Conformation-dependent evolution of copolymer sequences

Pavel G. Khalatur,[1,2] Viktor V. Novikov,[2] and Alexei R. Khokhlov[1,3]

[1]*Department of Polymer Science, University of Ulm, Ulm D-89069, Germany*

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

[3]*Physics Department, Moscow State University, Moscow 117234, Russia*

(Received 3 July 2002; revised manuscript received 8 January 2003; published 2 May 2003)

A ‘‘toy model’’ of molecular evolution of sequences in copolymers is proposed and implemented using a molecular-dynamics-based algorithm. The model involves coupling of conformation-dependent and sequence-dependent properties. It is shown that this model allows the realization of two main possibilities: ascending and descending branches of evolution (in terms of information content of a sequence), depending on the interaction parameters shaping the conformation of a polymer globule. The problem of adequate description of information complexity of copolymer sequences is studied. It is shown that Shannon’s entropy or compressibility of a sequence gives preference to random sequences and therefore cannot be applied for this purpose. On the other hand, the Jensen-Shannon divergence measure turns out to give the description of information complexity which corresponds to our intuitive expectations. In particular, this characteristic can adequately describe two branches of evolution mentioned above, exhibiting a singularity on the boundary of these regimes.

DOI: 10.1103/PhysRevE.67.051901 PACS number(s): 87.15.Aa, 87.15.Cc, 61.41.+e, 36.20.Ey

I. INTRODUCTION

The concept of evolution is one of the cornerstones of modern natural sciences: in cosmology, the evolution of the Universe is discussed; in geology, the evolution of the Earth; and in life sciences biological evolution (driven by selection) [1]. This concept can be also applied to polymer science [2,3]. The corresponding statement of the problem is very clear. The present day biopolymers (proteins, DNA, and RNA) possess complicated sequences of monomer units which encode their functions and structure (e.g., unique tertiary structure of globular proteins). Therefore, these sequences (in 20-letter alphabet for the case of proteins and in 4-letter alphabet for the cases of DNA and RNA) should be statistically very different from random ones and often exhibit significant correlation on different scales [4,5]. In other words, it is natural to expect that the content of information in these sequences is relatively high in comparison with random sequences (e.g., DNA sequences contain all genetic information) [4,5].

On the other hand, the formation of first copolymers at the very beginning of molecular prebiological evolution could lead only to random sequences or sequences with trivial short-range correlations. In other words, the information content of these sequences was practically zero. One can argue that in the course of molecular evolution, the copolymer sequences became more and more complex until they reached the stage of information complexity of present day biopolymers. The study of various possibilities of this evolution of copolymer sequences is just the area where the evolution concept can be used in the context of polymer science.

On the other hand, the formulated fundamental problem is extremely difficult due to the absence of direct information on the early prebiological evolution. Therefore, of particular interest are ‘‘toy models’’ of evolution of sequences, which show different possibilities for appearance of statistical complexity and long-range correlation in the sequences. Since by random mutations it is impossible to increase the information contents of a sequence, such toy models should take into account the coupling between polymer chain conformation (defined by the interactions between monomer units of different type) and evolution of sequence. In other words, we have to explore the possibilities of conformation-dependent evolution of copolymer sequences.

One of the variants of conformation-dependent design of copolymers that in one step leads to rather complicated statistical sequences has been recently considered in Refs. [6–12]. Following this approach, we start with a homopolymer globule stabilized by the attraction between monomer units. Then we introduce a ‘‘coloring’’ procedure. The units in the core of the globule remain ‘‘black’’ and are called hydrophobic ($\mathcal{H}$ units), while the units at the surface of the globule are colored in ‘‘white’’ and are called hydrophilic, or polar ($\mathcal{P}$ units). After that the uniform attraction between monomer units is removed, and we obtain an $\mathcal{HP}$ copolymer whose conformation depends on the interaction constants that we assign to $\mathcal{H}$-$\mathcal{H}$, $\mathcal{H}$-$\mathcal{P}$, and $\mathcal{P}$-$\mathcal{P}$ interactions.

Such a copolymer was called in Ref. [6] a proteinlike copolymer because it mimics one of the important features of real globular proteins: the possibility of formation of dense hydrophobic core stabilized by hydrophilic envelope in a globular conformation. It is because of this feature that proteins do not precipitate in the solution in the globular conformation, contrary to what would happen for statistically random copolymers. Of course, the presence of a hydrophilic envelope is a necessity, but not a sufficient condition for the absence of aggregation. Also, it should be mentioned that formation of a hydrophilic envelope is only one of the protein properties and therefore proteinlike copolymers have nothing to do with real proteins. Moreover, biological evolution forced proteins to be not only collapsed heteropolymers, but also to assume highly specific three-dimensional structures.

The procedure outlined above was first realized in computer experiments, and it was shown that the properties of proteinlike $\mathcal{HP}$ copolymers differ very significantly from the copolymers with random and random-block sequences [6–8]. Later, proteinlike $\mathcal{HP}$ copolymers were synthesized in real chemical experiments [13,14], and the predictions of computer experiments were confirmed. The role of coloring in real experiments is played by the reaction of a monomer unit with a reagent, which converts hydrophobic unit to a charged or polar group. In Ref. [13], hydrophilization was achieved by grafting of short poly(ethylene oxide) chains to the thermosensitive poly(isopropylacrylamide) backbone. It was shown that grafting to the more compact conformation (which occurs mainly from the surface leading to a kind of surface coloring) is more efficient than random grafting to a coil conformation. Recently, several papers by different groups described a new approach for obtaining proteinlike $\mathcal{HP}$ copolymers by copolymerization in a poor selective solvent [14]. This method automatically produced a core of thermosensitive weakly hydrophobic units surrounded by hydrophilic envelope. As a result, a solution of nonaggregating globules was obtained.

The statistical properties of the resulting proteinlike $\mathcal{HP}$ sequences were analyzed in Ref. [15]. It was shown that these sequences exhibit long-range correlations of Levy-flight type; therefore they acquire a certain degree of complexity as a result of a one-step coloring procedure [6–8].

The aim of the present paper is to introduce explicitly the concept of evolution of sequences into the scheme of generation of proteinlike copolymers. Namely, after the formation of initial $\mathcal{HP}$ sequence we will allow the macromolecule to undergo a coil-globule transition to a new globule, mainly induced by the strong attraction between $\mathcal{H}$ units, and then we will perform recoloring in the newly formed globular conformation. The rules of this recoloring will be the same as in Refs. [6–8]: the units that have maximum contacts with solvent molecules will be colored in white, while the units that are mainly in contact with other monomer units will be converted into black. In this way, we will obtain a macromolecule with a $\mathcal{HP}$ sequence. For this macromolecule, we can again perform globular folding induced by the attraction between $\mathcal{H}$ units in the new sequence, again perform recoloring, obtain new $\mathcal{HP}$ sequence, etc.

Following this procedure, the proteinlike $\mathcal{HP}$ sequences will undergo some evolution. The question is that whether this evolution leads to the increase of complexity or we will end up with some trivial sequence? We will show below that the answer to this question depends on the interrelation between $\mathcal{H}$-$\mathcal{H}$, $\mathcal{H}$-$\mathcal{P}$, and $\mathcal{P}$-$\mathcal{P}$ interaction constants. For some parameters, the sequences become more complex and long-range correlations more pronounced (model of the ascending branch of the evolution), while for the other parameters, the degree of complexity decreases and we come to a rather trivial $\mathcal{HP}$ block copolymer (model of descending branch of the evolution).

In the literature, several simple computer models describing the evolution of copolymer sequences were proposed (see, e.g., Refs. [16–27]). However, most of these models are aimed to resolve various problems of protein physics, e.g., the problem of generating specific amino acid sequences, which are thermodynamically stable in a target three-dimensional conformation and are able to fold fast into this conformation at a given temperature. As stated above, our approach in this paper is different. That is, we give the following approach.

(i) We propose a toy model of molecular evolution of sequences in copolymers that allow two main possibilities: ascending and descending branches of evolution (in terms of information content of a sequence), depending on the interaction parameters shaping the conformation of a polymer globule.

(ii) We investigate that which information characteristics of a sequence can be used to describe informational complexity. In particular, we will show that such well-known quantities such as Shannon's entropy or compressibility of a sequence cannot be applied for this purpose. On the other hand, the so-called Jensen-Shannon divergence measure seems to be a good candidate to describe the information complexity, at least for the sequences studied in this paper.

II. COMPUTATIONAL STRATEGY AND OBSERVED QUANTITIES

A. Model and simulation technique

First of all, we will define our model and algorithm employed for the simulation of evolutionary process. We will consider a ‘‘black-and-white’’ model of a copolymer chain that involves only two types of monomers, $\mathcal{H}$ (hydrophobic) and $\mathcal{P}$ (hydrophilic or polar). Since we are interested in describing the main general possibilities for the evolution of sequences, we will fix the $\mathcal{HP}$ composition at 1:1 for the sake of simplicity. Earlier, we have studied the effect of variation of $\mathcal{HP}$ composition on the properties of proteinlike copolymers (see, e.g., Fig. 6 in Ref. [9(b)]), and the general conclusion was that the main qualitative regimes of behavior are not significantly affected by this variation.

We consider a continuum space (bead-spring) model, as opposed to widely used lattice $\mathcal{HP}$ models [28,29], since the latter have an intrinsically discretized dynamics of a rather arbitrary nature and have slow relaxation for a dense globular state. The time evolution of the system is determined by Newton equations that are solved by using the method of molecular dynamics (MD). The monomers are linked by flexible bonds to form a linear $\mathcal{HP}$ copolymer chain of length N.

The Hamiltonian of the system is taken to be a pairwise type and it is given by

$$H=\sum_{i,j=i+1}^{N-1} H_b(r_{ij})+\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}[H_{\rm ev}(r_{ij})+H_a(r_{ij})]+\tfrac{1}{2}\sum_{i=1}^{N}p_i^2, \qquad (1)$$

where H_b is the bond potential, $H_{\rm ev}$ takes into account excluded volume, H_a characterizes attractive interactions between chain beads (monomer units), and i and j range from 1 to N. The distance between the beads is defined as $r_{ij}=|\mathbf{r}_i-\mathbf{r}_j|$, where $\mathbf{r}_i$ denotes the position vector of bead i in three-dimensional space. The last term in Eq. (1) is the classical kinetic energy of the chain, where the $\mathbf{p}_i$'s are the canonical variables conjugate to the $\mathbf{r}_i$'s.

Excluded volume between the all nonbonded beads is included via a repulsive Lennard-Jones potential

$$H_{\rm ev}(r_{ij})=\begin{cases} 4\varepsilon\left[\left(\dfrac{\sigma}{r_{ij}}\right)^{12}-\left(\dfrac{\sigma}{r_{ij}}\right)^{6}+\dfrac{1}{4}\right], & r_{ij}\leq r_0 \\ 0, & r_{ij}>r_0, \end{cases} \tag{2}$$

where $\sigma=\varepsilon=1$ for both $\mathcal{H}$ and $\mathcal{P}$ monomers and $r_0=2^{1/6}\sigma$ is the cutoff distance. The following quasiharmonic bond potential connects the beads of a chain:

$$H_b(r_{ij})=\begin{cases} C_b^{(1)}\left[\left(\dfrac{b_0}{r_{ij}}\right)^{12}-2\left(\dfrac{b_0}{r_{ij}}\right)^{6}+1\right], & r_{ij}\leq b_0 \\ C_b^{(2)}\left[\exp\left\{\left(\dfrac{r_{ij}}{b_0}\right)^{2}-1\right\}-\left(\dfrac{r_{ij}}{b_0}\right)^{2}\right], & r_{ij}>b_0, \end{cases} \tag{3}$$

where b_0 and r_{ij} are equilibrium and current bond lengths, respectively. This quasiharmonic term in the Hamiltonian (1), with the spring constants $C_b^{(1)}$ and $C_b^{(2)}$, couples the beads i and $j=i+1$ that are adjacent along the chain. The equilibrium bond length b_0 in Eq. (3) and other lengths are measured in units of $\boldsymbol{\sigma}$, the typical value of which for a real polymer is $\sigma=5$ Å. We set $b_0=1$ and $C_b^{(1)}=C_b^{(2)}=1$. The remaining term in Eq. (1) describes attractive interactions between nonbonded monomers

$$H_a(r_{ij})=\begin{cases} -\dfrac{\varepsilon_{\alpha\beta}\sigma}{r_{ij}}\left[1-\left(\dfrac{r_{ij}}{r_c}\right)^{2}\right]^{2}, & r_0<r_{ij}\leq r_c \\ 0, & r_{ij}>r_c. \end{cases} \tag{4}$$

The parameter $\varepsilon_{\alpha\beta}$ ($=\varepsilon_{\mathcal{HH}}$, $\varepsilon_{\mathcal{PP}}$, $\varepsilon_{\mathcal{HP}}$) sets the depth of the minimum of the nonlocal attraction and $r_c=2.8$ is the cutoff distance for attractive interactions. In Eq. (4), we adopt the simplest choice for the cross parameter: $\varepsilon_{\mathcal{HP}}=(\varepsilon_{\mathcal{HH}}\times\varepsilon_{\mathcal{PP}})^{1/2}$ [30]. For a globular conformation, the characteristic energy of $\mathcal{H}$-$\mathcal{H}$ interactions is fixed at $\varepsilon_{\mathcal{HH}}=2$ (all the energies are measured in units of k_BT). In our conformation-dependent evolutionary process described above, the value of $\varepsilon_{\mathcal{PP}}$ is considered as an only variable energy parameter. At $\varepsilon_{\mathcal{HH}}=\varepsilon_{\mathcal{PP}}=2$, we have in fact a homopolymer globule, and a well-compacted conformation emerges. At $\varepsilon_{\mathcal{HH}}=\varepsilon_{\mathcal{PP}}=0$, there is no attraction between nonbonded beads. For computational efficiency, the chain with relatively short length, $N=128$, is used for most of the calculations. Preliminary results with $N=512$ suggest that qualitative features do not differ much in the larger system.

No explicit solvent particles are included in the simulations. In order to simulate the solvation effects and time evolution of the system in contact with a heat bath of temperature T, we augment the equations of motion by the Langevin uncorrelated noise terms

$$m_i\ddot{\mathbf{r}}_i=\mathbf{F}_i-\Gamma_i\dot{\mathbf{r}}_i+\mathbf{R}_i,\quad i=1,2,...,N, \tag{5}$$

where $m_i=1$ is the mass of chain bead i, $\mathbf{F}_i=-\boldsymbol{\nabla}_{r_i}H(\mathbf{r})$ is the systematic force acting on the bead i, $\mathbf{R}$ describes the random force of the heat bath acting on each monomer, and Γ takes into account the viscosity of the solvent. The values $\mathbf{R}$ and Γ are connected through the fluctuation-dissipation theorem, $\langle R_{ai}(0)R_{\alpha i}(t)\rangle=2\Gamma_ik_BT\delta(t)$, $\alpha=x,y,z$, and ensures that the temperature is kept constant [30]. We take the parameter Γ to be dependent on solvent-accessible surface areas (SASA). To find the values of SASA for a given conformation, we perform an analytical computation of the surface areas $\mathcal{A}_i$ for each specified monomer [31]. Having $\mathcal{A}_i$, one can define Γ_i as $\Gamma_i=\Gamma_0\mathcal{A}_i/\mathcal{A}_{\rm max}$, where $\mathcal{A}_{\rm max}$ is the maximum solvent-accessible surface area of a monomer for the model under study and the reference value of Γ_0 is taken to be equal to unity. The weighting factor $\mathcal{A}_i/\mathcal{A}_{\rm max}$ represents the degree of exposure of the monomer i to the solvent. When the value of SASA for a given monomer is zero, the frictional and random forces are zero and the Langevin equation (5) reduces to Newton's equation of motion. Typically, this happens when the monomer is located in the core of a globule. On the contrary, a monomer located at the globular surface is strongly solvated; it means that $\mathcal{A}_i$ should be close to $\mathcal{A}_{\rm max}$ and, as a result, the value of Γ_i is close to its reference value Γ_0. In the following, the reference temperature is fixed at $T=\varepsilon/k_B$. The integration of the equations of motion is performed with the time step $\Delta t=0.01\sigma\sqrt{m/\varepsilon}$, using the Verlet leapfrog/central difference algorithm [30].

B. The model of molecular evolution

For the evolutionary process, the following algorithm is employed.

(i) For the initial generation $(G=0)$, a self-avoiding homopolymer chain with $\mathcal{H}$ units is randomly generated. This chain is considered as a "common ancestor" for a given run.

(ii) A swollen polymer coil is prepared by assigning zero to the parameters $\varepsilon_{\mathcal{HH}}$ and $\varepsilon_{\mathcal{PP}}$.

(iii) The folding of the chain is performed at $\varepsilon_{\mathcal{HH}}=2$ and at a given value of $\varepsilon_{\mathcal{PP}}$. Then this bare conformation is equilibrated for $\tau_1=4\times10^5$ integration time steps.

(iv) One half of the units, having the largest SASA and simultaneously well-separated from the center of mass of the globular core, are transformed into $\mathcal{P}$ type; the remaining units with the minimum values of SASA and occurring not far from the center of mass of the globular core are recolored in $\mathcal{H}$ type. Due to such recoloring, the sequence is mutated and passed on to the next generation, $G\rightarrow G+1$. The composition of the sequence is always constrained so that there are $N/2$ hydrophobic and $N/2$ hydrophilic units. Information-theoretic measures introduced below are used in detecting the statistical properties of the current sequence.

(v) To calculate structural and thermodynamic properties, we perform MD run and use averaging for the sufficiently large number of integration time steps, $\tau_2=4\times10^5$.

(vi) The last generation from step (v) is subjected to steps (ii)–(v). In other words, we iterate the procedure until it converges self-consistently. This gives a set of the globular copolymers with different primary $\mathcal{HP}$ structures.

In the present study, steps (i)–(vi) are independently repeated 20 times starting from the various random conformations and then all the results are averaged over these runs in order to gain better statistics. For each trajectory, we may interpret a set of different sequences generated in the course of our evolutionary process as different ‘‘species’’ originating from a common ancestor.

C. Conformation-dependent and sequence-dependent properties

The MD trajectories from step (v) are characterized by the total potential energy U and the mean-square radii of gyration that can be determined for overall macromolecule, R, as well as separately for $\mathcal{H}$ and $\mathcal{P}$ monomers, $R_{\mathcal{H}}$ and $R_{\mathcal{P}}$. These quantities depend both on current sequence and conformation and, as pointed out above, they are averaged over 4×10^5 time steps and considered as functions of a generation number G. The second type of the quantities studied here is related to the primary structure of a copolymer chain. After each mutation [step (iv)], we calculate the average length of $\mathcal{H}$ and $\mathcal{P}$ blocks (loops), $L_{\mathcal{H}}$ and $L_{\mathcal{P}}$, and various information-theoretic properties.

D. Shannon's entropy and related characteristics

A common approach for the analysis of complex systems is to use concepts from information theory and information-theoretic-based techniques [4,32–41]. Within this approach, copolymer sequences can be examined as messages written in two-symbol alphabet (letters $\mathcal{H}$ and $\mathcal{P}$). In general, the aim is to find a measure capable to indicate how far copolymer sequences generated during our evolutionary process differ from each other and from random or trivial (degenerate) sequences. In this study, we consider the statistical properties related to Shannon's entropy and the degree of complexity of copolymer sequences.

Let $s_1s_2,\ldots,s_n$ be the symbols of a given sequence $\mathbf{S}$ of length N. If $f_n(s_1s_2,\ldots,s_n)$ is the average frequency of a subsequence with the symbols $s_1s_2,\ldots,s_n$, i.e., of the ‘‘word’’ $s_1s_2,\ldots,s_n$ of length $n\leqslant N$, then Shannon's entropy of the whole sequence is given by [32]

$$h=-k_B\sum_{(\text{all words})} f_n(s_1s_2,\ldots,s_n)\ln[f_n(s_1s_2,\ldots,s_n)], \quad (6)$$

where k_B denotes the Boltzmann constant and the summation runs over all possible words. If k_B is replaced by $1/\ln 2$, then h quantifies the amount of information in units of bits [33]. Of course, Shannon's entropy depends on the definition of a set of words in the sequence. For our case of two-letter $\mathcal{HP}$ copolymer, we will adopt the following set of words (uniform blocks): $\mathcal{H},\mathcal{HH},\mathcal{HHH},\ldots,$ $\mathcal{P},\mathcal{PP},\mathcal{PPP},\ldots,$ i.e., word (block) is defined by its length n and type ($\mathcal{H}$ or $\mathcal{P}$). Then Shannon's entropy per monomer can be written as

$$h=-\frac{1}{2N}\sum_n\,[f_{\mathcal{H}}(n)\log_2 f_{\mathcal{H}}(n)+f_{\mathcal{P}}(n)\log_2 f_{\mathcal{P}}(n)], \quad (7)$$

where $f_{\mathcal{H}}(n)$ and $f_{\mathcal{P}}(n)$ are the frequencies of words of length n composed of letters $\mathcal{H}$ and $\mathcal{P}$, correspondingly, and $0\times\log_2 0=0$ is assumed by continuity. In a certain sense, $h(\mathbf{S})$ measures the diversity of events distributed in $\mathbf{S}$. Note that for the case of repeated or random sequences Shannon's entropy can be easily found analytically (e.g., for a uniform random sequence of *infinite* length written using an alphabet of two symbols, the exact result is $h=1$ bit; for any regular multiblock sequence, $h=0$ bit).

E. Compressibility

Additional measure of information complexity is sequence compressibility if sequence is represented as a set $\mathbf{S}$ of ASCII characters [35,36]. We use one bit to encode each character, i.e., 0 for $\mathcal{H}$ and 1 for $\mathcal{P}$. The definition of the compression ratio used in this work is the same as in Ref. [37], i.e., $\kappa_N=1-\eta/\eta_i$, where η_i is the length (number of bits) of the input sequence consisting of N units (characters) and η is the length (number of bits) of the output sequence. For a sequence written using an alphabet of two letters, one has $\eta_i=N$. A random sequence of sufficient length, where both symbols occur with the same probability, corresponds to $\kappa\approx0$. For a regular array consisting of two different characters, we expect that $\kappa\approx1$ if $N\gg1$. On the other hand, a sequence not arranged in a random or regular order would, in general, give $0<\kappa<1$; that is, in this case the compressibility is somewhere in between the two extreme cases considered above. To calculate η, we employ the UNIX Lempel-Ziv file compressor GZIP [38]. The length of empty sequence, η_0, is always extracted from η. It is clear that η_0 and η depend on the compression software applied for calculations. It appears, however, that the general trends are the same.

F. Jensen-Shannon divergence measure

An adequate definition of complexity of a copolymer sequence must be objective and consistent with our intuitive notion of what the complexity is about. Neither the algorithmic complexity [39] (maximum for random processes) nor other derived measures based on mutual information [40] or compressibility are completely satisfactory. As a measure of complexity, we will use the so-called Jensen-Shannon divergence measure [41,42] which is defined as follows.

Let $\mathbf{S}=\{s_1\ldots,s_N\}$ be a sequence of N symbols. For two subsequences $\mathbf{S}_1=\{s_1,\ldots,s_n\}$ and $\mathbf{S}_2=\{s_{n+1},\ldots,s_N\}$ of lengths n and $N-n$, the difference between the corresponding discrete probability distributions $f_1(s_1,\ldots,s_n)$ and $f_2(s_{n+1},\ldots,s_N)$ is quantified by the Jensen-Shannon (JS) divergence

$$h_{\mathrm{JS}}(\mathbf{S}_1,\mathbf{S}_2)=h(\mathbf{S})-\left[\frac{n}{N}h(\mathbf{S}_1)+\frac{N-n}{N}h(\mathbf{S}_2)\right], \quad (8)$$

where $\mathbf{S}=\mathbf{S}_1\oplus\mathbf{S}_2$ (concatenation) and $h(\mathbf{S})$ is Shannon's entropy of the empirical probability distribution obtained from block frequencies in the corresponding subsequences [see Eq. (7)]. The Jensen-Shannon divergence h_{JS} is zero for subsequences with the same statistical characteristics; it takes higher values for increasing differences between the statistical patterns in the subsequences. In particular, both random and any regular (multiblock) copolymer of infinite length show $h_{\rm JS}=0$; for a finite random-block (Poisson) copolymer, the $h_{\rm JS}$ value considered as a function of average block length goes though the maximum, thus giving a reasonably good measure of complexity for these sequences, corresponding to our intuition [we normally expect that a completely random sequence or a sequence with long uniform blocks contains less information than a sequence containing many different blocks (words) of medium length].

We will see below that the Jensen-Shannon divergence measure is indeed quite adequate for the description of information complexity of the sequences studied in the present paper. Of course, only by sequence analysis, we cannot unambiguously distinguish between what might be called quality and quantity of information.

III. RESULTS AND DISCUSSION: DIVERSITY VERSUS DEGENERACY

A. Conformational transitions

We start our discussion by estimating the reference energy parameter at which coil-globule transition takes place in a homopolymer chain. To this end, we calculated R for the 128-unit $\mathcal{H}$ homopolymer chain as a function of $\varepsilon_{\mathcal{HH}}$. An apparent transition energy ε^* can be identified with inflection point on the $R(\varepsilon_{\mathcal{HH}})$ curve. Note that in order to estimate ε^*, it is easier to use the radius of gyration than the specific heat [43,44]. The derivative $\partial R/\partial\varepsilon_{\mathcal{HH}}$ exhibits a pronounced peak, which signals a transition. After least-squares fitting of the simulation data and subsequent differentiation, we find $\varepsilon^*=0.42\pm0.08$. Practically, the same value of ε^* is obtained from the condition $R(\varepsilon_{\mathcal{HH}})/R_\Theta=1$, where R_Θ is the radius of gyration calculated for unperturbed chain, i.e., the chain without excluding volume and attractive interactions ($H_{\rm ev}=H_a=0$). At $\varepsilon_{\mathcal{HH}}<\varepsilon^*$, the chain is in a swollen coil state and collapses when $\varepsilon_{\mathcal{HH}}$ becomes significantly larger than ε^*. As has been noted, the value of $\varepsilon_{\mathcal{HH}}$ is chosen equal to 2 in order to ensure the collapse of hydrophobic units into a dense core shielded by hydrophillic units from the solvent.

We also studied the conformational behavior of a diblock 128-unit $\mathcal{HP}$ copolymer chain. In this case, the value of $\varepsilon_{\mathcal{HH}}$ was fixed at $\varepsilon_{\mathcal{HH}}=2$, while the energy parameter $\varepsilon_{\mathcal{PP}}$ describing the interaction between hydrophillic units was varied. For this system, we calculated the partial mean-square radius of gyration for hydrophillic block, $R_{\mathcal{P}}$. When $\varepsilon_{\mathcal{PP}}$ is increased, we observe in fact the adsorption of the hydrophillic block on the surface of globule formed by the hydrophobic block, resulting in a decrease in $R_{\mathcal{P}}$. The transition adsorption energy ε_a^*, estimated from the position of inflection point on the $R_{\mathcal{P}}(\varepsilon_{\mathcal{PP}})$ curve is found to be $\varepsilon_a^*=0.14\pm0.07$.

B. Evolutionary drift

Now we can ask that how two nearby generations, G and $G+1$, relate to each other as the system evolves. To answer this question, we define the following sequence mutation rate, which reflects the relevant features of diffusion (drift) in sequence space:

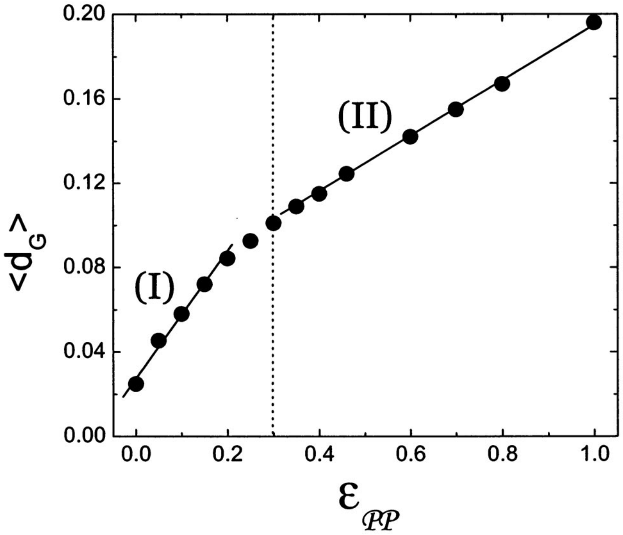

FIG. 1. Normalized Hamming distance averaged over the last 900 generations (3.6×10^8 time steps) of 20 independent trajectories as a function of $\varepsilon_{\mathcal{PP}}$. There are two different regions of $\varepsilon_{\mathcal{PP}}$ (region I located at $\varepsilon_{\mathcal{PP}}<0.3$ and region II located at $0.3\leq\varepsilon_{\mathcal{PP}}\leq1$), in which $\langle d_G\rangle$ shows an approximately linear behavior ($\langle\bullet\rangle$ means time averaging, and the time represents in our model the number of mutational events). Points present the simulation results. Solid lines are the best fit to the simulation data.

$$d_G=\frac{1}{N}\sum_{i=1}^{N}\left(1-\delta_{s_i(G),s_i(G+1)}\right). \qquad (9)$$

Here $\delta_{s_i(G),s_i(G+1)}$ is the Kronecker delta symbol. The d_G value is the normalized Hamming distance, which counts how many monomers are different between two sequences $\mathbf{S}(G)$ and $\mathbf{S}(G+1)$ of the same length N. Note that $d_G=\frac{1}{2}$ for any random process.

For all the cases studied in the present paper, we have observed that although for the first several generations, the sequence mutation rate changes, for the next generations, it reaches a steady state and seems to fluctuate more or less randomly around its average level $\langle d_G\rangle$ that depends on $\varepsilon_{\mathcal{PP}}$.

Figure 1 shows the $\langle d_G\rangle$ value, found for the range $10^2<G\leq10^3$ (where a steady state is already established) and averaged over 20 independent runs (7.2×10^9 time steps), as a function of $\varepsilon_{\mathcal{PP}}$. We conclude that, for the evolutionary model under study, the steady-state rate of mutations becomes approximately an order of magnitude larger when $\varepsilon_{\mathcal{PP}}$ increases from 0 to 1. For $\varepsilon_{\mathcal{PP}}=2$, the average Hamming distance between neighboring sequences is $\langle d_G\rangle=0.47\pm0.04$, that is, the rate of mutation approaches the value corresponding to a random process (as it should be) because the difference between $\mathcal{H}$ and $\mathcal{P}$ units disappears in this case. It is seen that there are two well-defined regions of $\varepsilon_{\mathcal{PP}}$ located at $\varepsilon_{\mathcal{PP}}<0.3$ and $0.3\leq\varepsilon_{\mathcal{PP}}\leq1$, in which $\langle d_G\rangle$ shows an

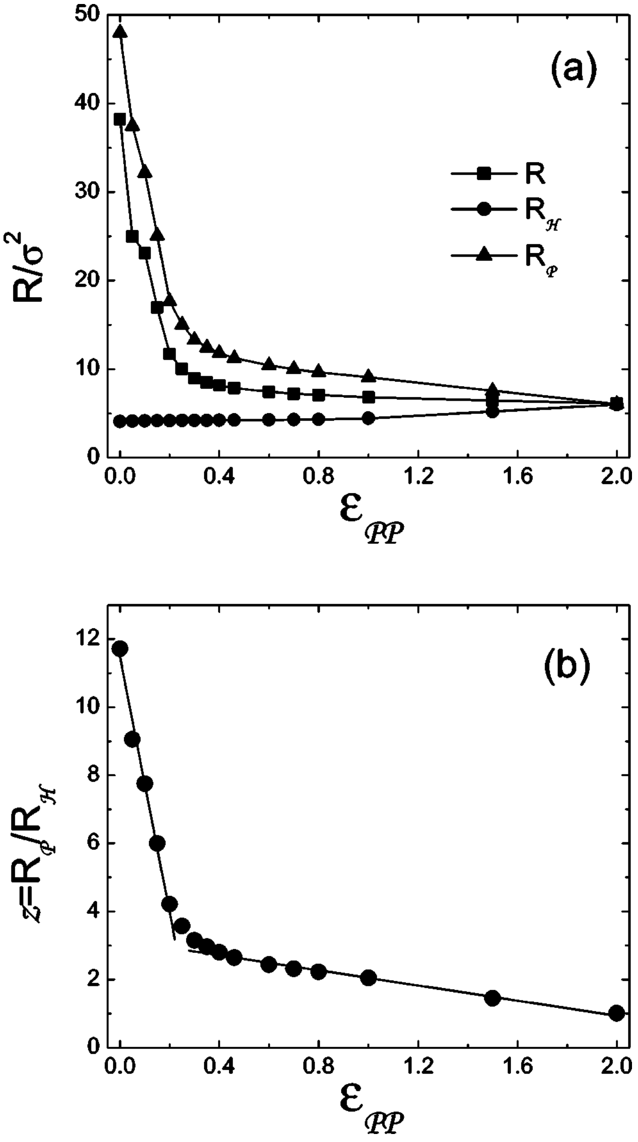

FIG. 2. (a) Mean-square gyration radius R and its components, $R_{\mathcal{H}}$ and $R_{\mathcal{P}}$, plotted vs $\varepsilon_{\mathcal{PP}}$. (b) Ratio $z = R_{\mathcal{P}}/R_{\mathcal{H}}$ as a function of $\varepsilon_{\mathcal{PP}}$.

approximately linear behavior as a function of $\varepsilon_{\mathcal{PP}}$. Thus, we can speculate that, depending on the interaction between hydrophilic monomers, there are two different regimes of mutation process. This observation is further supported by analysis of conformational properties.

C. The conformational properties

Figure 2 shows the total gyration radius R found for the whole macromolecule as well as partial gyration radii, $R_{\mathcal{H}}$ and $R_{\mathcal{P}}$, calculated for $\mathcal{H}$ and $\mathcal{P}$ components for stationary regime ($10^2 < G \leqslant 10^3$) at different $\varepsilon_{\mathcal{PP}}$. From $R_{\mathcal{H}}$ and $R_{\mathcal{P}}$, one can define the following characteristic ratio: $z = R_{\mathcal{P}}/R_{\mathcal{H}}$ that takes into account both the properties of compactness and solubility for a heteropolymer globule (compactness is directly related to the mean size of hydrophobic globular core, whereas solubility should depend on the size of hydrophilic corona, preventing the aggregation; thus the parameter z somehow describes both aspects). We find that $R_{\mathcal{H}}$ is a weakly increasing function of $\varepsilon_{\mathcal{PP}}$. This is due to the fact that, as $\varepsilon_{\mathcal{PP}}$ becomes larger, the attraction between $\mathcal{H}$ and $\mathcal{P}$ monomers increases, thus facilitating their compatibility and mixing in the globular core. On the other hand, $R_{\mathcal{P}}$ is a weakly decreasing function of $\varepsilon_{\mathcal{PP}}$ in the range $\varepsilon_{\mathcal{PP}} \geqslant 0.3$ and demonstrates a rapid growth when $\varepsilon_{\mathcal{PP}}$ decreases and becomes less than 0.3. Similar trends are found for R and z. For the parameter z introduced above, we observe rather distinctly the existence of two regions, located below and above $\varepsilon_{\mathcal{PP}} \approx 0.3$, where z can be approximated by linear relations. This fact is an indirect indication that there are two different regimes of evolutionary mechanism, leading to different final structures. We will call these regimes of evolutionary behavior regime I (for $\varepsilon_{\mathcal{PP}} < 0.3$) and regime II ($\varepsilon_{\mathcal{PP}} \geqslant 0.3$). Note that the corresponding critical value $\varepsilon^*_{\mathcal{PP}} \approx 0.3$ is found to be smaller than the critical energy ε^* at which coil-globule transition takes place in a homopolymer chain of the same length.

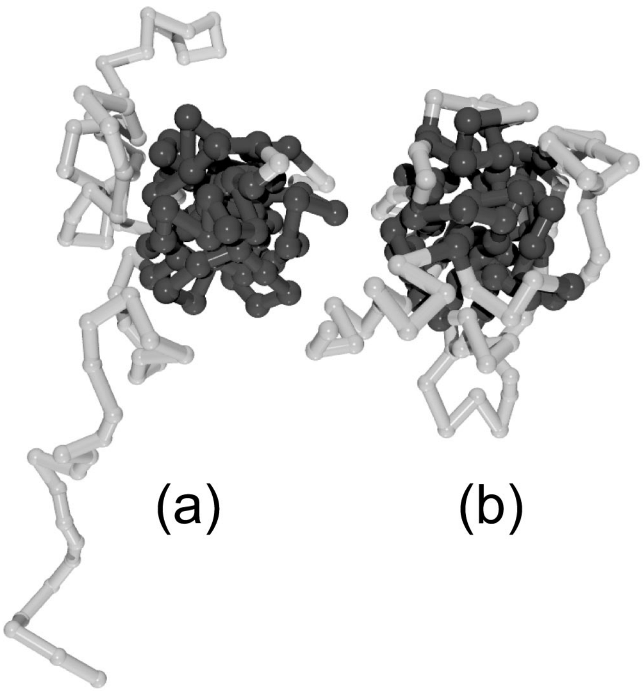

FIG. 3. Snapshots of two typical conformations of designed copolymers obtained after long evolution (3.6×10^8 time steps) of sequences. (a) Core-tail (tadpolelike) structure at $\varepsilon_{\mathcal{PP}} = 0$. (b) Core-shell structure at $\varepsilon_{\mathcal{PP}} = 0.3$.

In Fig. 3, we present two instantaneous pictures (snapshots) showing the typical globular conformations obtained after a long sequence evolution procedure for the case $\varepsilon_{\mathcal{PP}} = 0$ (regime I) and $\varepsilon_{\mathcal{PP}} = 0.3$ (regime II). From Fig. 3(a), it is seen that the evolution of sequences in regime I converted the typical globular conformation to a degenerated "core-tail" or "tadpolelike" structure. On the other hand, when the attraction between hydrophobic units is sufficiently strong (regime II), we observe "core-shell" structures having a dense hydrophobic core stabilized by hydrophilic envelope in a globular conformation [Fig. 3(b)]. One can intuitively

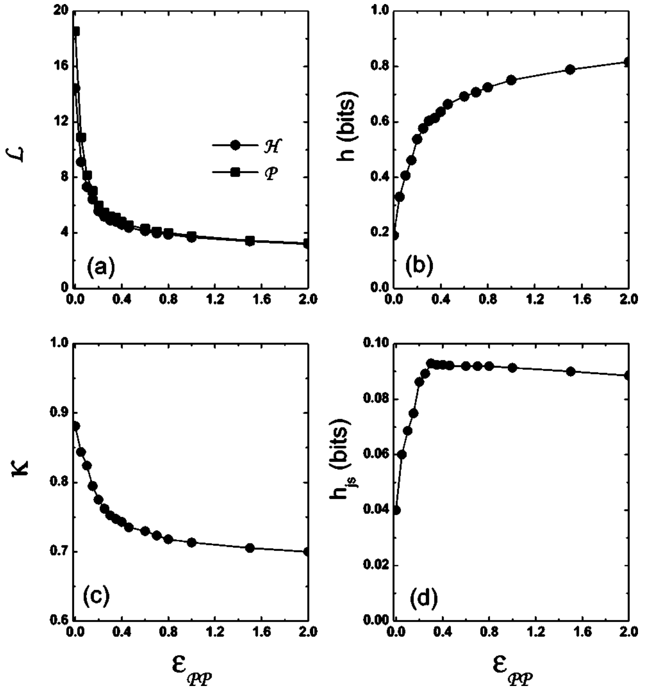

FIG. 4. Sequence-dependent parameters as functions of $\varepsilon_{\mathcal{PP}}$ (a) average length of $\mathcal{H}$ and $\mathcal{P}$ blocks, (b) Shannon's entropy, (c) compression ratio, and (d) the Jensen-Shannon divergence measure. The splitting parameter n in Eq. (8) is set to be $n=N/2$.

say that in the regime II the evolution preserved the sequence of high complexity, while in the regime I the information content of the sequence has degenerated in the course of evolution. In the following section, this intuitive feeling is supported by quantitative calculations.

D. The sequence-dependent properties

In Fig. 4, we present some of the information-theoretic-based parameters defined earlier. It is seen that the mean length of both $\mathcal{H}$ and $\mathcal{P}$ blocks increases as the energy of $\mathcal{P}$-$\mathcal{P}$ interactions decreases [Fig. 4(a)]. This increase becomes especially well pronounced in the range $\varepsilon_{\mathcal{PP}}<0.3$, indicating a similarity with the behavior found for the conformation-dependent characteristics (Fig. 2). Hydrophilic blocks are slightly longer as compared to hydrophobic blocks, the behavior expected for relatively short chains [15].

As compared to Shannon's entropy, two observations are noteworthy here: (i) it is always smaller than unity and (ii) it increases gradually with $\varepsilon_{\mathcal{PP}}$ [Fig. 4(b)]. The compression ratio κ is also smaller than unity and it is found to be a decreasing function of $\varepsilon_{\mathcal{PP}}$ [Fig. 4(c)], as can be expected taking into account the trends predicted for the average block lengths, see Fig. 4(a) (indeed, we expect that the κ value should increase as the block length is increased; in particular, for the $L=N/2$ limit, we have $\kappa=0.953$ at $N=128$). Also these results mean that the statistical properties of the sequences generated in the course of evolution deviate strongly from random ones. Indeed, for a random sequence containing 128 symbols, one has $h=0.920$ and $\kappa=0.672$ (these data were obtained by averaging over 10^5 random sequences).

From Figs. 4(b) and 4(c), we see that Shannon's entropy and compression ratio do not show any singularity when the energy $\varepsilon_{\mathcal{PP}}$ is varied. Such a behavior is in contrast with the results found for the average sequence mutation rate (Fig. 1) and for the parameter z [Fig. 2(b)], which demonstrate the existence of two different regimes.

Already from this observation it is possible to suspect that Shannon's entropy and compression ratio are not adequate measures of complexity for the generated sequences. Moreover, Shannon's entropy reaches maximum and compression ratio reaches minimum for a completely random sequence. It is clear that such a sequence cannot be regarded as having the maximum information complexity. Its information content is close to zero, similar to that of a completely regular sequence. Therefore, to make the mathematical information-related description consistent with our intuitive feeling of complexity, we have to choose the characteristics which (i) is equal to zero for both random and regular sequences, and reaches maximum for the sequences consisting of medium-size words with high degree of diversity (in our model, diversity is connected with the length of the word consisting of homogeneous symbols); (ii) another important feature of the sequence is connected with the fact that it encodes the spatial core-shell structure of a copolymer globule. This feature can be realized only if a statistical pattern is attributed to a sequence as a whole, and cannot be obtained by joining of independent statistical patterns of two subsequences of smaller length. In this respect, protein sequences are similar: sequence as a whole determines globular structure and hence biological function, while if this sequence is cut in two pieces, those pieces normally neither correspond to a soluble globule, nor have any biological function.

From this viewpoint, the Jensen-Shannon divergence measure introduced above [see Eq. (8)] seems to comply with both these requirements. Actually, it is designed in such a way that it takes zero value for the sequences which can be divided into two subsequences with similar statistical patterns. Therefore, we choose this characteristic for the description of information complexity.

Figure 4(d) shows the value of h_{JS} as a function of $\varepsilon_{\mathcal{PP}}$. The most important feature is that h_{JS} is a *nonmonotonous* function of $\varepsilon_{\mathcal{PP}}$, whereas Shannon's entropy and the compression ratio always change *gradually*. The value of h_{JS} reaches its maximum at $\varepsilon_{\mathcal{PP}}\approx 0.3$, i.e., just on the boundary of regimes I and II, see above. Considering h_{JS} as a measure of complexity, one can say that at $\varepsilon_{\mathcal{PP}}\approx 0.3$, the corresponding primary structure reaches its maximum complexity. In this case, $h_{\mathrm{JS}}=0.093$. This value is distinctly greater as compared to that found for a random 128-symbol two-letter sequence adjusted to achieve the 1:1 composition, $h_{\mathrm{JS}}=0.063$. Also, it is seen from Fig. 4(d) that in the region $\varepsilon_{\mathcal{PP}}\geqslant 0.3$, i.e., for regime II, the value of h_{JS} is rather high. Therefore, one can say that in this case our primitive evolutionary model imitates the ascending branch of the evolution, resulting in a considerable amount of information content remaining in the generated proteinlike sequences after long evolution. Moreover, this information content is increasing. On the other hand, at $\varepsilon_{\mathcal{PP}}<0.3$ (regime I), h_{JS} drops quickly, indicating a dramatic decrease in complexity for the primary structures with longer block lengths. This observation, together with the trend found for L [Fig. 4(a)] and typical snapshot conformations [Fig. 3(a)], suggest that for these conditions, designed sequences degenerate into trivial ones. The degenerated primary structure looks like a diblock or triblock sequence with a small amount of randomly arising "defects," and from this point of view such a sequence can be treated as trivial. Therefore, this is one more manifestation of the fact that, when the attraction between hydrophilic monomers is not sufficiently strong, we deal with the "downward" branch of the evolution, which leads to degenerate (nonproteinlike) sequences having low information content and low complexity. Such a behavior reflects the coupling between polymer chain conformation and sequence-dependent properties.

It is instructive to compare the behavior of designed sequences with that demonstrated by model sequences. To this end, it is pertinent to consider a two-letter sequence with the same composition in which the distribution of block length l is described by the Poisson law $P(l)=e^{-L}L^{l}/l!$. Thus, we generate the Poisson distribution adjusted to achieve the same 1:1 composition and the same "degree of blockiness" (average block length L) as for a proteinlike $\mathcal{HP}$ copolymer. We call such a sequence random-block one (cf. Ref. [8]).

In Fig. 5(a), we compare the Jensen-Shannon divergence measures calculated for random block and designed sequences of the same overall length, $N=128$, as a function of average block length (for designed sequences we take $L=(L_{\mathcal{H}}+L_{\mathcal{P}})/2$). From the results shown in Fig. 5(a), we can conclude that the Jensen-Shannon divergence turns out to be a suitable measure capable of indicating how far primary structures of the designed copolymer are from random-block ones. It is seen that the values of h_{JS} practically coincide for both sequences when $L>8$, i.e., for small $\varepsilon_{\mathcal{PP}}$'s (regime I). Therefore, in the course of evolution under corresponding conditions, the proteinlike sequence degenerates into something similar to a random-block sequence. On the contrary, in the range of shorter block lengths (regime II), which corresponds to higher $\varepsilon_{\mathcal{PP}}$'s, we observe quite visible differences. In this case, proteinlike sequences become more complex as compared to their random-block counterparts.

In Fig. 5(a), we also present the Jensen-Shannon divergence measure calculated for a model two-symbol sequence with Levy-flight statistics at $N=128$ (for more details, see Ref. [15]). As for the case of Poisson distribution, the Levy-flight distribution [15] was adjusted to achieve the 1:1 composition. It is seen that, similar to random-block sequences, the values of h_{JS} plotted as a function of average block length for the Levy-flight sequences goes though the maximum. For $N=128$, the maximum of h_{JS} is located at $L=4.31$. For all the block lengths considered, the value of h_{JS} found for the Levy-flight sequences is considerably greater as compared to that observed for the random-block sequences. On the other hand, the designed sequences for regime II have slightly higher values of h_{JS} in comparison with the Levy-flight sequences.

Figure 5(b), in which the sequence-dependent quantity h_{JS} is presented vs the conformation-dependent parameter z introduced above, illustrates the interplay between sequence and structure. Initially, in the region II, parameter h_{JS} increases with z, and then drops sharply in the region I. From these and other results we can see that sequence dictates structure and vice versa. A similar conclusion has been drawn by Dewey [45].

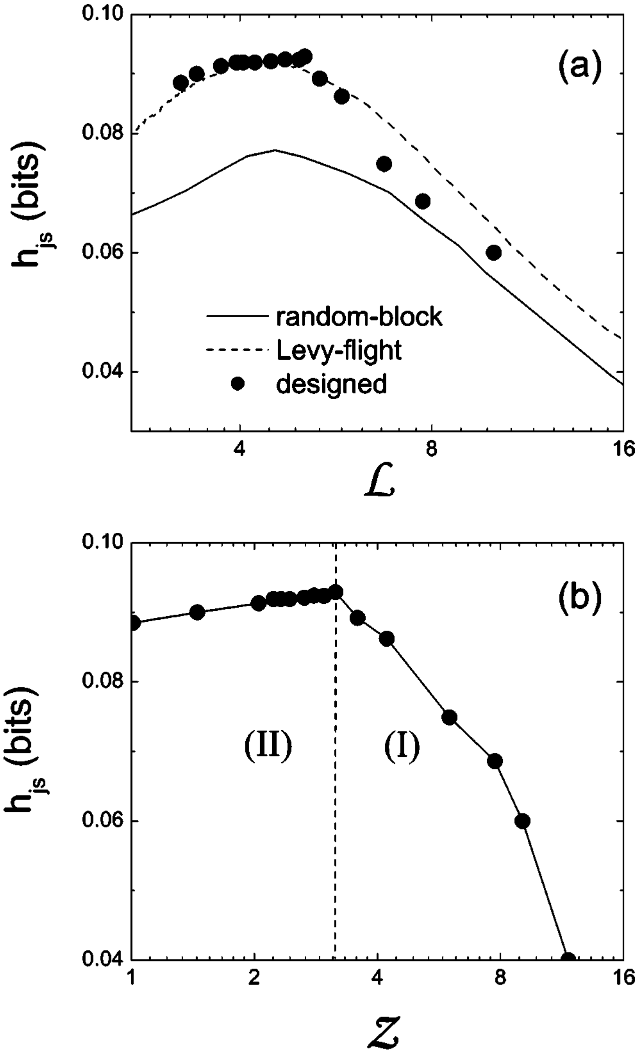

FIG. 5. (a) The Jensen-Shannon divergence for random-block, Levy-flight [15], and designed sequences of the same length, $N=128$, as a function of average block length L. (b) The Jensen-Shannon divergence for designed sequences as a function of ratio $z=R_{\mathcal{P}}/R_{\mathcal{H}}$. The splitting parameter n in Eq. (8) is set to be $n=N/2$.

E. Detrended fluctuation analysis

Having defined quantitatively the measure of information complexity of a sequence, let us now turn to the question of long-range correlations in the sequences generated via evolution mechanism described above. In Ref. [15], we have shown that already primary coloring procedure leads to long-range correlations of the Levy-flight type. The question is that whether these correlations are preserved after our multiple coloring evolution procedure?

Following Ref. [15], in order to monitor long-range statistical properties of generated sequences, we will employ the method similar to that used by Stanley and co-workers [46,47] in their search for long-range correlations in DNA sequences. In this approach, each $\mathcal{HP}$ sequence is transformed into a sequence of symbols 0 and 1. We choose the "window" of length $\mathcal{L}$, move it step by step along the sequence, and at each step count the number of $\mathcal{P}$ units inside the window. This number is a random variable with certain distribution and dispersion, $D_{\mathcal{L}}$. If the sequence is uncorrelated (normal random walk) or there are only local correlations extending up to a characteristic range (Markov chain), then $D_{\mathcal{L}}$ scales as $\mathcal{L}^{1/2}$ with the window width $\mathcal{L}$. A power law $D_{\mathcal{L}} \propto \mathcal{L}^{\alpha}$ with $\alpha > \frac{1}{2}$ would then manifest the existence of long-range (scale-invariant) correlations. However, due to large fluctuations, conventional scaling analyses of $D_{\mathcal{L}}$ cannot be applied reliably to the entire short sequence. To avoid this problem, we use the detrended fluctuation analysis, the method specifically adapted to handle problems associated with statistics of DNA sequences [46,47]. Following this approach, we calculate the so-called detrended fluctuation function $F(\mathcal{L})$, which characterizes the detrended local fluctuations within the window of length $\mathcal{L}$.

The functions $F(\mathcal{L})$ were averaged over the last 900 proteinlike sequences generated under stationary conditions (i.e., at $d_G^* \approx 0$) for $\varepsilon_{\mathcal{PP}} \geqslant 0.3$. For comparison, we calculated the same function for a purely random 1:1 sequence with $N = 128$ and found $F(\mathcal{L}) \propto \mathcal{L}^{1/2}$ scaling throughout the interval of $\mathcal{L}$ examined, as expected. Although the results for designed sequences do not fit accurately to any power low $F(\mathcal{L}) \propto \mathcal{L}^{\alpha}$ throughout the interval of $\mathcal{L}$ considered, the slope α observed for the dependence of $\ln[F(\mathcal{L})]$ on $\ln \mathcal{L}$ corresponds to a value significantly larger than $\frac{1}{2}$, up to about 1, thus indicating pronounced long-range correlations in a designed sequence. Moreover, we observe rather distinctly an increase in this initial slope when the parameter $\varepsilon_{\mathcal{PP}}$ is reduced from 2 to 0.3.

Therefore, we can conclude that the long-range correlations of Levy-flight type are preserved after evolution procedure in regime II. It should be emphasized that the most pronounced long-range correlations are observed near the singularity point, $\varepsilon_{\mathcal{PP}} = 0.3$.

IV. CONCLUDING REMARKS

Using a molecular-dynamics-based algorithm, we have simulated the conformation-dependent evolution of model two-letter ($\mathcal{HP}$) copolymer sequences. With this evolutionary process, structures and sequences are formed self-consistently. The following main questions were addressed: (i) whether this evolution can result in an increase in complexity of arising sequences or it ends up with some trivial (degenerated) sequence? and (ii) what is the interconnection between sequences and structures? To answer these questions, a 128-unit flexible-chain heteropolymer with the $\mathcal{HP}$ composition fixed at 1:1 has been simulated for conditions when hydrophobic $\mathcal{H}$ monomers strongly attract each other, thus stabilizing a dense globular core, while the attraction energy $\varepsilon_{\mathcal{PP}}$ between hydrophilic $\mathcal{P}$ monomers is considered as a parameter. For this model system, we have calculated various conformational-dependent and sequence-dependent properties, which are complementary to each other, including Shannon's entropy, Jensen-Shannon divergence measure, the compressibility of a sequence, the detrended local fluctuations characterizing $\mathcal{HP}$ distribution, etc. Using these quantities, we have found that, for the model under investigation, there are two regimes (branches) of evolution, depending on the energy parameter $\varepsilon_{\mathcal{PP}}$. If $\varepsilon_{\mathcal{PP}}$ is smaller than some crossover energy $\varepsilon^*_{\mathcal{PP}}$ ($\approx 0.3 k_B T$), the evolution can lead to a transition in sequence space from the sequences with proteinlike primary structures having relatively short $\mathcal{H}$ and $\mathcal{P}$ blocks to the degenerated (nonproteinlike) sequences having long uniform blocks, the length of which is close to half of the chain. The crossover energy $\varepsilon^*_{\mathcal{PP}}$ was found to be slightly smaller than the critical energy of coil-globule transition ε^* for a homopolymer chain of the same length. On the other hand, this energy parameter is greater than the critical adsorption energy of hydrophilic block on the surface of globule formed by hydrophobic chain section of a diblock $\mathcal{HP}$ copolymer. The degenerated primary structures correspond to diblock or triblock sequences with a small amount of randomly arising "defects," and from this viewpoint such a sequence can be treated as trivial. As can be seen from Fig. 3(a), the three-dimensional structures formed in this case look like a core-tail or a tadpole structure. Therefore, when the attraction between hydrophilic monomers is not sufficiently strong, we deal with the descending branch of the evolution, which leads to degenerate (nonproteinlike) sequences having low information content and low complexity. On the other hand, in the second regime (at $\varepsilon_{\mathcal{PP}} \geqslant \varepsilon^*_{\mathcal{PP}}$), the proteinlike structures have been found to be evolutionally stable (in the sense that initial sequence does not degenerate into a trivial one). The corresponding sequences are not random and exhibit strong correlations. For the sequences generated in this regime, it has been shown that the degree of complexity, as measured by the Jensen-Shannon divergence measure is considerably higher as compared to that observed for the first regime. The complexity increases with a decrease in $\varepsilon_{\mathcal{PP}}$ and reaches its maximum in the vicinity of $\varepsilon^*_{\mathcal{PP}}$. In addition, we have observed that as $\varepsilon_{\mathcal{PP}}$ approaches $\varepsilon^*_{\mathcal{PP}}$ long-range correlations become more pronounced. Taking into account all these results, one can attribute the corresponding evolutionary process to ascending branch of the molecular evolution, leading to more complicated structures. This conclusion is further supported by the analysis of the morphology of the proteinlike copolymer globules. In this case, we observe the formation of stable core-shell structures having dense hydrophobic core stabilized by hydrophilic envelope in a globular conformation.

ACKNOWLEDGMENTS

The authors acknowledge useful discussions with A. Yu. Grosberg and A. N. Semenov. The financial support from Alexander-von-Humboldt Foundation, Program for Investment in the Future (ZIP), INTAS (Project No. 01-607), and Russian Foundation for Basic Research is highly appreciated.

[1] J. Maynard Smith, *On Evolution* (Edinburgh University Press, Edinburgh, 1972); D. W. McShea, Evolution (Lawrence, Kans.) **50**, 477 (1996).

[2] A. Yu. Grosberg and A. R. Khokhlov, *Giant Molecules: Here and There and Everywhere*... (Academic Press, New York, 1997).

[3] A. Yu. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994).

[4] L. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, New York, 1972).

[5] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simon, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[6] A. R. Khokhlov and P. G. Khalatur, Physica A **249**, 253 (1998).

[7] P. G. Khalatur, V. I. Ivanov, N. P. Shusharina, and A. R. Khokhlov, Russ. Chem. Bull. **47**, 855 (1998).

[8] A. R. Khokhlov and P. G. Khalatur, Phys. Rev. Lett. **82**, 3456 (1999).

[9] (a) V. A. Ivanov, A. V. Chertovich, A. A. Lazutin, N. P. Shusharina, P. G. Khalatur, and A. R. Khokhlov, Macromol. Symp. **146**, 259 (1999); (b) A. V. Chertovich, V. A. Ivanov, B. G. Zavin, A. R. Khokhlov, Macromol. Theory Simul. **11**, 751 (2002).

[10] E. A. Zheligovskaya, P. G. Khalatur, and A. R. Khokhlov, Phys. Rev. E **59**, 3071 (1999).

[11] J. M. P. van den Oever, F. A. M. Leermakers, G. J. Fleer, V. A. Ivanov, N. P. Shusharina, A. R. Khokhlov, and P. G. Khalatur, Phys. Rev. E **65**, 041708 (2002).

[12] Yu. A. Kriksin, P. G. Khalatur, and A. R. Khokhlov, Macromol. Theory Simul. **11**, 213 (2002).

[13] J. Virtanen, C. Baron, and H. Tenhu, Macromolecules **33**, 336 (2000); J. Virtanen, H. Tenhu, *ibid.* **33**, 5970 (2000).

[14] V. I. Lozinsky, I. A. Simenel, E. A. Kurskaya, V. K. Kulakova, V. Ya. Grinberg, A. S. Dubovik, I. Yu. Galaev, B. Mattiasson, and A. R. Khokhlov, Rep. Russ. Acad. Sci. **375**, 637 (2000); P.-O. Wahlund, I. Yu. Galaev, S. A. Kazakov, V. I. Lozinsky, and B. Mattiasson, Macromol. Biosci. **2**, 33 (2002); M.-H. Siu, G. Zhang, and C. Wu, Macromolecules **35**, 2723 (2002).

[15] E. N. Govorun, V. A. Ivanov, A. R. Khokhlov, P. G. Khalatur, A. L. Borovinsky, and A. Yu. Grosberg, Phys. Rev. E **64**, 040903(R) (2001).

[16] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).

[17] E. I. Shakhnovich and A. M. Gutin, Protein Eng. **6**, 793 (1993).

[18] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **91**, 12972 (1994).

[19] E. I. Shakhnovich, Fold Des **3**, R45 (1998).

[20] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000).

[21] A. Irbäck, C. Peterson, F. Potthast, and E. Sandelin, Phys. Rev. E **58**, R5249 (1998); Structure (London) **7**, 347 (1999).

[22] Y. Iba, K. Tokida, and M. Kikuchi, J. Phys. Soc. Jpn. **67**, 3985 (1998).

[23] V. I. Abkevich, A. M. Gutin, E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **93**, 839 (1996).

[24] R. A. Broglia, G. Tiana, S. Pasquali, H. E. Roman, and E. Vigezzi, Proc. Natl. Acad. Sci. U.S.A. **95**, 12 930 (1998).

[25] P. Gupta, C. K. Hall, and A. C. Voegler, Protein Sci. **7**, 2642 (1998).

[26] S. Istrail, R. Schwartz, and J. King, J. Comput. Biol. **6**, 143 (1999).

[27] G. Giugliarelli, C. Micheletti, J. R. Banavar, and A. Maritan, J. Chem. Phys. **113**, 5072 (2000).

[28] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[29] K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 6388 (1990).

[30] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Claredon Press, Oxford, 1990).

[31] L. Wesson and D. Eisenberg, Protein Sci. **1**, 227 (1992); J. D. Augspurger and H. A. Scheraga, J. Comput. Chem. **17**, 1549 (1996).

[32] C. E. Shannon, *Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

[33] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); M. L. Rosenzweig, *Species Diversity in Space and Time* (Cambridge University Press, New York, 1995).

[34] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, J. Mol. Biol. **188**, 415 (1986).

[35] G. J. Chaitin, J. Assoc. Comput. Mach. **13**, 547 (1966).

[36] T. Alvager, G. Graham, D. Hutchison, and J. Westgard, J. Chem. Inf. Comput. Sci. **37**, 335 (1997).

[37] S. Grumbach and F. Tahi, J. Inf. Processing Management **30**, 875 (1994).

[38] J. Ziv and A. Lempel, IEEE Trans. Inf. Theory **23**, 337 (1977).

[39] M. Li and P. M. B. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications*, 2nd ed. (Springer-Verlag, New York, 1997).

[40] P. Grassberger, Int. J. Theor. Phys. **25**, 907 (1986).

[41] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).

[42] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).

[43] A. Irbäck and E. Sandelin, J. Chem. Phys. **110**, 12256 (1999).

[44] I. M. Lifshitz, A. Yu. Grosberg, and A. R. Khokhlov, Rev. Mod. Phys. **50**, 683 (1978).

[45] T. G. Dewey, Phys. Rev. E **60**, 4652 (1999).

[46] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).

[47] N. V. Dokholyan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, Phys. Rev. Lett. **79**, 5182 (1997); S. V. Buldyrev, N. V. Dokholyan, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, and G. M. Viswanathan, Physica A **249**, 430 (1998).