Genome organization in higher organisms.

Claire Gaillard and François Strauss

Université Pierre et Marie Curie, Paris
francois.strauss@snv.jussieu.fr

The complexity of higher organisms, which arises in the course of embryonic development from the much simpler fertilized egg, does not emerge by spontaneous generation nor by miracle. Such complexity must somehow be preexistent in the egg. Since the structure of organisms is genetically transmitted, it is DNA itself, the support of genetic information, that encodes this complexity. Here we propose a model of organization of the genome in DNA loops maintained by DNA crossings. In this model DNA sequences that do not encode proteins, which represent more than 98% of the human genome, are involved in the definition of DNA crossing points and can no longer be considered as junk, but instead play a fundamental part in the encoding of genetic information by modulating the transcriptional state of genome domains.

The structural and physiological complexity of organisms has long been known to be related to the amount of DNA per cell, the "c-value". Not that this amount is proportional to the intuitive, visible complexity, since the c-value can vary greatly between organisms that are very similar. However, considering that the structural complexity of organisms is compressed in their genomes in the same way as computer files can be compressed using appropriate algorithms, there is a lower limit to the size of the message, and therefore of the genome, coding for a given complexity[1]. As a consequence, the *minimal* amount of DNA required to encode organisms of a class increases with the complexity of organisms in that class. For example, no mammal can be found with as little DNA per cell as the fruitfly *Drosophila melanogaster*[2]. The c-value of mammals does not vary by a large extent from one species to another, and does not differ by much more than a factor of 2 from the ~$3{,}2 \times 10^9$ base pairs of the human genome. Therefore, at least about a gigabase of information seems required to encode genetically a mammal, using the mechanisms of embryonic development at work in mammals.

The question is to understand how this complexity is encoded in DNA. Before the advent of sequencing programs it was usually considered that all genomes were organized in the same way, on the model of the bacterial genomes, and comprised essentially genes plus their regulatory sequences interacting with specific proteins. A "central dogma" (actually incorrectly interpreted[3]) stated: "DNA makes RNA, RNA makes proteins". Repetitive DNA sequences, which in such a model could not contain any significant genetic information, were regarded as junk DNA with no real function. Similarly introns, as non-coding sequences present within genes, were often considered as useless sequences resulting from an incomplete optimization of the genomes. For many years it was assumed that the complexity of organisms would be reflected in the number of genes. The most striking result of the sequencing of complete genomes was thus to discover that the number of genes does not increase like the complexity of organisms. With 20000 to 25000 genes[4], the human genome does not contain many more genes than the fruitfly *D. melanogaster* (~13500 genes) or the worm *Caenorhabditis elegans* (~20000 genes).

The number of genes thus appears too small to take into account all the complexity of higher organisms, and the variant forms of proteins produced by alternative splicing of RNA rarely result in a multiplicity of function. Similarly, this complexity does not appear to rest on regulatory proteins, as few proteins able to interact very specifically with DNA sequences have been isolated in higher organisms. For example the homeodomain proteins, which play a fundamental role in development and were initially believed to act by binding specific sites on DNA, actually possess only a weak preference for short and degenerated sites ("the homeodomain is a highly conserved structure recognizing a six nucleotide consensus DNA sequence, NNATTA" [5]). Such results were not completely unexpected[6,7], as statistical mechanics makes it extremely difficult for a regulatory protein to find its specific binding site in, for example, the 3 $x10^9$ bp of the human genome, whereas this is possible with a 1000-fold smaller genome such as the *Escherichia coli* genome. This is probably one of the reasons why few DNA-binding proteins from higher eukaryotes have been found with an affinity for DNA and a specificity for their binding site comparable to the affinity and specificity of prokaryotic proteins such as the lac repressor or restriction enzymes.

While the gene number and the diversity of DNA regulatory proteins do not increase greatly with the complexity of organisms, in contrast the amount of DNA sequences that do not encode proteins increases dramatically, representing more than 98% of the human genome. Therefore it seems more and more certain that these 98% participate in the coding of the structure of the human body and contain genetic information encoded in a way that we are still unable to decipher. The model of genetic regulation based on genes and DNA regulatory proteins being insufficient for higher eukaryotes, new hypotheses are required. In this respect the field of non protein-coding RNA has been developing very actively during the last few years. Indeed, whereas less than 2% of the genome encode proteins, a much larger proportion is transcribed into RNA, and the hypothesis exposed in particular by Mattick[8,9] that untranslated RNA plays a major and critical role in regulatory mechanisms is being confirmed day after day by new discoveries.

Should DNA be considered as playing only the purely passive role of being transcribed into RNA that would possess the active regulatory functions? We believe that such a conclusion is premature, and we would like to suggest a new hypothesis concerning the organization of the genome of higher organisms, in which DNA plays a direct role in the regulation of the genetic information that it encodes.

In the course of our search for strong specificities among DNA-protein interactions in mammals, we came across an unexpected, novel DNA structure, DNA hemicatenanes[10], i.e. the crossing of two DNA duplexes in which one of the strands of one duplex passes between the two strands of the other duplex, and reciprocally (Fig. 1, (a), (c)). While little is yet known about DNA hemicatenanes, their extremely high affinity in vitro for nuclear protein HMGB1, one of the most abundant non-histone proteins in the nucleus of mammalian cells, is particularly striking[11,12]. This observation, which at first seemed difficult to fit into classical models, led us to consider the hypothesis that the genome might be organized in loops maintained at their bases by DNA crossings (Fig. 1). The first characteristic of this model is that non-protein-coding sequences are not considered as functionless, but instead play a fundamental role in the definition of crossing points. In addition this organization presents many original functional suggestions a few of which are discussed below.

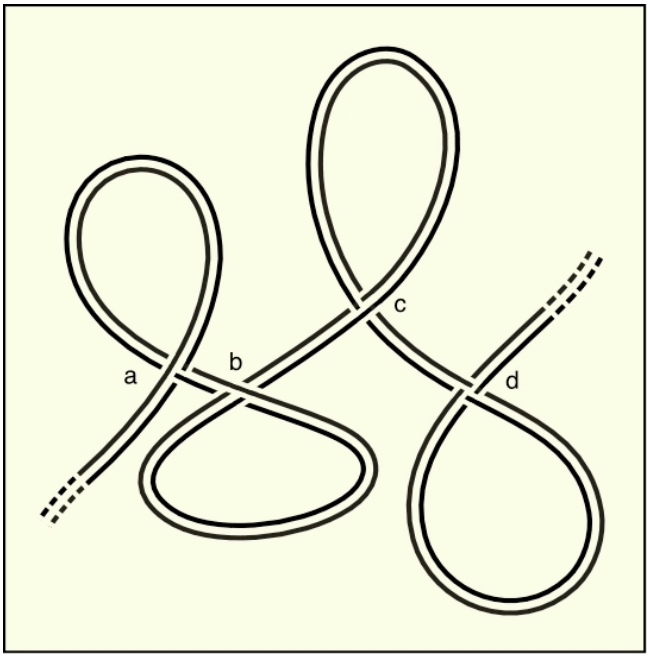

Figure 1. *A model of organization of the genome in higher organisms. DNA is organized in loops maintained by DNA crossings. Among the different DNA structure that can be considered at the bases of the loops the simplest are hemicatenanes (a) and (c). A pseudo-knot is also represented in (b), and a more complex knot in(d).*

A level of chromosomal compaction results automatically from such an organization in loops. While chromosomal loops are widely believed to exist, the actual structure present at their bases and responsible for this organization is still not known. The existence of a "nuclear matrix" to which DNA would be attached is an old hypothesis that fits many experimental results, but its exact nature has not yet been well established[13]. In contrast, in the present model the nature of the nuclear matrix becomes clear, it is DNA itself.

This model raises an important question concerning RNA synthesis : what would happen when an RNA polymerase molecule reaches a DNA crossing point in the course of transcription? would it be blocked? would it pass the obstacle and continue transcription on the same DNA strand? would it be able to leave the strand previously transcribed and continue RNA synthesis after switching DNA template, similarly to a train at a track switch on a railroad? (Fig. 2). The transcriptional activity of a given region of the genome could then be modulated as a function of the distribution and conformations of crossings along the DNA sequence, the combinatorial possibilities offered by such a system being almost unlimited.

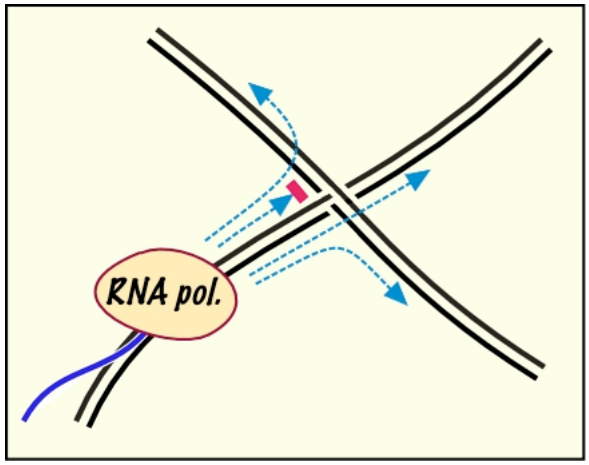

Figure 2. *RNA polymerase molecule reaching a DNA crossing in the course of transcription. Arrows indicate the different possibilities that can be considered. The enzyme could pass the crossing point, stop in front of it, or switch template and continue to the right or to the left according to the strand followed, in the respect of the polarity of RNA synthesis.*

In such a model introns are not useless DNA sequences, on the contrary they are extremely useful elements that allow the positioning of crossing points within genes and therefore increase combinatorial possibilities. For example, the problems posed by very large introns, namely how they can be correctly spliced and how genes of several megabases like Ultrabithorax can be transcribed during the short cell cycle in the early Drosophila embryo[14], no longer exist if large introns are actually only transcribed over a short portion of their length. In contrast, a gene with large introns fully transcribed will most likely result in premature termination of transcription or in transcript degradation by the mechanisms of nonsense-mediated RNA decay.

The regulatory efficiency of the model also rests on its flexibility. While the existence of a molecular mechanism to replicate and transmit the global genome organization from generation to generation is implicitly postulated, the precise location and the fine structure of crossing points should not be envisioned as absolutely fixed and identical in all cells, but instead as being controlled and modulated during development and differentiation, by the interplay of regulatory factors with DNA crossings and as a function of the nucleotide sequences involved.

Theoretical models of genomic organization based on the role of non-protein-coding sequences have been proposed previously[15-17]. An advantage of the present model is that it allows one to make many simple and precise hypotheses that should be amenable to experimental testing.

Jacques Monod used to say that "anything found to be true of E. coli must also be true of elephants", which was often misinterpreted as "there is no fundamental difference between E. coli and elephants". To the naked eye however the difference seems almost infinite, it must be reflected at the level of genome organization.

References

1. The mathematical theory of Kolmogorov Complexity describes the relationship between the size of a message and the informational complexity that it carries. See e.g. http://en.wikipedia.org/wiki/Kolmogorov_complexity ; also Li, M. and Vitanyi, P. (1997) An Introduction to Kolmogorov Complexity and Its Applications. Springer Verlag, New York.
2. Gregory, T.R. (2005) Animal Genome Size Database. http://www.genomesize.com.
3. Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561-563.
4. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.
5. Biggin, M.D. and McGinnis, W. (1997) Regulation of segmentation and segmental identity by Drosophila homeoproteins: the role of DNA binding in functional activity and specificity. *Development*, **124**, 4425-4433.
6. Lin, S. and Riggs, A.D. (1975) The general affinity of lac repressor for E. coli DNA: implications for gene regulation in procaryotes and eucaryotes. *Cell*, **4**, 107-111.
7. von Hippel, P.H. and Berg, O.G. (1986) On the specificity of DNA-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 83, 1608-1612.
8. Mattick, J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986-991.
9. Mattick, J.S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet.*, **5**, 316-323.
10. Gaillard, C. and Strauss, F. (2000a) DNA loops and semicatenated DNA junctions. *BMC Biochem.*, **1**, 1.
11. Gaillard, C. and Strauss, F. (2000b) High affinity binding of proteins HMG1 and HMG2 to semicatenated DNA loops. *BMC Mol. Biol.*, **1**, 1.
12. Jaouen, S., de Koning, L., Gaillard, C., Muselikova-Polanska, E., Stros, M. and Strauss, F. (2005) Determinants of specific binding of HMGB1 protein to hemicatenated DNA loops. *J. Mol. Biol.*, 353, 822-837.
13. Pederson, T. (2000) Half a century of "the nuclear matrix". *Mol. Biol. Cell*, **11**, 799-805.
14. Shermoen, A.W. and O'Farrell, P.H. (1991) Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell*, **67**, 303-310.
15. Scherrer, K. (1989) A unified matrix hypothesis of DNA-directed morphogenesis, protodynamism and growth control. *Biosci. Rep.*, **9**, 157-188.
16. Olovnikov, A.M. (1996) The molecular mechanism of morphogenesis: a theory of locational DNA. *Biokhimiia*, **61**, 1948-1970.
17. Zuckerkandl, E. (2002) Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine. *Genetica*, **115**, 105-129.