Origins of Fine Structure in DNA Melting Curves

Arevik V. Asatryan,[1, a)] Albert S. Benight,[2, b)] and Artem V. Badasyan*[3, c)]
[1)] *3D Printing Research Laboratory, A.B. Nalbandyan Institute of Chemical Physics, Yerevan, Armenia*
[2)] *Departments of Physics and Chemistry, Portland State University, Oregon, USA*
[3)] *Materials Research Laboratory, University of Nova Gorica, Nova Gorica, Slovenia*

(Dated: 11 April 2024)

With the help of one-dimensional random Potts-like model we study the origins of fine structure observed on differential melting profiles of double-stranded DNA. We assess the effects of sequence arrangement on DNA melting curves through the comparison of results for random, correlated, and block sequences. Our results re-confirm the smearing out the fine structure with the increase of chain length for all types of sequence arrangements and suggest fine structure to be a finite-size effect. We have found, that the fine structure in chains comprised of blocks with the correlation in sequence is more persistent, probably, because of increased sequence disorder the blocks introduce. Many natural DNAs show a well-expressed fine structure of melting profiles. In view of our results it might mean the existence of blocks in such DNAs. The very observation of fine structure may also mean, that there exists an optimal length for natural DNAs *in vivo*.

I. INTRODUCTION

Altered external conditions (such as increased temperature) in a system containing double stranded deoxyribonucleic acid (DNA) molecules may trigger a change of their conformation from ordered helical to disordered coil[1]. This transition is referred to as DNA melting. DNA is a double stranded heteropolymer, comprised of a sugar-phosphate backbone with side groups containing one of four nucleotide bases (cytosine [C], guanine [G], adenine [A] or thymine [T])[7,14]. Sequence complementarity rules for the double helix require that an A on one of the strands is always paired with T on the other, and G pairs with C. The A–T basepair is stabilized by two hydrogen bonds, and the G–C basepairs have three hydrogen bonds. Due to such compositional disorder, natural DNAs are treated as heteropolymers, and complementarity rules allow to describe DNAs as systems having binary disorder. Each particular sequence of basepairs along the double helix defines the genetic code responsible for the diversity of organisms in nature. Perhaps related to their biological relevance, different sequences give rise to different melting profiles, and display multiple peaks on differential melting curves (DMCs)[1]. The array of peaks observed along a DMC with increased temperature are referred to as *fine structure*[20,25]. Water-DNA interactions have been shown to play an important role for DNA conformations both experimentally[30], and theoretically[24]. How exactly is the fine structure affected by the presence of water, is not clear at this moment, and will be left for future studies.

In the mid-1960s and early 1970s at the beginning of the era of DNA melting, theoretical approaches were based on the two-component one-dimensional Ising model[16] formulated to predict DNA melting curves. When fitted to experimental melting curves, the Ising Model approach was able to predict, at least qualitatively, much of the fine structure experimentally observed on DMCs[17]. Perhaps the most successful approaches to predicting DNA DMC[3,19,34] were based on the Poland-Sheraga method[10] using the Fixman-Freire

[a)] Electronic mail: arevik.asatryan@ichph.sci.am
[b)] Electronic mail: abenight@pdx.edu
[c)] Corresponding author: artem.badasyan@ung.si

approximation[18]. These theories were formulated in terms of the Zimm-Bragg approach originally applied to analysis of the helix-coil transition in polypeptides[6,33]; but often used to describe the conformational changes in DNA as well[11,36]. The primary assumption underlying these approaches was that nearest-neighbour basepair interactions play a central role in the sequence dependence of DNA melting, and that peaks observed on DMCs correspond to the independent melting of lengthy sequence blocks differing in their percentage of G–C base pairs[17,19]. For long DNAs where the number of such blocks can be large (several hundred basepairs), the fine structure on DMCs appeared to get smeared due to the overlap of different peaks (see the recent review[3]). Therefore, it was surmised that clearly defined fine structure on DMCs should only be observed for DNA chains of moderate length where formation of a small number of loops is allowed.

Unfortunately, this classical viewpoint does not provide insight into the relationship between the number of blocks and the number of peaks observed on experimentally measured DMCs. While helical segments of a homopolymer have often been considered as blocks, the question of the range of correlations in heteropolymer systems is not a simple one and is closely related to the theory of correlations in disordered systems[9]. In 2021 the Nobel Prize was awarded to prof. Parisi "for the discovery of the interplay of disorder and fluctuations in physical systems from atomic to planetary scales"[22], and our problem belongs to the same category.

In order to improve the description of many-body effects, and to double-check the importance of system length for the appearance of fine structure, another approach has been put forth. It is based on a one-dimensional many-body Hamiltonian with Potts spins[37]. This approach is known as the Generalized Model of Polypeptide Chains (GMPC) and provides an alternative for describing melting of a double stranded DNA homopolymers in the limit of small loops[12]. The GMPC in the thermodynamic limit has been used to study the conformational transitions in both homopolymers[37] and annealed heteropolymers[23,29]. In order to separately consider the effect of frozen sequence and to depict the origins of the fine structure of DMCs, we have employed the constrained annealing method, which accounts for the existence of two types of degrees of freedom: annealed ones, which can rearrange in order to minimize the free energy and frozen ones, which do not change in time. Thus biopolymer conformations are considered as annealed degrees of freedom, while the sequence of repeating units as frozen[28,29].As a result of these considerations for infinite chain lengths, it was possible to estimate the melting temperature and melting interval[23,28,29]. The constrained annealing method also resulted in two large peaks on the DMC, but failed in reproducing any of fine structures on DMC[28,29]. Since multi-peak DMCs have been experimentally observed in a number of studies[25,35,38], there is an obvious necessity for more in depth theoretical investigation on the origins and accurate predictions of fine structure on DMCs of duplex DNAs.

In the present study, aiming to improve our understanding of correlations in helical regions of DNA, we directly generate binary random sequences of different finite lengths. The corresponding transfer-matrices[23] are directly multiplied to result in the partition function, which allows to calculate the degree of helicity for each of the generated sequences. Block structures of DNA heteropolymers are mimicked by merging sequences with different G–C content. Comparison of the results for random, correlated, block random and block correlated sequences provides new insights into the origin of DMC fine structure.

Specifically, numerical calculations addressed the following questions: (i) What is the difference between predicted DMCs for short versus long DNAs? (ii) How does the presence of sequence blocks affect DMC fine structure? (iii) What are the effects of correlations within sequence blocks?

II. MATERIALS AND METHODS

In this study we compare the effects of four types of sequence schemes onto the DNA melting curves. Considered schemes are: **random, correlated, random block, correlated**

block.

Random sequences with two-component heterogeneity were generated using Wolfram Mathematica[21]. Each repeating unit (r.u.) comprised of G – C basepair, assigned as type "A", enters the sequence with probability $P(A) = x$, and A–T basepair (type "B") with probability $P(B) = (1 - x)$[31]. Thus defined, x has the usual meaning of G–C fraction. In order to be able to observe strong disorder, and not just doping effects, x values from the middle of value interval should be taken. In this study two particular values for random sequences have been used: $x = 0.4$ and $x = 0.5$.

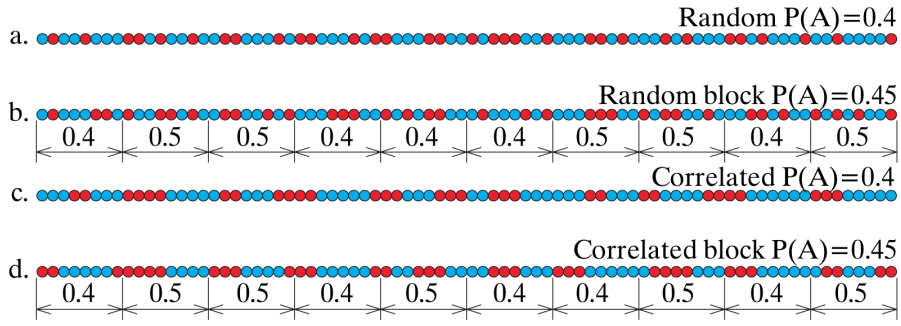


FIG. 1. Illustration of sequence generation: a) random, b) random block, c) correlated, and d) correlated block. Red spheres correspond to type "A", and the blue ones to type "B". Block length (in the figure, of 8 r.u) is actually 3000-r.u.-long, see text.

Correlated sequences are generated by increasing the probability of the neighboring basepair of the same type by Δx, and decreasing the probability for different neighboring basepairs by the same Δx. A detailed block-scheme of correlated sequence generation algorithm is provided in B. Introduction of short-ranged correlations in the sequence allows to reveal their effect on melting curve fine structure.

Block sequences for both random and correlated cases result from randomly merging different sequences with different G–C fractions, x. Mixing (merging) equal amounts of random sequences with $x = 0.4$ and $x = 0.5$, produces blocked sequences with $x = 0.45$. Sequence generation is illustrated in Fig.1.

As proposed by Parisi, reduced free energies at fixed disorder content (x, in our case) should tend to a certain limit at infinite system sizes. For a given parametrization, the length scale for self-averaging of the free energy (not shown) in our model is $N \geq 3000$ r.u.[26,32] both for random and correlated sequences. The type of a sequence is not affecting the self-averaging, since the self-averaging length scale is much larger than the sequence correlation, which affects the probability for the nearest neighbour only. Therefore, in the current study, the temperature dependence of the helicity degree is calculated for both random and correlated sequences in blocks of 3000 r.u.

For every sequence, the partition function is given by,

$$Z = \mathrm{Tr} \prod_{i=1}^{N} G_i, \tag{1}$$

where

$$G_i(\Delta \times \Delta) = \begin{bmatrix} e^{\frac{U_i}{T}} & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & Q_i - 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 & Q_i - 1 \end{bmatrix}, \tag{2}$$

is the transfer matrix of the GMPC model[23] for i-th r.u. of either "A" or "B"; $e^{\frac{U_i}{T}}$ is the energetic parameter where U_i is the hydrogen bond formation energy, and T is temperature; Q_i is the number of conformations. Dimensions Δ of the transfer-matrix are determined by the number of repeating units, affected by the formation of hydrogen bonds in one r.u. and reflects the single strand rigidity. The $\Delta = 2$ value corresponds to the Zimm-Bragg model[33], applied to polypeptides (see A for the detailed description on correspondence of parameters). To account for a larger rigidity of a single strand DNA, we use $\Delta = 4$ value throughout this study.

As opposed to treating the sequence disorder in partition function using approximate methods[23,29], applicable for the infinite chain length N, we use the straightforward multiplication of matrices for each generated sequence, which is exact and valid for any chain length. Also, the approach can be extended beyond the calculation of melting temperature T_m and melting interval ΔT. That is, the degree of helicity θ (fraction of intact basepairs) can be determined according to:

$$\theta = \frac{\mathrm{Tr}[E,O]\prod_{i=1}^{N}\hat{M}_i\begin{bmatrix}O\\E\end{bmatrix}}{N\mathrm{Tr}[E,O]\prod_{i=1}^{N}\hat{M}_i\begin{bmatrix}E\\O\end{bmatrix}};\hat{M}_i=\begin{pmatrix}\hat{G}_i & \hat{G}'_i\\ O & \hat{G}_i\end{pmatrix}. \tag{3}$$

where E is the unit and O is null matrix and $\hat{M}_i$ is the supermatrix. Here, $\hat{G}'_i = \partial\hat{G}_i/\partial J_0$, so the first element of the transfer matrix $\hat{G}'_{i,11} = e^{J_i}$ and all other elements are zero. For the details of the model definition please consult the C.

Calculations are performed by multiplication of transfer-matrices in Eq. (1) for each type of generated sequence. From Eq. (3) each supermatrix depends on the type of r.u., and the degree of helicity represents every particular sequence and is unique.

On a Macintosh HD running MacOS 14.1.1 with 16 GB of memory and an Apple M1 Pro chip, the process of calculating the helicity degree and gathering tensor data for each DNA sequence consisting of 3000 repeated units requires approximately 2.5 hours. Subsequently, the collected data is utilized for merging sequences, and the merging process along with helicity degree calculation varies in duration, ranging from several seconds to 3 minutes. The time required depends on the number of sequences being merged.

DMCs are obtained from the numeric derivation of Eq. (3). The estimated numerical error is of the order of 10^{-5}.

III. RESULTS AND DISCUSSION

A. Dependence of the DMC profile on length, block structure and sequence correlations

Equation (3) enables the examination of sequence dependent features such as length, block structure and sequence correlations, on the DMC. Obtained curves visually remind those familiar from the experiments. Thus, the dashed blue line of Fig.2a ($N = 6000$, correlated sequence) displays a profile visually similar to the experimental DMC of calf thymus DNA[35,38]. Since each particular DMC depends on sequence, different sequences are incomparable, and only general trends can be deduced from results of the calculations.

As seen in Fig.2, fine structure is present at shorter sequence lengths for all sequence arrangements considered (random, random block, correlated and correlated block), and DMC turns into a smooth curve, when sequences become longer, a trend first reported by Lyubchenko *et al*[25].

In absence of a block structure, random and correlated sequences (dotted and dashed lines, Fig.2) result in a smooth curve with a single maximum, while random block and correlated block sequences (dash-dotted and solid lines, Fig.2) give rise to a smooth curve with two well-expressed and wide maxima. Most persistent in terms of a fine structure is the correlated block scheme.

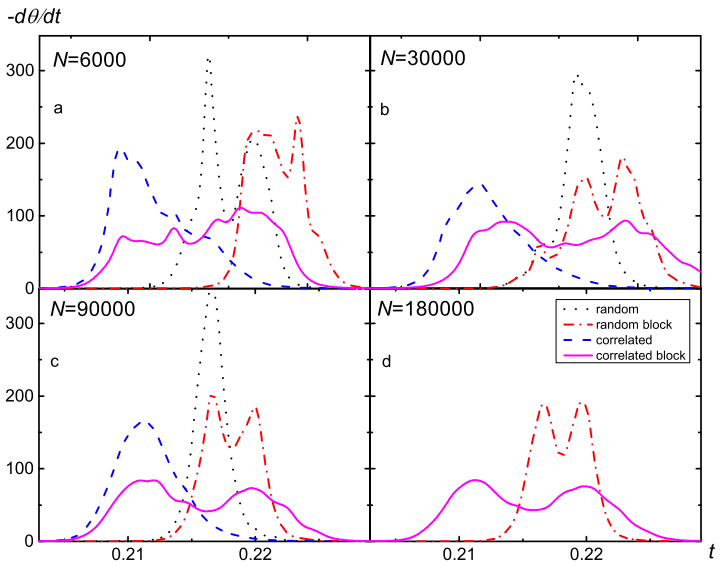


FIG. 2. Calculated DMCs for long heteropolymers with $N = 6000$ (a), $N = 30000$ (b), $N = 90000$ (c) and $N = 180000$ (d). Curves with $x = 0.4$ are compared with those for the joint block scheme generated with parameters $x = 0.4$, $x = 0.5$, $\Delta x = 0.3$. The other parameters, same for all curves, read: $U_A = 1$, $U_B = 0.8$, $Q_A = 71$, $Q_B = 51$, $\Delta = 4$. Temperature is given in units of $t = T/U_A$. The legend indicates the specific sequence arrangements.

It is tempting to suggest an explanation for the small peaks forming the DMC profile as appearing from individually melting regions, following Lyubchenko *et al*[25]. Using the expression for the transition interval of melting of a homopolymer, the same authors have qualitatively estimated the area of a peak of fine structure to be proportional to ν/N of the total area of the differential melting curve (ν is the average length of a helical region). The larger is the chain length N, the smaller will be the contribution from the peak, and eventually, peaks disappear for large N. The length of an individually melting region should be of the order of a spatial correlation length, which formally brings the problem to the calculation of the correlation function, which is not a trivial question for disordered systems, heteropolymer, in this case. Probably, this is the reason, why Lyubchenko *et al*[25] have to rely on the transition interval formula for a homopolymer. As it was reported before[1,23], the span of correlations within the secondary structure of a heteropolymer is certainly different from that of a homopolymer. This is obvious from the fact that the transition interval of a heteropolymer has a different functional dependence on model parameters than a homopolymer. Thus, while valuable as an idea, the concept of Lyubchenko *et al* needs to be justified, when it comes to the scale of conformational correlations for the heteropolymer.

With a very interesting numerical experiment, Lyubchenko *et al* have illustrated the relevance of the value of nucleation parameter σ for the fine structure. Fig.1 of Ref.[25] shows a calculated DMC for $N = 30000$ r.u. long random heteropolymer in a model that allows for loop formation, and apparently results in fine structure. Fig.3 of Ref.[25] is for exactly the same parameters, but without loops, and one can see the fine structure disappeared. What has changed? The presence of loops introduces additional cooperativity into the system, and significantly decreases the value of nucleation parameter σ. When the loop factor is removed from the model, the cooperativity decreases, and the value of σ increases. So, which factor is relevant *per se*, loop formation possibility, or the value of the

nucleation parameter σ? Lyubchenko *et al* gave a clear answer to the question in Fig.4 of Ref.[25], where they took the model without loops, but decreased the value of σ to the value it had in the case of loops, and have demonstrated the re-appearance of fine structure. So, according to Lyubchenko *et al*, it is all about the values of σ: at small values there is a fine structure, while at large ones there are no signs of fine structure. If recall the simple relationship between the parameter σ and the correlation length ξ, $\sigma = \xi_{max}^{-2}$ provided by the homopolymer version of the GMPC model (please see Eq. (S3) of SI and references therein), one can re-read the message: large conformational correlations (small σs) result in fine structures. Similar to the preceding paragraph, here again, the need to estimate the span of conformations for a heteropolymer becomes apparent. Let us say it clear, although we do not study the correlation function for the heteropolymer in this paper, the work of Lyubchenko *et al* inspires a clear working ansatz for the future study: fine structure is a finite size effect, related to the ξ/N ratio. When the ratio is comparable with unity, finite size effects, including the fine structure, are apparent, while at small values the system self-averages nicely, so that no fine structure is present.

As we illustrate in Fig.2, besides the abovementioned factors, the specific sequence arrangements can also seriously affect the DMC shapes. That is, our calculated DMCs of random sequences are smooth for $N = 30000$ (Fig.2b, dotted), while the presence of block structure results in fine structure (Fig.2b, dash-dotted). The significant difference between our approach and that of Lyubchenko *et al* is our utilization of sequence blocks. Pieces of random sequence of 3000 nucleotides each were linked to obtain the block structure in our study. In contrast, Lyubchenko *et al* generated random sequences, and considered chain pieces of one helical segment long as individual blocks.

To conclude this section, results presented in Fig.2, qualitatively support the view of fine structure as a finite size effect, and the expression of it depends on the disorder, encoded in the different schemes of sequence organization.

B. Effects of averaging over the sequences

DMCs for different sequences generated at the same value of x mimic melting curves for different DNA sequences with the same G–C content. The theory of systems with a random potential, one example being the model of heterogeneous sequence DNA melting[23], similar to that considered here, provides estimates of quantities, averaged over the disorder, in the limit of infinite system size[9]. Understanding the mechanisms behind this averaging may illustrate effects on the DMC, arising from particular sequence structure of DNA.

We start by generating ten sequences with $N = 3000$ at fixed x, calculating the DMC and then averaging the curves. As shown in Fig.3, the curves obtained do not follow a pattern. Some have well-defined fine structure, others do not, however, the fine structure is more expressed on DMCs for correlated sequences (see more curves in 7 of D). Interestingly, for both the uncorrelated (Fig.3a) and correlated (Fig.3b) sequences, results of averaging are similar: smooth curves with two weakly expressed maxima are obtained.

C. Comparison between block-averaging and sequence-averaging

Results in Fig.3 show, that averaging over large number of shorter sequences results in a smooth curve, in a way similar to the long sequence behavior. To check if it is true, we take 10 (30) sequences, $N = 3000$ r.u. each, calculate the average DMC (codename **average**), then compare it with the DMC of a **joint** sequence, made by gluing together exactly the same 10 (30) sequences. Both the random and correlated curves (Figs. 4a, c and 4b, d respectively) look quite similar, with a minimal number of peaks. The tendency to a shapeless DMC without fine structure is obvious. This demonstrates qualitatively similar DMCs result for sequences, made by gluing blocks into a large single chain and for those obtained by averaging over a large number of independent blocks. More joint curves are

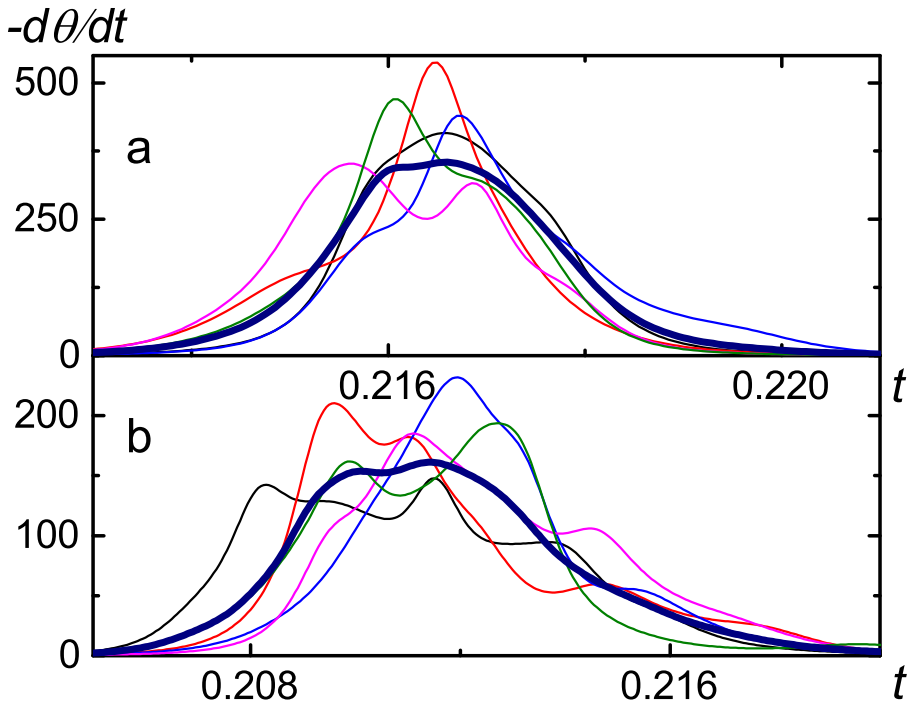


FIG. 3. Calculated DMCs for heteropolymers (colored thin lines) and their average (thick line). (a) random sequences: (b) sequences with correlation. $x = 0.4$, other parameters are as in Fig. 2. Averages were determined from 10 individually calculated curves (not all shown).

shown on 8 in D. Thus the *ansatz* is confirmed: whether shorter sequences are joint into a long chain, or results for short sequences are averaged, resulting DMC is same. And longer is the chain length (or larger is the number of shorter sequences), the better is agreement between the two (compare Figs. 4a,b with Figs. 4c,d).

Dashed thin vertical lines on Fig. 4 indicate the melting temperatures for random sequences, drawn according to the theoretically calculated value[4,23] for infinitely long sequences as $T_m = xT_A + (1-x)T_b$ (for $x = 0.4$ and $x = 0.5$ values). For random sequences, Fig. 4a,c, the maxima of DMCs are very close to the corresponding random heteropolymer values (dashed vertical lines), while the introduction of block structure in Fig. 4b,d, breaks the agreement with the theoretical values.

To study this question further, we plot the positions of maxima (considered as melting temperature) and transition interval vs the chain length in Fig. 5 (left). One can see that for random sequences the melting temperature is close to the theoretical value and for the correlated heteropolymers there is no agreement between the theoretical melting temperatures. As to the melting interval (Fig. 5, right), it is also strongly affected by the sequence organization. Thus, when the random and block sequences are compared, the melting interval of the second is about twice bigger. When the random and correlated sequences are compared, the melting interval is again twice bigger for correlated sequences. And finally, when the block sequences are compared with correlated block ones, the melting interval is again approximately twice bigger, thus it is about four times bigger than melting interval of random sequences. Additional info on melting temeperatures and intervals can be found in Table 1 of C.

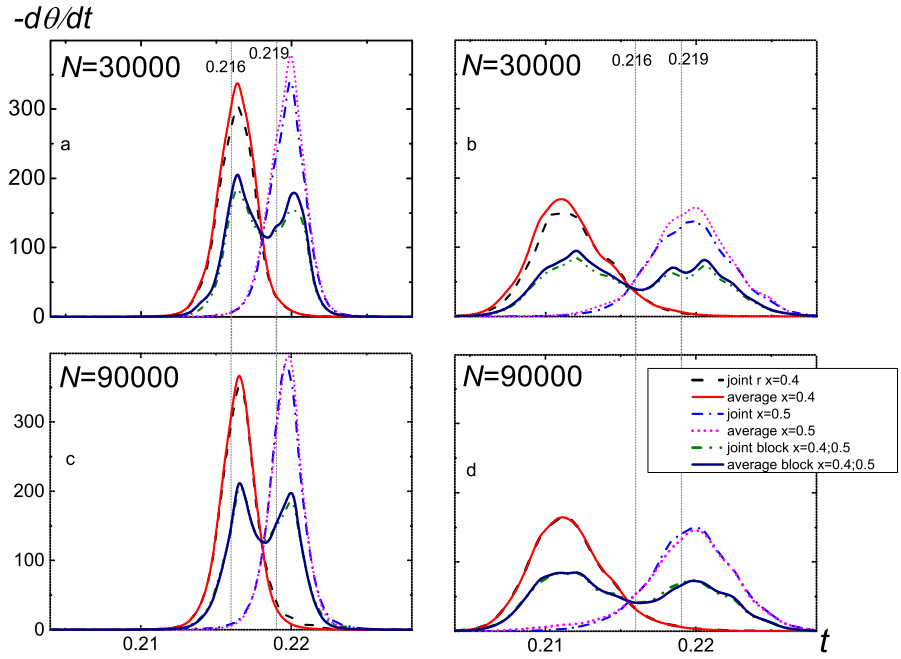


FIG. 4. Joint DMC of heteropolymers in comparison with corresponding average DMC: (a), (c) random sequences, (b), (d) sequences with correlation. The parameters are as in Fig. 2. Melting temperatures calculated from $T_m = xT_A + (1-x)T_b$ are indicated for joint heteropolymers as dashed thin vertical lines.

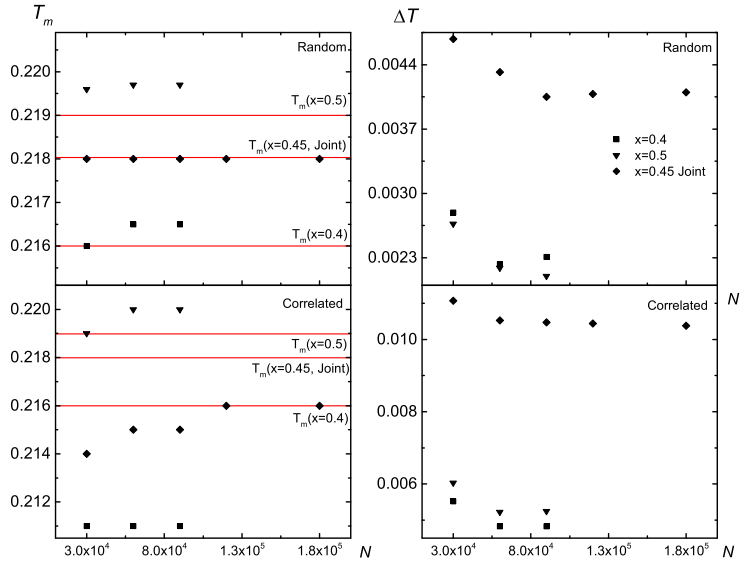


FIG. 5. The dependence of melting temperature (left) and melting interval (right) on the sequence length. The red line shows the theoretical value (see Ref.[23]).

IV. SUMMARY AND CONCLUSIONS

Our results indicate, that sequence organization strongly affects the presence or absence of fine structure (Fig.2). When different sequence schemes are compared, the expression of a fine structure (at fixed sequence length N) increases in the following order: **random** $\rightarrow$ **correlated** $\rightarrow$ **random block** $\rightarrow$ **random correlated**.

The fact, that increasing sequence length smears out the fine structure, means its presence on DMC curves is a finite-size effect, which depends on a ratio ξ/N, where the spatial scale ξ should be related to the correlation length of the system. The calculation of the correlation length of a system with disorder (heteropolymer) is related to the calculation of the second Lyapunov exponent[9], and is not a trivial task *per se*.

Introduction of blocks should decrease the physical cooperation between different parts of the system due to the additional disorder it creates as compared to the random sequence. The better expression of fine structure (see Fig.2) for the block sequences (at same N) can be understood as decreased value of the correlation length ξ of the system. However, since the calculation of the correlation length has not been done here, it is a speculation, for the moment.

Results of Fig. 2 showed that random and correlated sequences give rise to a single peak DMC curve at large N, while the sequences with blocks of different G–C content x result in two peaks. This confirms our earlier results regarding the presence of just two peaks, obtained using the method of constrained annealing (CA)[28,29]. Based on this observation we claim that the CA method is a good approximation for completely random heteropolymers. We notice correlated sequences resulting in more peaks on the DMCs than uncorrelated sequences. The number of peaks on the DMC tend to increase for block sequence structure. We also see the tendency to smear out all the peaks of fine structure at increased sequence length, in agreement with Lyubchenko et al[25]. The very fact, that many DNAs of living organisms show the presence of fine structure on their melting profiles, may be considered as a sign, that they are optimized not to exceed a certain length, the biological meaning of which has still to be clarified.

Our study suggests insights into the stability of DNA based on its primary structure and length. Targeted gene therapy crucially depends on DNA conformations to avoid biodegradaton during the delivery phase[2,8]. It makes our findings relevant for future development of drugs intended for use in gene therapy.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

FUNDING

A.A. acknowledges funding of Science committee RA in frame of scientific project 19YR-1F057 and the scientific project N22rl-012. A.B. acknowledges the partial financial support from Erasmus+ Project No. (2023-1-SI-KA171-HED-000122882).

ACKNOWLEDGMENTS

The authors would like to thank dr. V. Morozov and dr. Y. Mamasakhlisov for valuable discussions and advices.

SUPPLEMENTAL DATA

DATA AVAILABILITY STATEMENT

The datasets [GENERATED/ANALYZED] for this study can be found in the GitHub repository: https://github.com/AsatryanArevik/Publication-data.

Appendix A: The basic model for heteropolymers

A microscopic Potts-like one-dimensional model with δ-particle interactions is used to describe the helix-coil transition in polypeptides from early 1990s[27,37]. Later it was shown that the same approach could be applied to DNA if the large-scale loop factor is ignored[12]. This model, known as the Generalized Model of Polypeptide Chains (GMPC) has the Hamiltonian of the form:

$$-\beta\mathcal{H} = J\sum_{i=1}^{N}\delta_i^{(\Delta)}, \tag{A1}$$

where the summation is performed across all N repeated units, $\beta = 1/T$ is inverse temperature, $J = U/T$, U is the hydrogen bond formation energy, and $\delta_l^{(\Delta)} = \prod_{k=0}^{k-1}\delta(\gamma_l, 1)$, where $\delta(x, 1)$ is the Kronecker symbol. γ_l is a Potts spin, describing the conformational state of repeated unit, that can vary from 1 to Q values. $\gamma_l = 1$ is taken as helical state and all the other values of γ correspond to coil states. Thus, the presence of the Kronecker delta in the Hamiltonian guarantees that the energy J results only when all Δ adjacent repeated units are in the helical conformation. Consequently, this considers the limitations on backbone chain conformations imposed by the formation of hydrogen bonds. The transfer-matrix, corresponding to the Hamiltonian (A1) reads[37]:

$$G(\Delta\times\Delta) = \begin{bmatrix} W & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & Q \\ 1 & 1 & 1 & \cdots & 1 & 1 & Q \end{bmatrix}, \tag{A2}$$

where $W = e^{\frac{U}{T}}$, has the characteristic equation $\lambda^{\Delta-1}(\lambda-W)(\lambda-Q) = (W-1)(Q-1)$. Once the eigenvalues are found, the partition function of the model can be estimated (at large N) as $Z = \lambda_1^N$, and the spatial correlation length as $\xi = \ln^{-1}(\lambda_1/\lambda_2)$, where $\lambda_1,\ \lambda_2$ are the maximal and the second largest eigenvalues, correspondingly. The point of closest approach of two largest eigenvalues determines the transition point estimated from $W \approx Q$ as $T_m = U/\ln Q$. The correlation length is the distance, over which the conformations of repeated units are correlated. Since the range of interactions is finite ($\Delta < \infty$), the conformational transition in this 1D system is not a phase transition (Pierls-Landau theorem)[15]. Therefore, the correlation length does not tend to infinity and remains finite, but reaches its maximum $\xi_{max} \propto Q^{\frac{\Delta-1}{2}}$ at the transition point[5,37]. Nucleation parameter σ and equilibrium constant s from Zimm-Bragg model are in correspondence with the following parameters in GMPC[27]:

$$\sigma = \xi_{max}^{-2}, \quad s = \frac{W}{Q}. \tag{A3}$$

However, one should consider, that above results are derived for a homopolymer model, and are certainly different for a heteropolymer GMPC. DNA is a heteropolymer since the adenine-thymidine ($A-T$) and guanine-cytosine ($G-C$) base pairs differ in the number

of hydrogen bonds. We model the structural disorder of DNA through the dependence of W parameters on the position of r.u. in the sequence. Thus, the partition function for a given sequence of base pairs can be written as

$$Z = \mathrm{Tr} \prod_{i=1}^{N} G_i, \tag{A4}$$

where the transfer matrix Eq. (A2) is modified to read

$$G_i = \begin{cases} G_{AT} \text{ if } i \text{ -th r.u. is of A-T type} \\ G_{GC} \text{ if } i \text{ -th r.u. is of G-C type.} \end{cases}$$

Given that these matrices are not commutative, every sequence within the collection of sequences of length N and disorder concentration x possesses distinct statistical characteristics. Typically, each specific chain can be identified by a sequence-dependent free energy F_{seq}. Nevertheless, it is widely accepted that the free energy adheres to the principle of self-averaging. This principle asserts that the probability distribution of free energies for independent samples is highly constrained, resulting in the virtual alignment of the free energy with the mean free energy for nearly all sequences. This self-averaging tendency has been shown to come to saturation in the range of $2000-3000$ r.u. length of the sequence depending on the peculiarities of sequence generation[32]. In the past, we have treated general characteristics of such heteropolymer model, using the microcanonical[23] and the constrained annealing[28] methods. The essence of a microcanonical method is that the quenched averages can be substituted by the annealed average at fixed disorder concentration x, since the two quantities are equal to each other up to the fluctuations of x. The constrained annealing is a more sensitive method, and treats the conformations as annealed degrees of freedom, while the sequences of r.u. is considered frozen. Interested reader is addressed to Refs.[23,28], and references therein. Using these approaches the expressions for the transition temperature

$$T_m = xT_{GC} + (1-x)T_{AT}, \tag{A5}$$

where T_{GC} and T_{AT} are corresponding homopolymer melting temperatures, and the interval

$$\Delta T = 2x(1-x)\ln Q(T_{GC} - T_{AT})^2/T_m \tag{A6}$$

have been found. None of the approaches gave access to fine structure.

Appendix B: Algorithm of correlated sequence generation

When the parameters are defined: $x = 0.4$ (G-C fraction), $\Delta x = 0.3$ correlation parameter, and $N = 3000$ number of repeated units, there is random generation of a number between 0 and 1. The random number is compared with "x" and if it is smaller or equal to "x" then, sequence is filled with "A" type R.U. (corresponding to G-C bp) and in the same step "x" is increased with amount of Δx to increase the probability of the next R.U. being generated as "A" again, if no, the sequence is filled with "b" type R.U. (A-T bp) and "x" is decreased with amount of Δx to increase the probability of the next R.U. being generated as "B" again. The value of Δx is chosen comparable to the value of "x", so that the correlation is noticeable in the sequence and influences the DMCs.

Appendix C: Detailed calculation of helicity degree for heteropolymer

The Hamiltonian for a heteropolymer in frame of GMPC is as follows:

$$-\beta H = \sum_{i=1}^{N} J_i \delta_i^{(\Delta)}. \tag{C1}$$

Where $J_i = \frac{U_i}{T}$ and the energy of hydrogen bond U_i depends on the type of repeated unit, $\delta_j^{(\Delta)} = \prod_{k=\Delta-1}^{0} \delta(\gamma_{j-k}, 1)$, where $\gamma_i = 1, 2, ..., Q_i$. In this paper the model of DNA is taken as a sequence of repeated units with bimodal heterogeneity both by energies of helicity structure formation, and by number of conformations of repeated units. Thus a variable σ_i is introduced, to take value 1 with given probability x and -1 with probability $(1-x)$. Accordingly, x is the fraction of type A repeated unit in the system: $x_A = \frac{N_A}{N_A+N_B}$. Therefore, we deal with two-component heteropolymer and the intramolecular hydrogen bond's energies of A and B types of repeated units can be expressed as: $J_i = J_0 + \Delta J\sigma$

$$J_A = J_0 + \Delta J \text{ and } J_B = J_0 - \Delta J.$$

Where J_0 is energetic parameter to be changed accordingly with the type of repeated unit. According to[4,12,13], partition function determines as

$$Z = J^* \prod G_i J \quad \text{where} \quad J^* = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \quad J = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \cdots \\ 0 \\ 1 \end{pmatrix}. \tag{C2}$$

However, it has been shown[32] that for a heteropolymer longer than 30 nucleotides, eq. C2 can be replaced with the following one with high precision:

$$Z = Tr \prod_{i=1}^{N} G_i, \tag{C3}$$

While calculating the partition function according to Eq.(C3)) for long biopolymers, sometimes the values exceed the limits of the technical support (Wolfram mathematica). To avoid such disturbances we made the following transfiguration:

$$G_i = \lambda_{1i} g_i, \tag{C4}$$

where λ_{1i} is the principal eigenvalue for the transfer-matrix G_i. Thus, partition function for bimodal heterogeneity will appear as:

$$Z = \lambda_{1A}^{x} \lambda_{1B}^{1-x} Tr \prod_{i=1}^{N} g_i. \tag{C5}$$

Where λ_{1A} and λ_{1B} are principal eigenvalues of type A and type B repeated units correspondingly.

It is generally known, that the following expression can be used to calculate helicity degree of the system:

N		3000		30000		60000		90000		120000		180000	
T_m	$x = 0.4$	0.217	0.211	0.216	0.211	0.217	0.211	0.217	0.211				
	$x = 0.5$	0.220	0.220	0.220	0.219	0.220	0.220	0.220	0.220				
	$x = 0.45$			0.218	0.214	0.218	0.215	0.218	0.215	0.218	0.216	0.218	0.216
ΔT	$x = 0.4$	0.00212	0.00472	0.00279	0.00552	0.00223	0.00483	0.00231	0.00483				
	$x = 0.5$	0.00194	0.00521	0.00267	0.00603	0.00219	0.00522	0.00210	0.00525				
	$x = 0.45$			0.00468	0.01107	0.00432	0.01053	0.00405	0.01047	0.00408	0.01044	0.00410	0.01038

Table 1. Melting temperatures and intervals depending on random or correlated sequences and their lengths.

$$\theta_N = \frac{N_h}{N} = \frac{1}{NZ}\sum_i \delta_i^{(\Delta)} \exp^{-\beta H}. \tag{C6}$$

Where N_h is the number of repeated units in helical state. Taking into account the Hamiltonian expression (C1), for helicity degree we obtained:

$$\theta_N = \frac{1}{NZ}\frac{\partial Z}{\partial J_0}. \tag{C7}$$

With the help of partition function expression in terms of transfer-matrix, we will obtain:

$$\theta_N = \frac{1}{NZ}\sum_i Tr \prod_{k=1}^{i-1} G_k G_i' \prod_{k=i+1}^{N} G_k. \tag{C8}$$

For calculation of the latest equation we have used method of supermatrices. The supermatrix we have inserted has dimensions$(2\Delta \times 2\Delta)$ and can be expressed as:

$$\hat{M}_i = \begin{pmatrix} \hat{G}_i & \hat{G}_i' \\ O & \hat{G}_i \end{pmatrix}, \tag{C9}$$

where O is null matrix $(\Delta \times \Delta)$.

$$\hat{G}_i' = \frac{\partial \hat{G}_i}{\partial J_0}, \tag{C10}$$

consequently, the matrix has only one nonzero element: $\hat{G}'_{i,11} = \exp^{J_i}$, all the other elements of matrix are equal to zero. Therefore, the helicity degree takes the following form:

$$\theta_N = \frac{Tr[E,O]\prod_{i=1}^{N}\hat{M}_i \begin{bmatrix} O \\ E \end{bmatrix}}{NTr[E,O]\prod_{i=1}^{N}\hat{M}_i \begin{bmatrix} E \\ O \end{bmatrix}}. \tag{C11}$$

Where E is identity matrix $(\Delta \times \Delta)$.

Appendix D: Melting parameters and DMCs

From Fig. 8 various length of block systems were obtained through merging of heteropolymers each with 3000 base pairs. The shortest shown contains 30000 r.u. and the longest: 180000 r.u. It is clear, that with increasing number of r.u. the curves become smoother, however, the inclination toward smoothing is more pronounced for random blocks in contrast to correlated blocks.

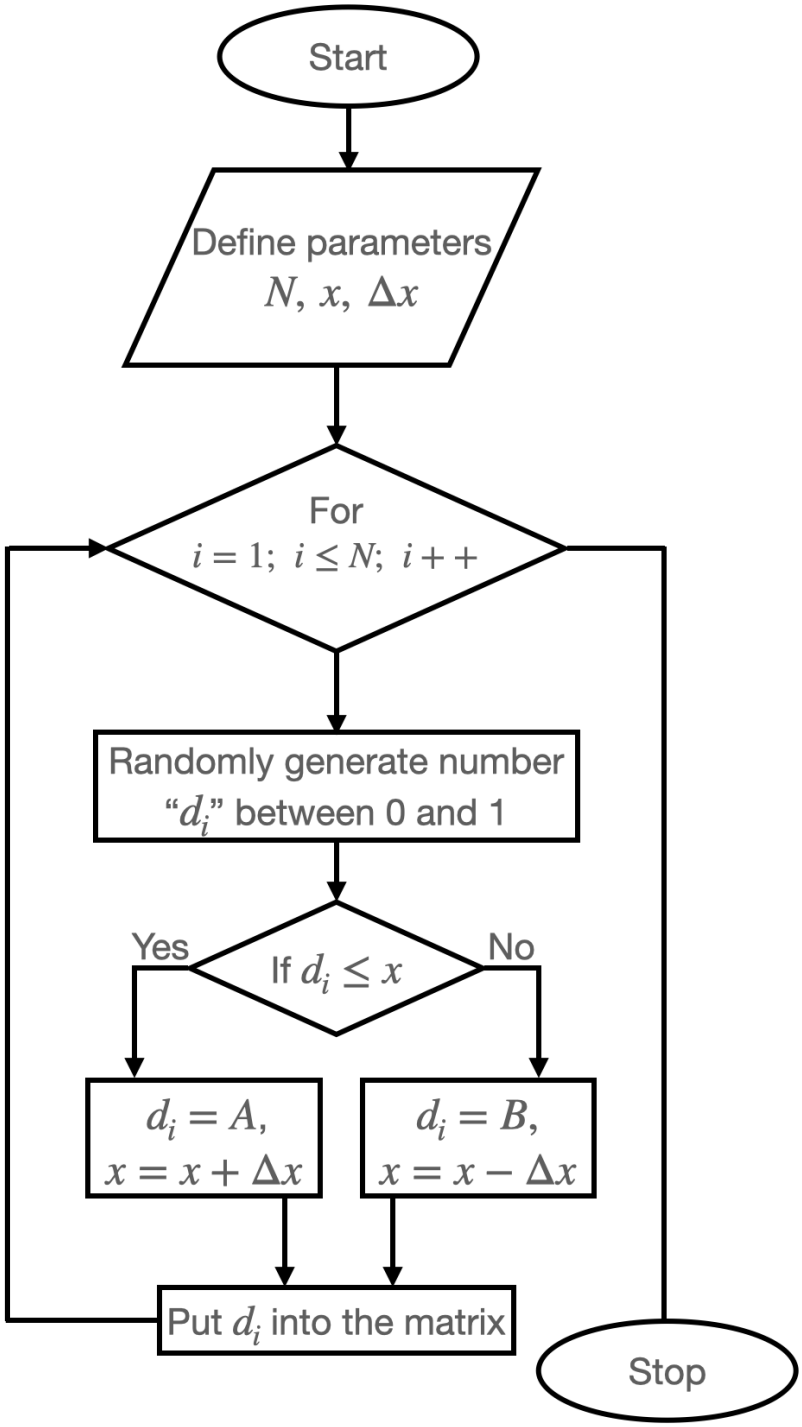


FIG. 6. Block-scheme for correlated sequence generation

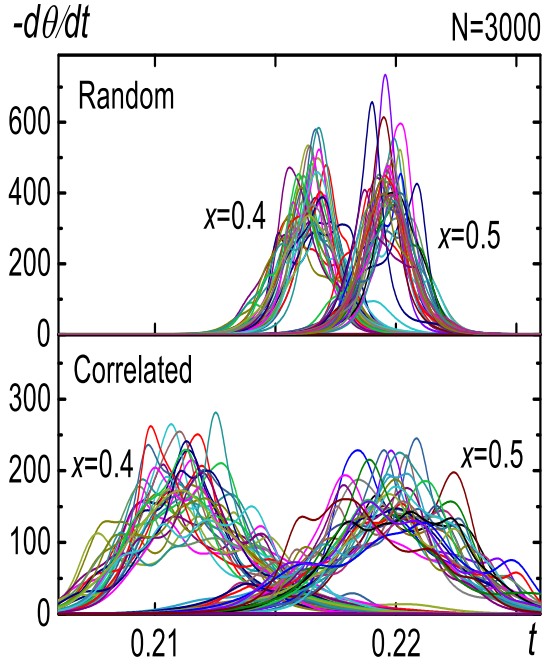


FIG. 7. DMCs for 30 sequences of each type.

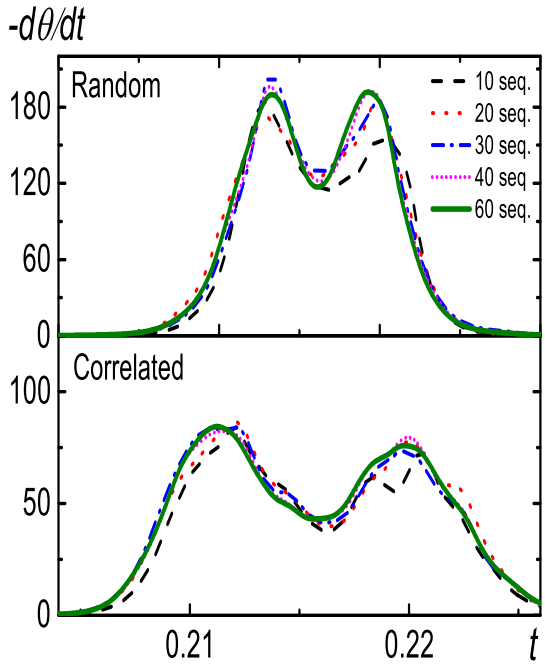


FIG. 8. From 10 to 60 merged sequences, each contains 3000 r.u. As a result block random and block correlated sequences are obtained.

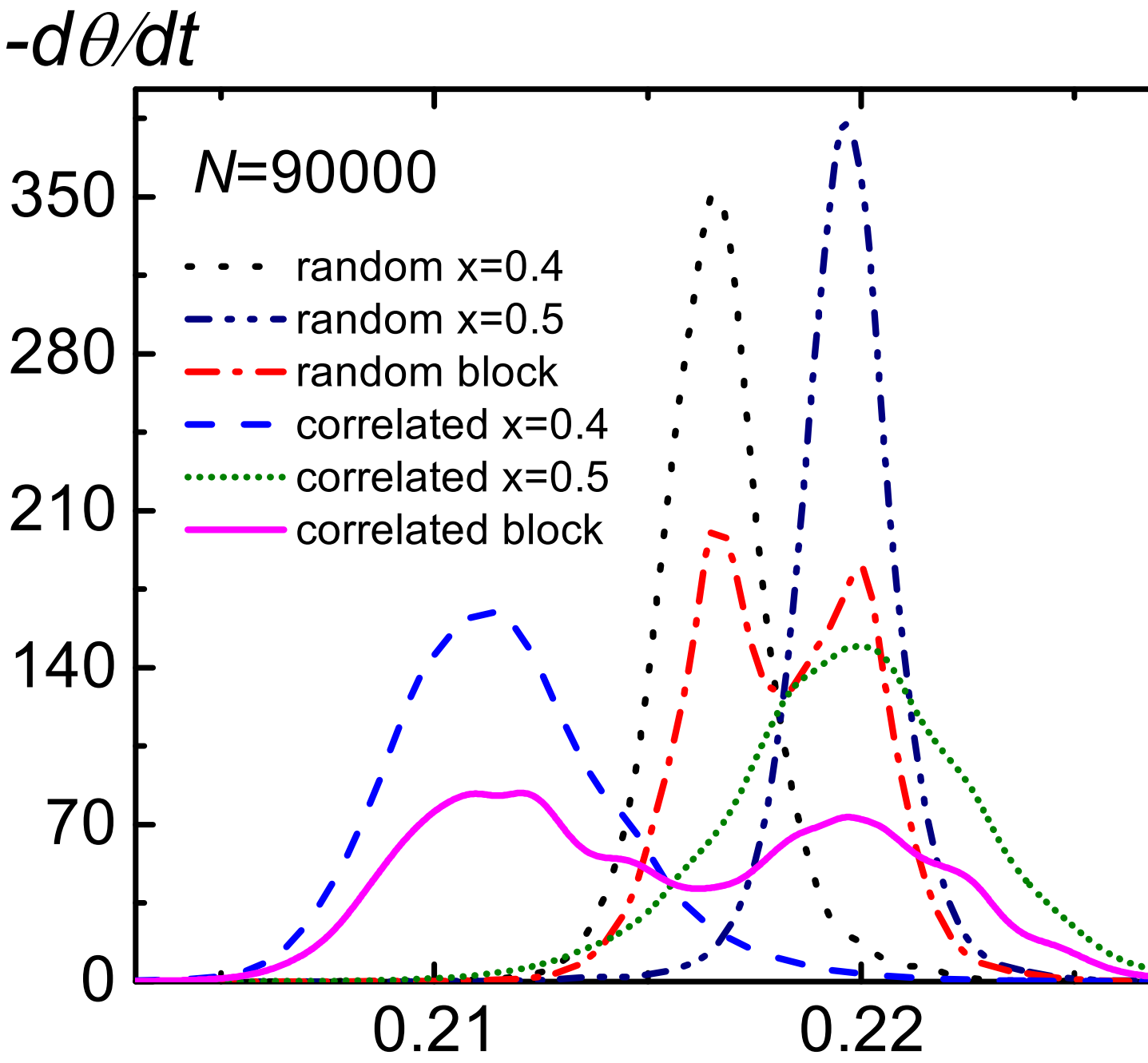
[1]Grosberg A. and Khokhlov A. *Statistical Physics of Macromolecules*. AIP Press, New York, 1994.
[2]Ibraheem D. Elaissari A. and Fessi H. Gene therapy and dna delivery systems. *Analytical biochemistry*, 459:70–83, 2014.
[3]Vologodskii A. and Frank-Kamenetskii M. D. Dna melting and energetics of the double helix. *Physics of life reviews*, 25:1–21, 2018.
[4]A. V. Badasyan. *Helix-coil transition in biopolymers: influence of inhomogeneities*. PhD thesis, Yerevan State University, Yerevan., 2004.
[5]Podgornik R. Mamasakhlisov Y. Morozov V. Badasyan A. Giacometti A. Helix-coil transition in terms of potts-like spins. *Eur. Phys. J. E*, 36:46–55, 2013.
[6]Zimm B.H. and Bragg J.K. Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.*, 31:526–535, 1959.
[7]Cantor C. and Schimmel T. *Biophysical Chemistry*. Freeman and Co., San-Francisco, 1995.
[8]Chi Q. Yang Z. Xu K. Wang C. and Liang H. Dna nanostructure as an efficient drug delivery platform for immunotherapy. *Frontiers in pharmacology*, 10:500239, 2020.
[9]Paladin G. Crisanti A. and Vulpiani A. *Products of Random Matrices in Statistical Physics*. Springer-Verlag, Berlin, 1993.
[10]Poland D. and Scheraga H. *Theory of Helix-Coil Transition in Biopolymers*. Academic Press, New York, 1970.
[11]Mamasakhlisov Y. Sh. Todd. B. A. Badasyan A. V. Mkrtchyan A. V. Morozov V. F. and Parsegian A. V. Dna stretching and multivalent-cation-induced condensation. *Phys. Rev. E*, 80:031915, 2009.
[12]Morozov V.F. Mamasakhlisov Y.Sh. Sh. Hayryan and Hu C.-K. Microscopical approach to the helix-coil transition in dna. *Physica A*, 281:51–59, 2000.
[13]Flory P. J. Determination of melting temperature and temperature melting range for dna with multi-peak differential melting curvese. *Analytical biochemistry*, 479:28–36, 1969.
[14]Watson J.D. and Crick F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.
[15]Landau L. and Lifshits E. *Statistical Physics*. Pergamon, Oxford, 1988.
[16]Azbel Ya M. Random two-component one-dimensional ising model for heteropolymer melting. *Phys Rev Lett*, 31:589–592, 1973.
[17]Azbel Ya M. Dna sequencing and melting curve. *Proceedings of the National Academy of Sciences*, 76:101–105, 1979.
[18]Fixman M. and Freire J.J. Theory of dna melting curves. *Biopolymers*, 16:2693–2703, 1977.
[19]Vedenov A. A. Dykhne A. M. and Frank-Kamenetskii M. D. The helix-coil transition in dna. *Sov. Phys. Uspekhi*, 14:715–736, 1972.
[20]Wartell R. M. and Benight A. S. Thermal denaturation of dna molecules: A comparison of theory with experiment. *Physics Reports*, 126:67–107, 1985.
[21]Wolfram research, inc. (2021), 2021.
[22]G. Parisi. Nobel lecture: Multiple equilibria. *Reviews of Modern Physics*, 95:030501, 2023.
[23]Badasyan A. V. Grigoryan A. V. Mamasakhlisov Y. Sh. Benight A. S. and Morozov V. F. The helix-coil transition in heterogeneous double stranded dna: Microcanonical method. *The Journal of Chemical Physics*, 123:194701, 2005.
[24]Badasyan A. V. Tonoyan Sh. A. Mamasakhlisov Y. Sh. Giacometti A. Benight A. S. and Morozov V. F. Competition for hydrogen-bond formation in the helix-coil transition and protein folding. *Phys Rev E*, 83:051903, 2011.
[25]Lyubchenko Y. L. Frank-Kamenetskii M. D. Vologodskii A.V. Lazurkin Y. S. and G. G. Gause. Fine structure of dna melting curves. *Biopolymers*, 25:1019–1036, 1976.
[26]Asatryan A. Mamasakhlisov Y. Sh and Morozov V. F. Differential melting curves in heterogeneous biopolymers. *Journal of Contemporary Physics (Armenian Academy of Sciences)*, 55:259–264, 2020.
[27]Hairyan Sh. A. Morozov V.F. Mamasahlisov Y. Sh. Helix-coil transition in polypeptides. a microscopic approach ii. *Biopolymers*, 35:75–84, 1995.
[28]Tonoyan Sh. A. Asatryan A. V. Andriasyan A. K. Mamasakhlisov Y. Sh. and Morozov V. F. Helix-coil transitions in heteropolymers: the constrained annealing approach. *Journal of Biomolecular Structure and Dynamics*, 33:126–126, 2015.
[29]Tonoyan Sh. A. Asatryan A.V. Mamasakhlisov Y. Sh. and Morozov V.F. Helix-coil transition in biopolymers with multicomponent heterogeneity of energy and number of conformations. *Journal of Contemporary Physics (Armenian Academy of Sciences*, 49:132–137, 2014.
[30]Cheng S. Singh A.K. Wen C. and Vinh N. Q. Long-range dna-water interactions. *Biophys J*, 120:4966–4979, 2021.
[31]Asatryan A. V. Algorithm of calculations of helix-coil transition parameters in heteropolymeric biopolymers on the bases of generalized model of polypeptide chain. *Armenian Journal of Physics*, 11:18–24, 2018.
[32]Asatryan A. V. *Influence of aqueous solutions of low-molecular ligands on helix-coil transition in heterogeneous biopolymers*. PhD thesis, Yerevan State University, Yerevan., 2020.
[33]Badasyan A. V. Giacometti A. Mamasakhlisov Y. Sh. Morozov V.F. and Benight A S. Microscopic formulation of the zimm–bragg model for the helix-coil transition. *Phys. Rev. E*, 81:1–4, 2010.
[34]Lehman G. W. and McTague J. P. Melting of dna. *The Journal of Chemical Physics*, 49(7):3170–3179, 1968.

[35]Monaselidze J. Majagaladze G. Barbakadze Sh. Khachidze D. Gorgoshidze M. Kalandadze Y. Haroutiunian S. Dalyan Y. and Vardanyan V. Microcalorimetric investigation of dna, poly (da) poly (dt) and poly [d (ac)] poly [d (gt)] melting in the presence of water soluble (meso tetra (4 n oxyethylpyridyl) porphyrin) and its zn complex. *Journal of Biomolecular Structure and Dynamics*, 25:419–424, 2008.

[36]Tonoyan Sh. Khechoyan D. Mamasakhlisov Y. and Badasyan A. Statistical mechanics of dna-nanotube adsorption. *Phys. Rev. E*, 101:062422, 2020.

[37]Ananikyan N.S. Hajryan Sh. A. Mamasakhlisov Y.Sh. and Morozov V.F. Helix-coil transition in polypeptides: A microscopical approach. *Biopolymers*, 30:357–367, 1990.

[38]Haroutiunian S. G. Dalian Y. B. Aslanian V. M. Lando D. Yu and Akhrem A. A. A method for determining the relative effect of ligands on at and gc base pairs in dna: applications to metal ions, protons and two amino acids. *Nucleic acids research*, 18:6413–6417, 1990.


-dθ/dt
N=90000
random x=0.4
random x=0.5
random block
correlated x=0.4
correlated x=0.5
correlated block
350
280
210
140
70
0
0.21
0.22