EUROPHYSICS LETTERS 10 August 1993

Europhys. Lett., 23 (5), pp. 373-378 (1993)

Crumpled Globule Model of the Three-Dimensional Structure of DNA.

A. GROSBERG (*) (§), Y. RABIN (**), S. HAVLIN (**) and A. NEER (***)

(*) *Institute of Chemical Physics - 117977 Moscow, Russia*
(**) *Department of Physics, Bar-Ilan University - Ramat-Gan 52900, Israel*
(***) *Department of Biology, Open University - Tel-Aviv 61392, Israel*

(received 12 February 1993; accepted in final form 24 June 1993)

PACS. 87.15B – Structure, configuration, conformation, and active sites at the biomolecular level.
PACS. 36.20 – Macromolecules and polymer molecules.

Abstract. – We argue that in order to maintain the biological function of DNA confined inside the cell nucleus, its spatial structure has to be unknotted, of the so-called «crumpled globule» type. The fixation of a particular realization of this non-equilibrium structure by attractive interactions between specific units imposes a connection between the spatial structure of DNA and the statistical distribution of these units along the chain contour. This suggests that both primary sequence and spatial structure of native DNA were formed simultaneously by a self-similar evolution process. The predictions of our model are compared with recent observations of long-range correlations in intron-containing genes and non-transcribed regulatory elements and further experimental tests are proposed.

The relation between the primary structure of DNA and its biological function is one of the outstanding problems in modern biology. There is increasing evidence that the functional role of the DNA sequence is not only to code for proteins but also to control the spatial structure of DNA. While it is generally believed that the biological function is extremely sensitive to complex molecular details and that, as a result, simple physical models based on universal considerations cannot provide useful guidelines for biologists, we will show that such considerations impose important connections between the primary sequence and the spatial structure of native DNA. Let us discuss what general statements can be made about this spatial structure, without making any specific assumptions about structural details.

Simple estimates based on packing considerations (packing a DNA polymer of length of up to 1 m into a cell nucleus of roughly 1 μm size or of length 10 μm into a virus head of 500 Å) show that in any real biological system, from virus to chromosome, the native spatial structure of DNA has to be of a dense globular type, rather than that of an expanded coil (for a review on the DNA behavior in the condensed globular phase, see ref. [1]). There are a large number of different globular structures which were investigated in polymer physics [2]. Spatial 3D structures of globules, either equilibrium or not quite equilibrium ones, are known to be controlled by volume interactions between chain monomers. As to the native DNA glo-

(§) Present address: Physics Department, MIT, Cambridge, MA 02139, USA.

bule, these volume interactions are of tremendous complexity, since they are mediated by proteins and include phenomena such as the recognition of particular sequences by proteins, etc.

It is known also that for a sufficiently long «simple» polymer, *i.e.* homopolymer, most of the conformations of any equilibrium globule contain a vast number of complicated knots, so that the number of entanglements is comparable to the number of chain segments. Since the set of conformations is roughly the same for homo- and for hetero-polymers (we mean, of course, the whole set of existing conformations, but not the set of thermodynamically relevant and/or kinetically available ones), this conclusion is correct for the heteropolymer globule, and, in particular, it is expected to be valid for the DNA globule as well. However, such complex knotted conformations cannot dominate the native state of a functioning biopolymer since entanglements will dramatically reduce its ability to respond to biochemical influences. Indeed, globular proteins, including complex ones which have quaternary structure, are free of any knots. Of course, a DNA globule is much more complex than a protein globule, since it involves dramatically larger length scales. Nevertheless, a similar conclusion has to be valid also for DNA since, if the number of entanglements in the globular structure of a high-molecular-weight polymer is comparable to the number of segments, the structure will become glasslike (*i.e.* kinetically frozen), with the result that many monomer units will be out of reach for any biological system involved in DNA processing. Therefore, we will assume that, *in a statistical sense, the* DNA *globule is practically unknotted.* This conclusion holds in spite of the existence of topoisomerases and other proteins which can cut DNA; being small compared to DNA dimensions they cannot even recognize the global topology of DNA and thus they cannot have a statistically significant effect on the number of entanglements in the globule.

For a sufficiently long polymer, the prohibition of knot formation leads to a non-trivial self-similar fractal spatial structure, the so-called «crumpled globule» [3] (see fig. 1). The key property of this structure is that each chain part of arbitrary length l has to be folded into a globular state, *i.e.* its spatial size should scale as $l^{1/3}$. It differs dramatically from the equilibrium (with respect to formation of knots) globule where any chain part which is small compared to the size of the whole globule looks like a chain in a polymer melt, *i.e.* its size is of order $l^{1/2}$. Such a

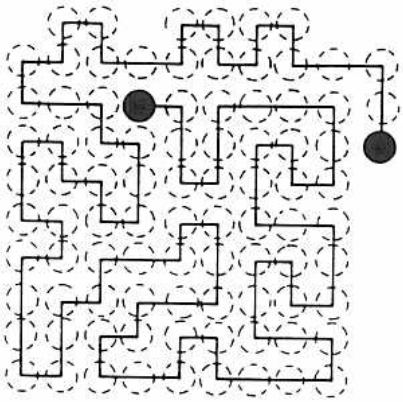

Fig. 1. – Schematic representation of a «crumpled globule» structure of an abstract polymer in 2D. The monomers are represented by dashed circles, the ends of the chain are given by solid circles and the chain contour is indicated by the solid line. Notice that while all random collapsed configurations of a polymer in two dimensions are of the crumpled-globule type, this does not hold, in general, in three dimensions where a typical configuration of a random globule will contain a large number of knots. Notice also that the above picture is scale independent and, therefore, the monomers can represent complexes of parts of DNA with the accompanying proteins. This is consistent with accepted notions about the chromatin structure since, on one of the length scales, these monomers can represent nucleosomes and, on larger scale, they can represent solenoids, etc.

globule of size R can be imagined as built of blobs formed by chain segments of length $(R/a)^2$, a being some microscopic scale; these blobs strongly penetrate each other and in fact are all placed in the same region of size R. In the crumpled globule, on the other hand, each crumple of arbitrary length l having a size of order $l^{1/3}$ does *not* penetrate other crumples of the same scale, which therefore remain segregated in space (as can be visualized in two dimensions, by considering a rope which is densely folded inside a slit between two flat surfaces).

Notice that the above-mentioned fractal structure cannot be realized for short DNA molecules which, because of their high rigidity, do not possess sufficiently many levels of self-similarity (corresponding to different spatial scales) and would form a different type of globule (as is probably the case in small viruses, plasmids and other biological systems).

In addition to these considerations there are several well-known observations in biology which support our claim that the native conformation of sufficiently long DNA is of the «crumpled globule» type. It is known that segments of DNA which are close to each other along the chain contour are also packed close to each other inside a chromosome, a feature which distinguishes the crumpled globule from other polymer structures. The self-similar nature of the folding of eukaryotic DNA into the cell nucleus is illustrated by the observation of hierarchy of structures on different length scales. Thus, two linear DNA chains form a double helix of 2 nm diameter which rolls into 10 nm beads (nucleosomes). The nucleosomes are arranged in 30 nm solenoids which, in turn, form 0.5 μm loops [4] (see also the work [5]).

All the above-described levels of structure maintain the linear-polymer character of DNA and can be thought of as a coarse-graining of the original linear chain, with an associated renormalization of the «monomer» unit. In this sense, there is an analogy between these intermediate-scale structures of DNA and the secondary structure of a protein globule. The crumpled-globule concept implies that the tertiary structure of DNA is the highest level of a hierarchy of secondary structures and that the chromosome as a whole is simply the largest crumple, of the same type as the smaller-scale ones.

A further constraint on our crumpled-globule model of eukaryotic DNA comes from «chromosome maps» which show the spatial position of specific genes inside the chromosome [4]. Since each such gene is a specific part of the DNA contour, we conclude that the spatial structure of native DNA is well defined, at least in a coarse-grained sense. This imposes a relationship between primary and spatial structure in eukaryotic DNA, since it implies that the shapes and relative positions of crumples must be well defined too. The need to stabilize a specific spatial structure leads to an important question: *how many special units in the primary structure of* DNA *are needed to fix a particular realization of crumpled spatial structure on some length scale l?* (Notice that this does not imply that the structure is completely fixed on a local scale; such a fixation would require order-l specific interactions between units of primary structure.) Since each chain part of contour length l forms a globular crumple of spatial size $l^{1/3}$, it has a surface area of order $l^{2/3}$. Larger-scale structures are informed as the result of surface interactions between the crumples (interpenetration does not occur!). Therefore, the number of units which participate in the fixation of spatial structure on any arbitrary scale l is of order $l^{2/3}$.

How does the above picture compare with experiment? Recently, there have been reports [6, 7] followed by a heated discussion [8-18] about the observation of long-range correlations in the primary structures of native eukaryotic DNA. These correlations were visualized and described quantitatively by means of the following «random walk» representation [6]. Let us denote the DNA sequence as B_t, B being the type of base at the position t along the DNA contour and map this sequence on the trajectory of a one-dimensional random walk in abstract space, $x(t)$, where x is the «coordinate» and t plays the role of «time». This random walk is defined as follows: the particle steps up, *i.e.* $x(t+1) = x(t) + 1$, if B_t is purine, and steps down, *i.e.* $x(t+1) = x(t) - 1$, if B_t is pyrimidine. The root-mean-square

displacement $r(l)$ of the trajectory during a «time» interval l,

$$r(l) = [\langle (x(t+l) - x(t))^2 \rangle]^{1/2} , \tag{1}$$

where $\langle ... \rangle$ means averaging over t, *i.e.* with respect to the position along the DNA contour, is then used as a quantitative measure of correlations along the DNA backbone.

It was found [6] that $r(l) \simeq l^{\alpha}$, where the «critical exponent» α is equal to 0.5 for intronless genes and 0.62 for intron-containing genes and non-transcribed regulatory elements (averaged over 24 sequences selected across the phylogenetic spectrum). The observation of a Gaussian random walk exponent $\alpha = 0.5$ for the intronless genes is not surprising, since these sequences are directly related to the primary structures of proteins which, in turn, are known to be nearly random (more precisely, each of them looks like a typical and only slightly «edited» [19] sample from the ensemble of random non-correlated sequences). The most important and unexpected discovery of ref. [6] is the non-trivial value of $\alpha = 0.62$ for intron-containing genes. The deviation from the «trivial» value of 0.5 reflects a long-range scale-independent (self-similar) property [20] of the primary structure. What could be the reason for these correlations in DNA sequences which appear not to code for proteins?

A plausible explanation is that the self-similarity of intron-containing parts of DNA is responsible for its spatial («tertiary») arrangement, *i.e.*, in our language, for the fixation of a particular realization of crumpled-globule–type structure. Although this hypothesis appears to contradict the traditional point of view which asserts the geometry and volume interactions of DNA double helix are almost independent of sequence, there is increasing evidence that DNA properties include many sequence-dependent phenomena, such as non-canonical structures, triplexes and even the recently described quadruplexes [21]. Since these parts can interact with each other, one has to conclude that the volume interactions of DNA parts and, therefore, the large-scale spatial structure of DNA are strongly influenced by the sequence.

Coming back to our crumpled-globule model of native DNA, we observe that in any part of length l of DNA there should be units of order $l^{2/3}$ of the primary structure which govern the fixation of spatial structure on this scale. To do so, they have to obey some rules and, therefore, they cannot be random. If they correspond to one particular direction of the $x(t)$ process (see eq. (1) above), then they would give a contribution of order $l^{2/3}$ to the value $r(l)$. Assuming a random distribution of the other monomers, we estimate their contribution to $r(l)$ in accordance with «square-root law» as $(l - l^{2/3})^{1/2}$. Since, for $l \gg 1$, this contribution is much smaller than the former, we obtain the result

$$r(l) \sim l^{2/3} . \tag{2}$$

This result is in excellent agreement with the reported value [6] $\alpha = 0.67 \pm 0.01$ for the human β-cardiac myosin heavy-chain gene, and in reasonable agreement with exponents (in the range $0.6 \div 0.7$) obtained for other sequences across the phylogenetic spectrum which contain a high percentage of introns.

We now proceed to analyse further consequences of the application of the crumpled-globule concept to DNA. Our model relates the displacement of a trajectory in an abstract space, $r(l) = x(t+l) - x(t)$, which characterizes the sequence, to the surface area $s(l)$ of a crumple formed in real space by a segment of DNA ranging from monomer number t to $t + l$:

$$|r(l)| = s(l)/\sigma , \tag{3}$$

where σ is the effective surface area per «important» (structure-controlling) contact between two units. It is possible to analyse the probability distribution of $r(l)$ and to predict the surface area and shape fluctuations of native DNA. Indeed, the surface area of a crumple can be written as

$$s(l) = [l/\sigma]^{2/3} \sigma , \tag{4}$$

where ρ is the density of DNA's spatial structure and ϕ is a shape factor. It is clear from the latter expression that the ratio $s(l)/l^{2/3}$ is scale independent and, therefore, that the probability distribution of $s(l)$ has to be of the scaling form

$$P_l(s) = P(s/l^{2/3}). \tag{5}$$

Furthermore, the form of the universal function $P(\xi)$ can be established using the following argument.

The spatial structure of the DNA globule, including the density ρ and shape factor ϕ distributions for different crumples, is governed by protein-mediated volume interactions of segments. The globular structure of any polymer is controlled by competition of attractive interactions of the two-body type and of three-body and higher-order repulsions. It is generally believed that the latter are non-specific and only play the role of the usual excluded-volume forces which prevent the catastrophic collapse of the system to an infinitely dense state, so that the only source of specificity are the two-body attractions (this conjecture becomes exact in the limit of large spatial scales [2]). Therefore, only binary interactions such as histone-mediated DNA folding around nucleosomes and the subsequent organisation of nucleosomes into solenoids are responsible for the fixation of the crumpled structure and, accordingly, only two-body correlations in the DNA sequences are necessary for the control of its spatial structure. Statistically, it means that all the higher correlations along the chain can be expressed in terms of the pair correlation function. It can be shown that the sum of *dependent* contributions, like $r(l)$ in our case, with *only pair* correlations is Gaussian, *i.e.* $P_l(r) \sim \exp[-(1/2)\cdot r^2/\langle r^2\rangle]$, where in the case of *long-range* correlations $\langle r^2\rangle$ appears to be a *non-linear* function of l. Taking into account relation (3) and comparing it with the scaling form (4), we obtain

$$P_l(r) = \text{const} \cdot \exp[-cr^2/l^{4/3}], \tag{6}$$

where c is a constant. A distribution of this form was recently found experimentally [17], with c ranging from 0.3 to 1, for the genes studied. The agreement between the predicted and the observed distributions is significant because of the unusual character of the Gaussian distribution which has a mean square that does not scale linearly with the number of steps l.

We would like to end with concluding comments which pertain to both parts of our work, *i.e.* to the crumpled-globule model of native-DNA 3D structure (1)-4) below) and to the relation between this model and the observed long-range correlations in DNA sequences (5) below):

1) Our claim that long native DNA has to be in an unknotted crumpled-globule state has important consequences from the evolutionary perspective. Since the self-similarity of the crumpled globule is dictated by geometry (rather than by biological function), this structure can evolve in a simple and highly efficient parallel manner following simple laws which are similar to the constructive algorithms used in computer generation of fractal pattern [20]. In this way, crumples on one scale will automatically combine into similar crumples on a larger scale, etc. This implies that primary sequence and spatial structure were created together, as a result of a self-similar evolution process.

2) There is an analogy between the present problem and that of protein folding [22]. The similarity between native DNA and proteins stems from the fact that both form globules with well-defined sequence-controlled spatial structure. However, unlike the folded protein, the DNA globule is not an equilibrium structure. Furthermore, while in proteins the primary sequence determine the spatial structure in a thermodynamic sense, we believe that the sequence and three-dimensional structure evolved in parallel in native DNA.

3) While direct measurements of fractal exponents ($n^{1/3}$ for crumpled *vs.* $n^{1/2}$ exponent for an equilibrium globule) may be too difficult, our claim that native DNA is practically unknotted

can be tested by carefully removing the DNA molecule from the cell into a *good* solvent (1) and monitoring its unfolding into a random coil configuration. Such a process cannot take place on experimentally accessible time scales if native DNA is a heavily knotted, glassy globule.

4) The value of the fractal exponent is also of importance for understanding the question of holes in chromosome structure [5].

5) Our model associates the observed long-range order in intron-containing eukaryotic genes with the fixation of the crumpled spatial structure. Accordingly, the absence of long-range order in intronless DNA should lead to a poorly fixed, fluctuatin structure of crumples, as is probably the case in prokaryotic DNA. Moreover, it appears plausible that the evolutionary role of introns, in general, is the control of the spatial structure of native DNA. If correct, this implies that the reported long-range correlations can be extrapolated to the presently unassigned parts of eukaryotic DNA which should be statistically similar to intron-containing genes. On the other hand, the observation of short-range correlations in presently unassigned parts could indicate that these parts either code for proteins or did so in earlier stages of evolution.

* * *

We have benefitted from helpful discussions with E. SHAKHNOVICH, A. GUTIN, S. BOULDYREV and G. STANLEY. This work was supported in part by Basic Research Foundation of the Israeli Academy grant 585/92 and by NIH grant GM39372.

(1) A good solvent is defined here as a solvent which will remove proteins such as histones which participate in fixation of the spatial structure of DNA.

REFERENCES

[1] SIKORAV J.-L. and CHURCH G. M., *J. Mol. Biol.*, **222** (1991) 1085.
[2] GROSBERG A. YU. and KHOKHLOV A. R., *Statistical Physics of Macromolecules* (Nauka Publishers, Moscow) 1989.
[3] GROSBERG A. YU., NECHAEV S. K. and SHAKHNOVICH E. I., *J. Phys. (Paris)*, **49** (1988) 2095.
[4] LEWIN B., *Gene 4* (Oxford University Press, Oxford) 1990; DARNELL J., LODISH H. and BALTIMORE D., *Molecular Cell Biology* (Scientific American Books, New York, N.Y.) 1990.
[5] TAKAHASHI M., *J. Theor. Biol.*, **141** (1989) 117.
[6] PENG C.-K., BULDYREV S. V., GOLDBERGER A. L., HAVLIN S., SCIORTINO F., SIMONS M. and STANLEY H. E., *Nature*, **356** (1992) 168.
[7] VOSS R. F., *Phys. Rev. Lett.*, **68** (1992) 3805.
[8] NEE S., *Nature*, **357** (1992) 450.
[9] MADDOX J., *Nature*, **358** (1992) 103.
[10] AMATO I., *Science*, **257** (1992) 747.
[11] PRABHU V. V. and CLAVERIE J.-M., *Nature*, **359** (1992) 782.
[12] LI W. and KANEKO K., *Nature*, **360** (1992) 635; *Europhys. Lett.*, **17** (1992) 655.
[13] MUNSON P. J., TAYLOR R. C. and MICHAELS G. S., *Nature*, **360** (1992) 636.
[14] CHATZIDIMITRIOU-DREISMAN C. A. and LARHAMMAR D., *Nature*, **361** (1993) 212.
[15] KARLIN S. and BRENDEL V., *Science*, **259** (1993) 677.
[16] GROSBERG A. YU., RABIN Y., HAVLIN S. and NEER A., *Biophysics (Moscow)*, **38** (1993) 75.
[17] PENG C.-K., BULDYREV S. V., GOLDBERGER A. L., HAVLIN S., SCIORTINO F., SIMONS M. and STANLEY H. E., to be published in *Phys. Rev. E*.
[18] BOROVIK A., GROSBERG A. YU. and FRANK-KAMENETSKII M. D., to be published.
[19] POROYKOV V. V., ESIPOVA N. G. and TUMANYAN V. G., *Mol. Biol. USSR*, **18** (1984) 541; PTITSYN O. B. and VOLKENSTEIN M. V., *J. Biomolec. Struct. Dynamics*, **4** (1986) 137; SHAKHNOVICH E. I. and GUTIN A. M., *J. Theor. Biol.*, **149** (1991) 537.
[20] MANDELBROT B. B., *The Fractal Geometry of Nature* (W. F. Freeman, San Francisco) 1982.
[21] FRANK-KAMENETSKII M. D., *Nature*, **356** (1992) 105.
[22] GUTIN A. M. and SHAKHNOVICH E. I., private communication.