Eur. Phys. J. E **13**, 15–25 (2004)
DOI: 10.1140/epje/e2004-00036-1

THE EUROPEAN
PHYSICAL JOURNAL E

Conformation-dependent sequence design: evolutionary approach

A.V. Chertovich[1,a], E.N. Govorun[1], V.A. Ivanov[1], P.G. Khalatur[2,3], and A.R. Khokhlov[1,3]

[1] Physics Department, Moscow State University, 119992 Moscow, Russia
[2] Department of Physical Chemistry, Tver State University, 170002 Tver, Russia
[3] Department of Polymer Science, University of Ulm, 89069 Ulm, Germany

Received 14 January 2003 / Received in final form 3 September 2003
Published online 2 March 2004 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2004

Abstract. A new modification of evolutionary approach to sequence design of copolymers has been proposed. A model of step-by-step evolution of a two-letter (HP) copolymer sequence has been studied by means of a coarse-grained Monte Carlo algorithm. The conditions for accepting a change in the primary sequence depend on the spatial conformation of HP-copolymer chain. This leads to a coupling between sequence and conformation and to formation of protein-like conformations and primary sequences (for some values of parameters of the model) independently of initial sequence and/or conformation. Simple theory describing these computer simulation observations is developed.

PACS. 87.15.Aa Theory and modeling; computer simulation – 87.15.Cc Folding and sequence analysis – 82.35.Jk Copolymers, phase transitions, structure – 82.35.Lr Physical properties of polymers

1 Introduction

The concept of evolution of primary sequences of biopolymers attracts large interest of biologists, chemists and physicists for a long time [1–4]. The present day biopolymers (proteins, DNA, RNA) possess complicated sequences of monomer units which encode their very sophisticated functions and unique spatial structure. These sequences are statistically very different from random ones and often exhibit significant correlations on different scales [5,6]. Therefore, it is natural to expect that the content of information in these sequences is relatively high in comparison with random sequences where it should be almost zero [5,6]. Presumably, the information complexity of early ancestors of present day biopolymers has been increased in the course of molecular evolution when the copolymer sequences became more and more complicated. 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. It is worthwhile to note that since the information content of a sequence can be represented as a mathematically defined quantity, the whole process of evolution of biopolymer sequences can be specified in exact mathematical terms. The formulated fundamental problem is extremely difficult because of 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 of long-range correlations in the sequences.

It is clear that information complexity cannot emerge just as a result of random mutations. Some coupling of mutations to other factors is necessary. It is most straighforward and natural to introduce the interrelation of mutations and conformations, i.e. to consider conformation-dependent sequence design in the context of evolution.

The idea of conformation-dependent sequence design of sequences in copolymers was first introduced in the papers [3,4] in order to solve some of the problems of protein physics.

Another variant of conformation-dependent sequence design for copolymers which in one step leads to rather complicated statistical sequences has been recently proposed in reference [7]. An important example of such design corresponds to a protein-like copolymer: first, a dense homopolymer globule is obtained (the so-called "parent conformation"), then all monomer units in the inner core are set to be hydrophobic (H-type), while units in the outer shell are set to be polar (P-type). Protein-like copolymers do not model directly proteins, but rather represent synthetic "functional" polymers mimicking some properties of biological macromolecules. One of the important features of globular proteins is the possibility of formation of a dense core stabilized by hydrophilic envelope. It is due to this feature that proteins do not precipitate in the solution, while taking a globular conformation.

Possible ways of experimental realization of conformation-dependent sequence design for protein-like

[a] e-mail: chertov@polly.phys.msu.ru

copolymers are described in references [8,9]. The conformation-dependent sequence design was realized in computer simulations for several schemes: protein-like copolymers, AB-copolymers tuned to adsorption on a plane surface, AB-copolymers which mimic membrane proteins, etc. [10,11]. For certain conditions, designed copolymers are able to memorize and then to reproduce some features of parent conformation. Several theoretical considerations of protein-like copolymers were proposed recently in references [12,13]. The statistical characteristics of the sequences were calculated assuming the picture of random walk for the chain conformation inside the dense globule. Such protein-like sequences were shown to correspond to the Levy-flight statistics. These statistical characteristics are in agreement with the results of computer simulations. Recently, the concept of evolution of sequences within the scheme of generation of protein-like copolymers has been introduced explicitly in reference [14] and a modification of this scheme has been investigated in reference [15].

The aim of the present paper is to propose a new modification and generalization of the evolutionary approach for this design method. In the present investigation we consider HP-copolymer molecules of 1:1 composition with interacting units and with evolutionary-designed conformation-dependent sequences. We assume that H-units strongly attract each other. In this case, the dense core is formed for any HP-sequence. The state of the system is characterized by conformation in the ordinary $3d$ space and by the specific sequence in the "space of sequences".

In the literature, some computer models describing the evolution of copolymer sequences have been proposed [3,4,16]. Most of them are based on a stochastic Monte Carlo (MC) optimization principle (Metropolis scheme) and aimed at the problems of protein physics. E.g. one of the problems is to design such protein sequences, which are thermodynamically stable in a target three-dimensional conformation and are able to fold fast into this conformation at a given temperature (see, e.g., Refs. [3,4,16–19]). Such optimization algorithms start with arbitrary sequences and proceed by making random substitutions biased to minimize relative potential energy of the initial sequence and/or to maximize folding rate of the target structure. Abkevich et al. [20] have considered a model of protein evolution where polypeptide sequences are mutated under influence of two competing factors: possibility of globular aggregation and hydrolysis of the protein globule. The competition between these two factors – compactness and solubility – can gradually shift the sequence distribution toward the more compact and more stable conformation with a unique primary structure [20]. The relevance of chain aggregation for the folding process has also been studied in references [21–24].

In the present work, the simultaneous evolution of sequences and conformations is studied. For sequences, we consider the information entropy and other statistical characteristics. The design procedure leads to the final state, which depends on the set of interaction parameters and on the rearrangements both in conformational space and in the space of sequences. These rearrangements are characterized by the usual thermodynamic temperature (T_{conf}) for conformational space, as well as effective sequence rearrangement temperature (T_{seq}) which is introduced in the present work.

Namely, after certain number of Monte Carlo steps in conformational space (in the course of this process the system equilibrates at temperature T_{conf}) we try the possibility of mutation of two randomly chosen monomer units: monomer unit H converts into P and vise versa. If this leads to the decrease in the energy of the globule this move is accepted. If the globular energy is increasing by the amount ΔE, this move is accepted with the probability $\exp(-\Delta E/kT_{seq})$.

In general, we can define evolution of sequences for any value of T_{seq}. However, the simulation for three characteristic cases can be most easily understood.

1. The inequality $T_{conf} \ll T_{seq}$ means that all the moves in the sequence space are accepted independent of conformations. This corresponds to random mutations, and the final sequence (after long evolution) will be that of a random HP-copolymer (no information complexity).

2. The inequality $T_{conf} \gg T_{seq}$ means that only the moves leading to the decrease of the globular energy E are accepted. Such evolution should lead to a sequence corresponding to a minimum of globular free energy in conformational space. For the case of absence of any attractive interactions between P-units (i.e. polar groups in the shell of protein-like globule; we will consider this case in the present paper) it was shown in reference [14] that the final sequence after evolution should have hydrophobic core with very few hydrophilic (or polar) loops and a long hydrophilic tail. This sequence is close to that of HP diblock copolymer, and should not exhibit any information complexity.

3. The case $T_{conf} = T_{seq}$ corresponds to annealed HP-sequence. This case is equivalent to the situation when monomer unit H can be converted into P by attaching some ligand $\mathcal{L}$:

$$P \rightleftarrows H + \mathcal{L}.$$

Number of ligands is fixed to maintain 1:1 H/P composition, however they can choose which monomer unit to bind. This defines, in particular, the chain sequence. Annealed HP sequences in the context of polymer globules were first considered by Grosberg in reference [25] (see also [26,27]). It was shown that non-trivial sequences leading to core-shell globular structures are possible. Our computer simulations (see below) support this conclusion. We will see that even in the absence of attraction of P-units the final sequences remain protein-like (i.e. corresponding to the core-shell structure of a globule) and therefore maintain certain information complexity.

In the general situation, when T_{seq} is not corresponding to any of the characteristic cases described above, the evolution of sequences coupled with conformations can be still defined in the same way. One has only to remember that if $T_{conf} \neq T_{seq}$ there is a flow of heat between conformational space and space of sequences, so full thermodynamic equilibrium is impossible. Still, we can be in a stationary regime corresponding to the sequences tending to a certain fixed point, and possessing (or not possessing) information complexity.

It should be emphasized from the very beginning that the problem which we address in the present study is somewhat different from that usually discussed in the context of protein physics. Unlike references [3,4,16–24] cited above, we are not aimed at searching for unique tree-dimensional (native) conformation with fast folding rate. On the contrary, we are interested in the non-unique conformations, i.e., with a large entropy. In general, our aim is to learn whether it is possible to make with synthetic copolymers a step along the same line as molecular evolution.

This paper is organized as follows. In Section 2 we introduce the model and describe our simulation technique. Then we present our computer simulation and analytical theory results in Sections 3 and 4. In Section 5 we make some concluding remarks.

2 Model and simulation technique

We will work in the statistical ensemble which we denote here as CS-ensemble (for the sake of brevity) which includes both a spatial conformation and a primary sequence of a given copolymer chain. The energy of a chain, $E(c,s)$, depends on its spatial conformation, c, which is defined by the position vectors of all monomer units of the chain, and on the primary sequence of units of different types, s. We consider the general case when there are two separate temperatures T_{conf} and T_{seq}, controlling the conformation and sequence in a separate manner. Thus, we combine sequence search for target conformation corresponding to the temperature in sequence space T_{seq} (cf., e.g., Refs. [3,4,16]) and conformation search for particular sequence with conformational temperature T_{conf}.

We will consider a lattice model of polymer chain. Simulations were performed using the bond fluctuation model [28] and the Monte Carlo (MC) scheme [29]. A single HP-copolymer chain is taken in an initial globular state, and then we start a procedure of the evolution (annealing) of the primary sequence together with the change of the chain conformation. The composition of the chain is fixed (1:1), P-units interact with P- and H-units only via excluded volume ($\varepsilon_{PP} = \varepsilon_{PH} = 0$), H-units are assumed to attract each other ($\varepsilon_{HH} < 0$). Such choice of interaction parameters leads to the formation of core-shell conformations (with H-core and P-shell). The iterative procedure for sequence evolution for the chain of a given length N consists of many mutation steps interchanged with partial equilibration of spatial conformation:

(i) Two monomer units are chosen randomly and, if they happen to be of different types, an attempt is made to exchange their types ($H \leftrightarrow P$, see Fig. 1). According to the Metropolis scheme the probability of such an exchange is set to 1 if $\Delta E \leq 0$ and to $\exp(-\Delta E/k_B T_{seq})$ if $\Delta E > 0$, where ΔE is the difference between the interaction energies of new and old configurations, T_{seq} is the "sequence design temperature" (see also the Introduction). This is a trial move in the space of sequences (mutation). We repeat this procedure N_{seq} times. N_{seq} characterizes the speed of mutation process, or how many point-mutations can be performed during one evolution step.

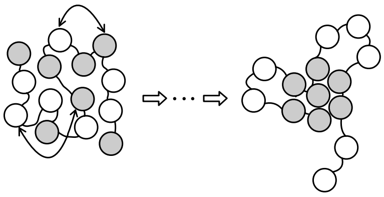

Fig. 1. The trial step of sequence design: two monomer units of different types are randomly chosen, the attempt to exchange their types is tried. The exchange probability is determined according to the Metropolis scheme.

(ii) Then we perform Monte Carlo simulations of chain motion at a fixed temperature T_{conf} in the conformational space during the period of time $\tau_{relax} = 1000$ MC steps. This period is not sufficient for complete relaxation of the system, but allows local structural rearrangements.

After that we return to step (i). The condition of microscopic reversibility in our MC algorithm is satisfied.

As initial primary structures, we take sequences with different statistics: protein-like copolymer sequences, statistically random and diblock-copolymer sequences. Several (about 5) realizations were tried in all cases. Interaction parameters were set to $\varepsilon_{PP} = \varepsilon_{PH} = 0; \varepsilon_{HH} = -1$. During simulations, we set $T_{conf} = 1.42$. This temperature is small enough to ensure the formation of globular structure from statistically random HP-sequence and high enough in order not to freeze spatial structure. The composition of H- and P units was chosen to be 1:1. We studied the chains of lengths $N = 1024$ and 256. All results concerning statistical characteristics of the sequences are shown below for 1024-unit chain, thermodynamic properties are presented for 256-unit chain. We have chosen the value $N_{seq} = 0.1N$. In other words, 10% of all monomer units were subjected to mutation procedure at each mutation step.

From the autocorrelation function for the end-to-end distance we estimated the relaxation time of the system at $T_{conf} = 1.42, N = 256$ to be about 10^4 Monte Carlo steps (MCS). Note that this time is larger than the time between two mutation attempts $\tau_{relax} = 1000$ MCS (see above). The energy histograms were gathered during 10^6 MCS after initial 10^6 MCS equilibration. All results on the system behavior depending on evolution time (number of evolution steps) were obtained for 4×10^3 evolution steps and averaged over 10 independent runs. Thus, $10 \times 4 \times 10^3 \times 10^3 = 4 \times 10^7$ MC steps were performed for each set of parameters.

A common approach for the analysis of complex systems is to use concepts from information theory and informational theoretic-based techniques [5,30,31]. Within this approach, copolymer sequences can be examined as messages written in two-symbol alphabet (letters H and 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 the Shannon entropy and the degree of complexity of copolymer sequences. Let $s_1 s_2 ... s_N$ be the symbols of a given sequence S of length N. If $f_n(s_1 s_2 ... s_n)$ is the average frequency of a subsequence with the symbols $s_1 s_2 ... s_n$, i.e., of the "word" $s_1 s_2 ... s_n$ of length $n < N$, then the Shannon's entropy of the whole sequence is given by [30]:

$$S = -k_B \sum_{all\ words} f_n(s_1 s_2 ... s_n) \ln[f_n(s_1 s_2 ... s_n)] \qquad (1)$$

where k_B denotes the Boltzmann constant and the summation runs over all possible "words". Note that the frequencies in (1) can be replaced by the corresponding probabilities if the sequence would have infinite length. If k_B is replaced by $1/ln2$, then S quantifies the amount of information in units of bits [31]. Of course, the Shannon entropy depends on the definition of a set of "words" in the sequence. For our case of two-letter HP copolymer, we will adopt the following set of "words" (uniform blocks): H, HH, HHH,..., P, PP, PPP,..., i.e., "word" (block) is defined by its length n and type (H or P). Then the Shannon entropy per monomer unit can be written as (see Appendix for accurate derivation of this formula)

$$S_{seq} = -\frac{\tilde{n}}{2N}\left[\sum_k p_H(k)\log_2 p_H(k) + \sum_k p_P(k)\log_2 p_P(k)\right] \qquad (2)$$

where $p_H(k)$ and $p_P(k)$ are the probabilities of "words" of length k [14,33], $\tilde{n}$ is the total number of "words" composed of letters H and P, and $0 \cdot log_2 0 = 0$ is assumed by continuity.

Thus statistical properties of the sequences for our system are characterized by the Shannon entropy (2) and by the dispersion of the number of $H(P)$-units in the segment of the length L within the sequence:

$$D^2(L) = \sum_{i,j=m}^{m+L} (\langle u_i u_j \rangle - \langle u_i \rangle \langle u_j \rangle), \qquad (3)$$

where u_i denotes the type of a monomer unit, $u_i = 1$ for H-unit and $u_i = 0$ for P-unit, (the number of H-units in the window L starting at the monomer unit m is equal to $\sum_{i=m}^{m+L} u_i$, brackets in (3) mean the averaging $\langle A \rangle = \frac{1}{N-L} \sum_{m=1}^{N-L} A_k$). The latter value characterizes the type of correlations along the chain, similar methods were used for the correlation analysis in DNA sequences [32].

Usually $D(L)$ can be approximated by the scaling law $D(L) \sim L^\alpha$. For completely random sequences, α is equal to 0.5; the value $\alpha > 0.5$ corresponds to long-range correlations in the sequence. The long-range correlations in protein-like sequences were predicted in reference [12]; for example, the value $\alpha \approx 0.78$ was obtained for a globule of length $N = 1024$.

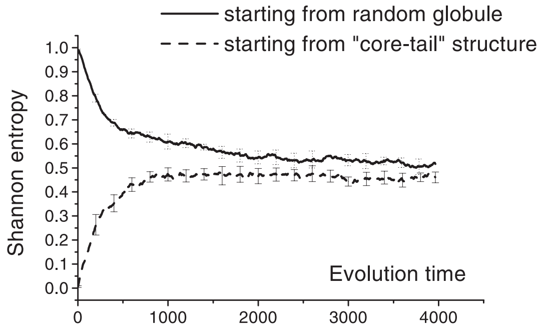

Fig. 2. Shannon's entropy vs. number of mutation steps (evolution time) for two different initial sequences and conformations, $N = 1024$.

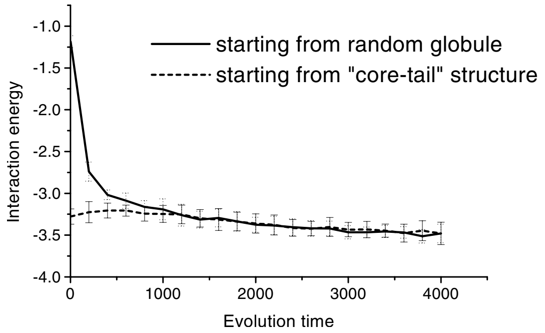

Fig. 3. System interaction energy vs. number of mutation steps (evolution time) for two different initial sequences and conformations, $N = 1024$. Asymptotic value (around −3.4) corresponds to dense H-core.

3 Computer simulation results

The first question is whether our scheme for evolution of sequences leads to some fixed point in the sequence space and how many mutations are needed for the system to come to this stationary state. In order to answer this question we studied statistical parameters of the sequence (Shannon's entropy (2), mean block length) and interaction energy (characterizing behavior in conformational space), as a function of the number of mutation attempts (evolution time). The case of equal temperatures ($T_{seq} = T_{conf} = 1.42$) was considered. Figures 2, 3 and 4 present the Shannon's entropy, system interaction energy, and mean block length as a function of the number of performed mutations. The curves are plotted for two initial sequences and conformations: a random copolymer globule and a "core-tail" structure corresponding to diblock copolymer. All considered values tend asymptotically to the same limit independently of the initial sequence. In this final state the average block length k_{av} is equal to 6.5 and the mean Shannon's entropy is $S_{seq} = 0.5$.

Let us now compare the properties of HP-copolymers obtained via our evolutionary procedure and via normal "coloring" procedure for protein-like copolymers originally proposed in [7]. The energy histograms for the systems corresponding to protein-like sequences obtained by instant coloring and by optimization via evolutionary annealing are presented in Figure 5. Data were obtained by

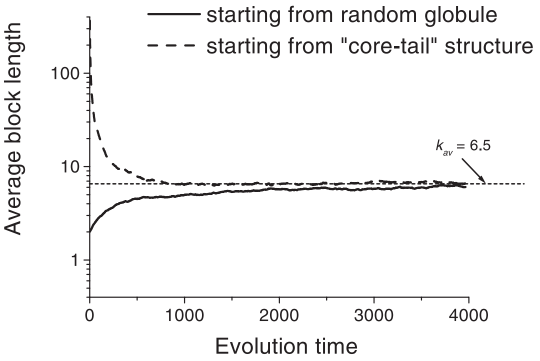

Fig. 4. Average block length vs. number of mutation steps (evolution time) for two different initial sequences and conformations, $N = 1024$. Asymptotically, $k_{av} \approx 6.5$.

equilibration of the system (prepared at $T_{conf} = 1.42$) at several different temperatures, for $N = 256$. Plot (a) corresponds to "core-shell" conformations for both cases. Plot (b) corresponds to the temperature close to coil-globule transition, when most of the instantly colored protein-like copolymers are in a coil state, while optimized sequences after evolutionary annealing are still in the "core-shell" conformation. Plots (c) and (d) correspond to coil conformations. The conformations obtained by evolutionary method have larger θ-temperature and are more stable: their histograms always have a bias to smaller energy values (in comparison with the histograms for instantly colored copolymers).

The variation of gyration radii for H- and P-units in the course of evolutionary procedure are presented in Figure 6 (for $N = 1024$). Again, the averaging over 10 independent runs was performed. After evolution, the values of $\langle R^2_{gyr} \rangle$ for core and shell units differ considerably from each other revealing the formation of stable "core-shell" structure. In spite of averaging we observe significant fluctuations in gyration radius of hydrophilic units. This effect is caused by pronounced fluctuations of polymer chains in a good solvent [34].

In Figure 7, we show the dependencies of $D(L)$ (see Eq. (3)) for protein-like sequences, calculated by evolutionary procedure and by instant coloring, as well as for completely random sequences. The curve for evolutionary coloring has larger slope than that for instant coloring, hence these sequences correspond to more pronounced long-range correlations. The upper line with the slope equal to 1 reflects the behavior of $D(L)$ for maximally "nonrandom" sequence, with $\alpha = 1$.

In addition to the case $T_{seq} = T_{conf} = 1.42$, we examined also the situation when $T_{conf} = const$ while T_{seq} is variable. We consider the case when all forces between monomer units are the same as before, i.e. correspond to $\varepsilon_{PP} = \varepsilon_{PH} = 0; \varepsilon_{HH} = -1$. On the other hand, evolutionary pressure due to the sequence variation can differ. In particular we studied the following two cases: $T_{seq}/T_{conf} = 10$ ($T_{conf} = 1.42$) and $T_{seq}/T_{conf} = 0.1$ ($T_{conf} = 1.42$). Results are presented in Figure 8. Like previously, we performed averaging over 10 independent runs, initial structures are random globule and core-tail structure. From the

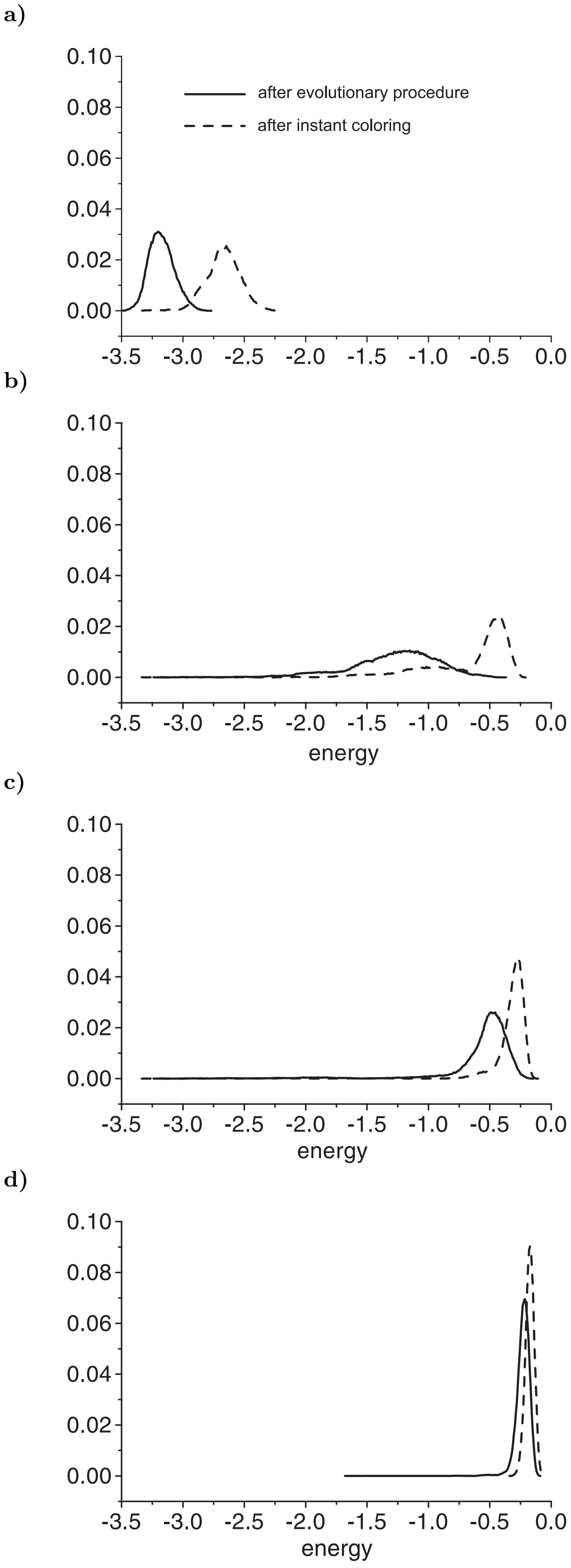

Fig. 5. Energy histograms for the macromolecules prepared at $kT = 1.42$ by two different methods (evolutionary procedure and instant coloring) and equilibrated at different temperatures: $kT = 1.42$ (a), 2.27 (b), 2.5 (c), 10.0 (d). ($N = 256$).

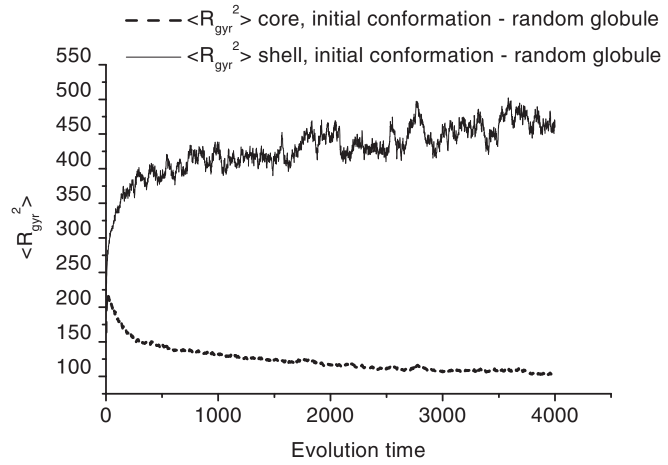

Fig. 6. Gyration radius of core (H-units) and shell (P-units) vs. number of mutations (evolution time), $N = 1024$.

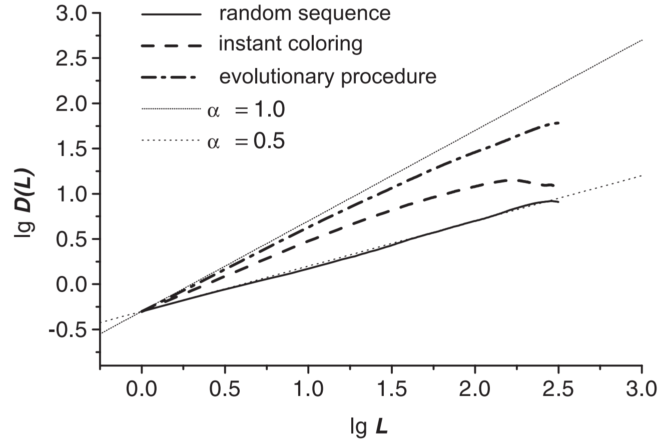

Fig. 7. The dependencies $D(L)$ for random, protein-like (instant coloring) and protein-like (evolutionary procedure) sequences in double logarithmic scale. The chain length was taken equal to $N = 1024$.

plots presented in Figure 8 it is seen that the sequence temperature T_{seq} plays an important role in the formation of core-shell structures. In the case of high T_{seq}, final sequence is statistically random, with mean block length $k_{av} = 2$. Shannon's entropy is close to 1 for this system. Initial sequences come to fixed point very quickly (in fact, random globule is almost in the state corresponding to the fixed point from the very beginning). Their conformations correspond to a coil, according to mean interaction energy values. Same conclusion can be drawn from the analysis of gyration radii (plots are not shown here).

Let us now consider an opposite case corresponding to low sequence design temperature, $T_{seq} = 0.142$. In this case, the interaction energies for both initial conformations after some evolution become close to each other; their values correspond to compact hydrophobic core (the same results follow from the analysis of gyration radius). This means that we achieve some stationary state in conformational space. However, it is much more difficult to achieve stationary state in sequence space. This is due to

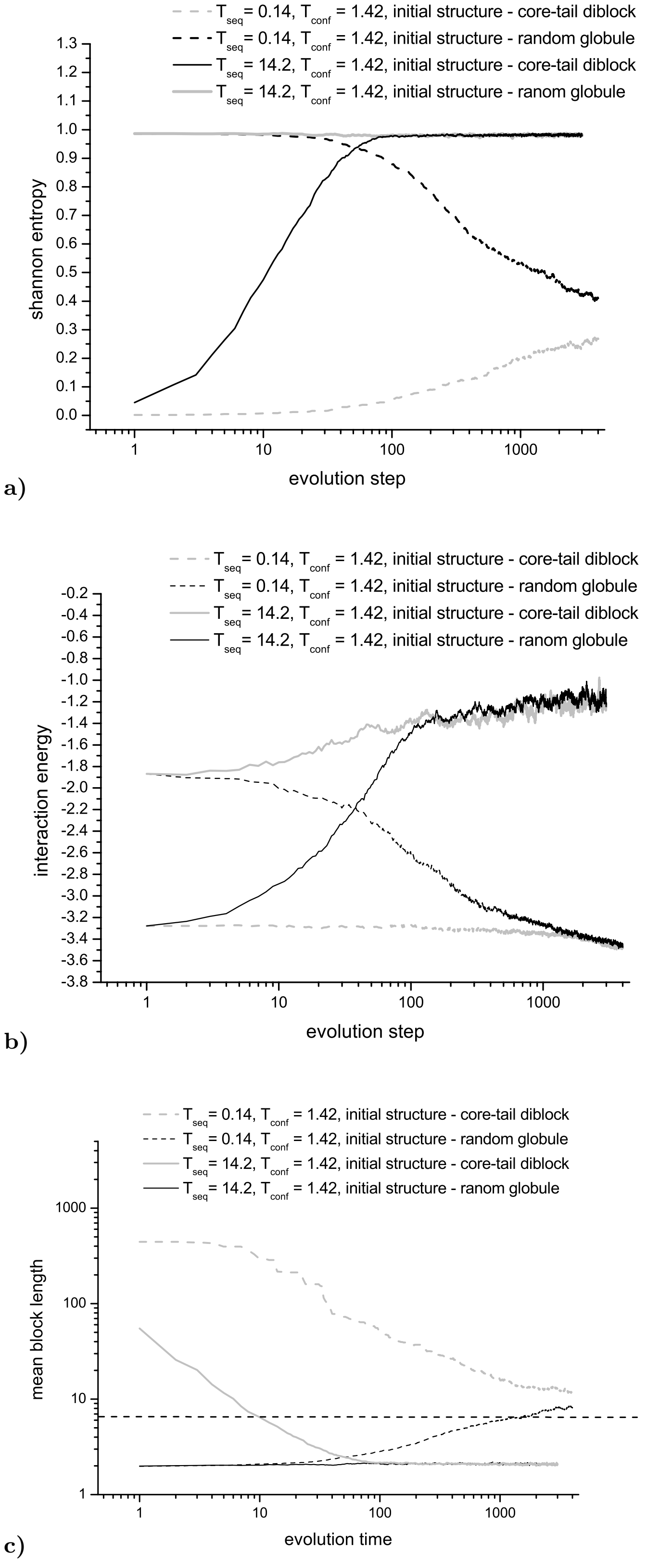

Fig. 8. Interaction energy (a), Shannon's entropy (b) and sequence mean block length (c) as a function of evolution time (logarithmic scale) for $T_{seq} \neq T_{conf}$. Horizontal dash line in Figure (c) corresponds to $T_{seq} = T_{conf} = 1.42$.

the fact that T_{seq} is low and all motions in sequence space are rather slow. Nevertheless, one can make assumption that there is an asymptotical convergence of statistical characteristics for the sequences obtained from different initial structures. Moreover, mean values of Shannon's entropy and block lengths considerably differ from the values corresponding to $T_{seq} = T_{conf} = 1.42$. A 10 times temperature decrease causes the increase of average block length from 6.5 up to approximately 10.5, while Shannon's entropy decreases from 0.5 to ca. 0.35. This means a trend to degeneration of the sequence, forming structures with dense core and few long loops. Analogous results were obtained recently in the paper [14]. Our procedure corresponds to "repeated coloring" considered in [14] if we set $T_{seq} = 0$ and $N_{seq} \to \infty$ (infinite number of mutation attempts per one evolution step).

Thus, we have demonstrated that by changing sequence temperature (i.e., mutation probability) the system can obey to different regimes of behavior: from evolution to random coil till emerging of degenerated "core-tail" structures. Besides, there are intermediate values of T_{seq} which allow the formation of core-shell structures. In this case the informational content of the sequence is maximally rich.

4 Theoretical description

In the present section we investigate theoretically the factors controlling the formation of the core-shell structure and describe the sequence characteristics corresponding to a final state.

The first obvious condition of the core-shell structure formation is the strong immiscibility of hydrophobic units with polar units and with the solvent. We assume that hydrophobic blocks form the dense core of radius R_c determined by the attraction between these units at the temperature which is lower than the coil-globule transition temperature [35].

Polar blocks in the shell are immersed in a solvent. We assume that polar units can weakly attract each other ($\varepsilon_{PP} \leq 0$), and their interaction is described by the second virial coefficient B_P. To estimate, whether they overlap or not, we should compare the area of the core surface per polar block with the squared block size in a good solvent. The area of the core surface per polar block is equal to $S_{block} = 4\pi R_c^2/\tilde{n}$, where $\tilde{n}$ is the number of polar blocks, while the average block length is $k_m = N/2\tilde{n}$. The block size in a good solvent R_0 can be estimated as [34]: $R_0 \approx a k_m^{3/5}(B_P/a^3)^{1/5}$, where a is the statistical segment length of the chain, $B_P \sim v$ for a chain in a good solvent, v is the volume of a monomer unit. For the values of the parameters related to the computer simulations: $N = 1024$, $R_c \approx 12b$, where b is the lattice step, $a \approx \sqrt{8}b$, $v \approx 8b^3$ (for the bond fluctuation model), we obtain that $S_{block} \approx R_0^2$ for all reasonable values of k_m. Hence, polar blocks immersed in a solvent do not overlap and their contributions to the free energy are additive.

If the core-shell structure still does not correspond to the stationary state, the sequence transforms slowly in the direction of increase or decrease of the number of blocks. We introduce the free energy of the system $F(R_c, k_m)$ as a function of the average block length k_m. The free energy F consists of the free energy of the H-core F_{core}, the contribution of H- and P-blocks F_{block}, and the sequence entropy contribution F_{seq}:

$$F(R_c, k_m) = F_{core}(R_c) + F_{block}(R_c, k_m) + F_{seq}(k_m). \quad (4)$$

The term $F_{core}(R_c)$ corresponds to the free energy of a hydrophobic globule of radius R_c [35]. The block contribution F_{block} is the free energy of the block end localization at the core surface and of the interaction of polar units. The sequence contribution is determined by the Shannon's entropy of the chain: $F_{seq} = -T_{seq}S_{seq}\log 2$.

Of course, writing the expression for the free energy for the situation when there are two temperatures (in the conformational space and in the sequence space) cannot be strictly justified from the viewpoint of statistical mechanics. However, introducing the addition F_{seq} to other "conformational" contributions to the free energy is: (a) definitely strictly valid for the case $T_{seq} = T_{conf}$ of annealed sequences (see the Introduction); (b) for the case $T_{seq} \neq T_{conf}$ it represents the correct trend of enhancement of randomization of sequences with increasing T_{seq}. We will look for the minimum of the constructed free energy (4) with respect to the average block length k_m, and we will see that the results obtained are quite reasonable and are in agreement with the computer simulations presented above.

The core free energy F_{core} dominates in the total free energy of a core-shell structure F, other terms are assumed to be of the order of the surface energy of a globule (for $k_m \gg 1$). On the other hand, the core free energy F_{core} is independent of k_m in the course of the evolution of a core-shell structure.

The block contribution into the free energy can be written as

$$F_{block}(R_c, k_m) = \tilde{n}\left(f_H(R_c, k_m) + f_P(R_c, k_m)\right), \quad (5)$$

where f_H and f_P are the free energies per hydrophobic and polar block, correspondingly.

The H-block free energy is determined by the statistical weight G_H of the block of length k_m with both ends located at the core surface. The P-block free energy is determined by the corresponding statistical weight G_P and by the polymer-solvent interaction energy f_{P-s}:

$$\begin{aligned} f_H &= -k_B T_{conf} \ln G_H(R_c, k_m), \\ f_P &= -k_B T_{conf} \ln G_P(R_c, k_m) + f_{P-s}. \end{aligned} \quad (6)$$

The ends of P- and H-blocks should be located within the surface volume $V_{surf} \sim R_c^2 D$, where D is the width of the globular surface layer. The probability that the end of a long H-block ($k_m \gg (R_c/a)^2$) returns to the surface is equal to $G_H \approx V_{surf}/V_c$, where V_c is the core volume: $V_c = 4\pi R_c^3/3$. The width of the surface layer D is approximately equal to $D \approx a/\phi$, where ϕ is the volume fraction of units in the core [35].

The probability that the end of a P-block is located at the surface is $G_P \approx V_{surf}/V_0$, where $V_0 = 4\pi R_0^3/3$. Following the Flory approach [36,34] one can write for the block free energy in a good solvent

$$f_{P-s} \approx k_B T_{conf} k_m^{1/3} (B_P/a^3)^{2/3}$$

for a block of size R_0.

The sequence entropy contribution F_{seq} can be taken in the form (2)

$$F_{seq} = k_B T_{seq}\ \tilde{n} \sum_k \left(p_P(k) \cdot \ln p_P(k) + p_H(k) \cdot \ln p_H(k)\right). \tag{7}$$

The characteristic width of the block length distributions is assumed to be of the order of k_m, in accordance with the dependencies $p(k)$ obtained in the computer simulations. Using a very rough estimation that all blocks of length $k < k_m$ can be found with the equal probability $p_P(k) \approx p_H(k) \approx 1/k_m$, we obtain

$$F_{seq} = -2k_B T_{seq} \tilde{n} \ln k_m. \tag{8}$$

As stated above, we assume that the sequence characteristics after the evolution correspond to the minimum of the total free energy F with respect to the mean block length k_m. The core radius and volume fraction are related via the obvious formula $\phi = Nv/2V_c$. Then, the expression for F can be rewritten using (5-8):

$$\begin{aligned} F(R_c, k_m) &= F_{core}(R_c) \\ &+ k_B T_{conf} \frac{N}{2k_m} \left(\left(\frac{9}{5} - 2\frac{T_{seq}}{T_{conf}} \right) \ln k_m + k_m^{1/3} \tilde{B}^{2/3} \right. \\ &\left. + \frac{3}{5} \ln \tilde{B} - \ln\left(\frac{9R_c}{16\pi^2 \phi^2 a} \right) \right), \end{aligned} \tag{9}$$

where $\tilde{B} = B_P/a^3$, $k_m \gg (R_c/a)^2$.

A core-shell structure which is stable in the course of evolution of sequences can be formed if the free energy minimum corresponds to a certain finite value of k_m^* ($k_m^* \gg 1$) depending on the core size and chain rigidity. The stability of the structure for long blocks is controlled by the increase of the block free energy with increasing the number of blocks and by the increase of the sequence free energy with decreasing that number.

If the factor before $\ln k_m$ in equation (9) is positive and, hence, the sequence design temperature T_{seq} is less than the threshold value T_{th} ($T_{th}/T_{conf} = 0.9$), then the free energy (9) decreases with the increase of k_m. In this case the structure with more long P-loops is favorable and, finally, it should contain a small number of loops or even only one polar block ("core-tail" structure). On the other hand, the free energy (9) can increase with k_m if $T_{seq} > T_{th}$ (see Fig. 9). In this case the structure with shorter blocks is more favorable and a core-shell structure can be formed, in agreement with the computer simulation results presented in Section 3.

Besides, the formation of the core-shell rather than core-tail structure is controlled by the interaction between P-units and by the core size, which, in turn, is determined by the chain length. If $T_{seq} > T_{th}$, then the free energy $F_{block} + F_{seq}$ increases or decreases with k_m for $N = const$ depending on the value of the second virial coefficient B_P. The higher values of the parameter B_P correspond to stronger repulsion between P-units leading to the higher free energy per polar loop. Hence, the decrease of the block number $\tilde{n}$ becomes favorable and, finally, a core-tail structure can be formed rather than a core-shell structure (see Fig. 10).

The influence of the attraction between polar units on the structure formation was studied by computer simulations in the work [14]. It has been demonstrated that such attraction can lead to the formation of a core-shell structure, whereas without attraction only a core-tail structure is formed, these conclusion being in agreement with the present consideration.

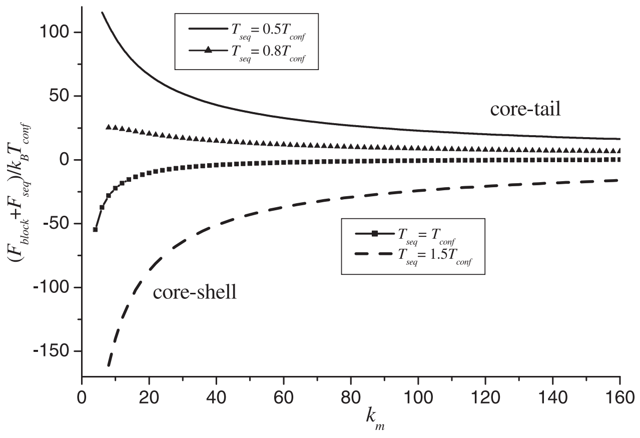

Fig. 9. The total free energy F, equation (9), vs. the mean block length k_m for different relations between temperatures T_{conf} and T_{seq}. The values of the parameters correspond to those used in the computer simulations: $N = 1024$, $\phi = 0.6$, $B_P/a^3 = v/2a^3 = 0.17$.

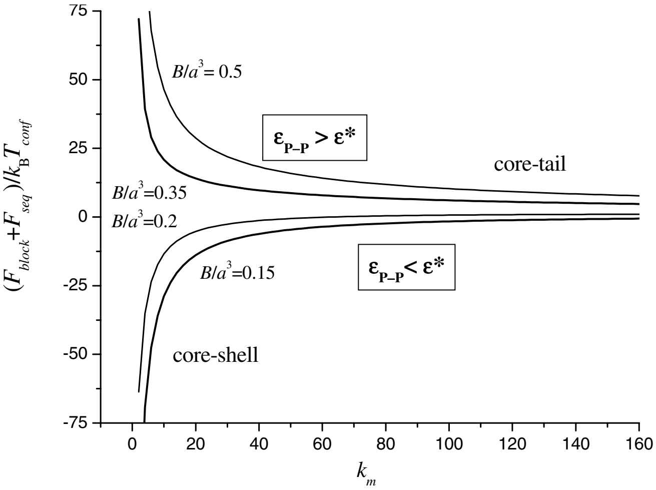

Fig. 10. The total free energy F, equation (9), vs. the mean block length k_m for different values of the second virial coefficient B_P describing the interaction between the polar units for $N = 1024$, $\phi = 0.6$, $T_{seq} = T_{conf}$.

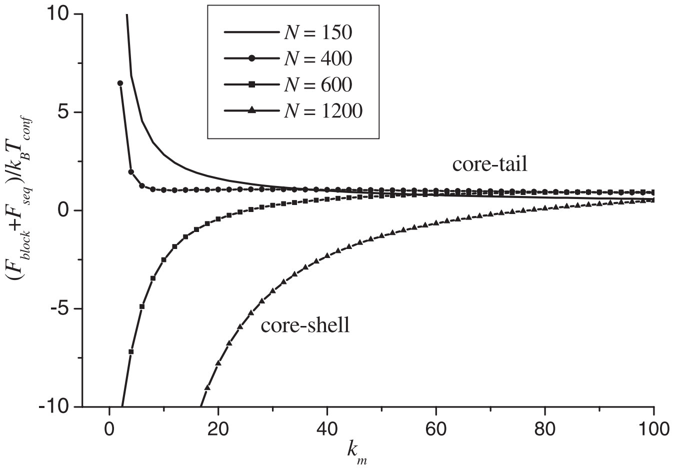

Fig. 11. The total free energy F, equation (9), vs. the mean block length k_m for different values of the chain length N for $\phi = 0.6$, $B_P/a^3 = 0.2$, $T_{seq} = T_{conf}$.

Now let us consider the influence of the chain length which is manifested mainly in the conformational contribution per P-loop. The free energy f_P is less for large values of $R_c \sim (N/\varphi)^{1/3}$ due to the increase of the surface volume accessible to a loop end. This contribution does not limit anymore the growth of the number of blocks which takes place in the course of the recoloring evolutionary procedure. As a result, a core-shell structure is formed more easily for longer chains (see Fig. 11).

The expression (9) and Figures 9–11 do not describe the behavior of the system for the mean block length $k_m < (R_c/a)^2$. For short blocks, a more detailed description of the structure should be used. However, we can obtain a simple estimation for the free energy F of a completely random chain with very short blocks ($k_m = 2$) and estimate whether this case can be realized or not. The sequence free energy of a completely random sequence has the value $F_{seq0} = -kT_{seq}N \ln 2$. Let us denote the sum of the interaction free energy and conformational energy by F_0: $|F_0| \ll |F_{core}|$ since a completely random copolymer forms either a globule of low density or a copolymer coil. The total free energy of this structure is $F_{rnd} = F_0 + F_{seq0}$. The total free energy of a core-shell structure F is approximately equal to F_{core} due to the strong immiscibility of hydrophobic units with the solvent and with polar units. Since $F_{rnd} - F_{core} \gg k_BT_{conf}$ for the considered values of the parameters, the formation of a structure with very short blocks is strongly unfavorable. Hence, the formation of a core-shell structure with the average block length $k_m \approx (R_c/a^2)$ is predicted.

However, if the sequence temperature T_{seq} is very high, then the sequence free energy dominates in the total free energy F (4) and a free energy minimum corresponds to a completely random sequence. The threshold sequence temperature T_0 can be roughly estimated by comparing the free energy of a core-shell structure $F \approx F_{core}$ and the free energy corresponding to a completely random sequence $F_{rnd} \approx F_{seq0}$. We estimate the core free energy using the Flory expression for the polymer-solvent interaction energy $F_{core}/V_c = k_BT_{conf}((1-\phi)\ln(1-\phi)+\phi-\chi\phi^2)$, where the Flory-Huggins parameter $\chi = -Z\varepsilon_H/2$, Z is the coordination number of the lattice. For the values of the parameters related to the computer simulation model ($\varepsilon_H = -1$, $\phi = 0.6$, $Z = 27$) we obtain $T_0/T_{conf} \approx 5.5$ for the threshold temperature of the transition from the core-shell structure to a completely random copolymer.

Therefore, we can conclude from the theoretical estimations that the core-shell structure can be formed if the sequence temperature corresponds to the range $0.9 < T_{seq}/T_{conf} < 5.5$ for the proper values of the chain lengths and the interaction energies between polar units. The upper threshold temperature T_0 decreases if the immiscibility between hydrophobic units and the solvent molecules and polar units becomes weaker. It is possible that the molecular parameters would correspond to the threshold temperature T_0 less than T_{th} estimated from (9). In this case, the formation of the core-shell structure can not be expected.

5 Conclusions

In the present study we have proposed a new method of conformation-dependent sequence design. The method has been tested by computer simulation and corresponding theoretical considerations. This method takes into account the randomness not only in the process of thermal motion but also in the process of sequence design. The rates of these processes are characterized by two different temperatures. The copolymers generated in this way form more stable “core-shell” structures than protein-like copolymers obtained by instant coloring. These differences are due to the process of simultaneous matching of sequences and conformations.

To describe the thermodynamic forces controlling the structure formation, sequence free energy F_{seq} was introduced. It is taken to be proportional to the Shannon’s entropy with the factor $-T_{seq}$ and (somewhat boldly) added to the normal conformational free energy. At very high $T_{seq} \gg T_{conf}$, the sequence free energy dominates and the sequence tends to be completely random, corresponding to the minimum of F_{seq}. At low T_{seq}, the evolution selects those sequences, which correspond to the low value of the conformation free energy (structure of the core-tail type). In the intermediate regime, the conformational thermodynamic force and the sequence contribution (evolution pressure) interplay. Therefore, the formation of non-trivial structures and sequences is possible. These theoretical predictions are in a good agreement with the results of computer simulations.

Main feature of our sequence design method is introduction of two different temperatures T_{seq} and T_{conf}, controlling sequence and conformational phase space independently. Unlike previous works, this approach means trend to simultaneous sequence and conformation selection from “expanded” CS-ensemble. The proposed method can be easily generalized (at least via computer simulation) to more complex external conditions and interaction potentials, however many aspects still need proper statistical-mechanical justification.

In conclusion, we would like to emphasize once more that a protein-like copolymer globule cannot be regarded as a model of real globular proteins in the native state: the H core is liquid-like (although compact), while the spatial structure of a native protein is unique. Our ultimate target was a synthetic "two-letter" copolymer with useful functional properties (e.g., solubility of globules), rather than biological protein.

The authors thank A. Yu. Grosberg for useful discussions. The financial support from Alexander-von-Humboldt Foundation, Program for Investment in the Future (ZIP), INTAS (project 2001-607), and Russian Foundation for Basic Research is highly appreciated.

Appendix: Primary sequence entropy

Let the sequence statistics be characterized by the distributions of length of hydrophobic and polar blocks $p_H(n)$ and $p_P(m)$. The normalization conditions

$$\sum_{n=1}^{\infty} p_H(n) = 1, \qquad \sum_{m=1}^{\infty} p_P(m) = 1$$

are satisfied.

The average sequence composition and the probabilities to find pairs of units HH, HP and PP in the sequence are determined by the average lengths of hydrophobic and polar blocks n^* and m^* correspondingly:

$$p_H = \frac{n^*}{n^*+m^*}, \quad p_P = \frac{m^*}{n^*+m^*},$$

$$p_{HP} = p_{PH} = \frac{1}{n^*+m^*}, \; p_{HH} = \frac{n^*-1}{n^*+m^*},$$

$$p_{PP} = \frac{m^*-1}{n^*+m^*},$$

where the average block lengths are equal to

$$n^* = \sum_{n=1}^{\infty} n p_H(n), \qquad m^* = \sum_{m=1}^{\infty} m p_P(m). \tag{A.1}$$

Let H_i, P_i denote the sequences of i units of one type, U_M denote a sequence of length M. For example, $U_M = H_{n_1}P_{m_1}H_{n_2}P_{m_2}\ldots H_{n_k}$ denote a sequence of length $M = \sum_{i=1}^{k} n_i + \sum_{i=1}^{k-1} m_i$ consisting of k H-blocks and $(k-1)$ P-blocks. The probability to find this sequence p_{U_M} is equal to

$$p_{U_M} = p_{HP}\tilde{p}_H(n_1)p_P(m_1)p_H(n_2)p_P(m_2)\ldots\tilde{p}_H(n_k), \tag{A.2}$$

where $\tilde{p}_H(n)$ is the probability to find a hydrophobic end block of length n. The end block can be considered as a part of a more long internal block, hence, this probability is equal to

$$\tilde{p}_H(n) = \sum_{j=n}^{\infty} p_H(j) = 1 - \sum_{j=1}^{n-1} p_H(j), \quad \tilde{p}_H(1) = 1. \tag{A.3}$$

The same expressions are valid for a polar end block:

$$\tilde{p}_P(m) = \sum_{j=m}^{\infty} p_P(j) = 1 - \sum_{j=1}^{m-1} p_P(j).$$

The probability of any other sequence U_M containing at least one HP-pair can be written in the similar way (A.2). The following relations are satisfied:

$$p_{U_iHP_m} = p_{U_iHP}\tilde{p}_P(m),$$
$$p_{U_iHP_m} = p_{U_iHP_{m+1}} + p_{U_iHP_mH}$$

The probability of a pure H- or P-sequence can be set by the recurrent formulae: $p_{H_{i+1}} = p_{H_i} - p_{H_iP}$ giving

$$p_{H_i} = p_H - p_{HP}\sum_{j=1}^{i-1}\tilde{p}_H(j). \tag{A.4}$$

The normalization condition for the sequence probabilities is satisfied:

$$\sum_{\{U_M\}} p_{U_M} = 1, \tag{A.5}$$

where $\{U_M\}$ is the total set of sequences of the length M.

The entropy of a sequence of two letter alphabet per 1 unit is determined by the relation [33]

$$S_{seq} = -\lim_{M\to\infty}\frac{1}{M}\sum_{\{U_M\}} p_{U_M}\log_2 p_{U_M}. \tag{A.6}$$

Let us introduce the entropy of sequences of length M $S_{seq}(M) = \sum_{\{U_M\}} p_{U_M}\log_2 p_{U_M}$. It can be written in the form

$$S_{seq}(M) = \sum_{i=1}^{M-1}\Delta S_i + S_{seq}(1), \tag{A.7}$$

where

$$\Delta S_i = S_{seq}(i+1) - S_{seq}(i) = \\ -\left(\sum_{\{U_{i+1}\}} p_{U_{i+1}}\log_2 p_{U_{i+1}} - \sum_{\{U_i\}} p_{U_i}\log_2 p_{U_i}\right),$$

$$S_{seq}(1) = -p_H\log_2 p_H - p_P\log_2 p_P.$$

Using the relations (A.2), (A.3), the difference ΔS_i can be written in the form

$$\begin{aligned}\Delta S_i = &\\ &- p_{HP}\left(\sum_{n=1}^{i-1} p_H(n)\log_2 p_H(n) + \sum_{m=1}^{i-1} p_P(m)\log_2 p_P(m)\right)\\ &- p_{HP}\log_2 p_{HP}\left(\tilde{p}_H(i)+\tilde{p}_P(i)\right)\\ &- 2p_{HP}\left(\tilde{p}_H(i)\log_2\tilde{p}_H(i) + \tilde{p}_P(i)\log_2\tilde{p}_P(i)\right)\\ &- \big(p_{H_{i+1}}\log_2 p_{H_{i+1}} - p_{H_i}\log_2 p_{H_i} + p_{P_{i+1}}\log_2 p_{P_{i+1}}\\ &- p_{P_i}\log_2 p_{P_i}\big).\end{aligned} \tag{A.8}$$

The probabilities of pure H- and pure P-sequences and the probabilities $p_H(i)$, $\tilde{p}_H(i)$ and $p_P(i)$, $\tilde{p}_P(i)$ tend to zero for long blocks: $i \gg n^*$, m^*. Hence,

$$\begin{aligned}\Delta S_i \to \Delta S = \\ - p_{HP}\left(\sum_{n=1}^{\infty} p_H(n)\log_2 p_H(n) + \sum_{m=1}^{\infty} p_P(m)\log_2 p_P(m)\right) \\ \text{for } i \to \infty.\end{aligned}$$

Then,

$$S_{seq}(M) \to M\Delta S \text{ for } M \to \infty \tag{A.9}$$

and, finally,

$$\begin{aligned}S_{seq} &= \Delta S \\ &= -p_{HP}\left(\sum_{n=1}^{\infty} p_H(n)\log_2 p_H(n) + \sum_{m=1}^{\infty} p_P(m)\log_2 p_P(m)\right).\end{aligned} \tag{A.10}$$

If we take into account that $p_{HP} = \tilde{n}/N$ we will get exactly formula (2) in Section 2.

References

1. M.V. Volkenstein, *General Biophysics*, I. and II. (Academic Press, 1983); A.L. Leninger, D.L. Nelson, M.M. Cox, *Principles of Biochemistry*, 2nd edn. (Worth Publishers, New York, 1993)
2. A.Yu. Grosberg, A.R. Khokhlov, *Giant Molecules: Here and There and Everywhere...* (Academic Press, New York, 1997)
3. E.I. Shakhnovich, A.M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993)
4. V.S. Pande, A.Yu. Grosberg, T. Tanaka, Proc. Natl. Acad. Sci. USA **91**, 12976 (1994)
5. L.L. Gatlin, *Information Theory and the Living System* (Columbia Univ. Press, New York, 1972)
6. C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simon, H.E. Stanley, Nature (London) **356**, 168 (1992)
7. A.R. Khokhlov, P.G. Khalatur, Physica A **249**, (1998) 253; Phys. Rev. Lett. **82**, 3456 (1999)
8. J. Virtanen, C. Baron, H. Tenhu, Macromolecules **33**, 336 (2000); J. Virtanen, H. Tenhu, Macromolecules **33**, 5970 (2000)
9. V.I. Lozinsky, I.A. Simenel, E.A. Kurskaya, V.K. Kulakova, V.Ya. Grinberg, A.S. Dubovik, I.Yu. Galaev, B. Mattiasson, A.R. Khokhlov, Rep. Russ. Acad. Sci. **375**, 637 (2000); P.-O. Wahlund, I.Yu. Galaev, S.A. Kazakov, V.I. Lozinsky, B. Mattiasson, Macromol. Biosci. **2**, 33 (2002); M.-H. Siu, G. Zhang, C. Wu, Macromolecules **35**, 2723 (2002)
10. E.A. Zheligovskaya, P.G. Khalatur, A.R. Khokhlov, Phys. Rev. E **59**, 3071 (1999)
11. V.A. Ivanov, A.V. Chertovich, A.A. Lazutin, N.P. Shusharina, P.G. Khalatur, A.R. Khokhlov, Macromol. Symp. **146**, 259 (1999); A.V. Chertovich, V.A. Ivanov, A.A. Lazutin, A.R. Khokhlov, Macromol. Symp. **160**, 41 (2000)
12. E.N. Govorun, V.A. Ivanov, A.R. Khokhlov, P.G. Khalatur, A.L. Borovinsky, A.Yu. Grosberg, Phys. Rev. E **64**, 040903(R) (2001)
13. L.V. Zherenkova, S.K. Talitskikh, P.G. Khalatur, A.R. Khokhlov, Dokl. Phys. Chem. **382**, 358 (2002); S.I. Kuchanov, A.R. Khokhlov, J. Chem. Phys. **118**, 4672 (2003)
14. P.G. Khalatur, V.V. Novikov, A.R. Khokhlov, Phys. Rev. E **67**, 051901 (2003)
15. A.V. Chertovich, V.A. Ivanov, B.G. Zavin, A.R. Khokhlov, Macromol. Theory Simul. **11**, 751 (2002)
16. V.S. Pande, A.Yu. Grosberg, T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000)
17. E.I. Shakhnovich, A.M. Gutin, Protein Eng. **6**, 793 (1993)
18. E.I. Shakhnovich, Fold. Des. **3**, R45 (1998)
19. A. Irback, C. Peterson, F. Potthast, E. Sandelin, Phys. Rev. E **58**, R5249 (1998); Structure **7**, 347 (1999)
20. V.I. Abkevich, A.M. Gutin, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA **93**, 839 (1996)
21. R.A. Broglia, G. Tiana, S. Pasquali, H.E. Roman, E. Vigezzi, Proc. Natl. Acad. Sci. USA **95**, 12930 (1998)
22. P. Gupta, C.K. Hall, A.C. Voegler, Protein Sci. **7**, 2642 (1998)
23. S. Istrail, R. Schwartz, J. King, J. Comput. Biol. **6**, 143 (1999)
24. G. Giugliarelli, C. Micheletti, J.R. Banavar, A. Maritan, J. Chem. Phys. **113**, 5072 (2000)
25. A. Yu. Grosberg, Biofizika **29**, 569 (1984)
26. T. Garel, L. Leibler, H. Orland, J. Phys. II France **4**, 2139 (1994)
27. A. Trovato, J. van Mourik, A. Maritan, Eur. Phys. J. B **6**, 63 (1998)
28. I. Carmesin, K. Kremer, Macromolecules **21**, 2819 (1988); W. Paul, K. Binder, D.W. Heermann, K. Kremer, J. Chem. Phys. **95**, 7726 (1991)
29. *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, 1995)
30. C.E. Shannon, *Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949)
31. C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948), 623; M.L. Rosenzweig, *Species Diversity in Space and Time* (Cambridge University Press, New York, 1995)
32. I. Grosse, H. Herzel, S.V. Buldyrev, H.E. Stanley, Phys. Rev. E **61**, 5624 (2000)
33. *Mathematical methods in contemporary chemistry*, edited by S. Kuchanov (Gordon and Breach, 1996)
34. A.Yu. Grosberg, A.R. Khokhlov, *Statistical Physics of Macromolecules* (American Institute of Physics, New York, 1994)
35. I.M. Lifshitz, A.Yu. Grosberg, A.R. Khokhlov, Rev. Mod. Phys. **50**, 683 (1978)
36. P.J. Flory, *Principles of Polymer Chemistry* (Cornell Univ. Press, Ithaca, N.Y., 1953)