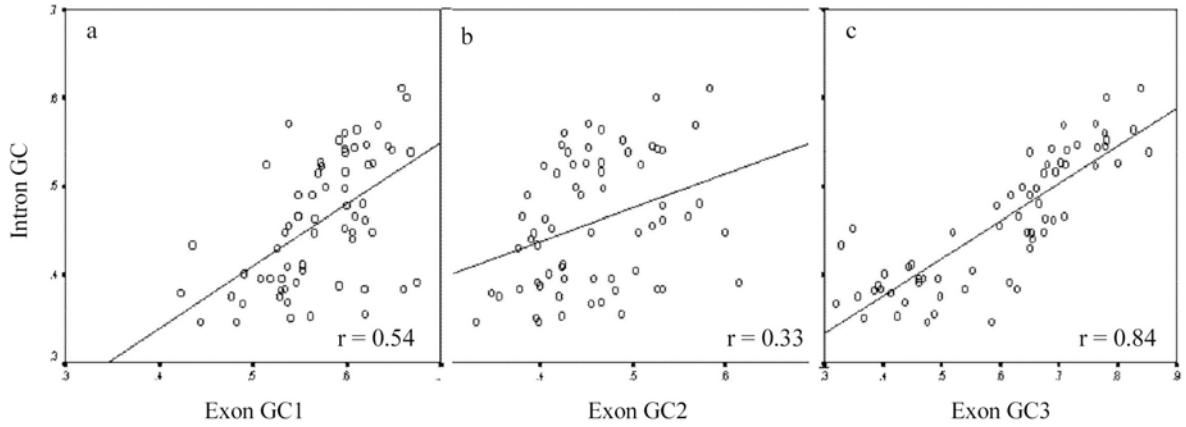

Fig. 5. Scatter plots of GC levels of introns versus GC_1 (panel a), GC levels of introns versus GC_2 (panel b) and GC levels of introns versus GC_3 (panel c).

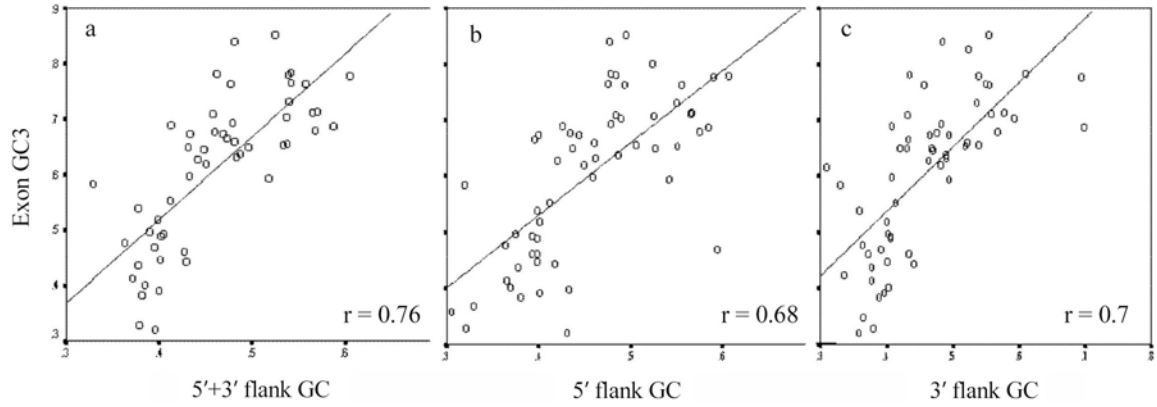

Fig. 6. Scatter plots of GC_3 of coding sequences versus the GC content of corresponding 5' flank (panel a), GC content of 3' flank (panel b) and GC content of 5' + 3' flank (panel c).

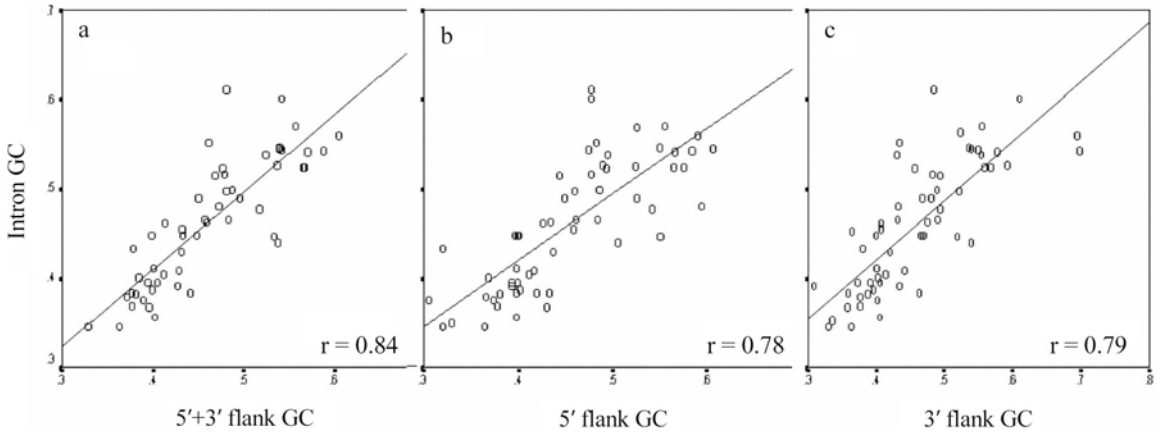

Fig. 7. GC levels of introns plotted against the corresponding GC levels of flanking (5' + 3') regions (panel a), 5' alone (panel b) and 3' (panel c) alone.

correlation values of the 5' and 3' flanking regions were separately analysed and shown in Fig. 6b and 6c. All the three cases are showing a strong correlation with high significance ($p < 0.01$).

It is important to note that the correlations of GC level of 5' and 3' flanking regions with GC_3 level of exons are almost similar to those obtained for human beings (0.65 and 0.61 respectively) from flanking regions of similar size (Jabbari *et*

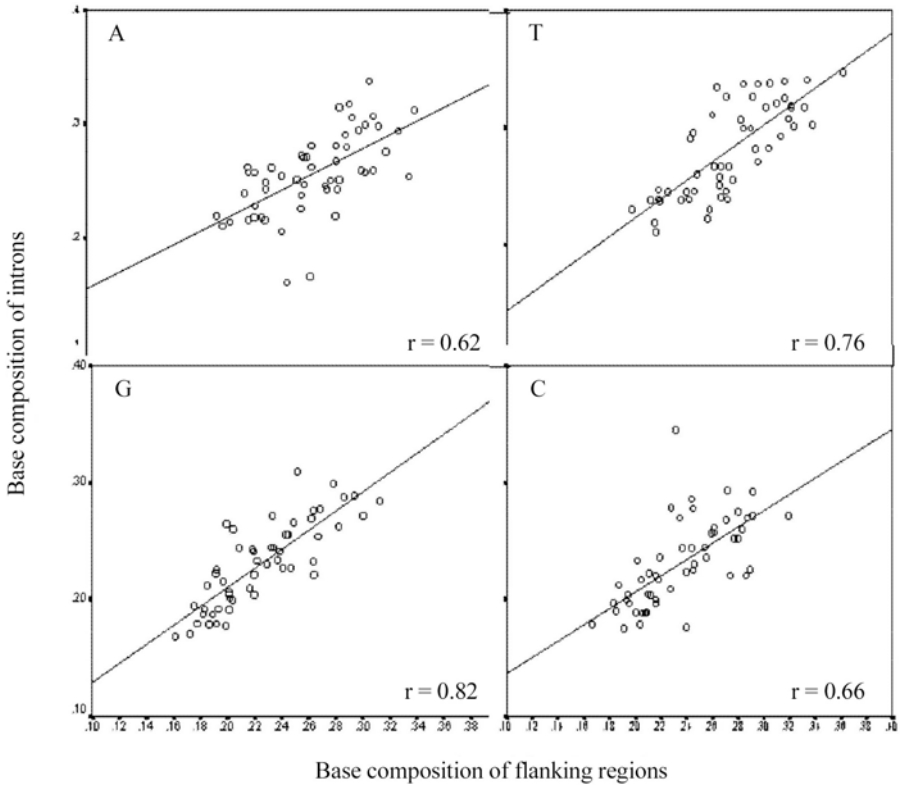

Fig. 8. Scatter plots of the frequencies of individual bases in introns plotted against those in flanking regions.

al., 2000). This indirectly suggests that the distribution pattern of genes in the human and canine genome may be similar. This revelations are not trivial when considering the fact that canine genome is being widely recommended to study the defects of human genes.

This result also points out that those changes in the composition of flanking DNA, which were considered as 'Junk DNA', is paralleled by similar changes in the functionally significant exons. This will welcome more research into the obscure roles of the so-called 'Junk DNA'.

The compositional correlations among the non-coding regions of the genes. Fig.7(a) depicts the correlations that hold between the non-coding regions of the canine genome, i.e., introns versus the flanking regions. The analysis was done with 5' and 3' flank regions separately in Fig.7(b) and (c). The correlation coefficients for all the three studies were almost equal with strong values. Indeed the r values are 0.84, 0.78 and 0.79 respectively for 5' + 3', 5' alone and 3' alone. All are significant with $p < 0.01$. As expected when the 5' GC was plotted against the corresponding 3' GC then a significant correlation was obtained ($r = 0.78$, $p < 0.01$).

We examined whether the correlations obtained between non-coding regions based on GC content extend to the level of individual bases. Fig. 8 displays the scatter plots of each individual base frequency in introns plotted against those in flanking sequences. We find a strong correlation for all the individual bases, as seen in case of exons and flanking regions.

Conclusion

Compositional analysis was conducted on a representative sample set of canine genes to study the correlations that exist between various regions of the genome. Our study has proved that there is a strong positive correlation among the different regions of the genome based on their base composition. Moreover the distribution of GC_3 values suggests that canine genome is having the predicted heterogeneous distribution of base composition similar to human (Bernardi, 2004). Significantly many correlations in canines are found paralleled by the correlation levels in human genome, indicating that the evolutionary forces that has acted over years on the human genome has done the same with similar amplitude on the genome of man's best friend. This has attributed canine the additional advantage as being the model for the study of human genetic diseases, especially since, the other mammalian models from rodentia has been reported to be deviating from the general compositional pattern of human beings.

Acknowledgments Authors wish to thank the Department of Biotechnology, Government of India, for the financial help.

References

Banerjee, T., Gupta, S. K. and Ghosh, T. C. (2005) Compositional transitions between *Oryza sativa* and *Arabidopsis thaliana* genes are linked to the functional change of encoded proteins. *Plant Science* **170**, 267-273.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953-958.

Bernardi, G. (1991) The distribution of genes in the human genome. *Gene* **100**, 181-187.

Bernardi, G. (2004) Structural and Evolutionary Genomics: Natural Selection in Genome Evolution, Elsevier, Amsterdam, USA.

Chase, K., Carrier, D. R., Adler, F. R., Jarvik, T., Ostrander, E. A., Lorentzen, T. D. and Lark, K. G. (2002) Genetic basis for systems of skeltal quantitative traits: principal component analysis of canid skelton. *Proc. Natl. Acad. Sci. USA* **99**, 9930-9935.

Clay, O., Caccio, S., Zoubak, S., Mouchiroud, D. and Bernardi, G. (1996) Human coding and noncoding DNA: compositional correlations. *Mol. Phylogenet. Evol.* **5**, 2-12.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. and Bernardi, G. (1991) Correlations between the compositional properties of human genes, codon usage and amino acid composition of proteins. *J. Mol. Evol.* **32**, 504-510.

D'Onofrio, G., Jabbari, K., Musto, H. and Bernardi, G. (1999) The correlations of protein hydropathy with the composition of coding sequences. *Gene* **238**, 3-14.

Federico, C., Saccone, S., Andreozzi, L., Motta, S., Russo, V., Carels, N. and Bernardi, G. (2004) The pig genome: compositional analysis and identification of the gene richest regions in chromosomes and nuclei. *Gene* **343**, 245-251.

Greer, K. A., Corgill, E. J., Cox, M. L., Clark, L. A., Tsai, K. L., Credille, K. M., Dunstan, R. W., Venta, P. J. and Murphy, K. E. (2003) Digging up the canine genome - a tale to wag about. *Cytogenet. Genome Res.* **102**, 244-248.

Gupta, S. K., Bhattacharya, T. K. and Ghosh, T. C. (2004) Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *J. Biomol. Struct. Dyn.* **21**, 1-9.

Jabbari, K. and Bernardi, G. (2000) The distribution of genes in the *Drosophila* genome. *Gene* **247**, 287-292.

Kirkness, E. F., Batna, V., Halpern, A. L., Levy, S., Remington, K., Rusch, D. B., Delcher, A. L., Pop, M., Wang, W., Fraser, C. M. and Venter, J. C. (2003) The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898-1903.

Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C. and Bernardi, G. (1991) The distribution of genes in the human genome. *Gene* **100**, 181-187.

Musto, H., Romero, H., Zavala, A. and Bernardi, G. (1999) Compositional correlations in the chicken genome. *J. Mol. Evol.* **49**, 325-329.

Ostrander, E. A., Galibert, F. and Patterson, D. F. (2000) Canine genetics comes of age. *Trends Genet.* **16**, 117-124.

Patterson, D. F. (2000) Canine genetic information system: a compositional knowledge base of genetic diseases in the dog. Mosby-Harcourt, St Louis, USA.