Dependence on temperature and GC content of bubble length distributions in DNA

Journal:	*Nano Letters*
Manuscript ID:	nl-2008-02461t
Manuscript Type:	Communication
Date Submitted by the Author:	12-Aug-2008
Complete List of Authors:	Kalosakas, George; University of Patras, Materials Science Ares, Saul; Max Planck Institute for the Physics of Complex Systems

Dependence on temperature and GC content of bubble length distributions in DNA

G. Kalosakas
Department of Materials Science,
University of Patras, 26504 Rio, Greece
S. Ares
Max Planck Institute for the Physics of Complex Systems,
Nöthnitzer Str. 38, D-01187 Dresden, Germany,
and Grupo Interdisciplinar de Sistemas Complejos (GISC)

August 12, 2008

Abstract

We report numerical results on the temperature dependence of the distribution of bubble lengths in DNA segments of various GC concentrations. Base-pair openings are described by the Peyrard-Bishop-Dauxois model and the corresponding thermal equilibrium distributions of bubbles are obtained through Monte Carlo calculations for bubble sizes up to the order of a hundred base-pairs. The dependence of the parameters of bubble length distribution on temperature and the GC content is investigated. We provide simple expressions which approximately describe these dependences.

Local openings of the DNA double helix are required in several biological functions, for instance transcription and replication. These local separations of the two DNA strands are mediated by a specific machinery in the cell. In order to deal with such complex processes, it is necessary first to understand the interactions keeping together the two complementary strands within a single DNA duplex, as well as the properties of fluctuating DNA openings in thermal equilibrium. This fact has been realized long time ago and base-pairing interactions and the thermal stability of the double helix have been conventionally probed by increasing temperature up to the DNA denaturation transition [1, 2].

Extended base-pair openings (bubbles) occur in DNA due to thermal fluctuations even at biological temperatures and, moreover, it has been speculated that they may play a role in the recognition of specific DNA sites by DNA-binding proteins [3, 4]. By increasing temperature these bubbles grow and more bubbles are nucleated, thus leading to the complete separation of the two strands at the denaturation transition. Therefore statistical properties of DNA bubbles in a wide temperature regime, extending from biological temperatures up to the melting transition, are of particular interest. The purpose of this work is the investigation of the distribution of bubble lengths and its temperature dependence for DNA sequences containing different percentages of guanine-cytosine (GC) base-pairs.

In a recent study we have presented the distribution of bubble lengths at 310K and we found that it can be described by a power-law modified exponential [5]. Increasing the GC content reduces the probability for bubbles of a fixed length and also diminishes the average bubble length. Anharmonic interactions between complementary bases forming base-pairs are responsible for the observed nonexponential distribution. The same form of bubble length distribution has been derived in the framework of the Poland-Scheraga (PS) model [6, 7, 8], which represents a completely different theoretical approach of DNA denaturation than the Peyrard-Bishop-Dauxois (PBD) model [9] that we use in our calculations. This distribution is also found for a primitive version of the PBD model, viz. the Peyrard-Bishop [10] model with linear stacking interactions, but in this case the characteristic values of the parameters of the distribution are different [11].

Here we examine how the bubble length distribution varies with temperature and present its complete dependence on both temperature and the GC fraction of the DNA segment. The PBD model [9] is used for the description of base-pair openings, where a set of continuous variables y_n represent the base-pair displacements from equilibrium distance, the index n labels the base-pairs along the DNA chain. The potential energy of the system consists of two parts: the on-site interaction $V(y_n)$ within each base-pair and the stacking interaction $U(y_n, y_{n+1})$ between adjacent base-pairs. A Morse potential is used for the on-site energy, $V(y_n) = D_n(e^{-a_n y_n} - 1)^2$, where the parameters D_n and a_n distinguish between GC and AT base-pairs ($D_{GC} = 0.075$eV, $a_{GC} = 6.9$Å^{-1} for a GC base-pair and $D_{AT} = 0.05$eV, $a_{AT} = 4.2$Å^{-1} for an AT pair), while a nonlinear potential describes the stacking interaction, $U(y_n, y_{n+1}) = \frac{K}{2}(1 + \rho e^{-b(y_n+y_{n+1})})(y_n - y_{n+1})^2$, with $K = 0.025$eV/Å^2, $\rho = 2$, and $b = 0.35$Å^{-1}. The values of the parameters have been obtained from previous works [12],[4, 13], where it has been found that these values are able to describe experimental situations. The efficiency of the rather simple PBD model to describe base-pair openings in DNA has led to its extensive use in the literature [14, 15, 12, 16, 4, 17, 13, 18, 19, 20, 21, 22, 23, 24]. In particular, our choice for the PBD model for the study of bubble length distributions is motivated by the success of the model in reproducing experimental measurements of bubble formation [13, 25] and the possibility to perform calculations with long sequences of up to tens of thousands of base pairs [24].

Considering a random DNA sequence of a given GC percentage, x_{GC}, at equilibrium at temperature T, we calculate the probability per base-pair for the formation of a bubble of length l, $P(l)$, by counting during Monte Carlo simulations the average occurrences of openings (base-pair displacements) larger than a fixed threshold y_{thres} at l successive base-pairs. The Monte Carlo details are as in Ref. [5] and the threshold value for the openings is $y_{thres} = 1.5$Å. We use periodic boundary conditions in DNA segments of length 1000 base-pairs (sufficiently larger than the studied bubble sizes) and therefore our results refer to internal bubbles in long DNA chains. We have repeated the calculation using different values for the threshold to consider a base pair open, finding no qualitative difference in the results. However, the quantitative values of the results do depend on this choice; this dependence will be described elsewhere.

In Figure 1 we show bubble length distributions at various temperatures for two values of GC percentage, $x_{GC} = 50\%$ and 87.5%. Similar plots have

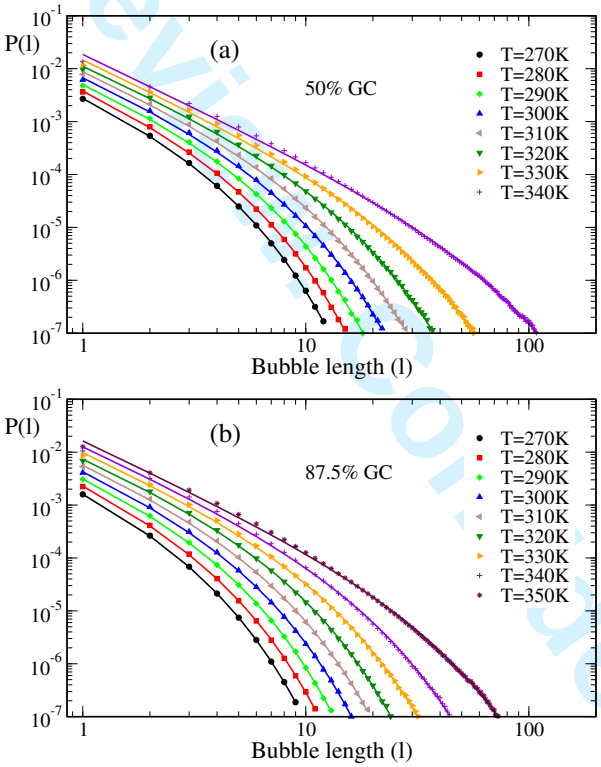

Figure 1: Distribution per base-pair of bubble lengths l (in number of base-pairs), $P(l)$, for different temperatures (points, as indicated in the plots) at random DNA sequences with a GC content of **(a)** 50 % and **(b)** 87.5 %. Continuous lines are fits with the distribution, Eq.(1). $y_{thres} = 1.5$Å.

been obtained for nine different values of x_{GC}: 0%,12.5%,25%,...,100%. In Ref. [5] we showed a plot of the distribution for all these GC fractions at a single temperature, 310K. Except for the cases of $l = 1$ at relatively higher temperatures (closer to the melting transition), the numerical results can be rather well described by the power-law modified exponential [5]

$$P(l) = W \frac{e^{-l/\xi}}{l^{\,c}}, \qquad \text{for } l > 1. \tag{1}$$

In the following we characterize the dependence of the parameters of the distribution on T and x_{GC} and provide approximate expressions for the relations $\xi(T, x_{GC})$, $c(T, x_{GC})$, and $W(T, x_{GC})$. The distribution parameters are obtained through fitting of plots like those of Figure 1 with Eq. (1), using a weight proportional to $1/P(l)^2$. We note that we are interested here in bubble lengths up to about a hundred, or a few hundred at most, base-pairs (which are relevant for any practical purpose) and not for the asymptotic behavior of the distribution, which may be described by different values of parameter c [23]. Therefore we are not concerned with the order of the melting transition and the exponent c presented here is not indicative of the kind of the transition [16, 7, 8, 23, 26], as it also depends on the somehow arbitrary value of the threshold chosen to consider a base pair open.

Figure 2 presents the dependence of the decay length ξ of Eq. (1). In 2a the variation of ξ with temperature is shown (points) for different values of x_{GC}. The T-dependence is accurately described by the divergent function

$$\xi(T) = \xi_0 + \frac{\xi_1}{T_c - T}, \tag{2}$$

where T_c is the denaturation temperature. Such a relation is also valid for the PS model [8]. Lines in Figure 2a show fittings of the $\xi(T)$ data with Eq. (2), using a weight proportional to $1/\xi^2$. The critical temperature T_c as obtained from the fitting at different values of x_{GC} is presented with circles in Figure 2b. A linear dependence of T_c on the GC content is found, in accordance with known experimental results [27] and calculations from simplified models [28]. A least square fitting of the $T_c(x_{GC})$ data results in the continuous line shown in Figure 2b. Regarding the homopolymer cases of poly(dA)-poly(dT) ($x_{GC} = 0\%$) and poly(dG)-poly(dC) ($x_{GC} = 100\%$), the critical temperatures for the transition can be independently calculated through the numerically exact transfer integral technique [29, 30, 31], and the results are $T_c(x_{GC} = 0) = 325.2$K and $T_c(x_{GC} = 100) = 366.0$K. These values are shown with squares joined by a dashed line in Figure 2b, the line lies inside the error interval of the Monte Carlo results. The other parameters ξ_1 and ξ_0 resulting from the fitting of the $\xi(T)$ data with Eq. (2) are shown with circles in Figures 2c and 2d, respectively. Their dependence on x_{GC} can be approximately considered as linear (continuous lines in 2c and 2d). Therefore, the relation

$$\xi(T, x_{GC}) = a_1 + a_2 x_{GC} + \frac{a_3 + a_4 x_{GC}}{a_5 + a_6 x_{GC} - T}\,, \tag{3}$$

where $a_1, a_2, \ldots, a_6$ are constants, can approximately provide the dependence of the decay length ξ on T and x_{GC}.

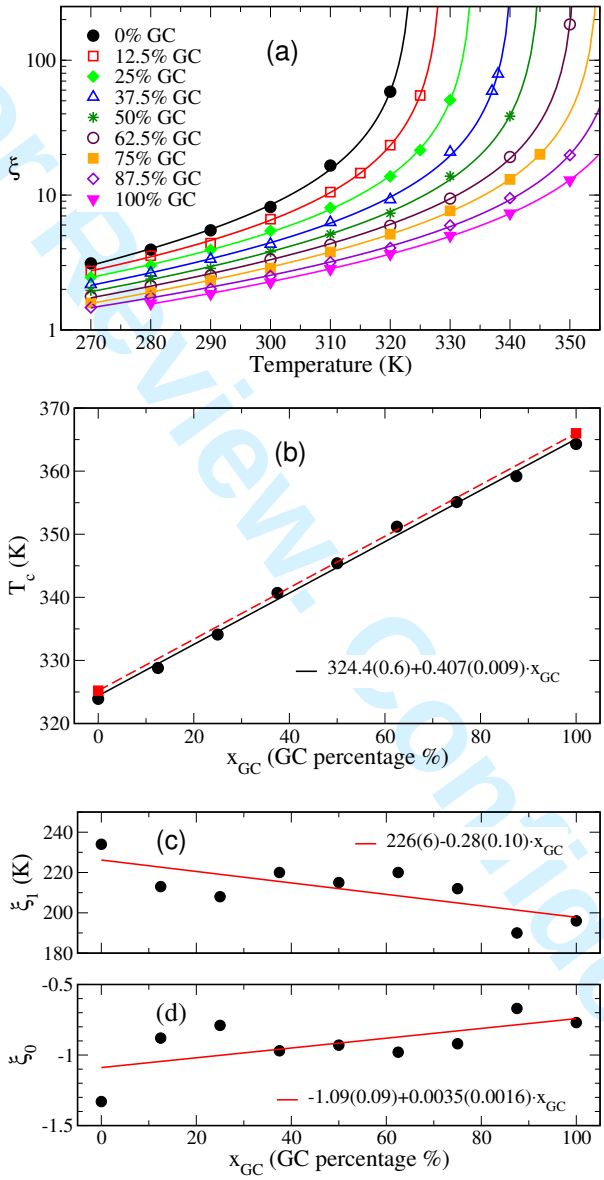

Figure 2: **(a)** Dependence of the decay length ξ of the distribution (1) on the temperature, for different values of the GC content of the DNA sequence (points, as indicated in the plot). Lines show fits with the function of Eq. (2). **(b)** Dependence of the critical temperature T_c, as obtained from the fitting of the $\xi(T)$ data with Eq. (2), on the GC content of the DNA sequence (circles). Solid line represents a least square fit according to a linear dependence. Squares show exact results of the critical temperatures for the homogeneous cases of 0% GC and 100% GC, obtained from transfer integral calculations, while the dashed line connects these two points. **(c)** and **(d)** Dependence of the parameters ξ_1 and ξ_0, respectively, of the fit of the $\xi(T)$ data with Eq. (2), on the GC content of the sequence (circles). Solid lines represent linear fits of the corresponding data. Equations of straight lines resulting from the corresponding fittings are shown in (b), (c), and (d), where the values in parentheses represent errors of the fitting parameters. $y_{thres} = 1.5$Å.

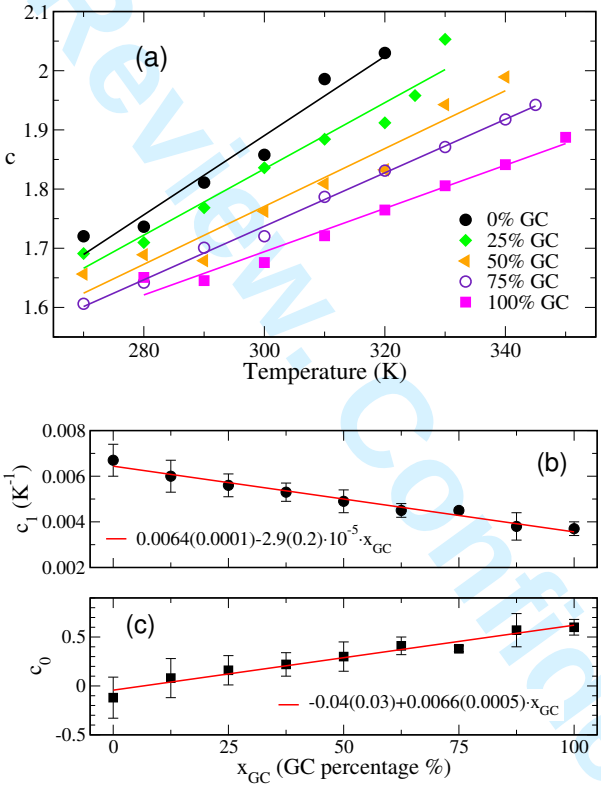

Figure 3: **(a)** Dependence of the exponent c of the distribution (1) on the temperature, for different values of the GC content of the DNA sequence (points, as indicated in the plot). Lines show linear fits with equation (4). **(b)** and **(c)** Dependence of the parameters c_1 and c_0, respectively, of the fit of the $c(T)$ data with Eq. (4), on the GC content of the sequence (circles). Error bars are standard erros resulting from the fitting procedure. Solid lines represent linear fits of the corresponding data and the resulting equations are shown (the values in parentheses represent errors of the fitting parameters). $y_{thres} = 1.5$Å.

The variation of the exponent c of the distribution (1) is presented in Figure 3. Points in 3a show the temperature dependence of c for different GC percentages (for clarity of the plot the corresponding results for $x_{GC} = 12.5\%, 37.5\%, 62.5\%, 87.5\%$ have been omitted). This dependence seems to be described by a linear function

$$c(T) = c_0 + c_1 T. \tag{4}$$

Lines in Figure 3a show fittings of the numerical results with the above formula. The parameters c_1 and c_0 obtained from the fitting at different values of GC content are shown with circles in Figures 3b and 3c, respectively, while the corresponding error bars are derived from the fitting procedure. The latter plots indicate a linear dependence of c_1 and c_0 on x_{GC}, implying the approximate relation

$$c(T, x_{GC}) = b_1 + b_2 x_{GC} + (b_3 + b_4 x_{GC})T, \tag{5}$$

where b_1, b_2, b_3, and b_4 are constants independent of T and x_{GC}.

In Figure 4 is shown the dependence of the coefficient W of the distribution (1). Points in 4a display the variation of W with T for various GC contents. A quadratic function

$$W(T) = W_0 + W_1 T + W_2 T^2 \tag{6}$$

can describe rather well the temperature dependence of W, at least in the studied temperature regime. The corresponding fittings with with Eq. (6) are shown by lines in Figure 4a. The resulting fitting parameters at different GC percentages are plotted by points in Figures 4b, 4c, and 4d, respectively. These plots can approximately be described by linear dependences of W_2, W_1, and W_0 on x_{GC} (solid lines). As a result the expression

$$W(T, x_{GC}) = d_1 + d_2 x_{GC} + (d_3 + d_4 x_{GC})T + (d_5 + d_6 x_{GC})T^2 \tag{7}$$

can approximate the dependence of W on T and x_{GC}, where $d_1, d_2, \ldots, d_6$ are constants.

The detailed results of this investigation do not confirm a bilinear dependence of the parameters of the distribution on the GC content at a fixed temperature, as was proposed in our previous work at $T = 310$K [5]. Instead, a linear dependence of the exponent c and the coefficient W on x_{GC} arises at constant T. However, we note that the results of Figure 2 of Ref. [5] can be described accurately by Eqs. (3), (5), and (7) for $T = 310$K and using the values of constants as provided in Figures 2b,c,d (for a_i), 3b,c (for b_i), and 4b,c,d (for d_i).

In conclusion, we have presented the dependence of bubble length distributions in DNA on temperature and the GC content. The investigated temperature regime was extended from below biologically relevant values up to the melting transition. Approximate expressions have been obtained for the parameters of the power-law modified exponential distribution. The exponent c behaves linearly both in temperature and GC content, Eq. (5), while the coefficient W shows a quadratic dependence on temperature and a linear dependence on the GC fraction, Eq. (7). The decay length ξ is described by the relatively simple equation (3). The constants a_i, b_i, and d_i appearing in these expressions depend on the amplitude y_{thres} of the considered base-pair openings, and for $y_{thres} = 1.5$Åare given by the values shown in Figures 2, 3, and 4.

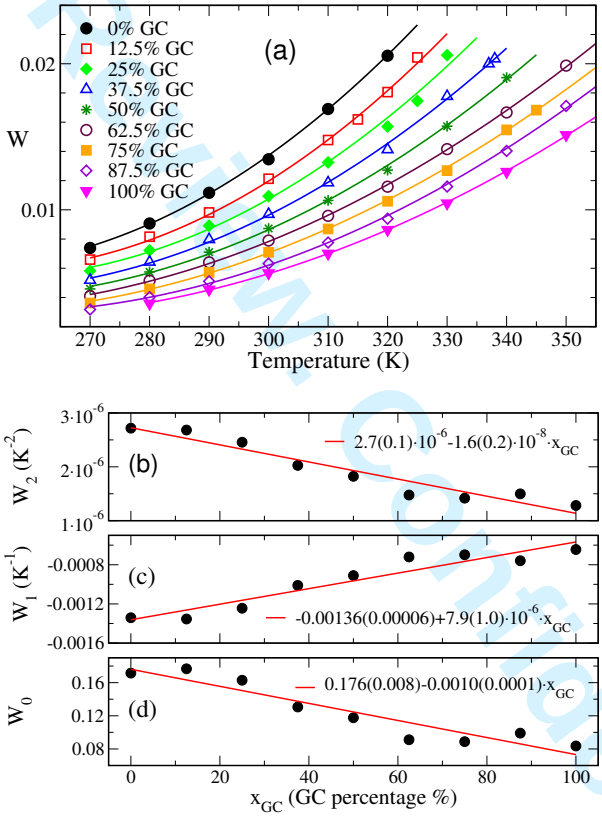

Figure 4: **(a)** Dependence of the preexponential coefficient W of the distribution (1) on the temperature, for different values of the GC content of the DNA sequence (points, as indicated in the plot). Lines show quadratic fits with equation (6). **(b)**, **(c)**, and **(d)** Dependence of the parameters W_2, W_1, and W_0, respectively, of the fit of the $W(T)$ data with Eq. (6), on the GC content of the sequence (circles). Solid lines represent linear fits of the corresponding data and the resulting equations are shown (the values in parentheses represent errors of the fitting parameters). $y_{thres} = 1.5$Å.

Finally, the functional form of the relations obtained in this paper can be used as guide in experimental work. As an example, when performing single molecule experiments, force-extension curves of double stranded DNA display hysteresis [32]. It has been recently shown that this hysteresis strongly depends on the value of the exponent c of the bubble length distribution (1) (see Figure 7 of Ref. [32]). The increase of c with temperature implied by our results leads to larger hysteresis loops at higher temperatures, as seen in these experiments [32].

Acknowledgments. We thank J. Bois for critical reading of the manuscript and J. Bois and N. Theodorakopoulos for enlightening discussions. G.K. acknowledges the hospitality of MPI-PKS in Dresden and support from the C. Caratheodori program C155 of University of Patras. S.A. acknowledges financial support from Ministerio de Educación y Ciencia (Spain) through grant MOSAICO.

References

[1] Wartell, R. M.; Benight, A. S. *Phys. Rep.* **1985**, *126*, 67.

[2] Poland, D.; Scheraga, H. R. *Theory of helix coil transition in biopolymers*, Academic Press (1970).

[3] Banerjee, A.; Sobell, H. M. *J. Biomol. Struct. Dyn.* **1983**, *1*, 253; Sobell, H. M. Proc. *Natl. Acad. Sci. USA* **1985**, *82*, 5328.

[4] Choi, C. H.; Kalosakas, G.; Rasmussen, K. Ø.; Hiromura, M.; Bishop, A. R.; Usheva, A. *Nucleic Acids Res.* **2004**, *32*, 1584; Kalosakas, G.; Rasmussen, K. Ø.; Bishop, A. R.; Choi, C. H.; Usheva, A. *Europhys. Lett.* **2004**, *68*, 127.

[5] Ares, S.; Kalosakas, G. *Nano Lett.* **2007**, *7*, 307.

[6] Poland, D.; Scheraga, H. A. *J. Chem. Phys* **1966**, *45*, 1464.

[7] Kafri, Y.; Mukamel, D.; Peliti, L. *Eur. Phys. J. B* **2002**, *27*, 135.

[8] Coluzzi, B; Yeramian, E. *Eur. Phys. J. B* **2007**, *56*, 349.

[9] Dauxois, T.; Peyrard, M.; Bishop, A. R. *Phys. Rev. E* **1993**, *47*, 44.

[10] Peyrard M.; Bishop, A. R. *Phys. Rev. Lett.* **1989**, *62*, 2755.

[11] Sung, W.; Jeon, J.-H. *Phys. Rev. E* **2004**, *69*, 031902; Jeon, J.-H.; Sung, W.; Ree, F. H. *J. Chem. Phys.* **2006**, *124*, 164905; Jeon, J.-H.; Park, P. J.; Sung, W.;*J. Chem. Phys.* **2006**, *125*, 164901.

[12] Campa, A.; Giansanti, A. *Phys. Rev. E* **1998**, *58*, 3585.

[13] Ares, S.; Voulgarakis, N. K.; Rasmussen, K. Ø.; Bishop, A. R. *Phys. Rev. Lett.* **2005**, *94*, 035504.

[14] Dauxois, T.; Peyrard, M. *Phys. Rev. E* **1995**, *51*, 4027.

[15] Cule, D.; Hwa, T. *Phys. Rev. Lett.* **1997**, *79*, 2375.

[16] Theodorakopoulos, N.; Dauxois, T.; Peyrard, M. *Phys. Rev. Lett.* **2000**, *85*, 6.

[17] Voulgarakis, N. K.; Kalosakas, G.; Rasmsussen, K. Ø.; Bishop, A. R. *Nano Lett.* **2004**. *4*, 629.

[18] van Erp, T. S.; Cuesta-López, S.; Hagmann, J.-G.; Peyrard, M. *Phys. Rev. Lett.* **2005**, *95*, 218104.

[19] Rapti, Z.; Smerzi, A.; Rasmussen, K. Ø.; Bishop, A. R.; Choi, C. H.; Usheva, A. *Europhys. Lett.* **2006**, *74*, 540; *Phys. Rev. E* **2006**, *73*, 051902.

[20] Alexandrov, B. S.; Wille, L. T.; Rasmussen, K. Ø.; Bishop, A. R.; Blagoev, K. B. *Phys. Rev. E* **2006**, *74*, 050901.

[21] Kalosakas, G.; Rasmussen, K. Ø.; Bishop, A. R. *Chem. Phys. Lett.* **2006**, *432* 291.

[22] Voulgarakis, N. K.; Redondo, A.; Bishop, A. R.; Rasmussen, K. Ø. *Phys. Rev. Lett.* **2006**, *96*, 248101.

[23] Theodorakopoulos, N. *Phys. Rev. E* **2008**, *77*, 031919.

[24] de los Santos, F.; Al Hammal, O.; Muñoz, M. A. *Phys. Rev. E* **2008**, *77*, 032901.

[25] Montrichok, A.; Gruner, G.; Zocchi, G. *Europhys. Lett.* **2003**, *62*, 452; Zeng, Y.; Montrichok, A.; Zocchi, G. *Phys. Rev. Lett.* **2003**, *91*, 148101; Zeng, Y.; Montrichok, A.; Zocchi, G. *J. Mol. Biol.* **2004**, *339*, 67; Zeng, Y.; Zocchi G. *Biophys. J.* **2006**, *90*, 4522.

[26] Everaers, R.; Kumar, S.; Simm, C. *Phys. Rev. E* **2007**, *75*, 041918.

[27] Marmur J.; Doty, P. *J. Mol. Biol.* **1962**, *5*, 109.

[28] Ares, S.; Sánchez, A. *Eur. Phys. J. B* **2007**, *56*, 253.

[29] Scalapino, D.J.; Sears, M.; Ferrell, R. A. *Phys. Rev. B* **1972**, *6*, 3409.

[30] Aubry, S. *J. Chem. Phys.* **1975**, *62*, 3217; Krumhansl, J. A.; Schrieffer, J. R. *Phys. Rev. B* **1975**, *11*, 3535.

[31] Ares, S.; Sánchez, A. *Phys. Rev. E* **2004**, *70*, 061607.

[32] Whitelam, S.; Pronk, S.; Geissler, P. L. *Biophys. J.* **2008** *94*, 2452.