Supporting Information

Simulating the lyotropic phase behavior of a partially self-complementary DNA tetramer

Silvia Cristofaro[1], Lorenzo Soprani[1], Lara Querciagrossa[1,3], Tommaso P. Fraccia[2], Tommaso Bellini[2], Roberto Berardi[1], Alberto Arcioni[1], Claudio Zannoni[1], Luca Muccioli[1], and Silvia Orlandi[1]

[1]Dipartimento di Chimica Industriale "Toso Montanari", Università di Bologna, Viale del Risorgimento 4, 40136 Bologna, Italy

[2]Dipartimento di Scienze Farmacologiche e Biomolecolari, Università di Milano, Via Balzaretti 9, 20133 Milano, Italy

[3]CINECA, Via Magnanelli 6/3, 40033 Casalecchio di Reno, Italy

1 Thermodynamics of melting transition

The melting temperature (T_M) of a DNA duplex is the temperature at which half of the oligonucleotide molecules are single-stranded and half are double-stranded, i.e., for every four DNA filaments, two of type 1 and two of type 2,

two are associated and form a double helix and two are single-stranded. For most DNA oligonucleotides, the thermodynamics of hybridization/denaturation is accurately described as a two-state process[1]: in this approximation, one neglects the possibility of the existence of intermediate partial binding states and of larger aggregates. The steep part of the melting curve reflects the single strand (ssDNA1 + ssDNA2) to double strand (dsDNA) to equilibrium.

$$\text{ssDNA1} + \text{ssDNA2} \rightleftharpoons \text{dsDNA} \tag{1}$$

The equilibrium constant of denaturation for this reaction is:

$$K_{\text{eq}} = \frac{[\text{dsDNA}]}{[\text{ssDNA1}][\text{ssDNA2}]} \tag{2}$$

where [ssDNA1], [ssDNA2], [dsDNA] are the molar concentrations.

To simulate the hybridization process for GCCG oligomers we ran MD simulations in the NVT ensemble using the coarse-grained oxDNA model[2] at constant DNA concentration (200, 380 and 450 mg/mL), for samples made of 64 tetramers. We carried out 500 ns trajectories of which 400 ns were for equilibration and 100 ns for production. We then calculated the fraction of paired bases $f(T)$: guanines and cytosines belonging to different tetramers were considered as paired if they were inside a distance range of 9.4 - 11 Å, that we determined by computing the radial distribution of Figure S3. Finally, we averaged it on the equilibrated MD trajectory for each temperature (melting curves are shown in the main article in Figure 1). We considered as melting temperature the temperature at which the fraction $f = 0.5$, i.e. when half of the helices are, on average, in the single strand form. Since GCCG single strands are indistinguishable in the case of self-association,

we can assume that [ssDNA1] = [ssDNA2] = [ssDNA], and the equilibrium to be 2ssDNA $\rightleftharpoons$ dsDNA.

One possible equation for the fraction f of aggregated strands is:

$$f = \frac{2[\text{dsDNA}]}{2[\text{dsDNA}] + [\text{ssDNA}]} \tag{3}$$

Introducing the total concentration of GCCG (either associated or isolated):

$$[\text{GCCG}] = 2[\text{dsDNA}] + [\text{ssDNA}] \tag{4}$$

one obtains

$$f = \frac{2[\text{dsDNA}]}{[\text{GCCG}]} \tag{5}$$

It follows that the concentration of associated GCCG pairs is:

$$[\text{dsDNA}] = \frac{[\text{GCCG}]f}{2} \tag{6}$$

Substituting in equation 4 one obtains:

$$[\text{GCCG}] = 2[\text{dsDNA}] + [\text{ssDNA}] = [\text{GCCG}]f + [\text{ssDNA}] \tag{7}$$

and finally:

$$[\text{ssDNA}] = [\text{GCCG}](1 - f) \tag{8}$$

The equilibrium constant can be then expressed in terms of f and [GCCG] as:

$$K_{\text{eq}} = \frac{[\text{GCCG}]f/2}{[\text{GCCG}]^2(1-f)^2} = \frac{f}{2[\text{GCCG}](1-f)^2} \tag{9}$$

Solving in term of f and using the definition of equilibrium constant $K_{\rm eq} = \exp\left[\Delta S/R - \Delta H/(RT)\right]$, one obtains the temperature dependence of the fraction of double strands that are actually present in the system:

$$2K_{\rm eq}[\mathrm{GCCG}]f^2 - (4K_{\rm eq}[\mathrm{GCCG}] + 1)f + 2K_{\rm eq}[\mathrm{GCCG}] = 0 \tag{10}$$

which has the only one acceptable solution in

$$f(T) = \frac{1 + 4K_{\rm eq}[\mathrm{GCCG}] - \sqrt{1 + 8K_{\rm eq}[\mathrm{GCCG}]}}{4K_{\rm eq}[\mathrm{GCCG}]} \tag{11}$$

that can be used to fit the experimental or simulation data of $f(T)$. To facilitate the fitting procedure we used the Arrhenius plot technique, calculating first the equilibrium constant as a function of $f(T)$ as:

$$K_{\rm eq} = \frac{f(T)}{2(1 - f(T))^2[\mathrm{GCCG}]} \tag{12}$$

Then we plotted $\ln K_{\rm eq}$ vs $1/T$ and we performed a linear fit to get the slope ΔH and the intercept to extrapolate ΔS. Finally, we estimated $T_{\rm M}$ values (Table 1 of the main text) as:

$$T_{\rm M} = \frac{\Delta H}{\Delta S + R\ln([\mathrm{GCCG}])} \tag{13}$$

For the same DNA concentrations, we calculated the aggregate average size in terms of the number of composing tetramers: as expected, temperature strongly affects the dimensions of double strands (see Fig S1, top left). The denaturation mechanism was investigated more in detail by counting the occurring number of coaxial stacking and pairing interactions as a function of temperature (see Fig

S1, at the bottom). The ratio between stacking and pairing interaction is ~ 2, meaning that there is no coaxial stacking without base pairing, thus only double strand aggregates are present. As a consequence, pairing and stacking interactions can be predicted starting from the average size ($\langle S \rangle$) and the average number of aggregates ($\langle N \rangle$) following the equation $\langle N \rangle * 2(\langle S \rangle - 1)$ and $\langle N \rangle (\langle S \rangle - 2)$ for pairing and stacking interactions, respectively (dotted line of Figure S1 (bottom)).

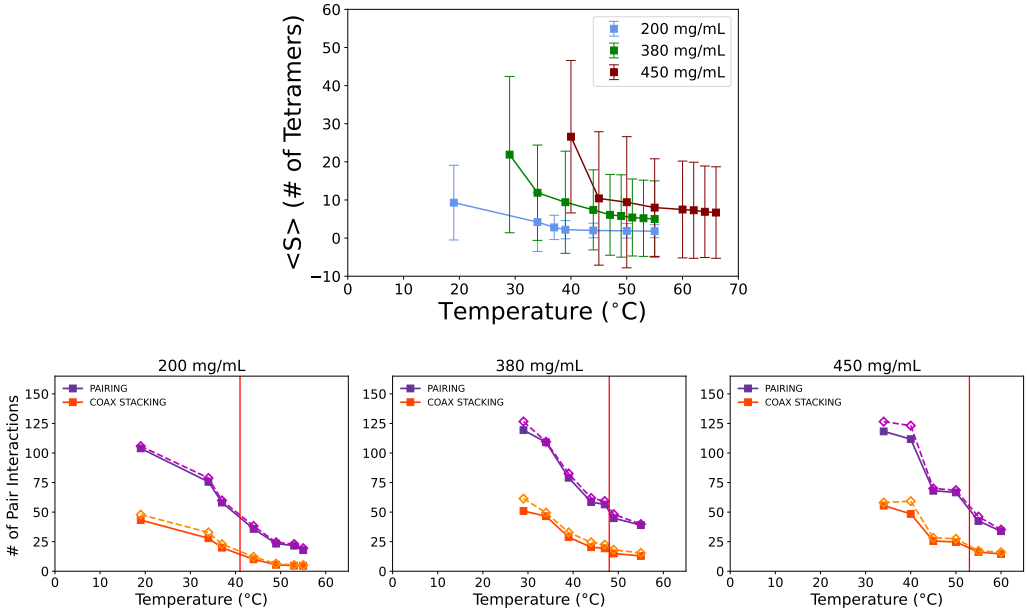

Figure S1: On the top panel, average double strand size $\langle S \rangle$ as a function of temperature for a sample made of N = 64 tetramers at $C_{\mathrm{GCCG}} = 200$ mg/mL (T_{M}=41 °C, light blue squares), 380 mg/mL (T_{M}=47 °C, green squares) and 450 mg/mL (T_{M}=53 °C, dark red squares). Lines are a guide to the eye. The largest size variations are directly correlated with those of the orientational order parameter (cf. Figure 2 in the main text) and not to the melting transition, and that isolated tetramers were not included in the calculation of the average size. On the bottom panel, the number of pairing (purple solid line) and coaxial stacking (orange solid line) interactions as a function of temperature, at concentrations of 200 mg/mL, 380 mg/mL, and 450 mg/mL. Dashed lines represent the predicted number of interactions computed from the average size ($\langle S \rangle$) and number ($\langle N \rangle$) of aggregates: $2 \langle N \rangle (\langle S \rangle - 1)$ and $\langle N \rangle (\langle S \rangle - 2)$ for pairing and stacking interactions, respectively. Vertical red lines correspond to the melting temperatures.

2 Size distribution as a function of concentration

In Figure S2 we show the aggregate size distribution histograms from which the average size double helix aggregates ($\langle S \rangle$) were calculated (cf. Figure 6 in the main text).

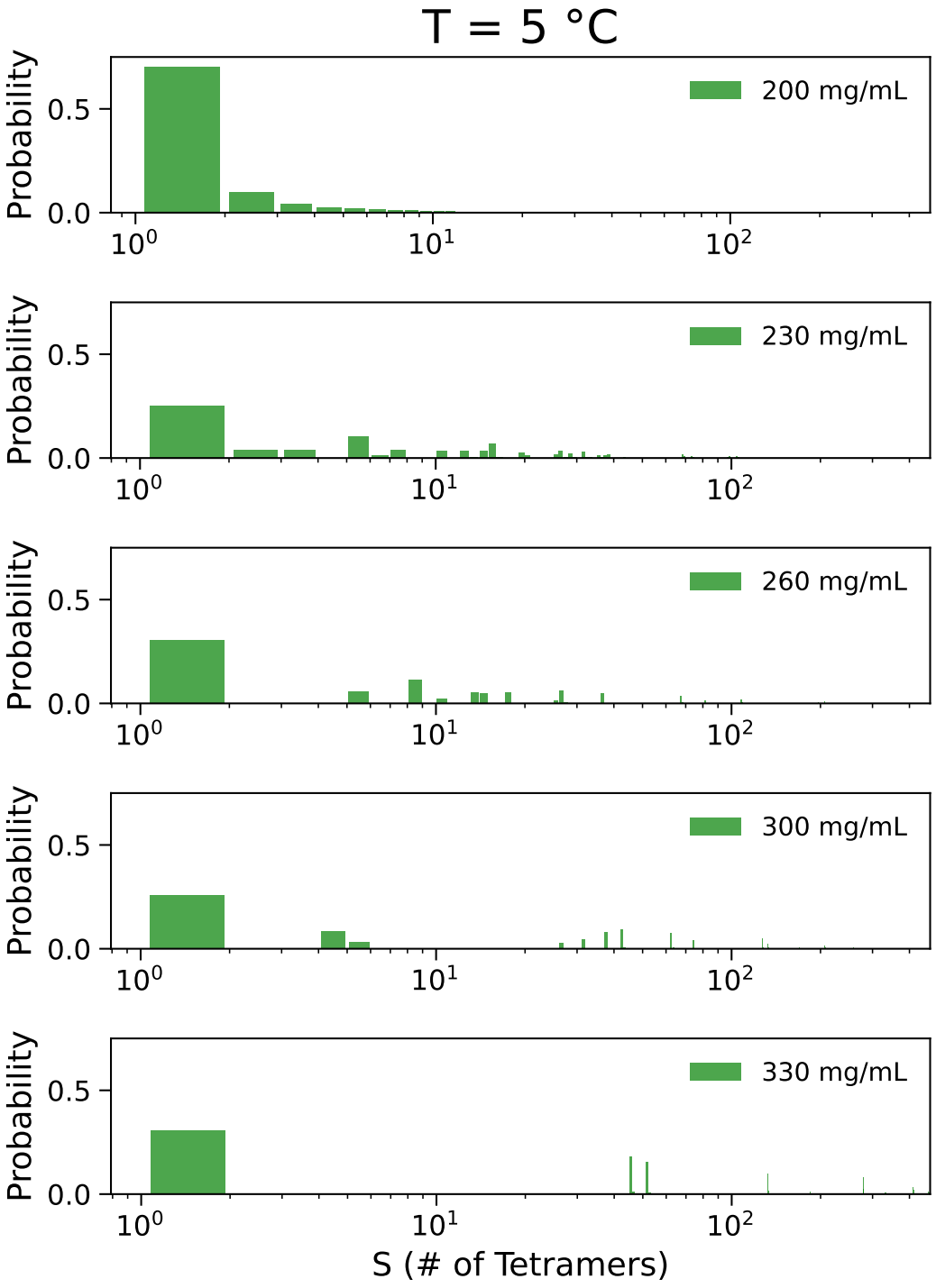

Figure S2: Aggregate size distribution histograms in logarithmic scale at C_{GCCG} = 200, 230, 260, 300 and 330 mg/mL and at T = 5 °C for a sample of N = 512 tetramers.

3 Double strand local structure

We inspected the local positional order through the calculation of the radial distribution function $g(r)$:

$$g(r) \;=\; \frac{1}{4\pi r^2 \rho_N} \left\langle \delta(\mathbf{r} - \mathbf{r}_{ij}) \right\rangle_{ij} \tag{14}$$

which provides the distribution of particle centers i and j as a function of their distance r (the modulus of the inter-particle vector $\mathbf{r}$) at a given number density of particles ρ_N.

In Fig S3 we show the radial distribution in the isotropic phase, obtained by using the center of mass of the coarse-grained sites, *i.e.* the nucleotides. The first peak is linked to the intra-strand distance between the first directly bonded nucleotides, which is found at larger distances in the case of the two guanines (GG) with respect to the guanine-cytosine and the cytosine-cytosine pairs (GC and CC), respectively at 5.0 Å for the first and at 4.7 Å for the latter two. The GG peak in fact arises from the π-stacking interaction between two quadruplets belonging to the same strand, while those for GC and CC are related to covalently bonded pairs inside the same quadruplet. The second and third peaks correspond to the distance between one C (or G) belonging to one strand and the closest couple of the same azo-base located on the complementary strand. Once again we observed that GG distances are longer, this time because of the higher volume occupied by guanine, composed of two condensed aromatic cycles instead of one as for cytosine. For the mixed correlation function CG, the second peak at about 9 Å is related to the second neighbor, either intra- or inter-quadruplet, between

bases belonging to the same strand. Concerning the prevalence of either intra-strand (π-stacking) or inter-strand (H-bond) interactions, we noticed that in our simulations the single strand structure is not stable in any phase, indicating that the strongest driving force for the self-assembly process is the hybridization into double helices by complementary base-pairing. A strong hint of this behavior is the presence, even in the isotropic phase at 200 mg/mL when the size of the aggregates is expected to be small, of a third CG $g(r)$ peak at 10.2 Å, arising indeed from H-bonded complementary bases located on two parallel strands.

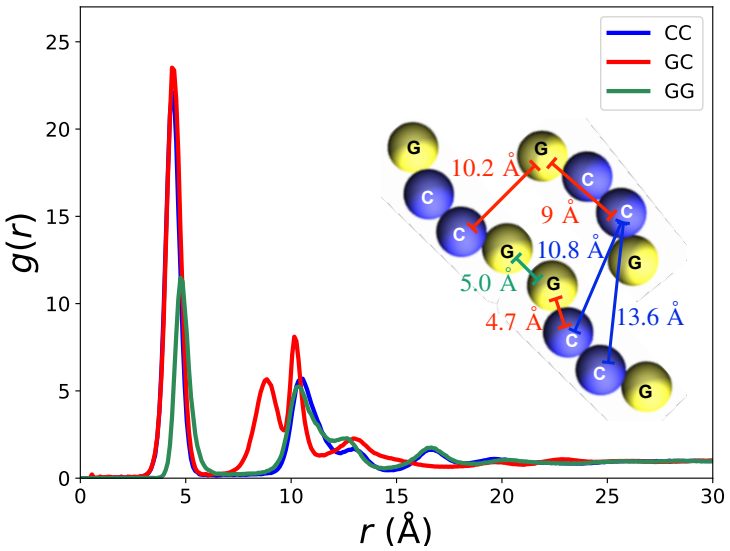

Figure S3: Radial distribution functions computed at $T = 5$ °C for a sample of $N = 512$ tetramers at $C_{\mathrm{GCCG}} = 200$ mg/mL. A schematic representation of the double helix primer (made by 3 GCCG quadruplets) is reported to illustrate the assignment of $g(r)$ peaks.

4 Comparison with Onsager's theory

In the literature, the appearance of LC phases in both long (> 100 bp) and nano (< 20 bp) DNA have been theoretically modeled by considering the B-DNA filaments as repulsive rigid rod-shaped elements[3] according to the well-established Onsager;s theory[4]. This theory accounts for only excluded-volume interactions between rods and therefore is especially suited for the description of lyotropic LC transition, where mesogen concentration is the main driving variable. According to the theory, no nematic ordering is predicted for rods with aspect ratio $L/D < 4$, meaning that the nematic phase should be found for volume fraction $\phi > \phi_{I-N} = 4D/L$, where D is the diameter of a DNA double helix ($D \sim 2$ nm) and L is the length of the DNA "rods". There are several examples in literature in which Onsager limit was confirmed through computer simulations [5, 6, 7], however most of them were conducted taking into account a value of $\langle L \rangle$ that was fixed *a priori*. Here, we propose a different approach which embodies the self-assembly process of DNA strands, so even though the overall number of nucleotides is kept constant, the value $\langle L \rangle$ is obtained from the simulation and it changes spontaneously with the concentration. For this analysis we observe that points in Figure S4, that from our simulations are labeled as isotropic, are situated well inside the isotropic Onsager region, while LC samples are reasonably placed in the region where the nematic phase is predicted to be stable. Nevertheless, the comparison between theoretical and computational results has to be considered in qualitative terms because, as it has been found in other LC systems, polydispersity of aggregates can strongly affect the theoretical phase diagram, leading to an expansion of coexistence regions [8], while intrinsic molecular flexibility can stabilize liquid crystal

phases with respect to Onsager theory predictions[9].

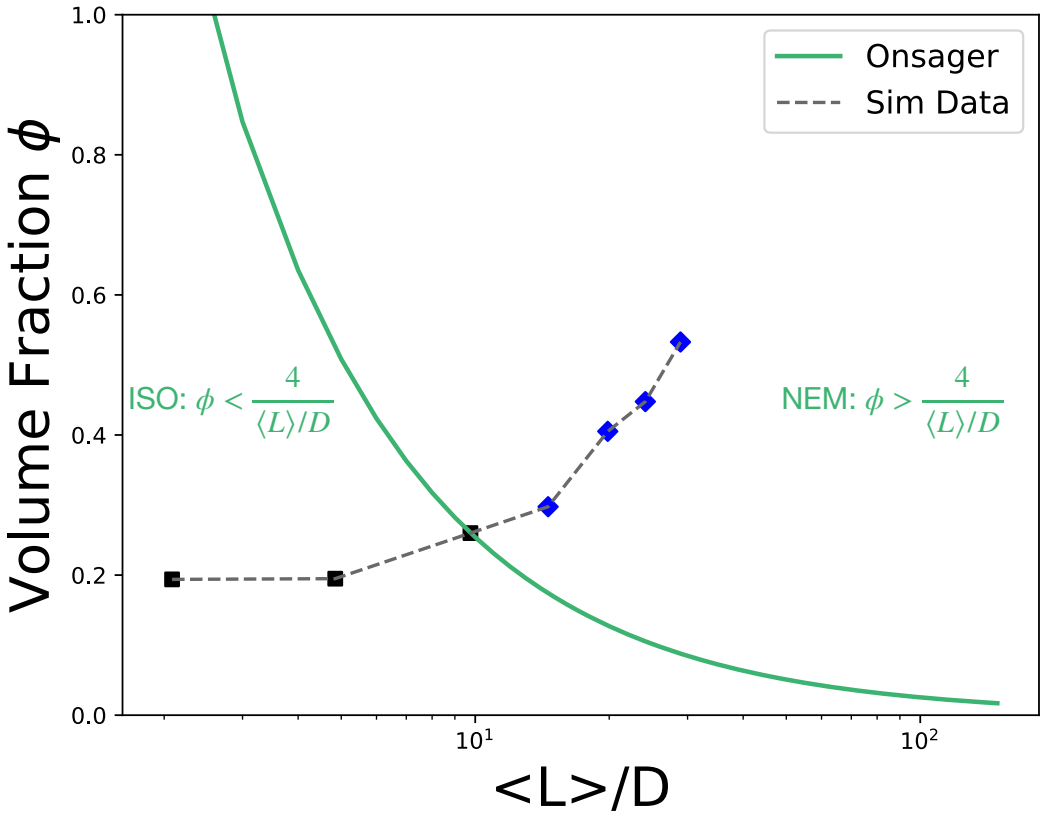

Figure S4: Volume fraction (ϕ) as a function of $\langle L \rangle / D$, where in our simulations $\langle L \rangle$ is the average length of the double helix aggregates and $D = 20$ Å is the DNA diameter. Points correspond to simulation results (black squares: isotropic phase, blue diamonds: LC phases). The nematic-isotropic phase boundary predicted by Onsager hard-rods theory [4] is shown as a green solid line. The average length is obtained as $\langle L \rangle = (\langle S \rangle + 1)/2 \cdot L_{\mathrm{GCCG}}$, where $\langle S \rangle$ is the average size of double stranded aggregates in terms of tetramers (Figure 5 in the main text), and $L_{\mathrm{GCCG}} = 16$ Å is the length of a single tetramer GCCG as taken from third peak of GG/CC RDF plots.

5 Mean squared displacement

We computed the Mean Squared Displacement (MSD) in the LC phases along the direction perpendicular and parallel to the phase director **n**, as a function of time,

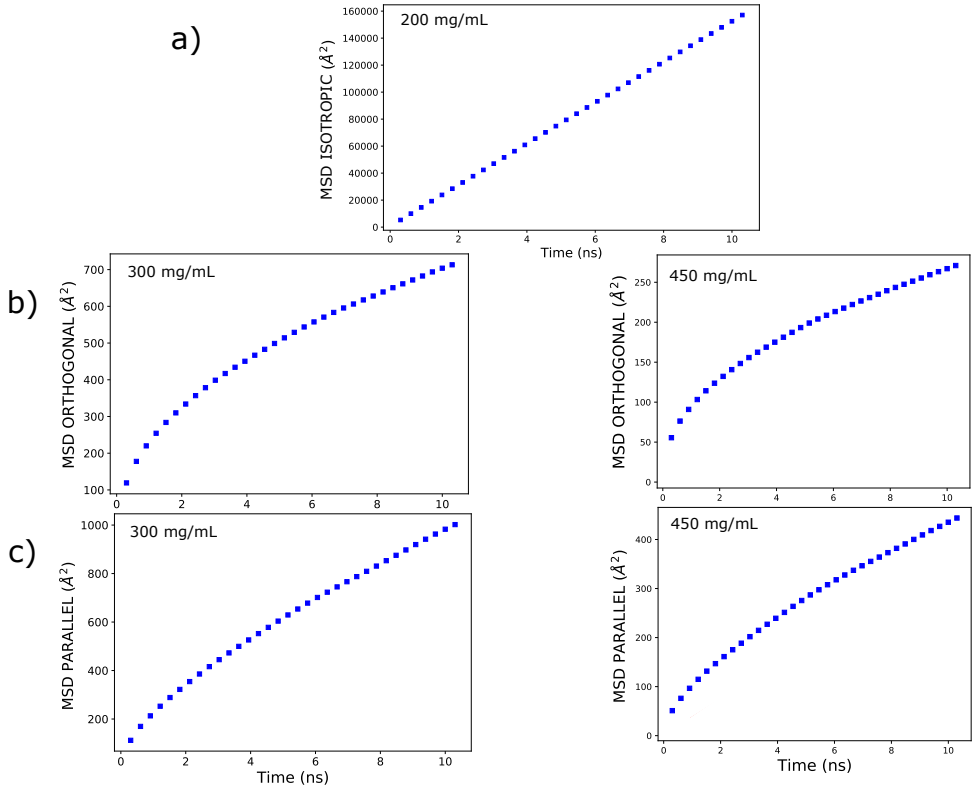

Figure S5: Mean squared displacement computed along 10 ns of MD equilibrated trajectory at T = 5 °C for a sample of N = 512 tetramers: (a) MSD_{ISO} at = 200 mg/mL (isotropic phase). (b) and (c) $MSD_{\perp}$ and $MSD_{\parallel}$ at C_{GCCG} =300 mg/mL (nematic) and C_{GCCG}= 450 mg/mL (columnar) phase.

while in the isotropic liquid only the isotropic MSD can be defined. Differently from what expected in a classical Brownian motion behaviour, LC phases (C_{GCCG}= 300 mg/mL and 450 mg/mL of Figure S5(b) and S5(c)) show a significant deviation from linearity within the investigated time interval. This trend is most likely due to the enhancement of polydispersity in the aggregates size at increasing DNA concentration. Moreover at high concentrations the simulated sample consists of just a few large aggregates that extend across the whole box. Conversely, when DNA concentration is low and the sample is isotropic, no deviation from linearity

is observed (see Figure S5).

References

[1] J. Santalucia and D. Hicks, "The thermodynamics of dna structural motifs," *Annual review of biophysics and biomolecular structure*, vol. 33, pp. 415–40, 02 2004.

[2] P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. Doye, and A. A. Louis, "Sequence-dependent thermodynamics of a coarse-grained DNA model," *J. Chem. Phys.*, vol. 137, no. 13, p. 085101, 2012.

[3] G. Zanchetta, "Spontaneous self-assembly of nucleic acids: Liquid crystal condensation of complementary sequences in mixtures of DNA and RNA oligomers," *Liquid Crystals Today*, vol. 18, no. 2, pp. 40–49, 2009.

[4] L. Onsager, "The Effects of Shape on the Interaction of Colloidal Particles," *Ann. N. Y. Acad. Sci.*, vol. 51, no. 4, pp. 627–659, 1949.

[5] X. Lü and J. T. Kindt, "Monte Carlo simulation of the self-assembly and phase behavior of semiflexible equilibrium polymers," *J. Chem. Phys.*, vol. 120, no. 21, pp. 10328–10338, 2004.

[6] E. S. Pyanzina, S. S. Kantorovich, and C. De Michele, "Nematic phase characterisation of the self-assembling sphere-cylinders based on the theoretically calculated RDFs," *Eur. Phys. J. E*, vol. 38, no. 7, pp. 2–7, 2015.

[7] R. Hentschke and J. Herzfeld, "Isotropic, nematic, and columnar ordering in systems of persistent flexible hard rods," *Phys. Rev. A*, vol. 44, no. 2, pp. 1148–1155, 1991.

[8] A. Speranza and P. Sollich, "Simplified Onsager theory for isotropic–nematic phase equilibria of length polydisperse hard rods," *J. Chem. Phys.*, vol. 117, pp. 5421–5436, 08 2002.

[9] Y. Olivier, L. Muccioli, and C. Zannoni, "Quinquephenyl: the simplest rigid-rod-like nematic liquid crystal, or is it? An atomistic simulation," *Chemphyschem.*, vol. 15, no. 19, pp. 1345–55, 2014.