DOI: 10.1002/masy. 200750604

Evolutionary Approach in Copolymer Sequence Design

Pavel G. Khalatur,[1,2] Alexei R. Khokhlov,[2,3] Maria K. Krotova[3]*

Summary: In this work, we discuss a simple evolutionary algorithm that introduces a "selection pressure" under which two-letter (AB) copolymer sequences can mutate and transform into the sequences tuned to microphase separation transition (MIST). In particular, we are interested in determining how a sequence of A and B units should be organized in order to reach maximum length scale for MIST at a given AB composition. It is found that such sequences are similar to those known for tapered or gradient copolymers exhibiting strong composition inhomogeneity along their chain. The problems of the evolution of copolymer sequences are considered from the viewpoint of emerging of information complexity in the sequences in the course of this evolution.

Keywords: copolymers; evolution; information complexity; phase behavior; sequence design; simulation

Introduction

In recent years, there has been intense interest in developing new types of polymeric materials via clever design of sequences of monomeric units in a copolymer chain. Broadly speaking, sequence design may be defined as an approach aimed at finding the optimum monomer sequence that provides desired (physicochemical, mechanical, etc.) properties of the resultant polymer. This requires a scoring function that may typically be based on physical principles, knowledge-based approaches, or a specifically designed function.

In a series of publications,[1–3] we have reported on several new design strategies allowing for synthesis of copolymers with a broad variation of their sequence distributions. The fundamental principle of these strategies is based on the conformational-dependent sequence design (CDSD), which takes into account a strong coupling between the conformation and primary structure of copolymers during their synthesis. Using computer simulation techniques and various theoretical approaches, we have shown that rather simple methods, such as polymer-analogous reactions and normal radical copolymerization, can lead to nontrivial chemical sequences, long-range correlations, and gradient structures, if they take place under unusual physical conditions.

In recent review articles,[4–9] we have discussed advances that have been achieved in the computer simulation and theoretical understanding of designed copolymers in solution and in bulk. The focus was on amphiphilic proteinlike (PL) copolymers the design of which was inspired by unique sophisticated functional performance of globular proteins-enzymes. Proteins and synthetic block copolymers have in common the property of both being self-organizing polymeric systems. While the former self-organize at a microscopic level, by folding to the native structure, the latter do it at a mesoscopic scale, by exhibiting a variety of different phases in melts, blends and solutions.

We have demonstrated that microphase separation transition (MIST) in a pure melt of designed PL copolymers occurs at lower

[1] Institute of Organoelement Compounds, Russian Academy of Science, Moscow 119991, Russia
E-mail: khalatur@germany.ru

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

[3] Physics Department, Moscow State University, Moscow 119899, Russia

segregation energies and exhibits higher characteristic length scale for microphase separation as compared to their random, random-block (RB), and regular multiblock (RM) counterparts.[4,10] In particular, the molecular dynamics simulations carried out for the melts of ‘‘two-letter’’ (AB) PL and RB copolymers showed that both systems demonstrate a common feature in that the small-angle peaks in the scattering functions $S(q)$ increase and their position q^* ($q^* \neq 0$), which is directly related to the scale of the concentration fluctuations, shifts to lower wave numbers q as A-B repulsion becomes stronger, indicating that the structure becomes better defined. What is most important is the fact that the spatial scale r^* of the segregated structure for PL copolymers is appreciably larger than that for RB copolymers with the same composition and the same average block length L. Also, MIST in the PL copolymer system occurs at a temperature higher than that of the RB system.

The polymer RISM calculations for long chains (up to $N = 4096$) also showed that in the $q^* \neq 0$ region, the phase behavior of PL copolymers sharply differs from that observed for RB and RM copolymers.[10] For example, the spinodal temperature T^* for an 1024-unit PL copolymer with the average block length $L \approx 6$ was found to be close to that of the RM and RB copolymers in which the block length is roughly tenfold (!) larger. At the same L’s, the spinodal temperatures for the PL copolymers are also several fold larger. This indicates that the processes of self-organization in the system of designed PL copolymers can proceed significantly more intense than in the other copolymer systems.

The reason behind this distinction is in that the behavior of random (quasirandom) copolymers with a quenched sequence is governed not only by the average block length L but also by the block length dispersion D_L. For correlated random copolymers, the structure formation is dominated by long blocks whose probability increases with an increase in D_L. Even at a relatively low fraction of these blocks in the chain, their effect can be decisive. The block length distribution in PL copolymers is described by a specific type of statistics, namely by the Lévy-flight statistics with a high dispersion and a slow decrease in the block length probability.[11] That is why the chains contain a significant fraction of long sections composed of chemically identical segments even at relatively small L values. This is precisely the major reason behind rather unusual behavior of such copolymers in self-organization processes.

In this work, we discuss a simple evolutionary algorithm that introduces a ‘‘selection pressure’’ under which random copolymer sequences can mutate and transform into MIST-tuned sequences. In particular, we are interested in determining how a sequence of A’s and B’s should be organized in order to reach maximum length scale for MIST at a given AB composition.

Sequence Design: Evolutionary Approach

Evolutionary computation approaches are optimization methods. They are conveniently presented using the metaphor of natural evolution: a randomly initialized population of individuals evolves following a crude parody of the Darwinian principle of the survival of the fittest. New individuals are generated using simulated evolutionary operations such as mutations. The probability of survival of the newly generated solutions depends on their fitness (how well they perform with respect to the optimization problem at hand): the ‘‘best’’ are kept with a high probability, the ‘‘worst’’ are rapidly discarded.

Shakhnovich and Gutin used evolutionary approach in protein science and showed that it is possible to design a model protein sequence in such a way that it will fold into a specific globular conformation.[12] To do this, they optimized the sequence, using a Monte Carlo (MC) method that randomly exchanged monomers within the sequence. Here, we employ an extension of this evolutionary approach to design a two-letter AB copolymer sequence that is capable of forming microphase-separated structures in melt with a required (micro)-domain spacing r^* for a given chain length N and chemical composition.

MC simulations are performed to search for "point mutations" that favor the microphase separation of copolymer melt. A random AB sequence is taken as an initial generation ($G = 0$), which is considered as a "common ancestor" of a given run. Then a procedure of the evolution (annealing) of the sequence starts. The iterative procedure consists of many mutation steps. At each MC step (with every "click of the evolutionary clock"), two monomers are chosen randomly and, if they happen to be of different types, an attempt is made to exchange their types (A↔B). This point mutation changes the copolymer sequence (S) and the corresponding sequence-dependent intrachain correlation function $\omega(q,\mathbf{S})$ calculated in the reciprocal $\mathbf{q}$-space. The $\omega(q,\mathrm{S})$ function is used as an input for some theory (see below) that gives the spinodal temperature T^* (the critical value of the Flory–Huggins parameter $\chi^* = 1/T^*$) and the wave vector of maximum instability q^*. These properties can be viewed as a scoring (*fitness*) function f. It describes how well a particular sequence is fit to perform its "duties". There should be a feedback in the system. In other words, to accept or reject current mutation, we should compare two fitness functions, old and new ones. This is done following the standard Metropolis ideology, using some design parameter, which we will call design (or sequence) temperature. In fact, this parameter characterizes the tolerance to mutations in sequence space or, in other words, an "evolution pressure".

If the spinodal temperature T^* is viewed as a fitness function $f = T^*(\mathbf{S})$, the resulting change in the spinodal temperature ΔT^* after an attempted mutation is found and the probability p to fix the mutation is guided by the Metropolis algorithm: $p = 1$ if $\Delta T^* \geq 0$, otherwise $p = \exp(\Delta T^*/T_s)$, where T_s is the fictitious temperature referred to as sequence design temperature. It is assumed that after N point mutations, the N-unit sequence is passed on to the next generation, $G \rightarrow G+1$. Such modifications, leading to changing in copolymer sequence, are repeated many times ($n_G \sim 10^6$). For each trajectory corresponding to a stationary regime, we may interpret the set of sequences generated in the course of our evolutionary process as n_G different "species" originating from a common ancestor. The sequence composition is constrained so that there are $N/2$ A and B units. To monitor the sequence-selection process, we employ the Hamming distance $h_G = N^{-1}\sum_{i=1}(1-\delta_{\mathbf{s}_i(G),\mathbf{s}_i(G+1)})$ that counts how many A/B units are different between two neighboring sequences. For a random process (at high design temperatures), the Hamming distance is ½; if the process is non-random, $h_G >$ ½. At low design temperatures, the relaxation time of the sequences may become pretty long. When h_G fluctuates near its equilibrium value, we can calculate averages.

Our evolutionary procedure is equivalent to the situation when monomer A is converted into B by attaching some ligand L: $\mathrm{A} + \mathrm{L} \rightleftarrows \mathrm{B}$. Depending on T_s, this "chemical equilibrium" can be gradually shifted. It is assumed that for any T_s, the number of ligands in the system is fixed to maintain 1:1 AB composition; however, they can choose which monomer to bind. This defines, in particular, the resulting copolymer sequence.

Computational Methods: RPA and pRISM

In the weak-segregation regime, the phase behavior of a melt composed of flexible-chain copolymers is described on the basis of the incompressible random-phase approximation (RPA)[13] and the polymer integral equation reference interaction site model (pRISM) theory[14,15] that allow finding the conditions under which the spatially homogeneous state of the system becomes unstable.

RPA. To analyze the stability of the ordered microphases, the simplest incompressible random phase approximation can be employed. Using this approach, the critical value of the Flory-Huggins parameter, χ^*, and the corresponding spinodal temperature, $T^* = 1/\chi^*$, can be determined by the condition that the scattering intensity $S(q)$ reaches its maximum value at a nonzero wave vector q^*. Within the RPA the scattering intensity is given by

$$S_{RPA}^{-1}(q) = \Re(q) - 2\chi \tag{1}$$

with

$$\Re(q) = \frac{1}{\delta\omega(q)}\left\{\frac{\omega_{AA}(q)}{\phi_B} + \frac{\omega_{BB}(q)}{\phi_A} + \frac{2\omega_{AB}(q)}{f_A f_B}\right\} \tag{2}$$

where $f_A = N_A/N$, $f_B = N_B/N$, ϕ_A and ϕ_B are the volume fractions of the corresponding species, and $\delta\omega(q)$ is defined as $\delta\omega(q) = \omega_{AA}(q)\omega_{BB}(q) - f_A^{-1} f_B^{-1} \omega_{AB}^2(q)$.

pRISM. For an AB copolymer melt, the polymer RISM (pRISM) equation is represented in the matrix form[15]

$$\mathbf{H}(r) = \int\limits_{(\mathbf{r}')}\int\limits_{(\mathbf{r}'')} \Omega(|\mathbf{r} - \mathbf{r}'|)\mathbf{C}(|\mathbf{r}' - \mathbf{r}''|)[\Omega(\mathbf{r}'') + \rho\mathbf{H}(\mathbf{r}'')]d\mathbf{r}'d\mathbf{r}''. \tag{3}$$

Here, **H** and **C** are symmetric matrices whose elements are the partial total $h_{\alpha\beta}(r)$ and direct $c_{\alpha\beta}(r)$ pair correlation functions (α, $\beta = $ A, B); Ω is the matrix of intramolecular correlation functions $\omega_{\alpha\beta}(r)$ that characterize the conformation of a macromolecule and its sequence distribution in direct space; and ρ is the average number density of units in the system. The pRISM Equation (3) is complemented by the closure relation that corresponds to the so-called high-temperature molecular Percus-Yevick approximation (HTA/PY)[15]:

$$\int\limits_{(\mathbf{r}')}\int\limits_{(\mathbf{r}'')} \Omega(|\mathbf{r} - \mathbf{r}'|)\mathbf{C}(|\mathbf{r} - \mathbf{r}''|)\Omega(\mathbf{r}'')\,d\mathbf{r}'d\mathbf{r}'' = = \int\limits_{(\mathbf{r}')}\int\limits_{(\mathbf{r}'')} \Omega(|\mathbf{r} - \mathbf{r}'|)[\mathbf{C}^{(0)}(|\mathbf{r}' - \mathbf{r}''|) + \Delta\mathbf{C}(|\mathbf{r}' - \mathbf{r}''|)]\Omega(\mathbf{r}'')\,d\mathbf{r}'d\mathbf{r}'', \quad r > \sqrt{\sigma_\alpha\sigma_\beta} \tag{4}$$

$$\Delta c_{\alpha\beta}(r) = [1 - e^{-u_{\alpha\beta}(r)/kT}][h_{\alpha\beta}^{(0)}(r) - 1], \quad r > \sqrt{\sigma_\alpha\sigma_\beta} \tag{5}$$

$$h_{\alpha\beta}(r) = -1, \quad r < \sqrt{\sigma_\alpha\sigma_\beta} \tag{6}$$

$$c_{\alpha\beta}^{(0)}(r) = 0, \quad r > \sqrt{\sigma_\alpha\sigma_\beta} \tag{7}$$

Here, $u_{\alpha\beta}(r)$ describes the interaction between nonbonded monomers and σ_α is the effective monomer size ($\sigma_A = \sigma_B = \sigma$). In this study, we use the repulsive Yukawa-type potential:

$$u_{\alpha\beta}(r) = \begin{cases} \varepsilon_{\alpha\beta}(\sigma/r)\exp[-2(r/\sigma - 1)], & r \geq \sigma \\ \infty, & r < \sigma \end{cases} \tag{8}$$

where $\varepsilon_{\alpha\beta}$ is an energy parameter which is directly related to the χ parameter ($\chi_{\alpha\beta} \propto \varepsilon_{\alpha\beta}/k_B T$). We consider the case when A and B units repeal each other ($\varepsilon_{AB} > 0$) while A-A and B-B interactions are purely excluded-volume type ($\varepsilon_{AA} = \varepsilon_{BB} = 0$). The functions $h_{\alpha\beta}^{(0)}(r)$ and $c_{\alpha\beta}^{(0)}(r)$ in Equation (4)–(7) correspond to the reference (athermal) system for which all $\varepsilon_{\alpha\beta} = 0$.

Although the phase-separated structures are not available from the pRISM theory, one can obtain the structural information for the disordered single-phase state at a certain distance from the phase separation point against the athermal reference system. Also, by varying temperature T and monomer density ρ, one can find the conditions under which the spatially homogeneous state of the system becomes unstable. This takes place on a spinodal line, which is determined by the set $\{T^*, \rho^*\}$ or parametrically as $T^* = T^*(\rho)$. In the mean-field approximation, the condition of spinodal instability is defined by[15]

$$S(q)|_{q=}q^* \to \infty \tag{9}$$

where $S(q) = \sum_{\alpha,\beta=A,B} \sqrt{x_\alpha x_\beta} S_{\alpha\beta}(q)$ ($x_\alpha = \rho_\alpha / \sum_\alpha \rho_\alpha$) and

$$\left\| \begin{matrix} S_{AA}(q) & S_{AB}(q) \\ S_{BA}(q) & S_{BB}(q) \end{matrix} \right\| = \left\{ \left\| \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \right\| - \left\| \begin{matrix} \omega_{AA}(q) & \omega_{AB}(q) \\ \omega_{BA}(q) & \omega_{BB}(q) \end{matrix} \right\| \times \left\| \begin{matrix} C_{AA}(q) & C_{AB}(q) \\ C_{BA}(q) & C_{BB}(q) \end{matrix} \right\| \right\}^{-1} \left\| \begin{matrix} \omega_{AA}(q) & \omega_{AB}(q) \\ \omega_{BA}(q) & \omega_{BB}(q) \end{matrix} \right\| \quad (10)$$

or by the following condition:

$$\Delta(q) \equiv \det[\mathbf{E} - \rho\Omega(q)\mathbf{C}(q)]_{q=}q^* \to 0 \quad (11)$$

where $\Delta(q)$ is the determinant of the matrix Equation (3), q^* represents the wave vector of maximum instability, and **E** is the unit diagonal matrix. In contrast to simple RPA, the pRISM takes into account thermal fluctuations and the compressibility of a polymer system. For every sequence generated in the sequence-selection process described above, we numerically solve the matrix pRISM integral Equation (3) and find the spinodal temperature T^*, the order-disorder transition temperature T_{ODT}, and the corresponding characteristic length scales.

To a first approximation, one can consider macromolecules on the basis of the highly simplified unperturbed model without intramolecular excluded volume interactions. This allows us to considerably simplify the problem by calculating the matrix elements $\omega_{\alpha\beta}(q)$ using the Gaussian intramolecular correlation function $\omega_{ij}(q) = \exp\left(-q^2\sigma^2|i-j|/6\right)$, which characterizes the distribution of segments i and j belonging to the species A and B inside an N-unit polymer.

In a pure melt of copolymer chains, a net repulsion between A and B segments (measured by the A-B segment-segment Flory-Huggins parameter, χ, or by the energy parameter $\varepsilon_{\mathrm{AB}}$) drives the segregation of the A and B segments. The chemical junction between chemically different segments restricts the process to occur locally, on the length scale of the chain's linear dimension. Transition from the $q^*=0$ regime (macrophase separation transition, MAST) to the $q^*>0$ regime (MIST) occurs at the Lifshitz point delimiting the MAST and MIST regions.[16]

Results and Discussion: MIST-tuned Copolymers

For the $T^*(\mathbf{S})$ fitness function used in our sequence-selection process, we find that if the sequence design temperature T_{s} is too high, the design can be viewed as a random walk in sequence space, and it yields random sequences, as expected. In contrast, when T_{s} is too low, A and B units tend to be separated within the chain and the evolutionary algorithm leads to trivial symmetric diblocks. It should be noted that the model does not include direct factors responsible for A/B separation within the chain.

It is instructive to explore a range of values for T_{s} that yields a compromise between the $T_s \to \infty$ and $T_s \to 0$ regimes. Average block length, L, is a good measure of order parameter. Figure 1 shows L as a function of T_{s} for different chain lengths N.

As seen, the reduced average block length (order parameter) $\xi = 2L/N$ increases with decreasing sequence temperature. The transition temperature, T_s^*, is about 0.8. Practically the same estimate for the transition temperature is obtained from the spinodal temperature T^* and the characteristic length scale r^* (=$2\pi/q^*$). These values are shown in Figure 2 as a function of T_s for different N. Note that both T^* and r^* are averaged over the ensemble of generated sequences ($\sim 10^6$). In the $T_s \to 0$ limit, when $\xi = 1$, we observe a typical mean-field behavior: $T^* \propto N$ and $r^* \propto N^{1/2}$ (for symmetric diblocks, $T^*/N = 0.096$).

What is rather unexpected here is that we find non-zero design temperature near which the r^* value has a maximum for sufficiently long chains, $N \gtrsim 200$. In other words, there are such sequences, which produce microphase-separated structures with larger characteristic length scales than simple symmetric diblocks (at least, in the weak segregation regime).

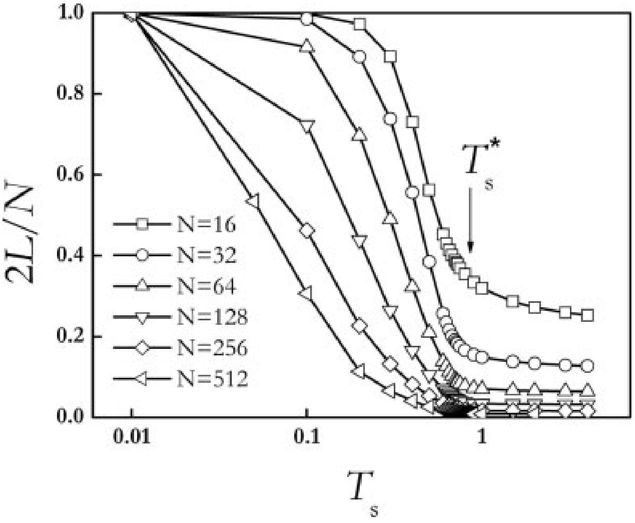

Figure 1.
RPA: Reduced average block length 2*L*/*N* as a function of the sequence design temperature.

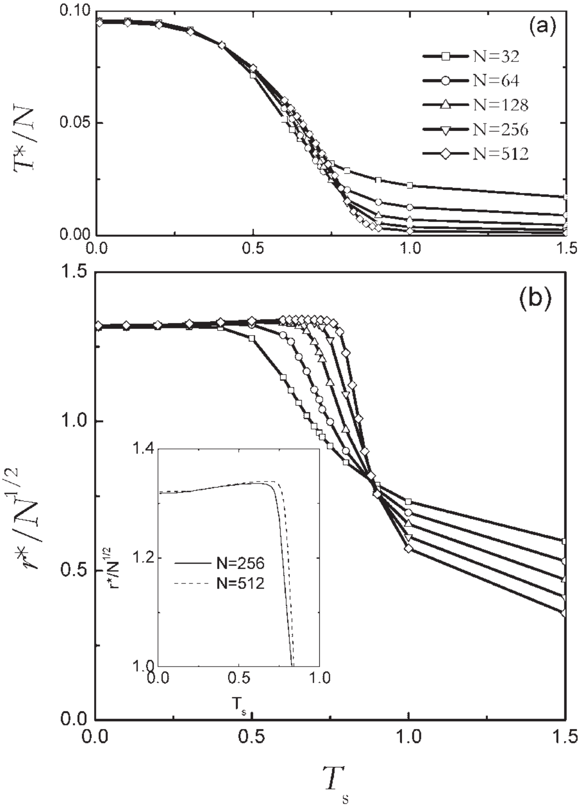

Figure 2.
RPA: (a) Spinodal temperature T^* and (b) characteristic length scale r^* ($=2\pi/q^*$) as a function of the sequence design temperature for different chain lengths *N*. The r^* value is measured in units of σ.

Another important observation is that the transition in sequence space from random to non-random sequences occurs as a first-order-like transition. Indeed, the calculations predict that near T_s^* the distribution function $W(r^*)$ for the ensemble of generated sequences demonstrates a clear bimodality (Figure 3).

Therefore, the transition in sequence space occurs as a first-order-like transition. We can speculate that, depending on the parameters characterizing the evolution pressure, there are two different regimes of mutation process.

The most intriguing observation that can be done for this transitory regime is that sufficiently long sequences ($N \gtrsim 200$) providing the maximum of r^* do not correspond to simple diblocks but rather they have a gradient shape, or it maybe better to say, an S-like shape (Figure 4).

Of course, among all generated sequences, symmetric diblocks have the highest spinodal temperature (that is, their critical χ parameter is lowest). However, S-like sequences can show larger characteristic length scales (see the insert in Figure 2).

The RPA and pRISM results are consistent with each other. As an example, we show in Figure 5 the composition profile calculated using the pRISM theory. pRISM calculations are much more time-consuming than those based on RPA, so that in this case the statistics is not very good. Nevertheless, again we observe the formation of S-like composition profiles.

In pRISM theory, the MIST process is directly reflected in the normalized static structure factor, $S_{\varepsilon=0}(q)/S_{\varepsilon}(q)$. As a copolymer system is cooled and microdomains are forming, the peak scattering intensity grows in a mean-field manner corresponding to the linear portion of the $S(q)$ curve in the coordinates $S^{-1}(q)$-χ or $S^{-1}(q)$-ε_{AB}. Extrapolation of this linear portion to divergent intensity defines an "apparent mean-field spinodal temperature", T_s.[15] However, no divergence actually occurs because there is no second-order phase

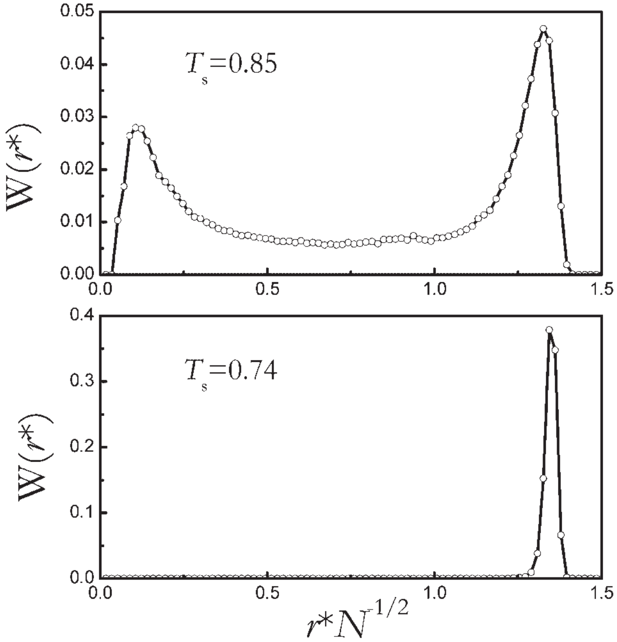

Figure 3.
Distribution function $W(r^*)$ for the ensemble of 512-unit sequences generated near and below the transition temperature T_s^*.

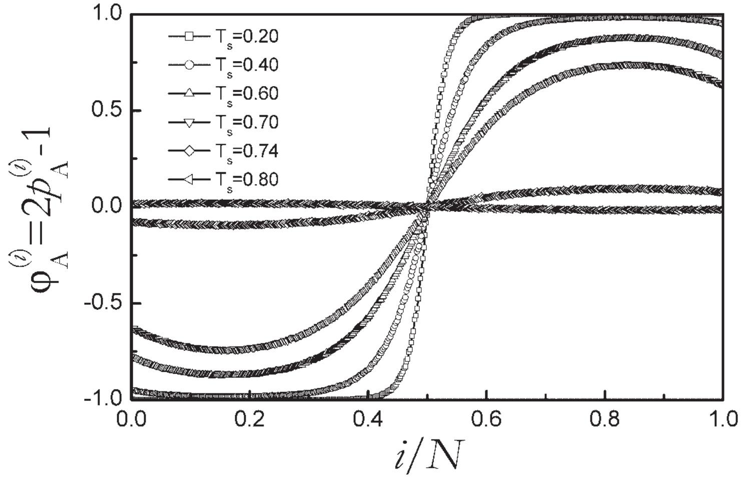

Figure 4.
RPA: Intramolecular composition profiles presented as a function of reduced monomer number *i*/*N* for 512-unit sequences having 1:1 AB composition. The present definition assumes that the A-type monomers are coded by symbol +1, whereas symbol −1 is assigned to the B-type monomers. For an ideal random copolymer in which chemically different units follow each other in a statistically random fashion, the probability p_A that monomer A is located at the i^{th} position in the chain is ½ for any *i*.

transition. Rather, fluctuation processes enter and this results in the nonlinear portion of the dependence of $S^{-1}(q)$ on χ (ε_{AB}). A lower temperature, T_{ODT}, which is a measure of the strength of this fluctuation stabilization process, can be extracted as shown in Figure 6a.

Inspection of the data presented in Figure 6b shows that in the vicinity of the transition temperature T^*, the microsegregation space scale r_{ODT}, as characterized by the value of $2\pi/q_{ODT}$, is larger for designed sequences than that observed for symmetric diblocks.

We now analyze our sequence design process with the value of r^* used as a fitness function. The maximization of the fitness function $f = r^*(\mathbf{S})$ with the sequence design

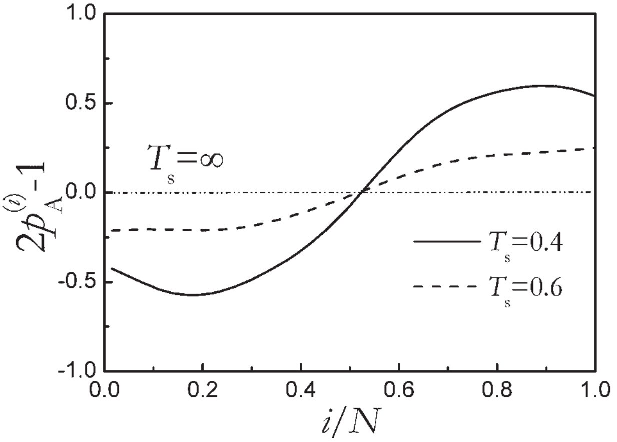

Figure 5.
pRISM: Intramolecular composition profiles presented as a function of reduced monomer number *i*/*N* for 128-unit sequences having 1:1 AB composition, at two sequence design temperatures.

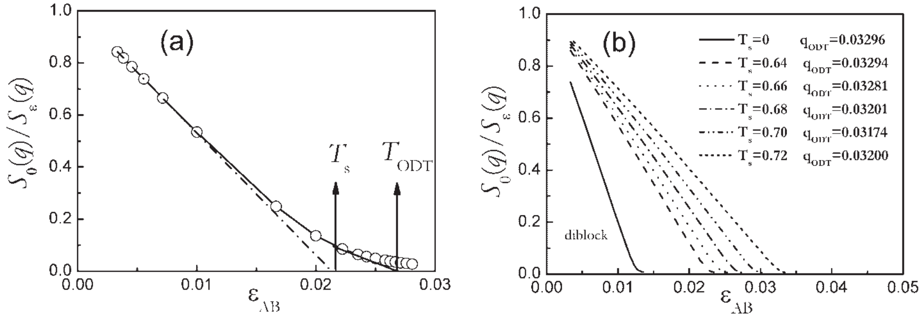

Figure 6.

(a) Inverse reduced peak intensity vs. energy parameter ε_{AB}. The figure shows how the apparent spinodal, T_s, and ODT, T_{ODT}, temperatures are defined. Dashed line corresponds to mean-field behavior. (b) Inverse reduced peak intensity vs. energy parameter ε_{AB} for different design temperatures. The figure shows the calculated values of q_{ODT}, which characterize microsegregation space scale, $r_{ODT} = 2\pi/q_{ODT}$. The q_{ODT} value is measured in units of σ^{-1}.

parameter r_s also leads to a gradient-like composition profile (Figure 7a). In this case, we deal with a copolymer whose primary structure is similar to that known for tapered or gradient copolymers exhibiting strong composition inhomogeneity along their chain. When $r_s \rightarrow 0$, the r^* value for MIST is significantly larger than that observed for diblock copolymers with the same AB composition and the same chain length (Figure 7b).

The concept of evolution of primary sequences attracts large interest also from the viewpoint of information content in the sequences. It is natural to expect that the content of information in the sequences of biopolymers (proteins, DNA, RNA) is relatively high in comparison with random sequences where it should be almost zero. 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 is just the area where the evolution concept can be used in the context of traditional 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 sequence

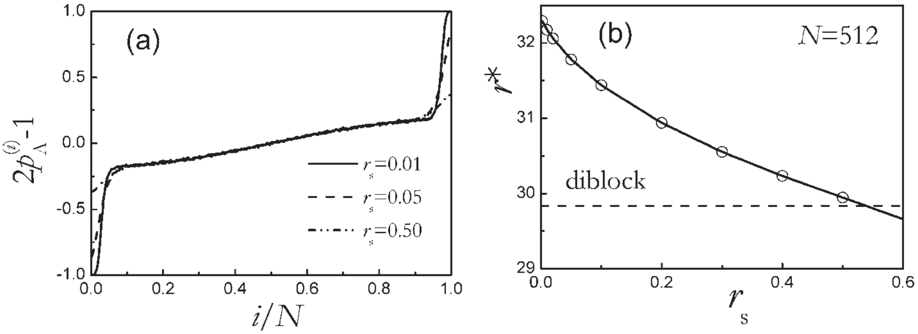

Figure 7.

RPA: (a) Intramolecular composition profiles presented as a function of reduced monomer number i/N for 512-unit sequences having 1:1 AB composition, at a few design parameters r_s. $p_A^{(i)}$ is the probability that the monomer of type A is located at the i^{th} position in a sequence. (b) Characteristic length scale r^* vs. design parameter r_s. Dashed line corresponds to r^* for diblock sequence.

evolution can be specified in exact mathematical terms. On the other hand, for biopolymers, 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 evolution models’’ like that considered in this study, 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.

A common approach to the analysis of the complexity of a system is to use concepts from information theory and information-theoretic-based techniques. In general, the aim here is to find a measure capable to indicate how far copolymer sequences generated during the evolutionary process differ from each other and from random or trivial (degenerate) sequences. It turned out that the usual measures of the degree of complexity (based, e.g., on Shannon's entropy and related characteristics) are nonadequate. To overcome this problem, it was proposed to use the so-called Jensen-Shannon ($\mathcal{JS}$) divergence measure.[17,18] Let us explain how it can be defined.

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 divergence

$$\mathcal{JS}(\mathbf{S}_1,\mathbf{S}_2)/N = h(\mathbf{S}) - \left[\frac{n}{N}h(\mathbf{S}_1) + \frac{N-n}{N}h(\mathbf{S}_2)\right] \quad (12)$$

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 subsequence. Of course, Shannon's entropy depends on the definition of a set of words in the sequence. For two-letter AB copolymers, one can adopt the following set of words (uniform blocks): A, AA, AAA,. . ., B, BB, BBB,. . .; that is, word (block) is defined by its length ℓ and type. In this case, Shannon's entropy per monomer can be written as

$$h = -\frac{N_w}{2N}\sum_{\ell}\left[f_A(\ell)\log_2 f_A(\ell) + f_B(\ell)\log_2 f_B(\ell)\right] \quad (13)$$

where $f_\mathrm{A}(\ell)$ and $f_\mathrm{B}(\ell)$ are the frequencies of words of length ℓ composed of letters A and

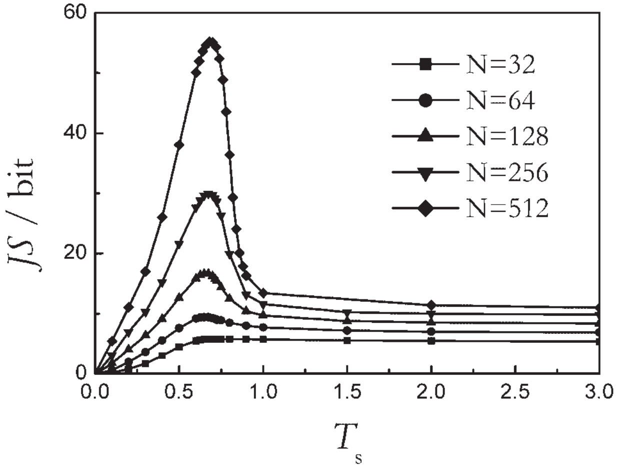

Figure 8.
The Jensen-Shannon divergence measure as a function of the sequence design temperature for different chain lengths N.

B, respectively, and N_w is the total number of words.

The Jensen-Shannon divergence $\mathcal{JS}$ is zero for subsequences with the same statistical characteristics; it takes higher values for increasing differences between the statistical patterns in the subsequences, and reaches its maximum value for a certain set of distributions. In particular, both random and any regular (multiblock) copolymers of infinite length show $\mathcal{JS}=0$. We normally expect that a completely random sequence or a sequence with long uniform blocks contain less information than a sequence containing many different blocks (words) of medium length.

Using the Jensen-Shannon divergence, $\mathcal{JS}$, as a measure of complexity for the generated sequences, one can obtain an interesting result (see Figure 8). The most important feature is that $\mathcal{JS}$ value is a *non-monotonous* function of the sequence design temperature T_s. Another intriguing result is that the Jensen-Shannon divergence measure shows a maximum in the vicinity of T_s*.

For the sequences generated in the evolutionary process described above, we find that at $T_s \approx T_s^*$ the degree of complexity, as measured by $\mathcal{JS}$ divergence, can be considerably higher as compared to that observed for very low and high T_s. The complexity increases with T_s decreasing, reaches its maximum just near T_s^*, and then sharply drops. Therefore, in the vicinity of T_s^*, the evolution preserved the copolymer sequence of high complexity, whereas at low T_s, the information content of the sequence has degenerated in the course of evolution. One can say that the most interesting and unusual sequences appear at the edge of chaos when T_s is close to the transition temperature T_s^*.

Conclusion

We have proposed a simple evolutionary algorithm that introduces a "selection pressure" under which two-letter (AB) random copolymer sequences can mutate and transform into the sequences tuned to microphase separation transition (MIST). In particular, we have interested in determining how a sequence of A and B units should be organized in order to reach maximum characteristic length scale for MIST at a given AB composition. It has been found that such sequences are similar to those known for tapered or gradient copolymers exhibiting strong composition inhomogeneity along their chain. The problems of the evolution of copolymer sequences were considered from the viewpoint of emerging of information complexity in the sequences in the course of this evolution.

Acknowledgements: The financial support from SFB-569 and RFBR is highly appreciated.

[1] A. R. Khokhlov, P. G. Khalatur, *Physica A* **1998**, *249*, 253.

[2] P. G. Khalatur, V. A. Ivanov, N. P. Shusharina, A. R. Khokhlov, *Russ. Chem. Bull.* **1998**, *47*, 855.

[3] A. R. Khokhlov, P. G. Khalatur, *Phys. Rev. Lett.* **1999**, *82*, 3456.

[4] A. R. Khokhlov, P. G. Khalatur, V. A. Ivanov, A. V. Chertovich, A. A. Lazutin, in: "*Challenges in molecular simulations*", G. Mac Kernan D, Eds., CECAM, Lyon, **2002**, vol. 4, pp. 79–100.

[5] A. R. Khokhlov, P. G. Khalatur, *Curr. Opin. Solid State Mater. Sci.* **2004**, *8*, 3.

[6] P. G. Khalatur, A. V. Berezkin, A. R. Khokhlov, *Rec. Res. Develop. Chem. Phys.* **2004**, *5*, 339.

[7] A. R. Khokhlov, A. V. Berezkin, P. G. Khalatur, *J. Polym. Sci. A: Polym. Chem.* **2004**, *42*, 5339.

[8] A. R. Khokhlov, P. G. Khalatur, *Curr. Opin. Colloid Interface Sci.* **2005**, *10*, 22.

[9] P. G. Khalatur, A. R. Khokhlov, *Adv. Polym. Sci.* **2006**, *195*, 1.

[10] L. V. Zherenkova, P. G. Khalatur, A. R. Khokhlov, *Dokl. Phys. Chem.* **2003**, *393*, 293.

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

[12] E. I. Shakhnovich, A. M. Gutin, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 7195.

[13] L. Leibler, *Macromolecules* **1980**, *13*, 1602.

[14] D. Chandler, H. C. Andersen, *J. Chem. Phys.* **1972**, *57*, 1930.

[15] K. S. Schweizer, J. G. Curro, *Adv. Chem. Phys.* **1997**, *98*, 1.

[16] G. H. Fredrickson, "*The Equilibrium Theory of Inhomogeneous Polymers*", Clarendon Press, Oxford **2006**.

[17] P. G. Khalatur, V. V. Novikov, A. R. Khokhlov, *Phys. Rev. E* **2003**, *67*, 051901.

[18] A. V. Chertovich, E. N. Govorun, V. A. Ivanov, P. G. Khalatur, A. R. Khokhlov, *Eur. Phys. J. E* **2004**, *13*, 15.