PHYSICAL REVIEW E, VOLUME 64, 040903(R)

Primary sequences of proteinlike copolymers: Levy-flight–type long-range correlations

Elena N. Govorun,[1] Victor A. Ivanov,[1] Alexei R. Khokhlov,[1] Pavel G. Khalatur,[2] Alexander L. Borovinsky,[3] and Alexander Yu. Grosberg[3,4]

[1]*Physics Department, Moscow State University, Moscow 119899, Russia*

[2]*Department of Physical Chemistry, Tver State University, Tver 170002, Russia*

[3]*Department of Physics, University of Minnesota, Minneapolis, Minnesota 55455*

[4]*Institute of Biochemical Physics, Russian Academy of Sciences, Moscow 117977, Russia*

(Received 20 November 2000; revised manuscript received 5 July 2001; published 25 September 2001)

We consider the statistical properties of primary sequences of two-letter *HP* copolymers (*H* for hydrophobic and *P* for polar) designed to have water soluble globular conformations with *H* monomers shielded from water inside the shell of *P* monomers. We show, both by computer simulations and by exact analytical calculation, that for large globules and flexible polymers such sequences exhibit long-range correlations which can be described by Levy-flight statistics.

DOI: 10.1103/PhysRevE.64.040903 PACS number(s): 87.15.Cc, 61.41.+e, 36.20.Ey

In Refs. [1,2], a new approach to the design of specific primary sequences for the *HP*-copolymers consisting of monomeric units of two types (hydrophobic, *H* and polar, *P*) has been proposed by some of the present authors. Unlike some other methods of sequence design known in the literature (see review [3] and references therein), the approach in question does not aim to mimic folding into one particular conformation. The goal is to model simpler and more robust property of proteins, such as their ability to stay dissolved and shield their hydrophobic monomers from water. The essence of this approach is illustrated in Fig. 1. We start with an arbitrary computer-generated globular conformation of a homopolymer chain [formed due to the strong attraction of monomer units, Fig. 1(a)] and perform a ''coloring'' procedure: monomer units in the core of the globule (having many neighbors) are set to be *H*-units while monomer units belonging to a globular surface (where the number of neighbors is smaller) are assigned to be of *P*-type, Fig. 1(b). Then the obtained primary sequence is fixed, uniform attraction of monomer units is removed and newly generated *HP*-copolymer is ready for the further investigation [Fig. 1(c)]. Thus, obtained macromolecules are *proteinlike* in the sense that they mimic segregation of globule into hydrophobic core and stabilizing hydrophilic envelope. The properties of proteinlike copolymers were examined in Refs. [1,2,4,5]; see also Refs. [6,7] for possible ways of experimental realization.

In this Rapid Communication, we address correlations between *H*- and *P*-units along the proteinlike sequences. This may shed light on the conditions which must be met by the sequence to provide for the water solubility of globules, the issue of great potential relevance to our understanding of early evolution. We show, both by computer simulations and by exact analytical calculation, that correlations have a long-range character. More specifically, for the simple model of flexible polymer, they belong to the so-called Levy-flight statistics.

To begin with, statistical properties of proteinlike *HP*-sequences can be assessed computationally by the method similar to that used by Stanley and co-workers [8,9] in their search for long-range correlations in DNA sequences. We choose the ''window'' of length ℓ, move it step by step along the generated *HP*-sequence, and at each step count the number of *H* units inside the window. This number, which we write as $\Sigma_{i=j}^{j+\ell} u_i$ is a random variable, depending on the position j of the window along the sequence; here u_i is the variable associated with every monomer i, such that $u_i=1$ if monomer i is *H* and $u_i=0$ if it is *P*. This random variable has certain distribution. Its average is determined by the overall sequence composition (total numbers of *H*- and *P*-monomers), and its dispersion is easy to calculate:

$$D_\ell^2=\sum_{i,j=k}^{k+\ell}(\langle u_i u_j\rangle-\langle u_i\rangle\langle u_j\rangle). \tag{1}$$

For a completely random *HP*-sequence, the value of D_ℓ scales as $\ell^{1/2}$ with the window width ℓ. The dependence $D_\ell\sim\ell^\alpha$ with $\alpha>1/2$ would then manifest the existence of long-range correlations.

The result of such calculation for averaging over 2000 independent proteinlike *HP*-sequences of $N=1024$ monomer units with 1:1 composition (obtained as in Ref. [2]) is presented by the squares in Fig. 2. For comparison, the data for two other types of sequences (averaged over 2000 independent species) are also shown. One of them is a purely random 1:1 sequence; it demonstrates $D_\ell\sim\ell^{1/2}$ scaling. Comparing this curve with Monte Carlo results we see immediately that the proteinlike sequence is not random and some correlations do exist in it. Thus, it is interesting to compare the squares in Fig. 2 and the dashed curve showing data for the sequence which was called ''random-block'' in Refs. [1,2,4,5]: the lengths of *H*- and *P*-blocks in a sequence

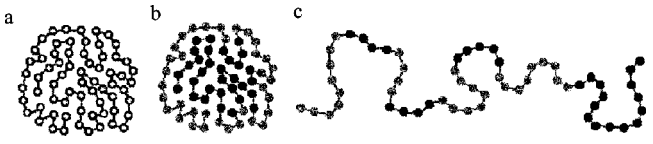

FIG. 1. Sequence design scheme for proteinlike copolymers: (a) homopolymer globule; (b) the same globule after the coloring procedure; (c) proteinlike copolymer in the coil state.

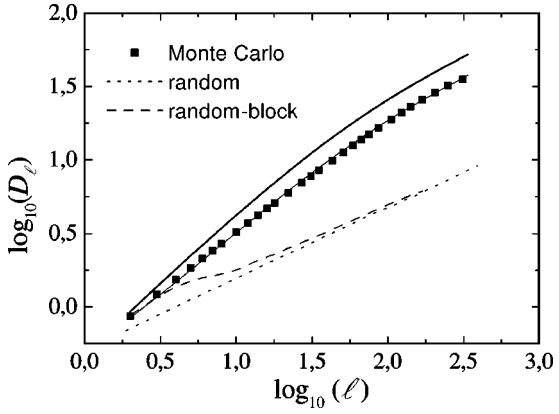

FIG. 2. Dispersion of the number of *H*-units in the fragment of sequence of size ℓ for the proteinlike *HP*-sequence, random copolymer and random-block copolymer. Results of analytical theory for proteinlike sequence are shown for both continuous approximation (thick solid line) and for discrete approximation (thin solid line) [see explanation around Eq. (11)]. Corresponding Monte Carlo results are presented by squares. There is no adjustable parameters involved in the fit, length scale a is uniquely determined by the geometry of the bond fluctuation model [13].

are determined by the Poisson distributions adjusted to achieve the same 1:1 composition and the same ''degree of blockiness'' (average block length) as for a proteinlike *HP*-copolymer. This sequence exhibits a somewhat more rapid variation of D_ℓ at small ℓ, but ultimately the law $D_\ell \sim \ell^{1/2}$ is obeyed for large values of ℓ. Nevertheless, this random-block model is also seen to be unsatisfactory for the statistical behavior of a proteinlike sequence throughout the interval of ℓ examined, $2<\ell<500$. Although the data do not fit accurately to any power law $D_\ell \sim \ell^\alpha$, the slope of the observed D_ℓ dependence corresponds to α significantly larger than 1/2, up to about 0.85, thus indicating pronounced long-range correlations in a proteinlike sequence. In what follows we present an analytical theory which produces curve D_ℓ in Fig. 2 in complete agreement with observations.

First of all, let us turn to the origin of long-range correlations in the primary sequences for *HP*-copolymers generated via the procedure illustrated in Fig. 1. Conceptually, this problem is fairly easy to address: since sequence in this scheme is uniquely determined by the parent conformation, the statistics of sequences reflects nothing but the statistics of parent conformations, which, in turn, is well understood. Indeed, the coloring procedure [Fig. 1(b)] operates in dense globular conformation. Since we consider very compact parent conformations, the statistics of polymer chain conformations inside the globule is ideal (Gaussian) according to the well-known Flory theorem [10]. Therefore, all the statistical properties of parental conformations, including the correlations in the primary sequences produced by coloring procedure, can be derived via the solution of diffusion equation for random walks with appropriate boundary conditions.

To understand the fractal aspect of the sequences, it is convenient to concentrate on their uninterrupted homocolored sections and on the points of connection between them. Our coloring procedure (Fig. 1) introduces a separation sphere of radius $R^*<R$, such that all the units which in the parent conformation are confined inside this sphere are of the *H*-type and the units belonging to the shell layer $R^*<r<R$ are of the *P*-type. Therefore, the homocolored section is produced in our model by the chain section of the parent conformation placed entirely in either internal or shell regions of the globule. The probability to have an uninterrupted succession of some k of *H*-monomer units in the sequence is equal to the probability that the Gaussian (due to Flory theorem) polymer has a loop of k monomer units entirely confined in the *H*-region with ends on the separation surface. Similarly, the probability to have an uninterrupted succession of some k of *P*-monomer units in the sequence is equal to the probability that ideal parental conformation has a loop of k monomer units confined within the shell *P*-region, again with ends on the separation surface.

To address probability distributions $P_H(k)$ and $P_P(k)$, we begin with simple physical arguments yielding

$$P_{H,P}(k) \simeq \begin{cases} k^{-3/2}, & 1<k<\left(\dfrac{d_{H,P}}{a}\right)^2; \\ \left(\dfrac{a}{d_{H,P}}\right)^3 e^{-\lambda_{H,P}k}, & k>\left(\dfrac{d_{H,P}}{a}\right)^2. \end{cases} \tag{2}$$

The upper asymptotic form is valid for short polymer loops, when neither curvature of the separating boundary nor overall globule shape play any role. In this regime, $P(k)$ is simply the probability for a random walk to start at the planar wall and to return to it for the first time after k steps. This is a classical probabilistic ''first return'' problem, for which the $\sim k^{-3/2}$ answer is well known [12]. This scaling is valid for loop sizes $a\sqrt{k}$ much larger than unity but smaller than the relevant characteristic length scale, $d_H=R^*$ for the *H*-loops inside the inner sphere, or $d_P=R-R^*$ for the *P*-loops in the spherical shell. The second asymptotic form in Eq. (2) indicates that for long polymer loops the function $P(k)$ decays exponentially. It is easier to explain this in terms of polymer statistics: to confine a polymer chain of k monomer units in a cavity costs some entropy ΔS, at $a\sqrt{k} \gg d$ this entropy goes linearly with k, making the probability, $\exp(\Delta S)$, exponential in k.

Let us now look closer at the cross-over values of k. In order to achieve the 1:1 composition, R^* must be chosen such that volumes of internal *H*- and shell *P*-regions are the same, which means $R^*=2^{-1/3}R\approx 0.8R$. The volume fraction of polymer units in a globule, ϕ, is controlled by the energy of interactions of monomer units used to prepare parental conformation. It is clear that $R\approx 0.6aN^{1/3}\phi^{-1}$. Therefore, *H*-loops remain in the power law long-range correlation regime up to the length $k<0.24N^{2/3}\phi^{-2}$, while for *P*-loops this cross-over occurs somewhat earlier: $k<0.015N^{2/3}\phi^{-2}$. Thus, we predict that there should be over a decade of length scales in which *H*-loops are still long-range correlated, while only short-range correlations remain in the *P*-loops.

Result (2) is sufficient to explain qualitatively correlations in proteinlike sequences, including the data shown in Fig. 2. Indeed, according to our discussion, a proteinlike sequence can be thought of as an alternating succession of *H*- and

P-stretches, with lengths of stretches taken independently from the corresponding distributions $P_H(k)$ and $P_P(k)$. This mathematical scheme is called a Levy flight [11]. We conclude that the long-range correlations in the primary sequences of proteinlike copolymers are described by Levy flight statistics. Furthermore, for the $k^{-3/2}$ behavior of $P(k)$, the averaged block length diverges, and, therefore, the value of D_ℓ in the power law regime is controlled by the longest block, yielding $D_\ell \sim \ell$, or $\alpha=1$. This is true as long as both *H*- and *P*-loops remain in fractal regime. On the other hand, when all loops cross-over to exponential distribution, D_ℓ crosses over to $\alpha=1/2$:

$$D_\ell \simeq \begin{cases} \ell, & \text{for } 1<\ell<0.015N^{2/3}\phi^{-2} \\ \ell^{1/2}, & \text{for } \ell>0.24N^{2/3}\phi^{-2}. \end{cases} \tag{3}$$

The cross-over region for D_ℓ is very broad, it corresponds to the situation in which *P*-loops are already ''large,'' while *H*-loops are still ''small.'' Both $\alpha=1$ and $\alpha=1/2$ limits and wide cross-over agree qualitatively well with computational data, Fig. 2. This motivates more careful theory, in which instead of scaling estimates (2) and (3), the expressions in terms of infinite series, suitable for numerical calculation, is obtained.

To develop full analytical theory, it is convenient to use the random walk terminology to describe parent conformation. In this language, for instance, $P_H(k)$ is the probability that the random walker enters a sphere of the radius R^* and then arrives back to the boundary *for the first time* after ''time'' *k*. Recall that the statistical weight of all random walk trajectories starting at the point $\vec{r}_0$ and arriving after *k* steps at the point $\vec{r}$, $G(\vec{r},k|\vec{r}_0)$, obeys the diffusion equation

$$\frac{\partial G(\vec{r},k|\vec{r}_0)}{\partial k}=\frac{a^2}{6}\Delta G(\vec{r},k|\vec{r}_0)+\delta(k)\,\delta(\vec{r}-\vec{r}_0), \tag{4}$$

where a^2 is the mean square length of one step (the squared size of one monomer unit along the chain). To introduce the condition of first return we have to say that the walker never touches the boundary, which is achieved by imposing the boundary condition

$$G(\vec{r},k|\vec{r}_0)|_{|\vec{r}|=R^*}=0. \tag{5}$$

The probability distribution of the ''first return times'' in terms of *G* is then given as the time-dependent flux of diffusing particles through the absorbing wall:

$$P_H(k)=\left| \oint d\sigma \frac{a^2}{6}\frac{\partial G}{\partial r}\bigg|_{r=R^*} \right|, \tag{6}$$

where $\partial/\partial r$ means the component of gradient normal to the surface, integration is performed over the closed separating surface, and the absolute value is written to avoid thinking about the direction of the flux. The normalization condition $\int P_H(k)dk=1$ is guaranteed by the fact that all diffusing particles eventually leave through the surface. As regards $\vec{r}_0$, it should be taken within a distance of order *a* from the separating R^* surface. The problem thus formulated, including Eqs. (4)–(6), is easy to solve: we write *G* in terms of bilinear expansion $G=\Sigma_n e^{k\lambda_n}\psi_n(\vec{r})\psi_n(\vec{r}_0)$ over the eigenfunctions ψ_n satisfying $(a^2/6)\Delta\psi_n=\lambda_n\psi_n$ with boundary condition (5). Upon spherical integration in Eq. (6), all angular dependent harmonics vanish, and we arrive at

$$P_H(k)=\frac{\pi a^2}{3R^* r_0}\sum_{n=1}^{\infty} n(-1)^{n+1}\sin\left(n\pi\frac{r_0}{R^*}\right) \times\exp\left[-\frac{a^2}{6}\left(\frac{n\pi}{R^*}\right)^2 k\right]. \tag{7}$$

The distribution $P_P(k)$ can be derived similarly, except that now we have to take care of the boundary condition at the outer surface of the globule. To this end, we argue that this condition must be taken in the form

$$\nabla_{\vec{r}} G(\vec{r},k|\vec{r}_0)|_{r=R}=0. \tag{8}$$

Indeed, formally this condition ensures the constant density of monomer units throughout the globule for large values of *k*, as well as breaking of correlations as soon as polymer chain is ''reflected'' by a globular boundary. Physically, this boundary condition reflects the fact that there is always a ''sticky layer'' (or depletion layer) formed self-consistently along the internal surface of the globule due to the effective attraction of monomer units to the outer region where polymer density is depleted and excluded volume effect is reduced. As long as we are not interested in the structure of surface layer of the globule, we can just replace this layer by effective boundary condition (8). After calculations for $P_P(k)$ we obtain

$$P_P(k)=\frac{a^2R^*}{3(R-R^*)^2 r_0}\sum_{n=1}^{\infty}\frac{\zeta_n^2\sin\left(\zeta_n\dfrac{r_0-R^*}{R-R^*}\right)}{\zeta_n-\sin\zeta_n\cos\zeta_n} \times\exp\left[-\frac{a^2}{6}\left(\frac{\zeta_n}{R-R^*}\right)^2 k\right], \tag{9}$$

where ζ_n satisfies $\zeta_n=(1-R^*/R)\tan\zeta_n$.

Finally, to compute the dispersion D_ℓ, we note that u_iu_j in Eq. (1) is the probability that both units *i* and *j* are of the *H* type, which happens if both are located inside the R^* region in the parental conformation. Thus,

$$\langle u_iu_j\rangle=\frac{1}{V}\int_{V^*}d^3r\int_{V^*}d^3r_0 G(\vec{r},|i-j||\vec{r}_0) =\frac{1}{V}\sum_{n=0}^{\infty}e^{\lambda_n|i-j|}\left(\int_{V^*}d^3r\,\psi_n(\vec{r})\right)^2, \tag{10}$$

where $G(\vec{r},k|\vec{r}_0)$ is the Green function satisfying Eq. (4) with boundary condition (8), ψ_n and λ_n are the corresponding eigenfunctions and eigenvalues. Plugging this into Eq. (1), one arrives at the cumbersome looking expression for D_ℓ which is easy to implement numerically; the result is

plotted in Fig. 2 and shows virtually a perfect fit to the Monte Carlo data [13]. In fact, this fit may even be somewhat fortuitous; indeed, along with diffusion equation (4), which is an approximation for the underlying bond fluctuation model, we can also switch from summation to integration in Eq. (1), yielding

$$D^2_{\ell} = 6\ell^2 \sum_{n=1}^{\infty} g\left(\frac{\xi_n^2 a^2}{6R^2}\ell\right)\alpha_n , \tag{11}$$

where ξ_n satisfies the equation $\xi_n = \tan \xi_n$, $\alpha_n = [(\xi_n R^*/R^*)\cos(\xi_n R^*/R^*) - \sin(\xi_n R^*/R^*)]^2(1+\xi_n^2)/\xi_n^6$, and the function g is defined as $g(x) = 2[x - 1 + \exp(-x)]/x^2$. [Note that the sum in Eq. (11) starts from $n=1$ and does not include the ground state, for which ψ_0 is a constant]. It is easy to check that Eq. (11) does indeed have asymptotic behavior in accord with Eq. (3), including a broad cross-over; similarly, Eqs. (7) and (9) agree with Eq. (2). Numerically, Eq. (11) fits pretty well to the data (Fig. 2), but, we repeat, best fit is achieved by the cumbersome discrete formula.

In conclusion, we have shown that proteinlike copolymers generated according to the coloring procedure proposed earlier [1,2] exhibit long-range correlations in the primary sequences. For the flexible polymers and large enough globules, these correlations belong to the Levy-flight statistics. This result, first observed in computer experiments, is confirmed and explained by analytical calculation. Analytical theory suggests that Levy-flight statistics, albeit with a broader cross-over region, is expected even if parental conformation is not maximally compact, but rather a globule somewhat closer to the θ-point. It becomes clear from our model that segregation of globule into a hydrophobic core and a hydrophilic peel, which is the necessary condition for water solubility, does impose severe restrictions on the sequence, and, therefore, must be manifested in certain correlations.

However, we would like to emphasize once more that the correlations in the primary sequences obtained above apply to designed synthetic copolymers rather than to real proteins. The chains in the core of real proteins do not obey Gaussian statistics (mainly due to the elements of secondary structure); therefore, deviations from randomness due to long-range correlations [14] and to regularity [15,16] were found in protein sequences. Besides, the compact folding of molecules was shown to favor regularity in unit sequences for small globules [16]. Nevertheless, identification of long-range correlations in protein sequences is an interesting task which promisses to shed light on the evolutionary criteria involved in the selection of proteins and the role of water solubility among them.

The work was supported by NATO (PST/CLG. 974956), INTAS (INTAS-OPEN-97-0678), and RFBR (98-03-33337a). The authors thank Dr. A. Irbäck for the useful correspondence.

[1] A.R. Khokhlov and P.G. Khalatur, Physica A **249**, 253 (1998).
[2] A.R. Khokhlov and P.G. Khalatur, Phys. Rev. Lett. **82**, 3456 (1999).
[3] V.S. Pande, A.Yu. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000).
[4] E.A. Zheligovskaya, P.G. Khalatur, and A.R. Khokhlov, Phys. Rev. E **59**, 3071 (1999).
[5] V.A. Ivanov, A.V. Chertovich, A.A. Lazutin, N.P. Shusharina, P.G. Khalatur, and A.R. Khokhlov, Macromol. Symp. **146**, 259 (1999).
[6] J. Virtanen, C. Baron, and H. Tenhu, Macromolecules **33**, 336 (2000).
[7] V.I. Lozinskii, I.A. Simenel, E.A. Kurskaya, V.K. Kulakova, V.Ya. Grinberg, A.S. Dubovik, I.Yu. Galaev, B. Mattiasson, and A.R. Khokhlov, Dokl. Chem. **375**, 273 (2000).
[8] N.V. Dokholyan, S.V. Buldyrev, Sh. Havlin, and H.E. Stanley, Phys. Rev. Lett. **79**, 5182 (1997).
[9] I. Grosse, H. Herzel, S.V. Buldyrev, and H.E. Stanley, Phys. Rev. E **61**, 5624 (2000).
[10] A.Yu. Grosberg and A.R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, Woodbury, NY, 1994).
[11] *Levy Flights and Related Topics in Physics*, edited by M.F. Shlesinger, G.M. Zaslavskii, and U. Frisch, Lecture Notes in Physics (Springer-Verlag, Berlin, 1996).
[12] W. Feller *An Introduction to Probability Theory and Its Applications*, 3rd ed. (Wiley, New York, 1970).
[13] For the bond fluctuation model [2], parameter a is given by $a = \sqrt{8}b$, where b is the lattice spacing. Also, the value $R = 27.8b$ was used to adjust to the actual radius of gyration of the parent globule at the coloring conditions.
[14] V. Pande, A. Grosberg, and T. Tanaka, PNAS **91**, 12 972 (1994).
[15] A. Irbäck, C. Peterson, and F. Potthast, PNAS **93**, 9533 (1996).
[16] A. Irbäck and E. Sandelin, Biophys. J. **79**, 2252 (2000).