Biophysical Journal, Volume 97

Supporting Material

DNA deformations near charged surfaces: electron and atomic force microscopy views

D.I. Cherny, F.G.A. Faas, B. Rieger, L.J. van Vliet

Online supporting material for:
DNA deformations near charged surfaces: electron and atomic force microscopy views
F.G.A. Faas, B. Rieger, L.J. van Vliet and D.I. Cherny

Sample preparation

Supercoiled pPGM1 plasmid DNA (2981 bp) was purified using Qiagen kits (Hilden, Germany) as described in (36). Linear DNA fragments were obtained after digestion with either EcoRI (2981 bp fragment), EcoRI/ ScaI (1444 bp and 1837 bp fragment) or PvuII (474 bp and 2507 bp fragments) and purified using gel filtration chromatography on a Superose 6 column (SMART system, Amersham Biosciences). A 474 bp fragment was obtained by additional purification through an anion exchange Waters Gen-Pak FAX column. All samples were stored in a buffer of 10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 0.1 mM Na_3EDTA at a concentration of 200 - 400 μg/ml.

Electron microscopy

Two different procedures were used for the deposition of DNA molecules: adsorption to glow-discharged carbon film and to polylysine film (36). In the first procedure, a stock solution of DNA was 100-200 fold diluted either in buffer (A) containing 1 mM Tris-HCl, pH 7.5, 0-100 mM NaCl or in buffer (B) containing 5 mM Hepes, pH 7.5, 5 mM $MgAc_2$, 10 mM KCl. The final DNA concentration was 0.2-1 μg/ml. A drop of this solution (6-8 μl) was placed onto the surface of carbon film mounted on an EM grid. Carbon films, 3-4 nm thick, were glow-discharged in the presence of pentylamine vapor (residual pressure 150 mTorr, discharge current 2-3 mA, duration of discharge 30 seconds) as described elsewhere (36). The adsorption was continued for one to two minutes, then the grids were rinsed with a few drops of 2% (w/v) aqueous uranyl acetate, blotted with filter paper and air-dried.

In the second procedure, adsorption to polylysine film was carried out as described in (36). In summary, carbon-coated EM grids were glow discharged in air (residual pressure (200 mTorr, discharge current 8-9 mA, duration of discharge 30 seconds, Bal-Tec MED 020 coater) and immediately coated with poly-L-lysine (molecular mass 2000, Sigma) by adding 8 μl of its aqueous solution at a concentration 3 μg/ml for one minute. The grids were then drained with a long tip connected to a vacuum-connected aspirator and air dried. A drop of DNA solution (6-8 μl) buffer (A) was

placed onto the polylysine film and DNA was allowed to adsorb for 1.5-2 minutes. The grid was rinsed with a few drops of 2% (w/v) aqueous uranyl acetate, blotted with filter paper and air-dried.

The samples were analyzed with a Philips CM12 electron microscope in a dark-field mode at a magnification of 28,000-35,000. The negatives were scanned with a DuoScan T2500 scanner (Agfa, Germany) at 600 - 1200 dpi. The sampling density of the digital images was 1.21 nm/pixel or 1.51 nm/px (See Fig. S.1 and 1). The absolute value for DNA rise was found to be 0.32 - 0.33 nm. In Fig. S.1 we show a panel of images of various DNA fragments mounted under different conditions. For printing, images were flattened using a high-pass filter with a radius of 250 pixels and subsequently adjusted for contrast/brightness with inverted grayscale using Adobe Photoshop^{TM}.

Atomic force microscopy

For AFM imaging samples were prepared as described (36). Briefly, a drop of DNA solution (10 μl) at a concentration of 2 μg/ml from buffer (B) was placed onto the surface of freshly cleaved mica (Muscovite, Plano GmbH, Germany) for two minutes, then rinsed with 1 ml of water, blotted with filter paper; the remaining water was blown away by the flow of compressed air. Samples were scanned using a Digital Instruments MultiMode scanning probe microscope Nanoscope IIIa (Veeco) operating in Tapping Mode. Scan rates varied from 3 to 5 Hz. Commercial silicon probes TESP-100 (Veeco) with a typical resonant frequency of around 300 kHz or ultrasharp NSC15 cantilevers (Mikromash, USA) were used throughout the experiments. The sampling density of the digital images was 7.81 nm/pixel (See Fig. S.1 and 1). The absolute value for DNA rise was 0.31 nm in accordance with previously published data (11).

Image analysis

The image analysis software was custom developed using in the Matlab toolbox DIPimage (40). The software is provided as Matlab (The Mathworks, USA) scripts which can be freely downloaded from
`www.diplib.org/home22266`. It provides the possibility to analyze images obtained from either EM or AFM imaging given sufficient image quality. A brief user manual and a few test images together with presegmented DNA strands are provided to make the software easily accessible to the field of experimental polymer studies on flat support.

Although the detection and tracing of the molecules was done fully au-

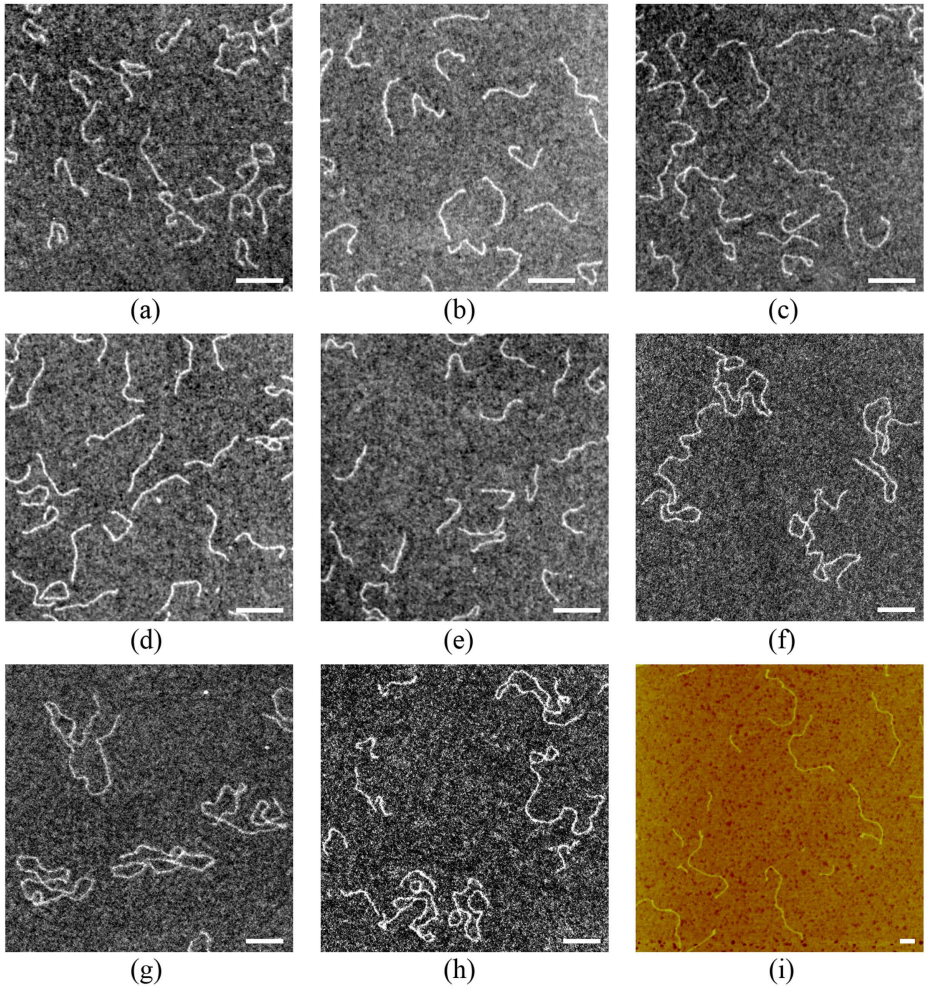


Figure S.1: Examples of EM (a-h) and AFM (i) images of DNA fragments deposited under various conditions and used for measurements. a) 474 bp fragment, 2.5 mM NaCl, carbon film, b) 474 bp fragment, 10 mM NaCl, carbon film, c) 474 bp fragment, 50 mM NaCl, carbon film, d) 474 bp fragment, 100 mM NaCl, carbon film, e) 474 bp fragment, 2.5 mM NaCl, polylysine, f) 2981 bp fragment, 50 mM NaCl, carbon, g) 2981 bp fragment, 5 mM MgAc2, 10 mM KCl, carbon, h) mixture of 474 bp and 2507 bp fragments, 5 mM MgAc2, 10 mM KCl, carbon, i) mixture of 474 bp and 2507 bp fragments, 5 mM MgAc2, 10 mM KCl, mica. Scale bars, 100 nm.

tomated, a human supervisor could reject erroneously segmented or traced molecules in an interactive step. After automated detection of DNA strands, a path through its centerline is extracted by an improved Fast Marching algorithm (28) resulting in x, y coordinate pairs tracing the molecule. They are resampled such that consecutive pairs are separated by 1 nm along the centerline of the molecule. The actual processing is divided into i) background subtraction, ii) coarse segmentation of the DNA molecules, iii) end point refinement and iv) tracing the centerline of the molecules.

First, a gradually changing background is removed by the subtraction of a low pass filtered version of the original image with a large filter kernel (Gaussian filter with $\sigma = 25$ nm). This is equivalent to high-pass filtering and results in a background corrected image I. See Fig. 1a) for an example before pre-processing.

The segmentation of the DNA strands is difficult due to the grainy structure caused by uranyl staining in combination with the high-resolution of the EM. As a result, the contrast along the molecules is far from constant and resembles a string of beads. To reduce the intensity variation along the contour the images are smoothed in by coherency enhancing diffusion step (41), resulting in I_{CED}. A single threshold is still not sufficient, therefore we use an anchor skeleton (42) to segment the molecules, i.e. a thinning operation in which the "anchors" are not allowed to be removed. The anchors are obtained by a relative low threshold, i.e. $I_{anchor} = I_{CED} < t_{low}$ and the image to be thinned by $I_{high} = I_{CED} < t_{high}$ with $t_{low} < t_{high}$. From this skeleton all branches, bifurcations and loops are removed such that the end points remain. In the following we obtain a better estimate of the end-points of the strand. First a region growing algorithm is applied to the pruned skeleton, i.e. the skeleton is allowed to grow into regions for which $I < t_{edge}$, where t_{edge} is halfway between fore- and background value of the flattened image I. From this generated mask we determine the center point by a skeleton operation which removes loose ends. The point in the mask with the largest distance, measured through the mask, from the mask center is the first end point E_1. The second end point E_2 is given by the point in the mask with the largest distance to E_1. Now a new anchor skeleton is made of the mask with E_1 and E_2 as the only anchors and the outcome is pruned. This skeleton is dilated to serve as the final mask for the tracking algorithm. In cases where the dilation would merge two regions or parts of one strand it is terminated at those specific positions.

The centerline of the DNA molecules is found using an improved Fast Marching algorithm (28). The centerline is the minimum-cost path between the two end points E_1 and E_2. The minimum arrival time T along all possible

paths P connecting the two points within a single mask is given by

$$T = \min_{\forall P_{E_1 \to E_2}} \int I_{cost}(P(s))\, ds \tag{S.1}$$

where I_{cost} is the cost map which is given by the inverse of the local travel speed. Here the cost map is derived directly from the flattened images by $I_{cost} = (\frac{I - min(I)}{max(I) - min(I)})^{\alpha} + 1$, with $\alpha = 3$. The minimum cost path from E_1 to E_2 is found by solving the Eikonal equation

$$|\nabla U| = I_{cost} \tag{S.2}$$

with U the arrival time map and initial condition $U(E_1) = 0$. We use a fast marching algorithm (43) for solving Eq.(S.2). From the arrival time map the estimated centerline can be extracted by descending along the opposite gradient direction starting at point E_2. Due to the smoothness of the integrated cost images one can obtain sub-pixel accuracy in the location of the minimum-cost path. At this point we have a sequence of points P (at spacing 1/3 of a pixel) tracing the DNA centerline from E_1 to E_2. Normal fast marching algorithms will always result in a path that is shorter and stiffer than the centerline of the underlying one-dimensional structure unless the structure is straight. To eliminate this problem the iterative scheme as presented in (28) is used. This scheme basically deforms the image space such that in the deformed space the underlying structure becomes straight (see Fig. 1c where red indicates the initial path and blue the path after 25 iterations upon convergence).
Finally, the endpoints are refined from the anchor skeleton with subpixel accuracy by extending the strand along the tangent of the path and finding the closest point with the value t_{edge}. Fitting a b-spline through this new path gives a smooth piecewise polynomial description of the DNA strand which is used in the actual data analysis. Note that the end parts, i.e. the first/last 3 pixels, of the estimated centerline are only used for the determination of the length of the DNA strands, for all other measurements the data from these parts is ignored.

Image analysis validation

The image analysis algorithm was validated on images generated by Monte Carlo simulations of the 2D WLC, i.e a homopolymer model with angle distribution according to Eq.(S.5). The evaluation was performed on noise free images and on images with added Gaussian and (correlated) Poisson

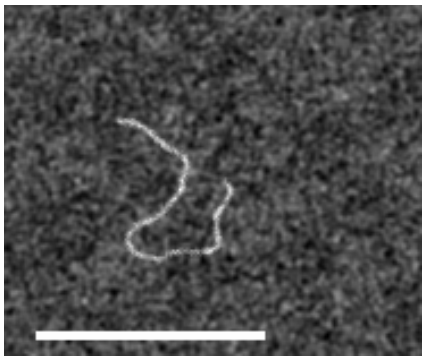


Figure S.2: Monte-Carlo simulation of a 474 bp long DNA homopolymer with $P = 55$ bp (18.7 nm) and added Gaussian and Poisson noise to match visually the experimental data. Scale bar, 100 nm.

noise to match visually the experimental data (see Fig. S.2). A DNA strand was generated by adding Gaussian blobs of $\sigma = 1$ nm at the simulated locations to a test image at spacings of 1 bp (image sampling pitch 1 pixel = 1 nm). In Fig. S.3 we depict the result of an analysis applied to 500 simulated images containing DNA molecules of 474 bp length with and without noise (only the non-self intersection molecules were processed). The algorithm permitted estimation of the underlying persistence length, the kurtosis and the bend angle histogram. We validated the image processing method for polymers with persistence lengths ranging from 18.7 nm to 52.7 nm (55 bp to 159 bp). The bias in the found persistence length values was about $\sim 5\%$, if the fit was performed in the range $L \in [0, 2P]$. The bias was always positive, i.e. the persistence length was too high. The kurtosis was found to be constant with value 3. The bending potential $G(\theta, L)$ was verified for $L = 4, 6$ and 8 nm and also here the fitted persistence length was within 5% of the ground truth. Generally, the method performed better on molecules with larger persistence length (stiffer), e.g. for $P = 52.7$ nm the error was <2%. Given the spread in the estimate of the persistence length from the true simulated coordinates of $\sim 2\%$ for 500 molecules and $\sim 4\%$ if only the non-self-intersecting molecules are considered, we conclude that the image processing retrieves the correct coordinates of the DNA centerline. In Fig. S.4 we show the extracted centerlines for simulated images. The red line connects simulated DNA positions, the blue line is the initial centerline found by the fast marching and the green line shows the centerline by the iterative procedure (28). The latter resembles the ground truth (red) very

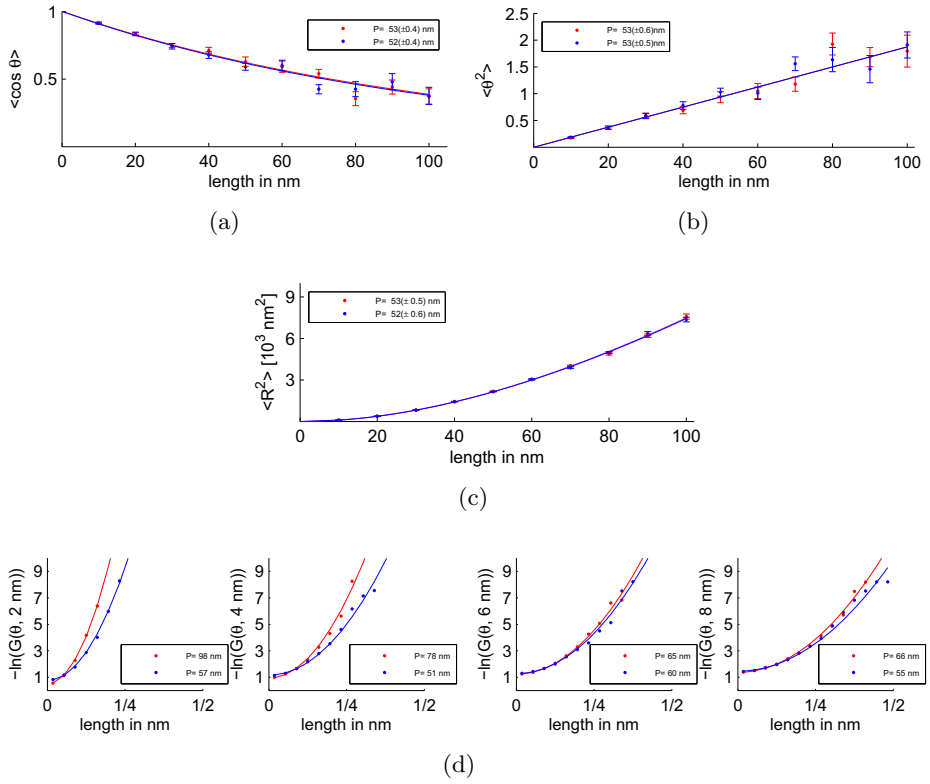
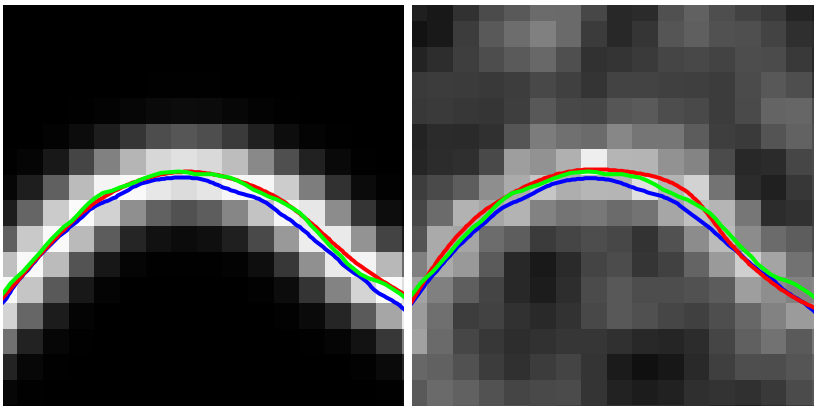


Figure S.3: Result for 500 Monte-Carlo simulated images of homopolymer 474 bp DNA with $P = 159$ bp (52.7 nm) with (red) and without noise (blue). Only non self-intersecting molecules are processed; a-c) show estimates of the persistence length based on $\langle \cos\theta(L) \rangle$ Eq. 5 and $\left\langle \theta^2(L) \right\rangle$ Eq. 6 and $\left\langle R^2(L) \right\rangle$ Eq. 4 respectively; d) shows the negative logarithm $-\ln G$ Eq. S.5 of the occurrence of the deflection angle θ for $L = 2, 4, 6$ and 8 nm.



(a) (b)

Figure S.4: Part of an image generated with Monte-Carlo simulation of 474 bp DNA with $P = 55$ bp (18.7 nm). The red line connects the simulated coordinates spaced at one bp of the DNA strand. In blue is depicted the initial found centerline and in green the finally used centerline; a) for the noise free case, b) for added Gaussian and Poisson noise to visually match the experimental data. Scale: 1 pixel corresponds to 1 nm.

well in the noise free case and even for the noisy case the correspondence is good.

We observed that the apparent length distribution of the imaged 474 bp DNA as shown in Fig. 2d) is narrow. Traditionally, the length of all molecules found is set to the a-priori know length of the fragments for the analysis (11, 12). However, the narrow distribution stimulated us to calibrate the magnification once for all images and keep the apparent length differences for the imaged ensemble of polymer. A comparison of both methods showed excellent agreement (found persistence length within 0.2 nm which equals the fitting error). Therefore we used a magnification calibration for all analysis. We emphasize that once the procedure is established, we can use segments of the molecules for the measurements instead of full-length. This greatly enhances the total scoreable length of DNA molecules in each data sets as overlapping molecules can be analyzed in parts.

Data analysis

Given a set of coordinates that trace the DNA backbone we compute the statistical quantities such as $\langle R^2 \rangle$ as a function of the contour length l. Extra

precaution was taken in sampling the strands to avoid correlation between points in the graph. Reusing all data for computing the quantities for each length along the DNA yields highly correlated points, hence a smooth curve. To avoid this, we divided each DNA strand into length segments randomly drawn from a predefined set of lengths, ℓ_i, such that no piece of strand is used twice and the whole strand is used. The set is given by $\ell_i = \frac{\ell_{max}}{n}\, i$ with $0 < i \leq n$, ℓ_{max} the maximum segment length and n the number of different segments lengths; here $n = 20$ and $\ell_{max} \approx 120$ nm. This yields a logarithmic length distribution of the segment lengths. In Fig. 2d) we show the number of segments for the different lengths. The same procedure is applied to compute the kurtosis (Figs. 2 and 4) and the bending angle histograms (Fig. 5). For the latter we used $n = 3$ and $\ell_{max} = L$. Furthermore, we investigated the influence of pre-selecting non self-intersecting molecules only for the analysis. For a persistence length of $P = 18.7$ nm less than 4% of 474 bp long molecules are expected to be self intersecting (for $P = 52.7$ nm less than 1%). That is why we restrict the fitting to $L \in [0, 2P]$. Without this restriction, i.e. fitting $\langle \theta^2 \rangle$ and $< \cos\theta >$ over the entire range, would yield a slight overestimation of P. The error bars given in the Figs.2 and 5 show one standard deviation which is computed from 20 different random draws ℓ_i from the same set of molecules. Thus the error bars are an indication for the statistical reliability of that measurement point. For the kurtosis and $\langle D^4 \rangle$ we applied error propagation to determine the standard deviation. For the bending angle distribution $-\ln G(\theta, l)$ we include the cumulative probability by evaluation of the error function for the bins in the deflection angle. The offset of the distribution is not neglected, see Eq. 12. Here a remark has to be made regarding the estimation of the bending angle. The DNA molecules are not imaged infinitely sharp but blurred due to the inherent width of DNA of ~ 2 nm and the point spread function of the imaging system. This effective blurring introduces correlation in the estimated tangents and thus angles over short separation ($L \approx \sigma$). From the evaluation on test images we found that this correlation introduces an apparent shorting of the strand when evaluating $\theta(L)$. The amount of apparent shorting was found to be equal to the standard deviation of the Gaussian spot placed at the simulated coordinate. As a consequence, a correction transformation of $L \rightarrow L + \sigma$ in the evaluation of $\theta(L)$ was applied. Obviously, this effect is strongest for small $L \approx \sigma$ and negligible for $L \gg \sigma$. From EM data we estimated the effective blurring to correspond to a Gaussian of $\sigma = 1$ nm. The evaluation of $R(L)$ does not suffer from this as the center point is not influenced by blurring.

The experimentally found kurtosis as a function of contour length (see

Fig. 2) was compared to Monte Carlo simulations with static bending θ_0 on a comparable number of molecules as presented in the experiments (see Fig. 4). Also for this simulations a non-overlapping logarithmic sampling of the polymers segments is essential to judge the statistical spread in the result. Non-uniform probability, i.e. $p \neq 0.5$, for left and right static bending was inspected in the same way, as well as an alternative hypothesis $p = 0, \theta = 0$.

Derivation of the expression for the forth order moments of the end-to-end distance in 2D

The energy required to bend a polymer molecule in two dimensions, i.e. a flexible rod, is given by (2, §17-18):

$$E = \frac{\Upsilon\, I\, \theta^2}{2L} \tag{S.3}$$

where θ is the angle between the tangent vectors to the rod separated by a distance l along the polymer. The macroscopic quantities Young's modulus Υ and area moment of inertia I of the molecule are related to the persistence length $P \equiv P_{3D}$ (2) by $\Upsilon\, I = k_B T P$ where k_B is the Boltzmann constant and T is the absolute temperature. The Boltzmann distribution of states is given by:

$$\mathcal{P}(E) = \frac{g(E)\, e^{-\frac{E}{k_B T}}}{\int_0^\infty g(E')\, e^{-\frac{E'}{k_B T}} dE'}\,. \tag{S.4}$$

In our case the density of states $g(E)$ is constant and as such can be dropped from the equation. Combining the above equations gives the normalized probability distribution function for the bend angle θ:

$$G(\theta, L) = \frac{e^{-\frac{\theta^2 P}{2L}}}{\int_{-\infty}^{\infty} e^{-\frac{s^2 P}{2L}} ds} = \sqrt{\frac{P}{2\pi\, L}} e^{-\frac{P\,\theta^2}{2\,L}}\,. \tag{S.5}$$

We must apply binning to calculate the histogram $-\ln G(\theta, L)$ from the measurements. For a bin size of $\Delta\theta$ around a bin center θ we obtain

$$G(\theta, L) = \operatorname{erf}\left(\sqrt{\frac{P}{2L}}\left(\theta + \frac{\Delta\theta}{2}\right)\right) - \operatorname{erf}\left(\sqrt{\frac{P}{2L}}\left(\theta - \frac{\Delta\theta}{2}\right)\right)\,. \tag{S.6}$$

As we investigate the magnitude of θ a factor of 2 enters the above equation. Neglecting the binning effect simplifies the formula to $-\ln G(\theta, L) \propto \frac{P}{L}\theta^2$

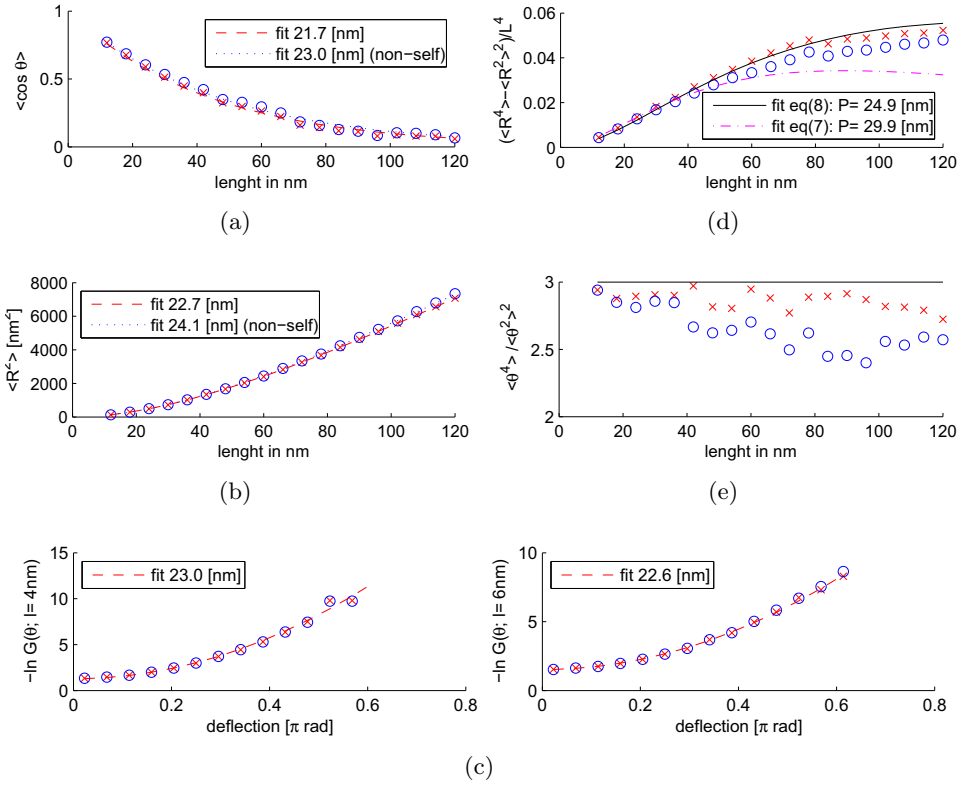


Figure S.5: Example of the data analysis of a Monte-Carlo simulation of 500 heteropolymer DNA molecules (474 bp long) with $P = 150$ bp (51.0 nm), $\theta_{stat} = 5.3°$ and $p = 0.5$. Data are shown for i) all simulated strands (×) and ii) only the non self-intersecting strands (○). a)-b) show estimates of the persistence length based on $\langle \cos\theta(L) \rangle$ Eq. 2 and $\langle R^2(L) \rangle$ Eq. 1 respectively; c) shows the negative logarithm $-\ln G$ Eq. S.5 of the occurrence of the deflection angle θ for $L = 4$ and $L = 6$ nm; d) shows the normalized difference $(\langle R^4 \rangle - \langle R^2 \rangle^2)/L^4$ Eq. 9 with fits for the 2D Eq. 8 and 3D Eq. 7 case; e) shows the kurtosis $\langle \theta^4 \rangle / \langle \theta^2 \rangle^2$ (which is equal to 3 for θ Gaussian distributed around zero).

which is most often seen in literature. However, such a plot cannot be used to judge the quality of the persistence length fit.

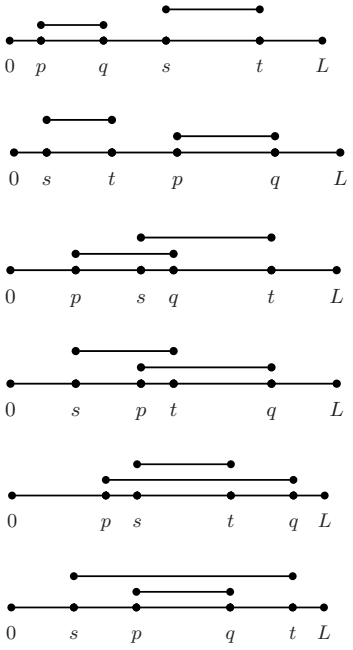
The Euclidean distance between points s and t on the curve separated by a distance L is given by $\vec{R} = \int_0^L \vec{u}(s)\, ds$ with $\vec{u}(s) = [\cos(\theta_s), \sin(\theta_s)]^T$ the tangent vector to the curve at contour position s. Now the average second order moment $\langle R^2 \rangle$ is given by:

$$\begin{aligned}
\langle R^2 \rangle &= \int_0^L \int_0^L \langle \vec{u}(s) \cdot \vec{u}(t) \rangle \, dt \, ds \\
&= \int_0^L \int_0^L \langle \cos\theta_s \cos\theta_t + \sin\theta_s \sin\theta_t \rangle \, dt \, ds \\
&= 2 \int_0^L \int_0^t \langle \cos\left(\theta_s - \theta_t\right) \rangle \, ds \, dt = 2 \int_0^L \int_0^t e^{-\frac{t-s}{2P}} \, ds \, dt \\
&= 4\,PL - 8\,P^2 + 8\,P^2 e^{-\frac{L}{2P}} \, .
\end{aligned} \tag{S.7}$$

Note that the argument of the average cosine is always positive, as the length is always positive. With the average operator defined as $\langle \cdot \rangle = \int_{-\infty}^{\infty} \cdot\, G(\theta, L) d\theta$ the average fourth order moment $\langle R^4 \rangle$ can be derived in

a similar fashion:

$$
\begin{aligned}
\langle R^4\rangle =& \langle \int_0^L \vec{u}(s)\,ds \cdot \int_0^L \vec{u}(t)\,dt \int_0^L \vec{u}(p)\,dp \cdot \int_0^L \vec{u}(q)\,dq \rangle \\
=& \int_0^L \int_0^L \int_0^L \int_0^L \langle \vec{u}(s)\cdot\vec{u}(t)\,\vec{u}(p)\cdot\vec{u}(q)\rangle\,ds\,dt\,dp\,dq \\
=& \int_0^L \int_0^L \int_0^L \int_0^L \langle \cos(\theta_s-\theta_t)\cos(\theta_p-\theta_q)\rangle\,ds\,dt\,dp\,dq \\
=& 4\Big(\int_0^L \int_0^t \int_0^s \int_0^q \langle\cdot\rangle_{A1}\,dp\,dq\,ds\,dt + \\
& \int_0^L \int_s^L \int_t^L \int_p^L \langle\cdot\rangle_{A2}\,dq\,dp\,dt\,ds + \\
& \int_0^L \int_0^t \int_s^t \int_0^s \langle\cdot\rangle_{B1}\,dp\,dq\,ds\,dt + \\
& \int_0^L \int_0^t \int_s^t \int_t^L \langle\cdot\rangle_{B2}\,dq\,dp\,ds\,dt + \\
& \int_0^L \int_0^t \int_t^L \int_0^s \langle\cdot\rangle_{C1}\,dp\,dq\,ds\,dt + \\
& \int_0^L \int_0^t \int_s^t \int_s^q \langle\cdot\rangle_{C2}\,dp\,dq\,ds\,dt \Big) \\
=& 8\Big(\int_0^L \int_0^t \int_0^s \int_0^q \langle\cdot\rangle_{A1}\,dp\,dq\,ds\,dt + \\
& \int_0^L \int_0^t \int_s^t \int_0^s \langle\cdot\rangle_{B1}\,dp\,dq\,ds\,dt + \\
& \int_0^L \int_0^t \int_t^L \int_0^s \langle\cdot\rangle_{C1}\,dp\,dq\,ds\,dt \Big)
\end{aligned}
$$



(S.8)

To handle the correlation between the different integration variables, the integration is separated in to six parts (see above). In case A we have non overlapping segments pq and st whereas in respectively case B these

segments are partially overlapping.

$$\begin{aligned}\langle\cdot\rangle_{A1} &= \langle\cos(\theta_s-\theta_t)\cos(\theta_p-\theta_q)\rangle \\ &= \langle\cos(\theta_s-\theta_t)\rangle\langle\cos(\theta_p-\theta_q)\rangle \\ &= e^{-\frac{t-s}{2P}}e^{-\frac{q-p}{2P}}\end{aligned} \tag{S.9}$$

$$\begin{aligned}\langle\cdot\rangle_{B1} &= \langle\cos(\theta_s-\theta_t)\cos(\theta_p-\theta_q)\rangle \\ &= \langle\cos(\theta_s-\theta_p)\rangle\langle\cos(\theta_q-\theta_s)\cos(\theta_q-\theta_s)\rangle\langle\cos(\theta_t-\theta_q)\rangle \\ &= \langle\cos(\theta_s-\theta_p)\rangle\langle\cos^2(\theta_q-\theta_s)\rangle\langle\cos(\theta_t-\theta_q)\rangle \\ &= e^{-\frac{s-p}{2P}}e^{-\frac{t-q}{2P}}\frac{1}{2}\left(1+e^{-\frac{2(q-s)}{P}}\right)\end{aligned} \tag{S.10}$$

$$\begin{aligned}\langle\cdot\rangle_{C1} &= \langle\cos(\theta_s-\theta_t)\cos(\theta_p-\theta_q)\rangle \\ &= \langle\cos(\theta_s-\theta_p)\cos(\theta_t-\theta_s)\cos(\theta_t-\theta_s)\rangle\langle\cos(\theta_q-\theta_t)\rangle \\ &= \langle\cos(\theta_s-\theta_p)\rangle\langle\cos^2(\theta_t-\theta_s)\rangle\langle\cos(\theta_q-\theta_t)\rangle \\ &= e^{-\frac{s-p}{P}}e^{-\frac{q-t}{P}}\frac{1}{2}\left(1+e^{-\frac{2(t-s)}{P}}\right)\end{aligned} \tag{S.11}$$

Combining these equations we obtain

$$\langle R^4\rangle = 32\,L^2P^2 - 240\,LP^3 + 696\,P^4 - \frac{320}{3}LP^3e^{-\frac{L}{2P}} - \frac{6272}{9}P^4e^{-\frac{L}{2P}} + \frac{8}{9}P^4e^{-\frac{2L}{P}}\,. \tag{S.12}$$

Derivation of the kurtosis for the heteropolymer model

For the purposes of analysis we make the following assumptions. First, we used the Schellman (3) approach for the description of a DNA chain. Second, Schellman also suggested that a bending potential can still be quadratic for the non-zero value of static bend. This idea was later used by Cognet, leading to the following formula of the bending potential

$$p(\theta_i) = N^{-1/2}e^{-\frac{g_i''\theta_i^2}{2RT}} = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left(\theta_i-\theta_{i,stat}\right)^2}{2\sigma^2}}\,. \tag{S.13}$$

Third, Schellman showed that a good approximation to the shape of real polymers can be achieved by using the following formula for the persistence length $P \sim 1/(1-\langle\cos\theta_{i,dyn}\rangle\langle\cos\theta_{i,stat}\rangle)$ (5), where $\langle\cos\theta_{i,dyn}\rangle$ accounts for the dynamic behavior of polymer chains and $\langle\cos\theta_{i,stat}\rangle$ for local static

bending. The latter means averaging of all static bending angles, i.e. along DNA position. In another words, it requires double averaging. First, over an ensemble of molecules and second along the length of the molecules. In view of small values for both $\langle\theta_{i,dyn}\rangle$ and $\langle\theta_{i,stat}\rangle$ (usually less than about 5 degrees) we use the following formula (13)

$$P \approx 2/\left\langle\theta^2\right\rangle\,, \tag{S.14}$$

where $\left\langle\theta^2\right\rangle = \left\langle\theta^2_{dyn}\right\rangle + \left\langle\theta^2_{stat}\right\rangle$ (5). Next, we have to keep in mind that according to the measurement procedure all values for $\langle\cos\theta\rangle$ etc. are obtained using double averaging, i.e. configurational and positional. To start the analysis let us consider a segment of polymer chain with fixed first and last points (fixed length along the DNA). This is configurational averaging similar to (10) with the difference that static bending angles in plane are included. In the following we use short $\theta^0 \equiv \theta_{stat}$ and $\sigma \equiv \theta_{dyn}$. The angel θ_n over a length along the DNA n is

$$\theta_n = \sum_1^n \theta_i = \sum_1^n (\theta_i - \theta_i^0 + \theta_i^0) = \sum_1^n (\theta_i - \theta_i^0) + \sum_1^n \theta_i^0 = \sum_1^n \Delta_i + \sum_1^n \theta_i^0\,, \tag{S.15}$$

with $\Delta_i = \theta_i - \theta_i^0$ and since $\left(\sum_1^n \Delta_i\right)^2 = \sum_1^n \Delta_i^2 + \sum_{i\neq j} \Delta_i\Delta_j$. Using the harmonic potential of $(\theta_i - \theta_i^0)$ as in eq.(S.13) we have $\left\langle\Delta_i^2\right\rangle = \sigma^2$ and $\left\langle\Delta_i^4\right\rangle = 3\sigma^4$. From that follows

$$\left\langle\left(\sum_1^n \Delta_i\right)^2\right\rangle = \sum_1^n \left\langle\Delta_i^2\right\rangle + \sum_{i\neq j} \left\langle\Delta_i\right\rangle\left\langle\Delta_j\right\rangle = n\sigma^2\,, \tag{S.16}$$

$$\left\langle\left(\sum_1^n \Delta_i\right)^4\right\rangle = \sum_1^n \left\langle\Delta_i^4\right\rangle + 3\sum_{i\neq j} \left\langle\Delta_i^2\right\rangle\left\langle\Delta_j^2\right\rangle + 4\underbrace{\sum_{i\neq j} \left\langle\Delta_i\right\rangle\left\langle\Delta_j^3\right\rangle}_{=0} \tag{S.17}$$

$$= 3n\sigma^4 + 6\frac{n(n-1)}{2}\sigma^4 = 3n^2\sigma^4\,. \tag{S.18}$$

Let us introduce $A_n^0 = \sum_1^n \theta_i^0$ for the intrinsic bending over length n. Note that this is identical for all molecules for a fixed length along the DNA. Thus

we can write for the moments

$$(\theta_n)^2 = \left(\sum_1^n \Delta_i\right)^2 + \left(A_n^0\right)^2 + 2A_n^0 \sum_1^n \Delta_i \tag{S.19}$$

$$(\theta_n)^4 = \left(\sum_1^n \Delta_i + A_n^0\right)^4 \tag{S.20}$$

and finally

$$\left\langle (\theta_n)^2 \right\rangle = n\sigma^4 + \left(A_n^0\right)^2 \tag{S.21}$$

$$\left\langle (\theta_n)^4 \right\rangle = 3n^2\sigma^4 + 6\left(A_n^0\right)^2 n\sigma^2 + \left(A_n^0\right)^4 . \tag{S.22}$$

It means for a fixed length along the DNA we get for the kurtosis

$$k = 3 - \frac{2(A_n^0)^4}{(n\sigma^2 + (A_0^n)^2)^2} . \tag{S.23}$$

It is clear that the kurtosis is always smaller than 3 for any sequence exhibiting static bending with $A_n^0 \neq 0$.

For our purposes we have to make the second averaging over the position, which means that we have to average the higher moments of static angles, this leads to the following formulas

$$k = \frac{\left\langle\left\langle \theta_n^4 \right\rangle\right\rangle}{\left\langle\left\langle \theta_n^2 \right\rangle\right\rangle^2} \tag{S.24}$$

$$\left\langle\left\langle \theta_n^2 \right\rangle\right\rangle = n\sigma^4 + \left\langle (A_n^0)^2 \right\rangle \tag{S.25}$$

$$\left\langle\left\langle \theta_n^4 \right\rangle\right\rangle = 3n^2\sigma^4 + 6\left\langle (A_n^0)^2 \right\rangle n\sigma^2 + \left\langle (A_n^0)^4 \right\rangle . \tag{S.26}$$

To simplify the calculations let us make the assumption that all static bends occur in plane and behave independent of each other. Moreover, we consider that all static bends are equal in value, i.e. $|\theta_i^0| = |\theta^0|$. Let us also assume that the choice of sign is random (plus or minus direction have equal probability of $p = 0.5$), implying that the behavior of a "frozen" polymer chain (no dynamic fluctuations) resembles a one dimensional walk in terms of angles. From this it follows that

$$\left\langle (A_n^0)^2 \right\rangle = n(\theta^0)^2 \tag{S.27}$$

$$\left\langle (A_n^0)^4 \right\rangle = (\theta^0)^4(3n^2 - 2n) . \tag{S.28}$$

Inserting eqs.(S.27,S.28) into eqs.(S.25,S.26) yields

$$\begin{aligned} \langle\langle \theta_n^2 \rangle\rangle &= n\sigma^2 + n(\theta^0)^2 && \text{(S.29)} \\ \langle\langle \theta_n^4 \rangle\rangle &= 3n^2\sigma^4 + 6n(\theta^0)^2 n\sigma^2 + (\theta^0)^4(3n^2 - 2n) && \text{(S.30)} \end{aligned}$$

and finally for the kurtosis k

$$k = 3 - \frac{2}{n}\frac{(\theta_0)^4}{(\sigma^2 + (\theta^0)^2)^2}\,. \qquad \text{(S.31)}$$

Additional Supplementary References

40. Luengo Hendriks, C., B. Rieger, M. v. Ginkel, G. v. Kempen, and L. v. Vliet, 1999-. DIPimage: a scientific image processing toolbox for MATLAB. Delft University of Technology, `http://www.diplib.org`.

41. Weickert, J., B. ter Haar Romeny, and M. Viergever, 1998. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Transactions on Image Processing* 7:398–410.

42. Vliet, L. v., and B. Verwer, 1988. A contour processing method for fast binary neighbourhood operations. *Pattern Recognition Letters* 7:27–36.

43. Adalsteinsson, D., and J. A. Sethian, 1995. A Fast Level Set Method for Propagating Interfaces. *Journal of Computational Physics* 118:269–277.