BIOCHEMISTRY AND BIOLOGY OF DNA METHYLATION

Proceedings of a Fogarty International Center Conference
held in Bethesda, Maryland
April 17–19, 1985

Editors

Giulio L. Cantoni
Laboratory of General
and Comparative Biochemistry
National Institute of Mental Health
Bethesda, Maryland

Aharon Razin
Department of
Cellular Biochemistry
The Hebrew University
Hadassah Medical School
Jerusalem, Israel

ALAN R. LISS, INC. • NEW YORK

Biochemistry and Biology of DNA Methylation, pages 3–10
© 1985 Alan R. Liss, Inc.

THE ORGANIZATION OF THE VERTEBRATE GENOME AND THE PROBLEM OF THE CpG SHORTAGE

Giorgio Bernardi

Laboratoire de Génétique Moléculaire,
Institut Jacques Monod
2, Place Jussieu, 75005 Paris, France.

Investigations carried out in our laboratory (see Bernardi et al., 1985 for a recent report) have shown that nuclear DNA from warm-blooded vertebrates can be separated, by using density gradient centrifugation in the presence of DNA ligands, into a small number of major components and into several satellite and minor components. The major components comprise : (i) two light components, L1 and L2, poorly or not resolved in some genomes; and (ii) two or three heavy components, H1, H2 and H3; the latter represent about one third of main-band DNA from warm-blooded vertebrates and account for its strong heterogeneity and CsCl band asymmetry. In contrast, main-band DNAs from most cold-blooded vertebrates show weak heterogeneities, only slightly skewed CsCl bands, and major components having buoyant densities that are only or mainly in the range of the light components of warm-blooded vertebrates. The families of DNA molecules (30-100 kb) forming the major components are derived, by the unavoidable breakage which accompanies DNA preparation, from corresponding families of much longer segments, the isochores, which have an average size well above 200 kb and a fairly homogeneous base composition.

The heavy DNA components of warm-blooded vertebrates : (i) have mainly arisen by regional increases in GC, which accompanied the transition from cold-blooded to warm-blooded vertebrates; these increases in GC seem to be mainly due to structural requirements at the DNA, chromatin and chromosome level; (ii) contain over half of the genes tested so far, a particularly high concentration being found in the heaviest components which only represent 4-8 % of nuclear DNA; and (iii) are responsible for an alternating pattern of heavy and light isochores which underlies the Giemsa and Reverse chromosomal patterns.

Major components can be used to study the genome distribution of any sequence which can be probed. This approach has revealed : (i) that the distribution of genes, of families of interspersed repeated sequences, and of integrated viral sequences is highly non-uniform in the genome of warm-blooded vertebrates; (ii) that the GC content and CpG/GpC ratio of both coding and non-coding sequences, as well as the GC levels of codon third positions show a direct linear relationship with the GC contents of the isochores harboring the sequences.

Here I would like to discuss in more detail the distribution of CpG doublets in the vertebrate genome and the problem of the CpG shortage. It is well known that vertebrate DNAs show a characteristic deficiency in the CpG doublet (Josse et al., 1961; Swartz et al., 1962). The C residue in this doublet is 55-90 % methylated in vertebrate DNAs (Van der Ploeg and Flavell, 1980; Gruenbaum et al., 1981; Kunnath and Locker, 1982).

CpG shortage and polypeptide-specifying DNA

The first explanation proposed to account for this outstanding feature of the vertebrate genome (Subak-Sharpe et al., 1966) was based on the observation that a CpG shortage is also exhibited by the genomes of small vertebrate viruses (like polyoma and SV40), which use essentially all of their genetic information for directing protein synthesis. It was proposed (Subak-Sharpe, 1967; 1974) that the genomes of these viruses derived from polypeptide-specifying DNA stretches of the ancestral host cells and therefore exhibited the same CpG shortage, which was visualized as reflecting constraints from the translation apparatus. In agreement with this proposal, CpG shortage was absent in genes not coding for polypeptides, like the tRNA, 5S RNA and rRNA genes. The lack of a CpG shortage in the genomes of intermediate and large vertebrate viruses (like adenovirus and herpes symplex, respectively) was attributed to their capacity of modifying the host translation apparatus either by virus-encoded tRNAs, or by modification of pre-existing cell-specified tRNAs through virus-encoded enzymes, or by differential stimulation of the host cell to produce those tRNAs needed to redress the balance. Since only a small proportion of vertebrate DNA is thought to be actively involved in protein coding and since the same CpG shortage was found in fractions of guinea pig DNA, as obtained by density gradient centrifugation, by differential renaturation, or from dispersed and condensed chromatin, it was further assumed that "the bulk of vertebrate DNA derives from and maintains the gross sequence characteristics of polypeptide-specifying DNA " (Russell et al., 1976).

Two subsequent observations from other laboratories did not support the explanation just outlined, namely (i) the presence of a large number of CpG doublets in rabbit (Salser, 1977) and human (Forget et al., 1979) α-globin genes; (ii) the uniform levels of CpG in both coding and non-coding sequences from the human α- and β-globin gene clusters (Lennon and Fraser, 1983). Since these observations concerned such a minute fraction of the genome, it could be argued, however, that the sequences studied were just exceptions to the rule.

Our investigations (Bernardi et al., 1985) have provided an unequivocal demonstration that CpG shortage is not dependent upon polypeptide coding. In fact (see Fig. 1), (i) CpG shortage usually is strong in the "light" genes from warm-

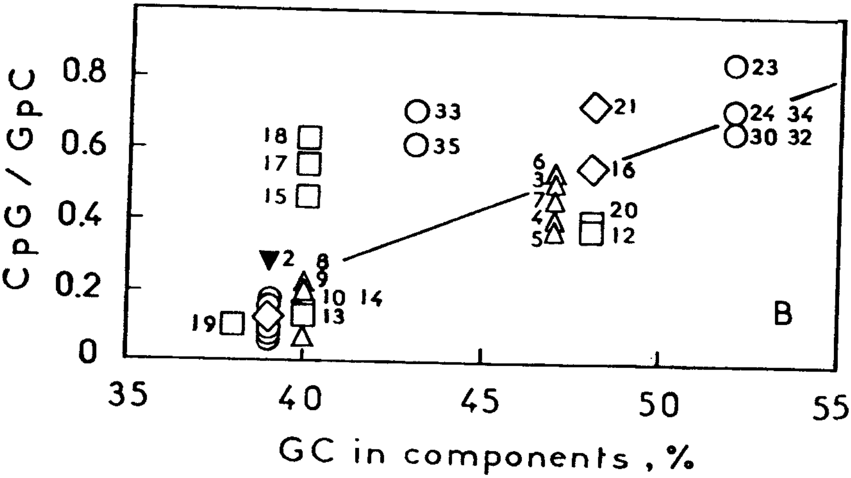

Fig. 1. Plot of CpG to GpC ratios for vertebrate genes and exons against the GC contents of DNA components harboring the genes. From Bernardi et al.(1985); the cluster of unnumbered genes comprises gene 22 and genes 25 to 29 (the key for the genes investigated can be found in the reference).

blooded vertebrates and, in all likelihood, in most genes from cold-blooded vertebrates, but becomes increasingly weaker and almost disappears in "heavy" genes from warm-blooded vertebrates; (ii) CpG shortage affects both coding and non-coding (intronic and intergenic) sequences from a given isochore family essentially to the same extent. In other words, CpG levels in both coding and non-coding sequences from warm-blooded vertebrate depend upon the GC levels of the corresponding isochores. This conclusion relies not only on the set of genes shown in Fig. 1, but also on a larger sample from data banks (not shown; Bernardi et al., 1985). In fact, our findings not only contradict the explanation proposed by Subak-Sharpe and co-workers for the CpG shortage, but also demonstrate that CpG shortage is not uniform throughout the genome of at least warm-blooded vertebrates in contrast with the conclusion of Russell et al. (1976). Finally, as far as the different CpG levels exhibited by the genomes of small and large vertebrate viruses are concerned, we have shown that these differences are not related to the size, but simply to the GC levels of the corresponding genomes. Indeed, a plot of CpG versus genome GC (not shown) exhibits a direct linear relationship, as previously found for vertebrate genes (Fig. 1).

CpG shortage and methylation

An alternative explanation (Salser, 1977) for the CpG shortage was that this doublet is highly methylated in vertebrates and that mCpG is a hot spot for mutation since it can be deaminated to TpG. Indeed, m5C residues in single stranded DNA are deaminated at 95°C at three times the rate of C residues and thermophylic bacteria with optimal growth temperatures equal to or higher than 60°C generally avoid having m5C in their genome (Ehrlich et al., 1985).

As far as this explanation is concerned, it is of interest to consider our observations : (i) that the transition from cold-blooded to warm-blooded vertebrates is accompanied by the appearance, or by a strong increase, of GC-rich segments in the genome (the "heavy" isochores) and of "heavy" genes; and (ii) that "heavy" isochores and "heavy genes" show a much decreased discrimination against the doublet CpG.

The latter point generalizes the previous observations that the rabbit and human α-globin genes show a much reduced CpG shortage. This situation was explained by assuming that in these cases CpG was not counterselected because required for mRNA structure and function (Salser, 1977). This explanation is, however, contradicted by the fact that CpG levels are comparable in both genes and pseudogenes from the human α-globin cluster (Lennon and Fraser, 1983) and, more generally, by our observation that the same relationship holds between CpG levels of coding and non-coding sequences and GC levels in corresponding isochores (Fig. 1).

Our findings may indicate : (i) that methylated CpG doublets are more resistant to deamination when present in a GC-rich environment which protects them from DNA breathing; or (ii) that CpG doublets in GC-rich sequences are much less methylated, than CpG doublets in GC-poor sequences .

Experiments under way in our laboratory should allow us to decide which one of these alternative explanations is the correct one.

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

Ehrlich M, Gama-Sosa MA, Carreira LH, Ljungdahl LG, Kuo KC, Gehrke CW (1985). DNA methylation in thermophylic bacteria : N4-methylcytosine, 5-methylcytosine and N6-methyladenine.

Gruenbaum Y, Stein R, Cedar H, Razin A (1981). Methylation of CpG sequences in eukaryotic DNA. FEBS Lett 124 : 67-71.

Josse J, Kaiser AD, Kornberg A (1961). Enzymatic Synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J Biol Chem 236 : 864-875.

Kunnath L, Locker J (1982). Characterization of DNA methylation in the rat. Biochem Biophys Acta 699 : 264-271.

Lennon GC, Fraser NW (1983). CpG frequency in large DNA segments. J Mol Evol 19 : 286-288.

Russell GJ, Walker PMB, Elton RA, Subak-Sharpe JH (1976). Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 108 : 1-23.

Salser W (1977) Globin mRNA sequences : analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp Quant Biol 40 : 985-1002.

Subak-Sharpe H, Burk RR, Crawford LV, Morrison JM, Hay J, Keir AM (1966). An approach to evolutionary relationship of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequence. Cold Spring Harbor Symp Quant Biol 31 : 737-738.

Subak-Sharpe JH (1967). Doublet patterns and evolution of viruses. British Med Bull 23 : 161-168.

Subak-Sharpe JH, Elton RA, Russell GJ (1974). Evolutionary implications of doublet analysis. Symp Soc Gen Microbiol 24 : 131-150.

Swartz MN, Trautner TA, Kornberg A (1962). Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxribonucleic acid. J Biol Chem 237 : 1961-1967.

Van der Ploeg LHT, Flavell RA (1980). DNA methylation in the human β- globin locus in erythroid and non-erythroid tissues. Cell 19 : 947-958.