

Universality in a DNA Statistical Structure

Mark Ya. Azbel^{*}

Department of Physics, Harvard University, Cambridge, Massachusetts 02138

(Received 3 November 1994)

A DNA autocorrelation function is a slightly and slowly modulated autocorrelation function of a random sequence. Its coarse grained root-mean-squared fluctuations are approximately homogeneous and equal to those of a random sequence. A DNA may be decomposed into a random sequence of white noise domains, which have different lengths and nucleotide concentrations, but a universal length scale. No long-range correlations are found in any of the studied DNA sequences.

PACS numbers: 87.10.+e, 05.40.+j, 07.05.-t, 72.70.+m

It is well known that a DNA consists of two types of base nucleotide pairs: adenine-thymine and guanine-cytosine. Denoting them by 0 and 1, one obtains a binary sequence. Early attempts at DNA structure studies preceded its sequencing and were related to DNA double-helix unwinding in a helix-coil transition [1]. When temperature increases, DNA unwinds stepwise in a hierarchy of specific domains [2]—"phases" with different concentrations of zeros, which form a DNA. Their relation to the DNA nucleotide structure allows one [3] to accurately determine their composition and situation in relatively short DNAs, to establish the DNA statistical structure in long DNAs, and to prove that the boundaries of the domains are close to those of genes and *mRNA* start. The study of the root-mean-squared fluctuation (RMSF) in the DNA nucleotide sequence at a segment of a given length yields the short-range order in DNA³ (see also Ref. [4]) and suggests a quantitative measure of randomness in DNA sequences and linguistic texts [3].

Successes in DNA sequencing allowed for a direct statistical analysis of DNA and the discovery [5–7] of the $1/f$ -like spectral components in the DNA correlation structure. Three groups [5–8] claimed the existence of long-range power law correlations in DNA. Peng *et al.* [8] suggested the dependence of RMSF on the segment length ℓ as a quantitative measure of the long-range correlation in DNA sequence and claimed its discovery in introncontaining genes and nontranscribed regulatory DNA sequences only—contrary to the Voss observations [7]. Some authors explained their results with the DNA repetitiveness [5] and patchiness [9,10]; others [11] emphasized its fractality.

In an infinite system the RMSF $\propto \ell^\alpha$ (when $\ell \rightarrow \infty$) with $\alpha > \frac{1}{2}$ does indeed imply long-range correlations. However, all sequenced DNAs and their regions are $\sim 10^5$ base pairs at best [5–11]. They consist of the domains whose length, according to the DNA coiling curves, is $\sim 10^3$ base pairs [2–4]. Their concentrations of zeros are different from the DNA average. I demonstrate that each of these domains is (apart from possible short-range correlations) a white noise sequence, with accuracy allowed by its finite length. The domain lengths and concentrations are consistent with their random choice at

each stage of coiling, e.g., in a way studied in Ref. [11]. I speculate that qualitatively such statistical structure is universal for all DNAs, and is in fact characteristic of any informative text, e.g., a book (cf. Refs. [3,8]).

To elucidate the impact of DNA finite length and domain structure, consider a transparent case of a random sequence of homo-0 and homo-1 domains of the same length $\Lambda \sim 1000$ (base pairs). Clearly, their infinite sequence yields strong correlations at the length $\ell \leq \Lambda$, which exponentially vanish with ℓ/Λ when $\ell \gg \Lambda$. In a finite sequence of the total length ℓ_{\max} the autocorrelation function slightly fluctuates. Its fluctuations are $\sim \sqrt{\Lambda/\ell_{\max}} \ll 1$. They slowly change at the (large) distance $\sim \Lambda$, do not vanish at ℓ_{\max} , and thus simulate long-range correlations. They do not dominate only when $\ell_{\max} \gg \Lambda \exp(2\ell/\Lambda)$. If $\Lambda \sim 10^3$, $\ell \sim 10^4$, this implies that only when $\ell_{\max} \gg 10^{11}$ the RMSF asymptotic at $\ell \gg \Lambda$ allows one to rule out true long-range correlations. When the domains are white noise ones with different zero concentrations and characteristic length $\Lambda \sim 10^3$, the reasoning is similar. Then the autocorrelation function also reduces to a slightly ($\sim \sqrt{\Lambda/\ell_{\max}} \ll 1$) and slowly ($\sim \Lambda$) modulated autocorrelation function of a random sequence for all sequenced DNA lengths; the "scaling analysis" at $\ell \leq 1000$ may be misleading and a different approach (which is free of an *a priori* assumption and does not draw conclusions beyond the accuracy consistent with the finite length of a sequence) is called for.

Start with the longest (48 502 base pairs) "white noise" DNA sequence of bacteriophage λ (an intronless virus) from Ref. [8]. To eliminate the singularity of sites close to edges, (here and on) impose periodic boundary conditions: $\alpha(j) = \alpha(j + \ell_{\max})$, where $\alpha(j)$ is the digit at the j th site and ℓ_{\max} is the total number of sites (of course, this does not eliminate finite size effects). The DNA autocorrelation function $C(\ell)$,

$$C(\ell) = \left[\ell_{\max}^{-1} \sum_{j=1}^{\ell_{\max}} \alpha(j)\alpha(j + \ell) - (1 - x)^2 \right] / x(1 - x), \quad (1)$$

where x is the concentration of zeros at ℓ_{\max} , is presented in Fig. 1(A0). In virtue of periodicity $C(\ell_{\max} - \ell) =$

$C(\ell)$, and it is sufficient to consider $\ell \leq \ell_{\max}/2$. Clearly, $C(\ell)$ exhibits a long-range correlation which does not vanish with ℓ . (This naked eye observation is verified by the calculation of the coarse grained \bar{C} .) However, the value of C is very small (~ 0.01) and, except for $\ell \leq 100$, it is $\sim 1/\sqrt{\ell_{\max}}$, i.e., asymptotically equal to its fluctuation. The local coarse grained mean squared fluctuation $\Delta C(\ell)$ of C ,

$$\Delta C(\ell) = \left[\overline{(C - \bar{C})^2} \right]^{1/2}, \quad \bar{C} \equiv \sum_{j=\ell-s}^{\ell+s} C(j)/(2s + 1) \quad (2)$$

(I chose $s = 100$, which is $\gg 1$ but $\ll \ell_{\max}$), for $\ell \gg 100$ is approximately homogeneous and equal to that of a random sequence ($\approx 1/\sqrt{\ell_{\max}}$); see Fig. 2(A). Thus, the DNA $C(\ell)$ is a relatively slightly $\sim 1/\sqrt{\ell_{\max}}$ and slowly (at $\ell \gtrsim \sqrt{\ell_{\max}}$) modulated autocorrelation function of a white noise sequence. To understand the origin of such $C(\ell)$, plot [2,3] a DNA walk $y(\ell) = 2\ell(x_\ell - x)$, where x_ℓ is the purine concentration at the length ℓ , $1 \leq \ell \leq \ell_{\max}$; see Fig. 3(A0). The global minimum at $\ell = 21\,931$ separates two “first generation” domains with the concentration higher and lower than x . The larger domain [Fig. 3(A1)]

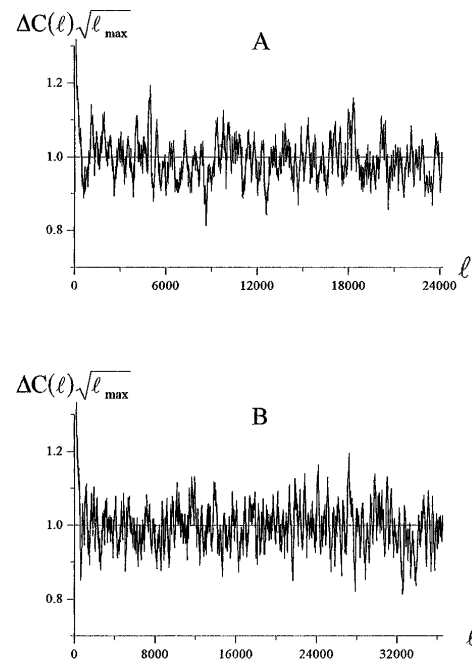


FIG. 2. Reduced coarse grained mean squared fluctuation $\Delta C\sqrt{\ell_{\max}}$ for bacteriophage λ DNA (A) and a human β -globin (B).

is separated by its global maxima and minima into the “second generation” domains [12]. The walks for the largest domains of the second, third, and fourth generations [Figs. 3(A2–A4)] look self-similar. The autocorrelation functions of the successive domain genera-

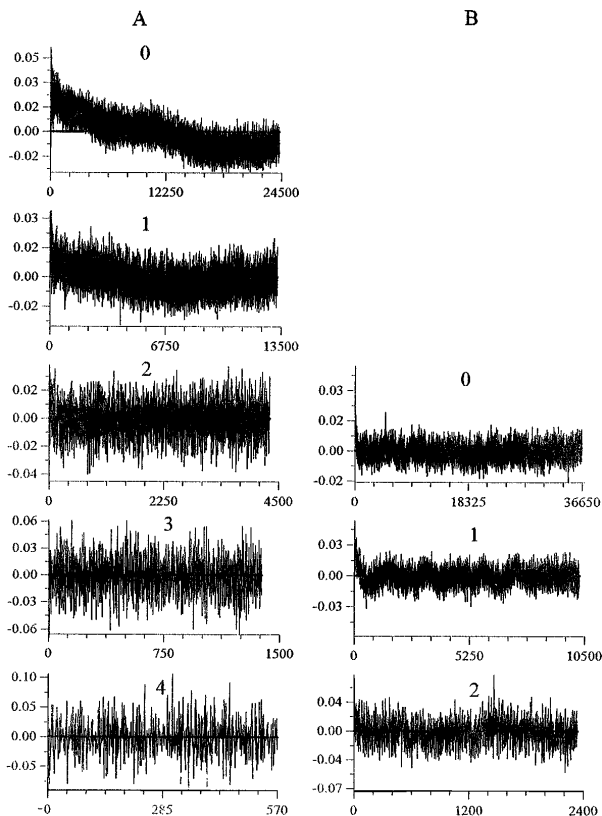


FIG. 1. DNA autocorrelation function $C(\ell)$ for (A) a bacteriophage λ (48 502 base pairs), (B) a human β -globin (73 323 base pairs), and their largest domains in successive generations [1–4 in (A), 1–2 in (B)]. The base pair numbers are (A1) 21 932–48 502, (A2) 38 388–47 031, (A3) 41 162–43 926, (A4) 41 913–43 053; and (B1) 47 274–66 774, (B2) 57 000–60 853.

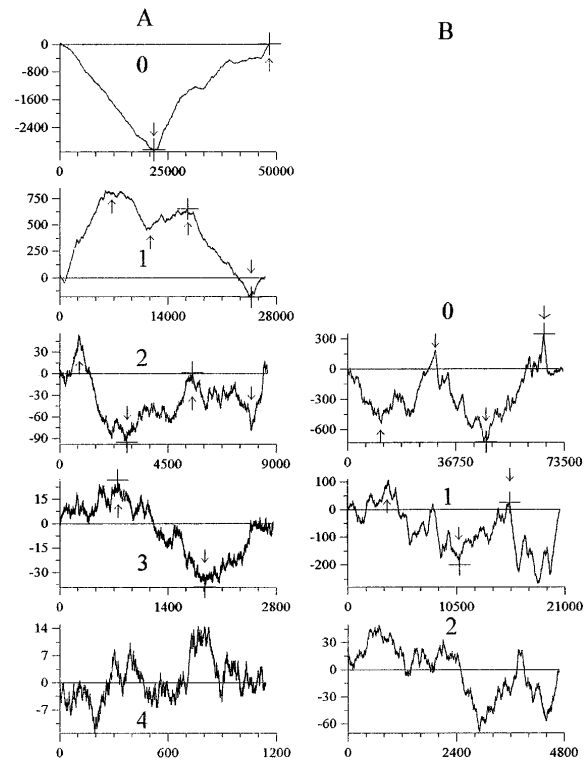


FIG. 3. The walks $y(\ell)$ for DNAs and domains from Fig. 1. Arrows denote the next generation domains; crosses are the largest domain boundaries.

tions converge to that of a white noise sequence; see Figs. 1(A1–A4)]. To describe the convergence quantitatively, reduce the noise in $C(\ell)$ by considering the RMSF [3] $F(\ell) = \{\ell_{\max}^{-1} \sum_{j=1}^{\ell_{\max}} [y(j+\ell) - y(j)]^2\}^{1/2}$ of a DNA ℓ -step walk. Accounting for $C(0) = 1$ and $C(-\ell) = C(\ell) = C(\ell + \ell_{\max})$, one obtains

$$[4x(1-x)]^{-1} F^2(\ell) = \ell + 2 \sum_{j=1}^{\ell-1} (\ell-j) C(j) \\ \equiv \ell + \ell(\ell-1) C^*(\ell-1), \quad (3)$$

where $C^*(\ell) = 2 \sum_{j=1}^{\ell} \sum_{q=1}^j C(q) / \ell(\ell+1)$ is the twice averaged $C(\ell)$. Since $C(\ell) \leq 0.01$ in Fig. 1(A), the second term may dominate when $\ell \gg 100$ or never [13]. So, consider $C^*(\ell)$. A white noise sequence yields $C^*(\ell) \leq 1/\sqrt{\ell\ell_{\max}}$, which slowly oscillates with ℓ . In the successive generations of domains the maximal value of $C^*(\ell)\sqrt{\ell\ell_{\max}}$ decreases [see Fig. 4(A)]. In the fourth generation it is ~ 1 , and in this sense the domains are indistinguishable from white noise ones. Their zero concentrations are relatively close [for instance, in Figs. 1(A0–A4), correspondingly $x = 0.5, 0.56, 0.51, 0.52, 0.5$]. The autocorrelation function of a DNA, composed of such long (~ 1000 base pairs) white noise sequences with close concentrations and thus weak boundary effects [14], looks like a modulated white noise $C(\ell)$. The lengths and concentrations in the fourth generation domains are consistent

with a random choice (for their given mean squared fluctuations) in each generation. This agrees with Ref. [11], whose authors generated examples of random fractal sequences which fit experimental “long-range correlation” plots. (They also studied patchiness, but only for homogeneous distributions of lengths and concentrations, with relatively large mean quadratic fluctuations.)

Now consider the longest (73 323 base pairs) and simultaneously the “highest long-range order” DNA from Ref. [8]—a human β -globin chromosomal region. Its $C(\ell)$, $y(\ell)$, $C^*(\ell)$ are presented in Figs. 1(B), 3(B), and 4(B). They are *closer* to a white noise sequence than those of a (white noise, according to Ref. [8]) bacteriophage λ : The walks are self-similar from the very beginning, $|C^*(\ell)\sqrt{\ell\ell_{\max}}|$ in the same generation is less, and white noise domains are those in the second (rather than the fourth) generation. (This conclusion is consistent with Voss [7].) There is no regularity in the slopes of $\log C^*(\ell)$ vs $\log \ell$ for different domains and DNAs; see Fig. 4. Moreover, $C^*(\ell)$ often has an extremum at $\ell \sim 100$. (This is related to an insufficient averaging for such ℓ_{\max} and usually occurs in white noise sequences also.) Sometimes [in Figs. 4(A2), 4(A3), and 4(B2)] $C^*(\ell)$ changes its sign in the vicinity of $\ell_{\max}/2$ [where the finite value of ℓ_{\max} is very pronounced and where $C^*(\ell) \leq 1/\ell_{\max}$].

The application of the above analysis to the “second highest order” DNA from Ref. [8] (human β -cardiac

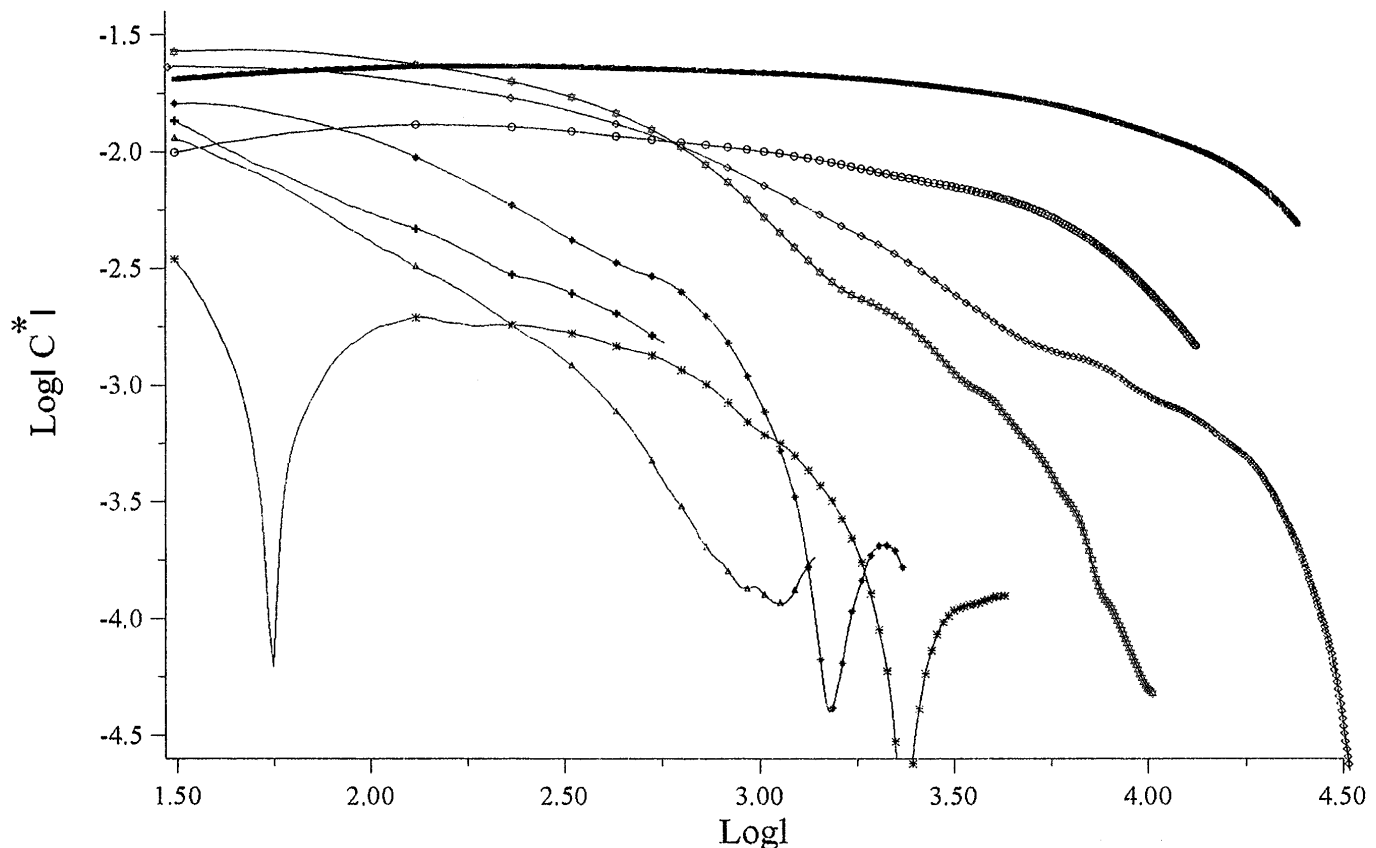


FIG. 4. The averaged autocorrelation function $C^*(\ell)$ vs $\log \ell$ for A1 (—), A2 (○), A3 (*), A4 (△), A5 (+) and B1 (◇), B2 (*), B3 (*).

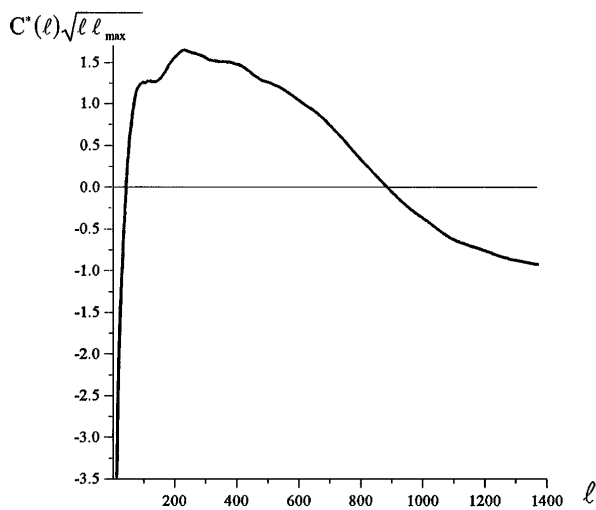


FIG. 5. $C^*(\ell)\sqrt{\ell_{\max}\ell}$ vs ℓ for a white noise domain 7900–12 3000 of the human β -cardiac (28 437 base pairs) DNA.

DNA, 28 437 base pairs), as well as to all other DNAs with $\ell_{\max} \gg 1000$, yields similar results. A typical $C^*(\ell)\sqrt{\ell_{\max}\ell}$ plot for a white noise domain (the sites 7900–12 300) of the β -cardiac DNA is presented in Fig. 5; note $C^*\sqrt{\ell_{\max}\ell} \leq 1$.

The main problem is the number (~ 50) of domains. Only sequencing of much longer DNA sequences may comprehensively establish the randomness of their lengths and concentrations. Then the study may be extended to a sequence of nucleotides rather than of their pairs. Also, very long DNAs may demonstrate true long-range correlations, significantly higher than the fluctuations [3].

To summarize, all studies DNA sequences exhibit a certain universality in their statistical structure. They are composed of random length and concentration white noise domains of the length ~ 1000 base pairs. The DNA autocorrelation functions reduce to a slightly and slowly modulated autocorrelation function of a white noise sequence. This modulation is related to the domain inhomogeneity of the DNA sequence rather than to long-range correlations. It is different in different DNAs and may be quantitatively related to the asymptotics of certain DNA statistical characteristics with length ℓ (provided that the ℓ scale is adequately chosen to indeed yield asymptotics rather than a linear interpolation of a relatively short region of, for example, Fig. 4 on a log-log scale), which were studied in Refs. [5–8,11].

I am grateful to T. Vasilevsky, L. Garber, M. Kheifits, and G. Braverman for numerical calculations and to Professor B.I. Halperin for his hospitality at Harvard. Financial support from J. and R. Meyerhoff chair is appreciated.

*On a sabbatical from Tel-Aviv University, Israel

- [1] D. Poland and H.A. Scheraga, *Theory of Helix-Coil Transition in Biopolymers* (Academic, New York, 1970); R.W. Wartell and E.W. Montroll, *Adv. Chem. Phys.* **22**,

129 (1972); Yu.S. Lazurkin, M.D. Frank-Kamenetskii, and E.N. Trifonov, *Biopolymers* **9**, 1253 (1970).

- [2] M. Ya. Azbel', *Phys. Rev. Lett.* **31**, 589 (1973); **31**, 1592(E) (1973); I.M. Lifshitz, *Sov. Phys. JETP* **38**, 545 (1974). At low temperatures, DNA is a helix. Then some domains coil. At higher temperatures, certain domains at helix regions coil, while few domains at coiled regions become helix again. This is followed by the hierarchy of the third, fourth, etc. generation domains, whose concentrations of zeros progressively increase at coiled and decrease at helix domains. Finally, DNA is completely coiled at sufficiently high temperature. At any temperature domains are well defined [3] because the nucleotide pair binding energy is ~ 3000 K, the coiling temperature is ~ 300 K, the difference between 0- and 1-homopolymer cooling temperatures is ~ 30 K, and its change related to nearest neighbor interactions is ~ 3 K. In fact, the helix-coil transition may still be helpful for a statistical analysis of very long unsequenced DNAs [3].
- [3] M. Ya. Azbel', *Phys. Rev. A* **20**, 1671 (1979); *Biopolymers* **19**, 61 (1980); **19**, 95 (1980); **19**, 1311 (1980); *Proc. Natl. Acad. Sci. U.S.A.* **76**, 101 (1979); M. Ya. Azbel' *et al.*, *Biopolymers* **21**, 1687 (1982); A. Vilenkin and L. Verkh, *Biopolymers* **21**, 1691 (1982).
- [4] A. Wada, H. Tachibana, O. Goton, and M. Takenami, *Nature (London)* **263**, 439 (1976); M.D. Frank-Kamenetskii, *Nature* **269**, 729 (1977); S. Tavaré and B.W. Giddings, in *Mathematical Methods for DNA Sequences*, edited by M.S. Waterman (CRC Press, Boca Raton, 1989), p. 117.
- [5] W. Li and K. Kaneko, *Europhys. Lett.* **17**, 655 (1992); *Nature (London)* **360**, 635 (1992).
- [6] W. Li, *Int. J. Bifurcation Chaos* **2**, 137 (1992).
- [7] R. Voss, *Phys. Rev. Lett.* **68**, 3805–3808 (1992); **71**, 1777 (1993).
- [8] C.K. Peng *et al.*, *Nature (London)* **356**, 168 (1992); S.V. Buldyrev *et al.*, *Phys. Rev. E* **47**, 3730 (1993); *Phys. Rev. Lett.* **71**, 1776 (1993); C.K. Peng *et al.*, *Phys. Rev. E* **49**, 1685 (1994); **49**, 3730 (1994); **49**, 4514 (1994); R.N. Mantegna *et al.*, *Phys. Rev. Lett.* **73**, 3169 (1994); S.V. Buldyrev *et al.*, *Biophys. J.* **65**, 2673 (1993); **67**, 64 (1994).
- [9] S. Nee, *Nature (London)* **357**, 450 (1992); W. Li, T.G. Marr, and K. Kaneko, *Physica (Amsterdam)* **75D**, 392 (1994), and references therein.
- [10] S. Karlin and V. Brendal, *Science* **259**, 677 (1993).
- [11] A.S. Borovik, A.Yu. Grosberg, and M.D. Frank-Kamenetskii, *J. Biomol. Struct. Dynam.* **12**, 655 (1994).
- [12] The domains depend on the chosen minimal value of the domain $y(\ell)$ change. The biologically meaningful hierarchy of coiling domains [2,3] is close to that in Fig. 3.
- [13] Thus, the approximation $\ln F \propto \ln \ell$ for $\ell \leq 1000$ in Ref. [8] is misleading; see also Ref. [10].
- [14] Consider, for instance, ℓ small compared to the white noise domain size. Neglecting boundary effects,

$$C(\ell) \approx \frac{\sum \lambda_m x_m^2 / \ell_{\max}}{\ell_{\max}^{-1} \sum \lambda_m x_m^2} - \frac{(\sum \lambda_m x_m / \ell_{\max})^2}{(\ell_{\max}^{-1} \sum \lambda_m x_m)^2},$$

where λ_m and x_m are the m th domain length and purine concentration; $\ell_{\max} = \sum \lambda_m$.