Volume 217, number 2, 184–186 FEB 04772 June 1987

Hypothesis

Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells

J. Filipski

Institut Jacques Monod, Paris, France

Received 6 April 1987

The vertebrate genome is built of long DNA regions, relatively homogeneous in GC content, which likely correspond to bands on stained chromosomes. Large differences in composition have been found among DNA regions belonging to the same genome. They are paralleled by differences in codon usage in genes differently localized. The hypothesis presented here asserts that these differences in composition are caused by different mutational bias of α and β DNA polymerases, these polymerases being involved to different extents in the repair of DNA lesions in compact and relaxed chromatin, respectively, in germline cells.

Molecular clock; Chromosome banding; Codon usage; Repair fidelity

The earliest demonstration of compartments in the genetic material of vertebrates came with the observation of bands on stained chromosomes. Later it was found that the DNAs in these bands differ in composition and in timing of replication. The dark Giemsa bands contain AT-rich DNA and replicate in late S phase of the cell cycle. The GC-rich, dark Reverse bands replicate in early S [1–3]. On the other hand, studies at the molecular level by density gradient centrifugation [4], electron microscopy [5] or sequence analysis [6,7] revealed considerable uniformity of composition of DNA molecules. Usually segments of 3 kb or longer, randomly chosen from a DNA region of several dozen or even several hundred kilo base pairs, all have similar GC contents. At the same time there are large differences in composition between different DNA regions in the same genome. It was also found that the GC-rich exons are located mostly in GC-rich DNA regions while the AT-rich exons are found in the AT-rich regions. Understandably, the codon usage is correlated with the composition of exons: the GC-rich genes are coded preferentially by GC-rich codons and AT-rich genes by the AT-rich codons [4]. An important feature of the GC-rich genes is the presence of compositional islands rich in CpG dinucleotides close to their 5′-ends [8–10]. This is an indication that these genes are partially undermethylated and available for transcription in the germline [10]. Studies on localization of genes and their time of replication provide a link between the results obtained at the molecular and ultrastructural levels. GC-rich exons carried by GC-rich DNA regions are located in GC-rich Reverse chromosomal bands, while AT-rich exons carried by AT-rich DNA regions are located in the AT-rich Giemsa bands [11].

A hypothesis was put forward that these two distinct genomic compartments, AT-rich and GC-rich, contain tissue-specific genes and housekeep-

Correspondence address: J. Filipski, Institut J. Monod, 2 Place Jussieu, 75251 Paris Cedex 05, France

Published by Elsevier Science Publishers B.V. (Biomedical Division)

ing genes, respectively, and that they differ in the mechanism of the regulation of transcription [11]. An accumulation of GC base pairs in the DNA through a positive Darwinian selection which would compensate for the effect of higher body temperature on the DNA and RNA secondary structure might explain the increase in GC content in coding sequences in warm-blooded vertebrates as compared with cold-blooded vertebrates [4]. It does not, however, account for the high GC level in nontranscribed sequences in these organisms. It seems likely that the compositional compartments containing either GC-rich or AT-rich DNA diverged because of an accumulation of mutations compositionally biased but neutral for the most part in their selective value. The differences in composition may have resulted from the fact that the relaxed and compact chromatin loops carrying transcriptionally competent and incompetent genes (see [12] for recent review), respectively, are controlled in germline cells by different repair systems. The error-prone repair system acts on DNA in relaxed chromatin regions which in the germline is partially undermethylated, is GC-rich and carries CpG-rich islands. The DNA in compact chromatin in the germline cells is AT-rich and is controlled by a less error-prone repair system, differently biased.

Experimental data which support this model are the following.

(i) It was found that active genes and the sequences flanking these genes are repaired 5-times faster than the lesions in the transcriptionally inactive DNA in which they persist unrepaired for hours [13].

(ii) Although the details of the repair processes in the higher organisms are far from being understood, it is fairly well established that the enzyme responsible for DNA repair-related polynucleotide synthesis is the β DNA polymerase [14]. The α DNA polymerase is a main replicating enzyme (review [15]). Both enzymes, however, are involved in repair of DNA lesions to different extents depending on the type of lesion [16]. The α and β DNA polymerases are characterized by a very different transcriptional fidelity, the β being much more error-prone than the α (at least in vitro [17]). According to the hypothesis discussed here, the repair of DNA lesions in relaxed chromatin structures is assured mainly by the error-prone β polymerase, while at least some of the lesions in the condensed chromatin persist long enough to be repaired by the α DNA polymerase during DNA replication. Two observations corroborate this assumption. One is a very low level of α polymerase activity before the beginning and after the end of the S phase of the cell cycle [18]. The other is the total absence of the α polymerase activity in cells in which DNA does not replicate but in which it is transcribed [14]. The repair of DNA lesions in transcriptionally active chromatin is thus assured by β DNA polymerase, at least in the cases where α polymerase is inactive. On the other hand, the condensed, transcriptionally inactive chromatin is generally less accessible than relaxed, transcriptionally active to nucleolytic enzymes, thus it is probably also less accessible to the β polymerase. Different involvement of the α and β polymerases in the repair of relaxed and compact chromatin, respectively, might also be caused by different kinds of lesions to which the DNA is exposed in these two chromatin compartments.

(iii) Comparison of the point mutations caused by chicken and rat β DNA polymerases has been carried out by Kunkel [19]. Examination of his data reveals that the rat enzyme when making mistakes replaces AT base pairs by GC base pairs as frequently as the reverse while the chicken enzyme causes point mutations biased in the ratio 3:2 towards the enrichment in GC content of the newly synthesized DNA strand. Different β polymerases are thus differently biased and these differences may explain why the rat genome shows a lower percentage of very GC-rich DNA sequences than the chicken genome [20,21].

(iv) Accumulation of mutations during evolution (the 'molecular clock ticking') is not uniform throughout the genome even in DNA regions which do not seem to carry any specific genetic information. The noncoding sequences in the AT-rich β-globin gene are evolutionarily conserved while the noncoding sequences in the α-globin gene cluster have diverged almost entirely in mammals [22]: the α-globin type genes in mammals are GC-rich and contain CpG-rich islands which correlates with their partial undermethylation, availability to transcription and relaxed conformation of the chromatin carrying this gene cluster in the germline cells. This gene cluster thus seems to be under the control of the error-prone repair system while the repair of the AT-rich β-globin gene cluster is

probably assured mainly by the error-proof α polymerase. Similar high conservation of non-coding sequences in another AT-rich gene was reported by Marinaga et al. [23].

(v) Different codon usage in different genes in the same organism is quite common [24]. Among several constraints and biases shaping the coding sequences during evolution, the bias caused by the exposure to various repair systems in the germline is likely to have considerable influence on the codon usage. An example of genes belonging to the GC-rich α-globin gene cluster and much less GC-rich β-globin gene cluster in mammals is, also in this case, a very informative one. Genes belonging to these clusters are expressed in the same cells so there are no differences in the tRNA pools which might influence the codon usage. The proteins encoded by these genes do not seem to be so different as to impose any differential functional constraint on the codon usage. Accumulation of the GC-rich codons in the α-globin gene and less GC-rich codons in the β-globin gene resulted from an exposure to different repair systems in germline cells according to the presented hypothesis. It is worthwhile to note that the α-globin gene in chicken has more Gs and Cs in the codon third positions than the corresponding gene in mouse which is predicted by the differences in the mutational biases of the rodent and avian β polymerases.

Verification of the presented hypothesis can be made by studying the chromatin structure-related repair processes in the germline cells. The idea, however, that different fidelity and different mutational bias of the two main DNA polymerases is responsible for most of the heterogeneity of the vertebrate DNA appears to fit well the available data and also the generally accepted model of the chromatin structure. It is therefore important to examine very carefully estimates of genetic distances between species which have been based on nucleotide substitution data. This caution is especially necessary when comparing GC-rich genes with an AT-rich one or when comparing genes from very diverse organisms [25] such as plants (which have neither β DNA polymerase nor a germline), insects (which do not have the β polymerase) and vertebrates (which have both).

REFERENCES

[1] Holmquist, G., Gray, M., Porter, T. and Jordan, J. (1982) Cell 31, 121–129.
[2] Kornberg, J.R. and Engels, W.R. (1978) Proc. Natl. Acad. Sci. USA 75, 3382–3386.
[3] Tobia, A.M., Schildkraut, C.L. and Maio, J.J. (1970) J. Mol. Biol. 54, 499–515.
[4] Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science 228, 953–958.
[5] Moreau, J. (1986) Thesis, Inst. J. Monod, Paris.
[6] Aota, S. and Ikemura, T. (1986) Nucleic Acids Res. 14, 6345–6355.
[7] Filipski, J., Salinas, J. and Rodier, F. (1987) DNA 6, 109–118.
[8] Adams, R.L.P. and Eason, R. (1984) Nucleic Acids Res. 12, 5869–5874.
[9] Bird, A.P. (1986) Nature 321, 209–213.
[10] Tykocinski, M.L. and Max, E.E. (1984) Nucleic Acids Res. 12, 4385–4396.
[11] Goldman, M.A., Holmquist, G.P., Gray, M.C., Caston, L.A. and Nag, A. (1984) Science 224, 683–692.
[12] Eissenberg, J.C., Cartwright, I.L., Thomas, G.H. and Elgin, S.C.R. (1985) Annu. Rev. Genet. 19, 485–536.
[13] Bohr, V.A., Smith, C.A., Okumoto, D.S. and Hanawalt, P.C. (1985) Cell 40, 395–396.
[14] Waser, J., Hubscher, U., Kunzle, C.C. and Spadari, S. (1979) Eur. J. Biochem. 97, 361–368.
[15] Loeb, L.A., Liu, P.K. and Fry, M. (1986) Prog. Nucleic Acid Res. Mol. Biol. 33, 57–110.
[16] Miller, M.R. and Chinault, D.N. (1982) J. Biol. Chem. 257, 10204–10209.
[17] Kunkel, T.A. and Loeb, L.A. (1981) Science 213, 765–767.
[18] Jackson, D.A. and Cook, P.R. (1986) J. Mol. Biol. 192, 65–67.
[19] Kunkel, T. (1985) J. Biol. Chem. 260, 5787–5796.
[20] Cortadas, J., Olofsson, B. and Bernardi, G. (1979) Eur. J. Biochem. 99, 179–186.
[21] Thiery, J.P., Macaya, G. and Bernardi, G. (1976) J. Mol. Biol. 108, 219–235.
[22] Hardison, R.C. and Gelinas, R. (1986) Mol. Biol. Evol. 3, 243–261.
[23] Marinaga, T., Sakai, M., Wegeman, T. and Tamaoki, T. (1983) Proc. Natl. Acad. Sci. USA 80, 4604–4608.
[24] Perrin, P. (1984) Nucleic Acids Res. 12, 5515–5537.
[25] Wilson, A.C., Carlson, S.S. and White, T.J. (1977) Annu. Rev. Biochem. 46, 573–639.