Supporting Information: Dinucleotides as simple models of the base stacking-unstacking component of DNA 'breathing' mechanisms

Eric R. Beyerle,[†,1], Mohammadhasan Dinpajooh[†,1,4] Huiying Ji,[2,3] Peter H. von Hippel,[3] Andrew H. Marcus[2,3] and Marina G. Guenza[1,2,3,*]

[1.] Institute for Fundamental Science and Department of Chemistry and Biochemistry, University of Oregon, Eugene, Oregon 97403, USA

[2.] Center for Optical, Molecular and Quantum Science and Department of Chemistry and Biochemistry, University of Oregon, Eugene, Oregon 97403, USA

[3.] Institute of Molecular Biology and Department of Chemistry and Biochemistry, University of Oregon, Eugene, Oregon 97403, USA

[4.] Present address: Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

* To whom correspondence should be addressed. Tel:001-541-3462877; Fax:001-541-3464643; Email:mguenza@uoregon.edu

[†]These authors contributed equally to this work

I. CALCULATIONS OF CIRCULAR DICHROISM (CD) SPECTRA FROM MOLECULAR CONFIGURATIONS

We applied the standard methods developed by Schellman and others to model the delocalized electronic states of the dApdA dinucleotide as a function of base stacking conformations.[1,2,3] In the formalism that follows, we consider only the contribution to the CD spectrum that emerges from the exciton interactions between the component adenine bases of the dApdA dinucleotide, and we neglect the minor contribution to the CD from the non-interacting adenine monomer, which provides a relatively weak signal for the fully unstacked conformation. When light of frequency ν interacts with a solution of optically active molecular chromophores, the left and right circularly polarized components are absorbed to different extents. The frequency (or wavelength) dependence of the differential extinction between left and right circular polarizations, $\Delta\varepsilon(\nu) = \varepsilon_L(\nu) - \varepsilon_R(\nu)$, is called the CD spectrum. The CD spectrum can be understood in terms of the rotational strength R_{if} of an electronic transition from an initial state $|\Psi_i\rangle$ to a final state $|\Psi_f\rangle$, which is defined by the Rosenfeld equation

$$R_{if} = \mathrm{Im}\left[\langle\Psi_i|\hat{\boldsymbol{\mu}}|\Psi_f\rangle \cdot \langle\Psi_f|\hat{\boldsymbol{m}}|\Psi_i\rangle\right] . \tag{s1}$$

Here $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{m}}$ are the electric and magnetic dipole transition moment operators, respectively. The states $|\Psi_i\rangle$ and $|\Psi_f\rangle$ are electronic eigenstates resulting from a chiral arrangement of coupled electric dipole transition moments (or EDTMs), which are each localized to a nucleic acid base residue. Equation (s1) shows that the rotational strength depends on the chirality of the coupled EDTMs, and its sign indicates the handedness (left versus right) of the chiral arrangement.

The Hamiltonian of the coupled system is given by

$$\hat{H} = \hat{H}_1 + \hat{H}_2 + \hat{V}_{12} \tag{s2}$$

where $\hat{H}_1$ and $\hat{H}_2$ are the Hamiltonian operators of monomers 1 and 2, respectively and $\hat{V}_{12}$ is the coupling between electronic transitions localized to each monomer as defined in the Main Text. The matrix element $V_{a1b2} = \langle\psi_{a1}|\hat{V}_{12}|\psi_{b2}\rangle$ defines the coupling between monomer excited electronic states $|\psi_{a1}\rangle$ (labeled a on monomer 1) and $|\psi_{b2}\rangle$ (labeled b on monomer 2). The electronic coupling is calculated using the extended-dipole model (EDM),[4] which has been applied previously to cyanine dyes in self-assembled tubular J-aggregates,[5] to cyanine dimers in DNA,[6,7] and to canonical nucleic acid bases in short segments of DNA.[8] In our current studies, the EDM accounts for the physical length of the adenine base by including for each monomer electronic transition a one-dimensional displacement vector, $\boldsymbol{l}$, that is oriented parallel to the EDTM direction. Each transition dipole moment is represented as two-point charges of equal magnitude and opposite sign $(\pm q)$ separated by distance l. The coupling matrix element is given by

$$V_{a1b2} = \frac{|\boldsymbol{\mu}_{a1}||\boldsymbol{\mu}_{b2}|}{4\pi\epsilon\epsilon_0 l_{a1} l_{b2}} \left[\frac{1}{r_{12}^{a+b+}} - \frac{1}{r_{12}^{a-b+}} - \frac{1}{r_{12}^{a+b-}} + \frac{1}{r_{12}^{a-b-}} \right] \qquad \text{(s3)}$$

In Eq. (s3), $\boldsymbol{\mu}_{a1} = q_{a1}\boldsymbol{l}_{a1}$ and $\boldsymbol{\mu}_{b2} = q_{b2}\boldsymbol{l}_{b2}$ are the EDTMs of the transitions a and b on monomers 1 and 2, respectively, and the four distances $r_{12}^{a\pm b\pm}$ are those between the positive and negative point charges on monomers 1 and 2. The vacuum permittivity of free space is given by ϵ_0, and ϵ is the local dielectric constant. For all of our calculations we used the value of the dielectric constant, $\epsilon = 2$, in accordance with prior conventions.[9]

In principle, further improvements to the accuracy of our calculations could be achieved by using more detailed, quantum chemical calculations of the electronic transition charge densities. Nevertheless, the favorable comparison between our calculations and experimental data presented below suggests that the EDM provides a reliable estimate of the electronic couplings between adjacent bases for present purposes.

We write the Hamiltonian on a monomer-site basis, such that singly-excited state wave functions are given by tensor products according to

$$|\Phi_{a1}\rangle = |\phi_{a1}\rangle|\phi_{g2}\rangle \text{ and } |\Phi_{a2}\rangle = |\phi_{a2}\rangle|\phi_{g1}\rangle \qquad \text{(s4)}$$

In Eq. (s4), $|\phi_{a1}\rangle$ and $|\phi_{a2}\rangle$ denote the a^{th} electronic excited states of monomers 1 and 2, respectively, and $|\phi_{g1}\rangle$ and $|\phi_{g2}\rangle$ are the electronic ground states. The number of distinct electronic transitions local to monomer 1 (2) is given by $n_{1(2)}$, such that the total number of site-localized transitions is $n_{tot} = n_1 + n_2$. The Hamiltonian of Eq. (s2) may thus be written on this site basis as a $n_{tot} \times n_{tot}$ matrix with diagonal elements representing the single site excitations (with energies E_{a1} and E_{b2}) and off-diagonal elements representing the couplings V_{a1b2} between monomer sites. Note that our formalism neglects the contribution from the isolated adenine monomer, which provides a signal for the fully unstacked conformation. In our calculations, however, this contribution is much smaller than the contribution due to the degenerate coupling of the adenine transitions.

Diagonalization of the Hamiltonian provides the eigen-states $|\Psi_k\rangle$ and eigen-energies E_k of the electronically coupled dinucleotide. In the so-called 'exciton' basis, the kth singly-excited state $|\Psi_k\rangle$ may be written

$$|\Psi_k\rangle = \sum_{m=1}^{2} \sum_{a} C_{ma}^{k} |\Phi_{ma}\rangle \tag{s5}$$

where C_{ma}^{k} is the expansion coefficient corresponding to transition a local to monomer m. In the exciton basis, the ground state of the dinucleotide is given by $|\Psi_g\rangle = |\phi_{g1}\rangle|\phi_{g2}\rangle$. Using Eq. (s1), we may calculate the rotational strength R_{gk} (=R_k) for the kth electronic transition, where we assign the initial and final states to $|\Psi_g\rangle$ and $|\Psi_k\rangle$, respectively, and the total electric and magnetic dipole transition moment operators are given by vector sums $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{a1} + \hat{\boldsymbol{\mu}}_{b2}$ and $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}_{a1} + \hat{\boldsymbol{m}}_{b2}$.

For a given transition k, the rotational strength depends on the relative orientation of the monomer EDTMs. For the case of coupled degenerate transitions (i.e. $E_{a1} = E_{b2}$ and $E_k = E_{a1} + V_{a1b2}$), the rotational strength is given by

$$R_k = \frac{E_k}{4\hbar} [\boldsymbol{r}_{12} \cdot (\boldsymbol{\mu}_{b2} \times \boldsymbol{\mu}_{a1})] \tag{s6}$$

For the case of non-degenerate coupled transitions (i.e. $E_{a1} \neq E_{b2}$ and $E_k \approx E_{a1}$), the rotational strength is given by

$$R_k(E_{a1}) = -\frac{E_{a1}E_{b2}}{\hbar(E_{b2}^2 - E_{a1}^2)} [\boldsymbol{r}_{12} \cdot (\boldsymbol{\mu}_{b2} \times \boldsymbol{\mu}_{a1})] \tag{s7}$$

We note that Eq. (s7) is written such that $E_{a1} > E_{b2}$.

To calculate the CD spectrum, we consider the relationship between the rotational strength and the integrated area of the CD spectrum within a finite spectral range $\nu' \rightarrow \nu''$:

$$R = A \int_{\nu'}^{\nu''} \frac{\Delta\varepsilon(\nu)}{\nu} d\nu \tag{s8}$$

where A = 7.659 × 10^{-54} C^2 m^3 s^{-1}. The CD spectral line shape is obtained by summing over all contributions from individual transitions according to

$$\Delta\varepsilon(\nu) = \sum_{k=1}^{n_{tot}} \Delta\varepsilon(\nu_k) \tag{s9}$$

For each of the k electronic transitions, we approximate the CD spectral line shape as a Gaussian function $\Delta\varepsilon(\nu_k) = \Delta\bar{\varepsilon}_k exp\{-[(\nu_k - \bar{\nu}_k)^2/2\sigma_k^2]\}$, where σ_k is the Gaussian standard deviation, $\bar{\nu}_k$ $(= E_k/h)$ is the mean transition frequency, and $\Delta\bar{\varepsilon}_k$ is the magnitude. Upon substitution of the above Gaussian function, Eq.(s8) is approximated by considering the frequency in the denominator to be constant over the width of the k spectral line, $\nu \approx \bar{\nu}_k$, and by extending the limits of the integral, $\nu' \approx \bar{\nu}_k - \Delta\nu_k$ and $\nu'' \approx \bar{\nu}_k + \Delta\nu_k$ with $\Delta\nu_k \gg 0$. This is a standard approximation used in the calculations of the CD spectra.[10] Solving the Gaussian integral, it follows that we may write the magnitude in this approximation as $\Delta\bar{\varepsilon}_k = R_k\bar{\nu}_k/A\sqrt{2\pi}\,\sigma_k$. The whole spectrum is then calculated from Eq.(s9) by including the EDM modeling of the rotation strength for each k spectral line.

II. SELECTING THE PARAMETERS FOR THE CALCULATION OF THE CD SPECTRUM

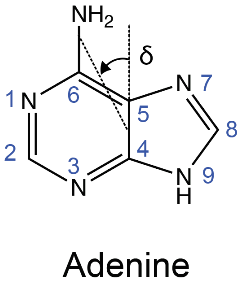

Figure S1. The angle δ defines the direction of the electric dipole transition moment (EDTM) used in the CD calculations for the adenine bases of the dApdA dinucleotide monophosphate.

For the majority of our CD calculations, we used as input parameters to Eqs. (s6) and (s7) the EDTM data for 9-methyladenine obtained by Holmén et al. (Table S1) [11] and the dielectric constant ϵ = 2. In Table S1 we list for each transition the values we have used for the EDTM magnitude $|\boldsymbol{\mu}|$, orientation δ, transition frequency ν, and extended transition dipole charge q and displacement l (see Fig. S1). In addition, to model the spectral line width of all monomer electronic transitions we assumed the Gaussian standard deviation σ_k = 0.2 eV. Our selection of these parameters was based on comparisons between the experimental CD spectrum of dApdA at room temperature in buffer at pH 7.2 containing 0.01 M $NaPO_3$ and 0.1 M $NaClO_4$, and CD calculations for which we assumed initially that the dApdA dinucleotide adopts only the B-form.

Table S1. Experimental values for the magnitudes and molecular frame orientations of the electric dipole transition moments (EDTMs) for 9-methyladenine obtained by Holmén et al,[11] and which we have used to model adenine mononucleotide in this work. All transitions are in-plane $\pi \rightarrow \pi^*$, and are listed in order of increasing transition frequency. The angle δ specifies the counter-clockwise rotation of the EDTM vector within the plane of the adenine base relative to the C_4-C_5 bond axis (see Fig. S1). The partial charges for the extended dipole model were derived using the relation $|\boldsymbol{\mu}| = q|\boldsymbol{l}|$, and by representing the adenine base as an ellipse with major diameter (a) 4.6 Å and minor diameter (b) 2.6 Å, such that $l = 2ab/[a^2\cos^2(\delta) + b^2\sin^2(\delta)]^{\frac{1}{2}}$.

Transition	ν (cm^{-1})	λ (nm)	$\lvert\boldsymbol{\mu}\rvert$ (D)	δ (°)	l (Å)	q (e)
I	36 710	272.4	1.65	+66 ± 7	3.96	0.09
II	38 820	257.6	3.63	+19 ± 7	2.70	0.28
III	43 370	230.6	1.15	−15 ±6	2.66	0.09
IV	46 840	213.5	2.52	−21 ±7	2.72	0.19
V	48 320	207.0	2.30	−64 ±10	3.87	0.12

For comparison, we present in Table S2 the empirical parameters from Williams et al.[12]

Table S2. Empirical spectroscopic parameters from Ref. 12 for the Adenine monomer.

Transition	ν, cm^{-1}	μ, Debye	δ, deg
I	37037	1.1	-87
II	38022	4.0	-3
III	42553	1.0	-87
IV	46296	3.7	-87
V	51282	3.7	-3
VI	53476	4.2	-87

For all of the parameters that we tested (see Tables S1 and S2), we obtained moderately favorable agreement between experiment and theory. We note that the sensitivity of the calculated CD to the choice of input parameters was greatest at the shortest wavelengths (200 – 250 nm) and least at the longer wavelengths (250 – 300 nm).

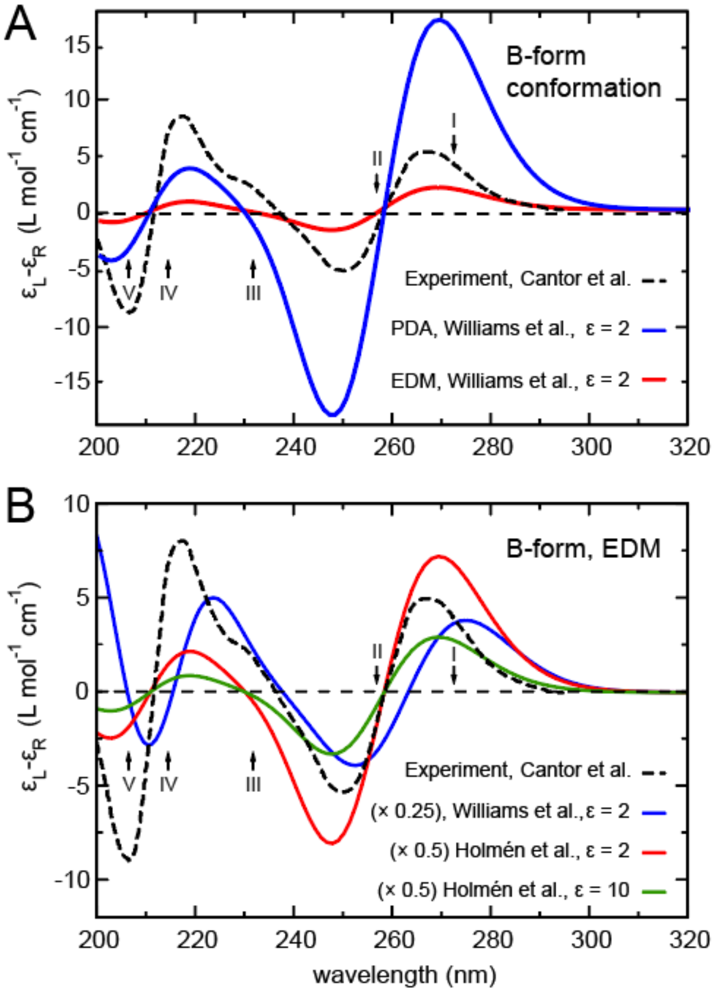

Figure S2. (***A***) comparison of the CD spectrum theoretically predicted for the Watson-Crick B-form of dApdA and the experimental data by Cantor at al.[13] We show both the spectra calculated using the Point Dipole Approximation (PDA) and the Extended Dipole Model (EDM). (***B***) Comparison of the CD spectrum theoretically predicted for the Watson-Crick B-form of dApdA and the experimental data, using either the empirical parameters from Holmén et al. (Ref. 11) or from Williams et al. (Ref. 12). The effect of varying the dielectric constant (from ε = 2 to ε = 10) is also shown. In both panels, vertical arrows indicate the positions of the uncoupled transitions of the Adenine monomer listed in Table S1.

To demonstrate the sensitivity of the CD theoretical predictions to the choice of the empirical parameters selected in the CD modeling, we report first, in Fig. S2*A*, a study of the CD spectrum for the Watson-Crick B-form of dApdA calculated using two different models: (i) the simple Point Dipole Approximation (PDA); and (ii) the Extended Dipole Model (EDM). The spectrum of the B-form, predicted by the theory is similar in both approximations, and shows a good agreement with experiments in the low energy part of the spectrum. Figure S2*B* shows, instead, a study of the sensitivity of the calculations to the choice of the parameters. It reports results for the Point Dipole Approximation (PDA) calculation of the CD spectrum for the Watson-Crick B-form, while adopting either the empirical spectroscopic parameters from Holmén et al.[11] or those from Williams et al.[12] As can be seen, the positions and heights of the peaks change significantly at low wavelengths (high excitation energies), while they agree reasonably well at high wavelengths (low excitation energies). Thus, the results are qualitatively consistent with a slightly better agreement with the experimental values when using parameters from Table S1. Finally, the figure shows a study of the variation of the empirical dielectric constant, which is found to produce just a change in the intensity of the spectrum, without affecting the positions of the peaks.

III. DECOMPOSITION OF THE CD SPECTRUM OF THE dApdA DINUCLEOTIDE MONOPHOSOPHATE

Figure S3 displays the contributions to the CD spectrum of dApdA from each MSM macrostate. We assume only single-excited electronic states and five electronic transitions per adenine monomer, following the conventions used by Holmén et al (Ref. 11, Table S1). Proceeding on this basis we obtain ten total degenerate-pair contributions, given by five symmetric and five anti-symmetric transitions. To keep things simple we neglect here the contributions to the spectral decomposition from non-degenerate transitions, because degenerate contributions are dominant in the total CD spectrum (note that both degenerate and non-degenerate contributions are included in the CD calculations reported in the Main Text).

In Fig. S3, different colors represent different transitions, while the dotted and the dashed curves represent the contributions from the symmetric and the anti-symmetric transitions, respectively. Due to the chirality, or lack of chirality, of the average structure in each macrostate, only the states S3 and S4 provide significant contributions to the final CD spectrum. Of particular interest is the spectral decomposition of the two chiral macrostates with the largest stationary probabilities, S3 and S4. The first macrostate includes in its conformational distribution the Watson-Crick B-form, while the second includes the Hoogsteen form.

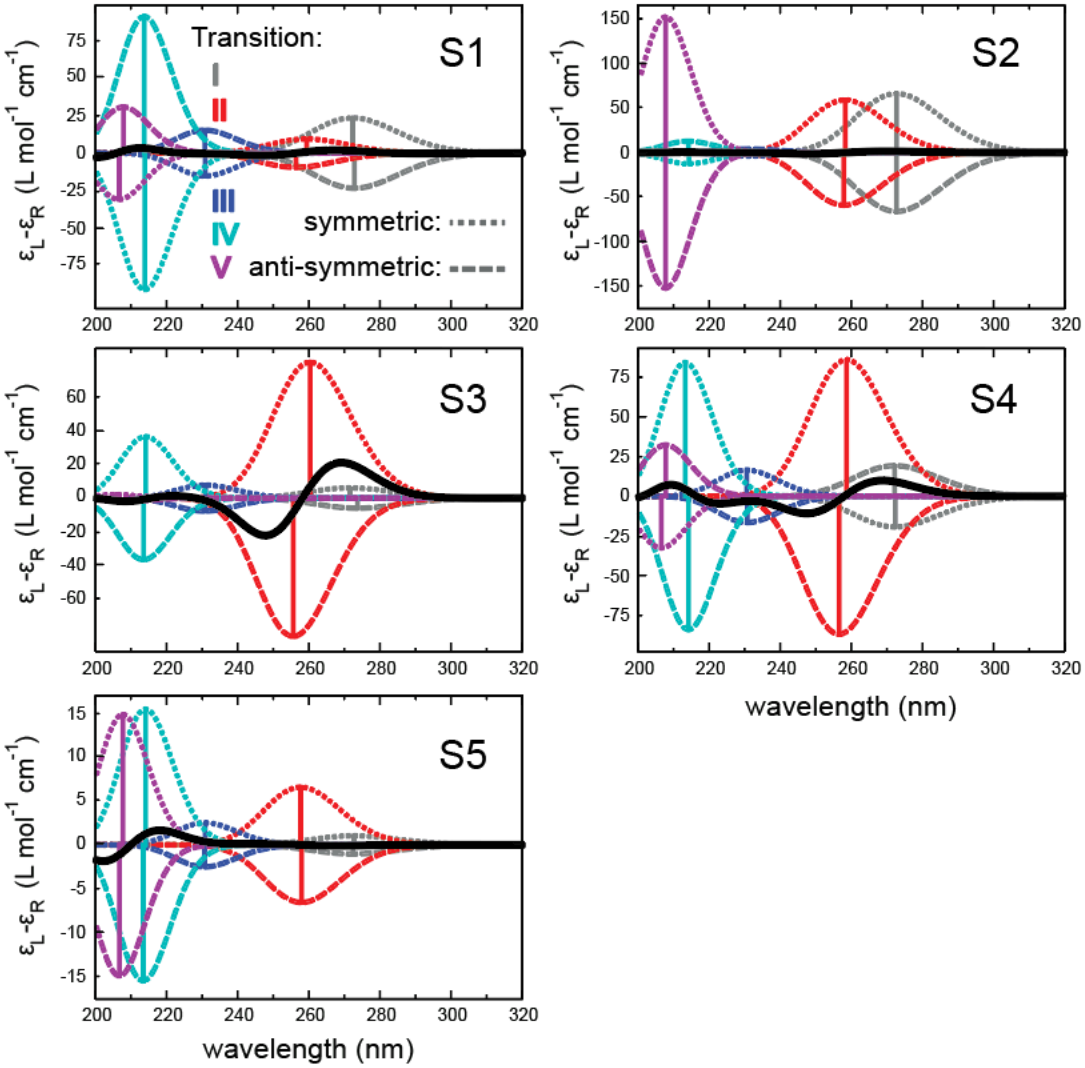

Figure S3. Spectral decomposition of the degenerate CD for the average structures of each of the five macrostates (S1 – S5) for dApdA. The contributions due to transition I (gray), II (red), III (blue), IV (cyan) and V (magenta) are decomposed into signals reflecting the symmetric (dotted) and anti-symmetric (dashed) transitions. The total CD spectrum for each average structure is given by the solid (and thick) black line in each plot.

The average structures of each of these macrostates show that for the dominant low-energy transition (II, red), the decomposition into symmetric and anti-symmetric contributions are of the same sign for both the S3 and the S4 states, while the contributions to the first low-energy transition (I, black) have opposite signs in the two macrostates. The average S3 and S4 macrostate structures, while of opposite handedness, are not mirror images of each other; rather, one of the adenine monomers in the average structure of S4 is 'flipped' relative to the same monomer in the average structure of S3, which is compatible with the Hoogsteen structure. It is the flipped base in the S3 macrostate that is responsible for its right-handed CD signal in the long wavelength region.

IV. MODELING THE BASE STACKING OF dApdA USING MARKOV STATE MODELING PROCEDURES

To analyze the base stacking of dApdA, we adopted a two-site description for each base, which defines vectors within the plane of each nucleotide. The first site is positioned between the C_4 and the C_5 carbon atoms in Adenine (see Fig. S4), while the second site was positioned 0.1 nm along a line perpendicular to the vector connecting those atoms.

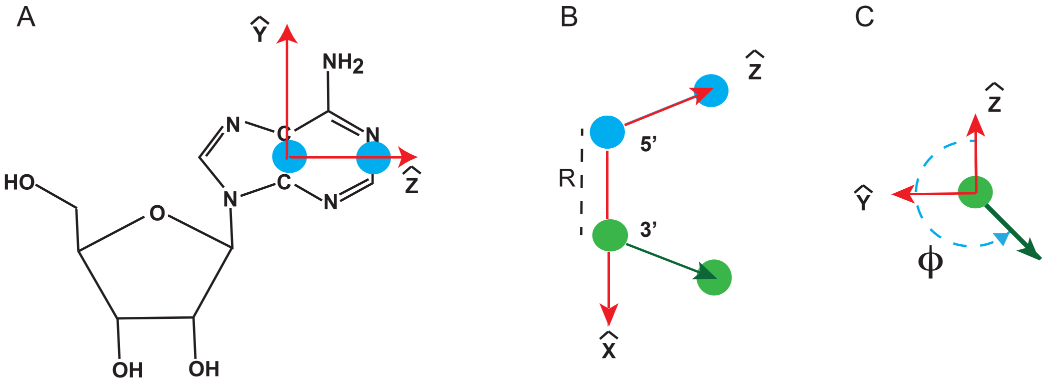

Figure S4. (***A***) The two sites are placed within the plane of the Adenine base. The independent free energy parameters used in the MSM analysis are the radial separation between C_4 and the C_5 midpoints within each adenine monomer, and (***B***) the dihedral angle, ϕ, between the in-plane vectors, here represented in an overhead view (***C***).

The relative orientation of the four-bead model captures the relative distance between the adenine bases, and their relative torsion angle. The parameters reported in the Main Text for the free energy maps are thus defined in Fig. S4: the stacking torsional angle, ϕ, is defined by the dihedral between the in-plane vectors, while the distance, R, between nucleotides is given by the distance between the C_4 and the C_5 midpoints of each base (see also Fig. 2 in the Main Text) .

V. PARTITIONING THE FREE ENERGY SURFACE IN MACROSTATES USING THE MARKOV STATE MODEL PROCEDURE

We used the k-means++ algorithm[14,15] to construct a kinetically-relevant, balanced clustering of the trajectories (using the Euclidean criterion) by partitioning the 10^7 conformations into 100 initial microstates. A transition rate matrix was constructed for these microstates and then diagonalized into eigenvalues and eigenvectors. From the eigen-spectra of the transition probability matrix, we constructed five macrostates by implementing a minimum error propagation version of the Perron-cluster cluster analysis (PCCA+). We justified our choice for these five macrostates by considering the related conformational landscape and the implied interconversion time scales.

Rapidly interconverting molecular conformations were assigned to the same macrostate, while slowly interconverting conformations, which are separated by large barriers, occur between conformations that lie within different macrostates. By identifying and separating slowly interconverting conformations from rapidly interconverting ones, the MSM ensures that the slow processes obey Markovian statistics. To sample slow transitions, we adopted a lag-time of 500 ps, and confirmed that under these conditions Markovian behavior was satisfied by checking that the Chapman-Kolmogorov condition applies[16,17,18] (see Fig. S5).

We started by assigning the simulation trajectory to W=100 discrete microstates using the k-means++ clustering algorithm as implemented in PyEMMA[19] software program (the comparison with the case of assuming W=1000 is reported at the end of this section and shows no substantial difference in their predictions). Once the clustering of the simulation trajectory into microstates was completed, we defined a lag time τ and calculated the transition matrix, $T(\tau\ )$, by counting the number of transitions occurring between two given microstates during the defined lag time. The transition matrix, therefore, models the evolution of the probability vector, $P^T(t+\tau\ ) = P^T(t\ )T(\tau)$, which gives the probability of finding the system in a final state at time t+τ, given that the system was in an initial state at time t, and had a probability of transition between states, during lag time τ, that is given by the transition matrix, $T(\tau\ )$.

The diagonalization of the transition matrix gives eigenvalues and eigenvectors, which contain important information on the stable states in the free energy landscape and the kinetics of transitions between states. The eigenvalues (λ) define the timescale of a transition i during the lag time τ as $t_i = -\tau/[\ln \lambda_i(\tau)]$, while the eigenvectors define the partitioning of the free energy surface, and its 100 microstates, into a smaller number of macrostates. Because the transition matrix is a regular stochastic matrix, the Perron-Frobenius theorem guarantees that the first eigenvalue is equal to one, which corresponds to an infinite transition time t_1. Thus, the corresponding first eigenvector has only positive entries, and defines the equilibrium ($t_1 = \infty$) population of the macrostates, which in our calculation of the CD spectrum gives the percentage contribution of each macrostate to the final CD function (see Fig. 7 in the Main Text). By inspecting the eigenvalues, it is possible to identify a gap, corresponding to a gap in the transition times, which separates slow from fast transitions. The eigenvector that corresponds to that eigenvalue defines, with the number of its "nodes", the number of macrostates into which the free energy landscape needs to be partitioned to separate fast transitions (inside one macrostate) from slow, and biologically relevant, transitions (between macrostates). In our case this condition is fulfilled when the free energy map is partitioned into five macrostates (see Fig. 6*A* in the Main Text). The lag time, τ, at which the kinetics of the stacking-unstacking fluctuations become uncorrelated, is identified by testing at which τ the transitions between states become Markovian. To test the Markovian nature of those transitions, we adopted the standard procedure based on fulfilling the Chapman-Kolmogorov condition. When dynamical processes are Markovian (i.e. uncorrelated), the transition matrix sampled at a multiple, n, of the lag time, τ, is equal to the transition matrix at lag time τ to the n power: $T(n\tau) = T(\tau)^n$, which implies that the eigenvalues $\lambda(n\tau) = \lambda(\tau)^n$. As a consequence, the timescale of a transition becomes independent of the time used to sample the simulation trajectory. In fact $t_i(n\tau) = -\frac{n\tau}{[ln\lambda_i(n\tau)]} = -\frac{n\tau}{n[ln\lambda_i(\tau)]} = \frac{\tau}{[ln\lambda_i(\tau)]} = t_i(\tau)$, which is the Chapman-Kolmogorov condition.

To test this condition, Fig. S5 reports the transition time as a function of the lag time and identifies the time at which the process becomes Markovian as the time where $t_i(n\tau)$ becomes constant. Fig. S5 shows that in our system there are four slow processes that are fully Markovian when transitions are sampled at a lag time longer than 500 ps. Thus, a lag time of 500 ps has been selected for the calculations reported in the Main Text. The number of slow

Markovian processes detected in this plot, which is defined as the number of lines that becomes constant while fulfilling the necessary condition that $t_i(\tau) \geq \tau$, is consistent with a number of macrostates equal or higher than five, thus supporting the number of macrostates selected in our analysis.

As can be seen in Fig. S5, beyond 0.5 ns, the transition time (implied timescale) is almost level. This means that starting at 0.5 ns, and at longer time lag, the coarse graining of the free energy map into five macrostates gives kinetics transitions between macrostates that are Markovian, indicating that a reasonable separation of timescales exists in the spectral decomposition of the transition matrix. The fulfillment of the Chapman-Kolmogorov condition ensures that the transition time is independent of the number of uncorrelated steps that are used to model the process.

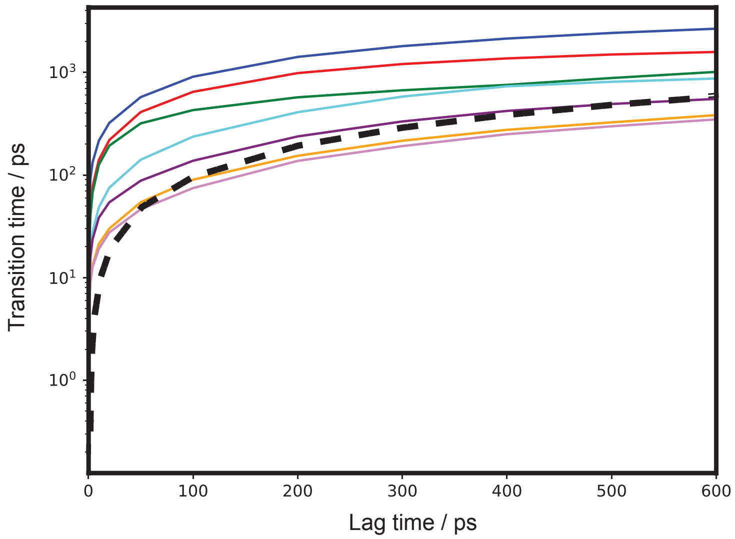

Figure S5. Transition time measured as a function of increasing lag time, for the simulation of dApdA in solution at 0.1 M salt concentration, where the free energy landscape is partitioned into five Markov states. Different colors display different transition processes. For the four slowest processes the transitions become Markovian around 500 ps. The black dashed line defines the condition for which the transition time is equal to the sampling lag time: processes that occur in a time faster than the sampling time cannot be sampled and are discarded.

VI. IDENTIFYING THE OPTIMAL NUMBER OF MICROSTATES.

The number of microstates to be used in the MSM analysis depends on the precision we want to achieve in the MSM calculations, while avoiding underfitting and overfitting. The optimal number is roughly related to the number of structural parameters and the number of groups/residues in the molecule. Here, we studied a relatively small system (dApdA) using simple structural parameters, which suggests that a small number of microstates could be sufficient. Note that when the number of microstates is not optimum, the kinetic information and/or the positions of the barriers and the border between metastable states can be inaccurately predicted because of underfitting or overfitting. In this Section, we investigate the sensitivity of the CD spectra predictions and the related kinetic information to the number of microstates and compare the results for 100 and 1000 microstates, while in the Main Text we reported the results for 100 microstates.

Table S3. Stationary normalized probability distributions of PCCA+ macrostates when the MSM contains either 100 or 1000 microstates.

	S1	S2	S3	S4	S5
100 microstates	0.026	0.028	0.090	0.373	0.483
1000 microstates	0.024	0.027	0.095	0.372	0.482

Table S3 reports the stationary distribution (i.e. the first eigenvector of the transition matrix) for each of the macrostates in the 100- and 1000-microstate MSM models (see also Fig. 6 in the Main Text). As it can be seen, there is no significant difference between them, indicating that selecting either number of microstates does not entail significant changes in the borders between macrostates. Thus, the slight change in border resolution does not significantly affect the long-time properties of the Markov chain, whether 100 or 1000 microstates were used.

In Table S4 we show the mean first passage times (MFPTs) between the macrostates when 100 microstates are used to build the transition matrix for the MSM. When compared to the analogous table for 1000 (see Table S5), it shows that there is no significant difference between the MFPTs calculated using 100 or 1000 microstates. Finally, we compare the behavior of the implied timescales as a function of the lag time for the 100- and the 1000-microstate MSMs, and we find that the transition time fulfills the Chapman-Kolmogorov condition at a lag time consistent in the two cases (data not shown).

Table S4. Mean first passage times (MFPTs) in nanoseconds between the five macrostates of the dApdA system at [NaCl] = 0.1 M. The MFPTs were calculated using the Markov state model analysis of the free energy landscape (Fig. 6*A* in the Main text) and 100 microstates. Note that the unit of time is ns and rows and columns indicate initial and final macrostates, respectively.

i \ *f*	S1	S2	S3	S4	S5
S1	0.0	35.0	13.1	5.9	3.2
S2	56.4	0.0	12.6	4.5	3.2
S3	58.7	37.0	0.0	6.1	2.4
S4	59.1	36.4	13.4	0.0	4.6
S5	58.1	36.9	11.5	6.5	0.0

Table S5. Mean first passage times (MFPTs) in nanoseconds between the five macrostates of the dApdA system at [NaCl] = 0.1 M. The MFPTs were calculated using the Markov state model analysis of the free energy landscape (Fig. 3*A* in the Main text) and 1000 microstates. Note that the unit of time is ns and the initial and target macrostates are in rows and columns, respectively.

i \ *f*	S1	S2	S3	S4	S5
S1	0.0	36.2	13.0	6.3	2.9
S2	68.4	0.0	12.3	4.6	3.3
S3	70.9	38.3	0.0	6.2	2.5
S4	71.5	37.6	13.1	0.0	4.7
S5	69.9	38.2	11.2	6.6	0.0

VII. SIMPLIFIED CD SPECTRAL CALCULATION USING MSM AVERAGED MACROSTATE CONFORMATIONS

In the Main Text, we showed that the CD spectrum of the dApdA system can be decomposed into contributions from five different MSM macrostates, which are defined based on their bounded positions within the free energy landscape $G(R,\phi)$ of 10 million MD-sampled microstate configurations (see Fig. 3*B*). Of the five macrostates, only S3 and S4 contribute significantly to the CD spectrum (see Fig. 7). Our results suggest that we may apply a relatively simple model to achieve the structural interpretation of the dApdA CD spectrum. Having identified the key macrostates relevant to the CD observable (see above), we next determine the smallest number of structural parameters necessary to characterize these macrostates. Given a reduced set of conformations for the various macrostates, in addition to specification of their relative weights, it is possible to greatly speed up the computation time needed to simulate the CD spectrum. In principle, such structural models can be used for the general interpretation of any spectroscopic measurement performed on the dApdA system.

For each of the five macrostates we determined an average conformation with mean and standard deviation of the inter-base separation defined according to

$$\bar{R}_A = \frac{1}{N_A}\sum_{n=1}^{N_A} R_{A,n} \tag{s10}$$

and

$$\sigma_{A,R} = \left[\frac{1}{N_A}\sum_{n=1}^{N_A}\left(R_{A,n} - \bar{R}_A\right)^2\right]^{1/2} \tag{s11}$$

We similarly defined the means and standard deviations of the inter-base twist, tilt and roll angles according to

$$\bar{\theta}_A = \frac{1}{N_A}\sum_{n=1}^{N_A} \theta_{A,n} \tag{s12}$$

and

$$\sigma_{A,\theta} = \left[\frac{1}{N_A}\sum_{n=1}^{N_A}\left(\theta_{A,n} - \bar{\theta}_A\right)^2\right]^{1/2} \tag{s13}$$

where $\theta_{A,n} \in \left\{\phi_{A,n}, \alpha_{3'(5')A,n}, \beta_{3'(5')A,n}\right\}$ are the *n*th twist, tilt and roll angles, respectively, of the N_A configurations contained within the *A*th macrostate (see Fig. 2 in the Main Text and Fig. S4 for coordinate definitions). In Table S6, we list the mean and standard deviation parameters for each macrostate, in addition to the associated number fractions p_A $(= N_A/N_{tot})$ calculated from the MSM analysis.

In Figs. S6*A* – S6*E* we show our CD calculations for macrostates S1 – S5, respectively, which were determined from the mean macrostate conformations (shown as red curves). We compare these to the calculated CD spectra by summing over all of the configurations contained within each macrostate (blue curves).

Table S6. Means and standard deviations of the structural parameters corresponding to the five macrostates *A*, for dApdA, which are labeled S1 – S5. The mean inter-base separation $\bar{R}_A$, mean twist angle $\bar{\phi}_A$, mean tilt angles $\bar{\alpha}_{3'(5')A}$, and roll angles $\bar{\beta}_{3'(5')A}$ are defined by Eqs. (s10) – (s13), which are based on the structural coordinates defined in Fig. 2 of the Main Text and Fig. S4. The values for the equilibrium population, p_A, are used as weights for computing the CD spectrum from the mean macrostate conformations, as shown in Fig. S6.

A	p_A	$\bar{R}_A$	$\sigma_{A,R}$	$\bar{\phi}_A$	$\sigma_{A,\phi}$	$\bar{\alpha}_{3'(5')A}$	$\sigma_{A,\alpha_{3'(5')}}$	$\bar{\beta}_{3'(5')A}$	$\sigma_{A,\beta_{3'(5')}}$
		(Å)	(Å)	(°)	(°)	(°)	(°)	(°)	(°)
S1	0.026	4.5	1.1	145.0	6.0	155.3	9.8	33.2	39.7
						(28.6)	(10.6)	(-28.9)	(51.4)
S2	0.028	12.2	1.4	12.0	183.6	160.1	5.3	-58.6	61.9
						(23.8)	(6.3)	(-4.1)	(59.8)
S3	0.090	4.0	0.3	44.0	1.8	139.0	8.3	-28.6	82.2
						(43.5)	(8.6)	(-14.3)	(72.2)
S4	0.373	4.7	0.9	-91.7	15.2	148.5	9.5	-76.2	28.9
						(33.2)	(9.8)	(21.3)	(56.1)
S5	0.483	5.0	0.9	5.9	10.0	159.0	7.4	-69.3	29.4
						(22.7)	(7.6)	(32.5)	(38.9)

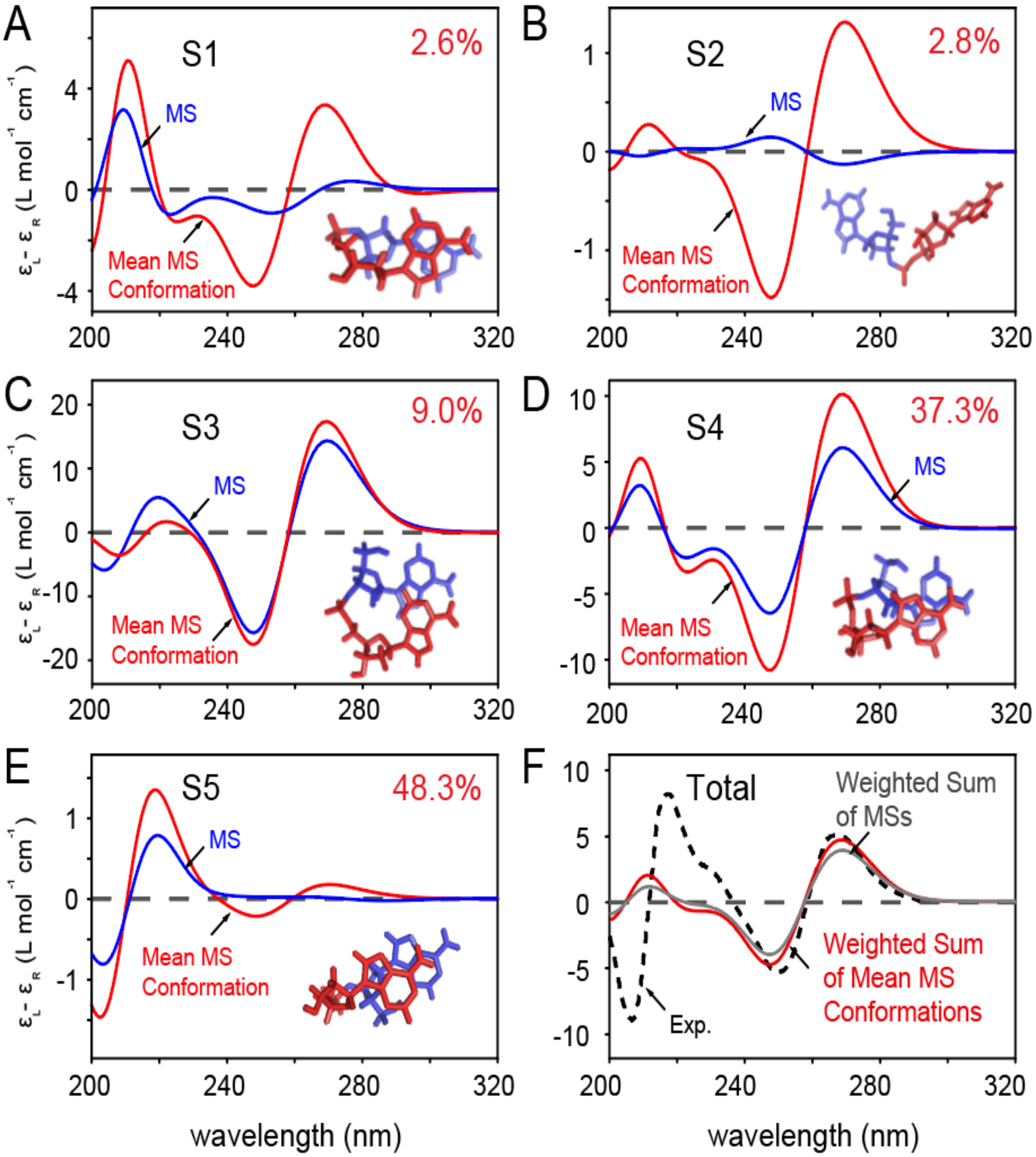

Figure S6. (***A***) – (***E***) Each panel shows, for each macrostate, the comparison between the contribution to the CD spectrum from all the conformational states in the macrostate (blue curve) and the contribution from the averaged macrostate structure (red curve), with structural parameters listed in Table S6. The molecular models representative of the averaged dApdA structures are shown as insets within each panel. The 5' nucleotide is shown in blue, and the 3' nucleotide and phosphate are shown in red. (***F***) The weighted sum of the macrostate contributions to the total CD are shown in gray, and the weighted sum from the averaged structures in red. The experimental CD spectrum[13] is shown as a dashed black curve.

The insets in Figs. S6*A* – S6*E* show the molecular models of the dApdA dinucleotide that represent the corresponding mean macrostate conformations. For macrostates

S1, and S3 – S5, the calculated CD spectrum of the mean conformation are similar in shape to those of the full macrostates, although the magnitude of the CD in each case is somewhat overestimated by the mean conformation. This agreement is less favorable for macrostates S2, which is likely due to the broad dispersion of conformations contained within this macrostate.

As discussed in the Main Text, the two macrostates that contribute most significantly to the CD spectrum are S3 and S4. Our determination of the mean conformations allows us to apply a structural interpretation to these contributions. Macrostate S3 exhibits, on average, a stacked and right-handed conformation, while macrostate S4 exhibits a stacked and left-handed conformation. The S3 mean conformation resembles that of flanking bases in B-form duplex DNA, which gives rise to the 'right-handed' Cotton effect observed in the long wavelength region of the CD spectrum. The relative roll angle of the S4 conformation is large ($\bar{\beta}_{5'S4} - \bar{\beta}_{3'S4}$ = 97.5°), such that the 3' base is 'flipped' relative to the 5' base. This conformation also gives rise to the right-handed long wavelength CD spectrum. A spectral decomposition analysis of macrostate S4 shows that the right-handed long wavelength CD is consistent with this flipped base conformation (see Fig. S3 above). Although the S5 macrostate represents 48.3% of the total population, it contributes very little to the CD spectrum due to its predominantly achiral symmetry and correspondingly low rotational strength. Finally, the S1 and S2 macrostates do not contribute significantly to the CD spectrum due to their small populations.

VIII. DISTRIBUTIONS OF THE THETA ANGLE AT INCREASING WATER SEPARATION FROM THE PHOSPATE OXYGEN FOR dApdA

In Fig. S7, we report the distributions of the angle θ that defines the orientation of the water dipole moment $\vec{\mu}_{\mathrm{H_2O}}$ relative to the vector $\overrightarrow{\mathrm{PO}}_{\mathrm{H_2O}}$, which connects the water oxygen to the central phosphorous atom P of the dApdA (see Fig. 5*A* of the Main Text). The distributions are reported for the dApdA dinucleotide at monovalent salt concentration [NaCl] = 0.1 M. Each distribution is defined over a narrow range of distances corresponding to a given hydration shell relative to the P atom. With the exception of the first distribution, which displays an irregular structure due to the exclusion of water molecules from the nearest distances around the P atom,

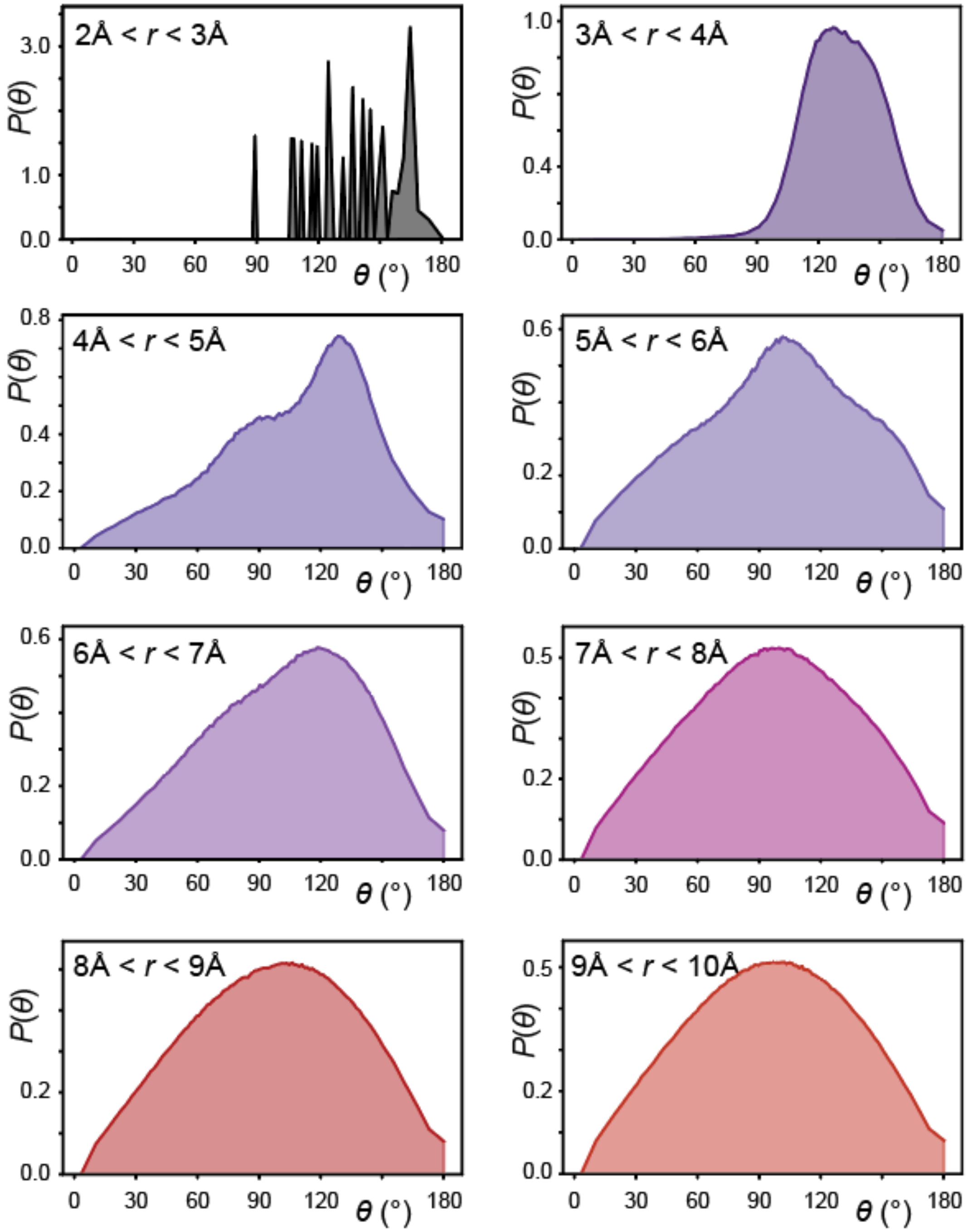

Figure S7.Distributions of the angle θ between the water dipole moment $\vec{\mu}_{H_2O}$ and the vector $\overrightarrow{PO}_{H_2O}$, as a function of distance between the water oxygen and the central phosphorous atom P. The distributions are shown as a function of increasing distance from the P atom, and for the salt concentration [NaCl] = 0.1 M. The non-uniform broadening of the distributions indicate the presence of hydrogen bonding between successive hydration shells and loss of orientational correlation with the P atom with increasing distance.

all of the remaining distributions appear as continuous and smoothly varying functions of the angle θ. The distributions exhibit a symmetric shape only for the hydration layers that are separated from the P atom by more than 7 Å. For shorter distances, we observe well-defined non-uniform distributions of the water dipole orientations, which indicate the presence of hydrogen bonding between successive hydration shells. From our calculations of the average angle, we see that at short distances $< \langle \cos\theta \rangle \approx \cos\langle\theta\rangle$, as shown in Fig. 5 of the Main Text.

IX. DECAY OF THE TIME AUTOCORRELATION FUNCTION OF THE INTER-BASE SEPARATION, R(t).

To ensure that the system is converged and that the FES displayed in Fig. 3*A* in the main text represents the system at equilibrium, we report in Fig. S8 the decay of the autocorrelation function of the fluctuations away from the equilibrium value of the inter-base distance, $\Delta R(t) = R(t) - \langle R \rangle$.

The autocorrelation function is defined as $C_R(t) = \frac{\langle \Delta R(t) \Delta R(0) \rangle}{\langle \Delta R(0)^2 \rangle}$. The autocorrelation function decays to zero, for all the salt concentrations, on a timescale of a few nanoseconds. Note that the decay is similar at all salt concentrations up to [NaCl]=1. M, but becomes faster in the high salt concentration regime ([NaCl]=1.5 M), where bases are largely unstacked.

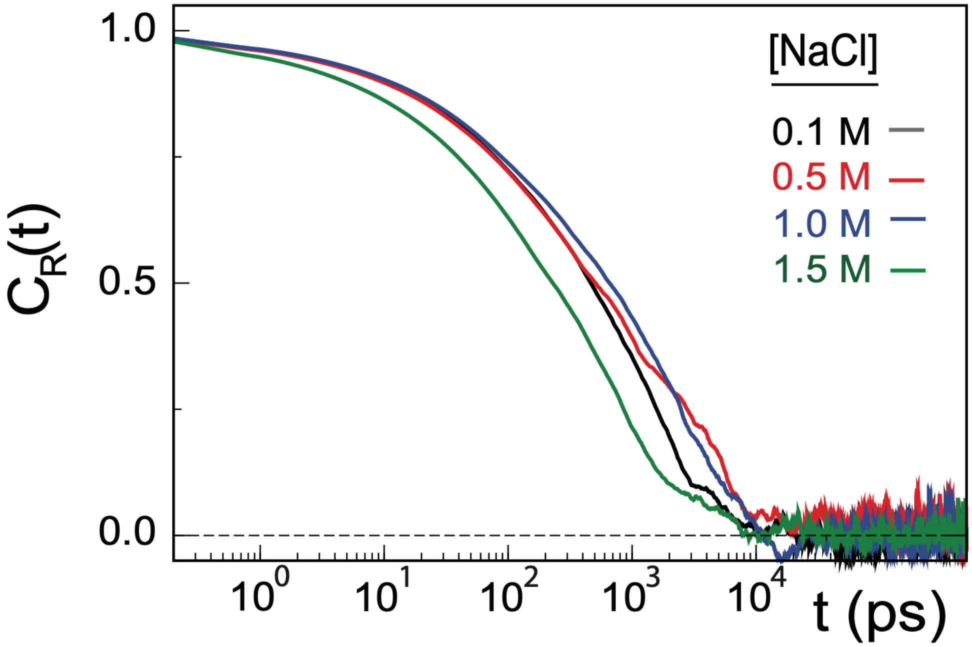

Figure S8. Decay of the time autocorrelation function of the fluctuations of inter-base separation for samples at increasing salt concentration. The system reaches equilibrium in approximately 10 ns, with the sample at salt concentration higher than 1.0 M relaxing faster than any of the other salt concentrations. In contrast, the samples at salt concentrations below 1 M decay at very similar rates.

References.

[1] Rizzo V., Schellman J. A. (1984) Matrix-Method Calculation of Linear and Circular Dichroism Spectra of Nucleic Acids and Polynucleotides. Biopolymers. 23:435–70.

[2] Rodger A., Nordén B. (1997) Circular Dichroism and Linear Dichroism, Oxford Chemistry Masters (Oxford ; New York: Oxford University Press).

[3] Ji H., Johnson N. P., von Hippel P. H., Marcus A. H. (2019) Local DNA Base Conformations and Ligand Intercalation in DNA Constructs Containing Optical Probes. Biophys. J. 117: 1101–15.

[4] Czikklely V., Forsterling H. D., Kuhn H. (1970) Extended Dipole Model for Aggregates of Dye Molecules. Chem. Phys. Lett. 6:207–10.

[5] Didraga C., Pugzlys A., Hania P. R., von Berlepsch H., Duppen K. J., Knoester J. (2004) Structure, Spectroscopy, and Microscopic Model of Tubular Carbocyanine Dye Aggregates. J. Phys. Chem. B 108:14976–85. (2004).

[6] Heussman D., Kittell J., Kringle L., Tamimi A., von Hippel P. H., Marcus A. H. (2019) Measuring Local Conformations and Conformational Disorder of (Cy3)2 Dimer Labeled DNA Fork Junctions Using Absorbance, Circular Dichroism and Two-Dimensional Fluorescence Spectroscopy. Faraday Discuss 216:211–35.

[7] Cannon B. L., Kellis D. L., Patten L. K., Davis P. H., Lee J., Graugnard E., Yurke B., Knowlton W. B. (2017) Coherent Exciton Delocalization in a Two-State DNA-Templated Dye Aggregate System. J. Phys. Chem. A 121:6905–16.

[8] Bouvier B., Gustavsson T., Markovitsi D., Milli P. (2002) Dipolar Coupling between Electronic Transitions of the DNA Bases and Its Relevance to Exciton States in Double Helices. Chem. Phys. 275:75–92.

[9] Dinpajooh M., Matyushov D. V. (2016) Dielectric Constant of Water in the Interface. J. Chem. Phys. 145:014504(1-7).

[10] Schellman J. A. (1975) Circular Dichroism and Optical Rotation. Chem. Rev. 75: 323-331.

[11] Holmén A., Nordén B., Albinsson B. (1997) Electronic transition moments of 2-aminopurine. J. Am. Chem. Soc. 119:3114-3121.

[12] Williams A. L., Cheong C., Tinoco I., Clark L. B. (1986) Vacuum Ultraviolet Circular Dichroism as an Indicator of Helical Handedness. Nucleic Acids Res.14:6649–6659.

[13] Cantor C. R., Warshaw M. M., Shapiro H. (1970) Oligonucleotide Interactions. 3. Circular Dichroism Studies of the Conformation of Deoxyoligonucleotides. Biopolymers 9(9): 1059–77.

[14] Arthur D., Vassilvitskii S. (2007) k-means++: The Advantages of Careful Seeding Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Philadelphia, PA; pp 1027–35.

[15] Celebi, M. E., Kingravi H. A., Vela P. A. (2013) A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. Expert Syst Appl 40:200–210.

[16] Prinz J.-H., Wu H., Sarich M., Keller B., Senne M., Held M., Chodera J., Schütte C., Noé F. (2011) Markov Models of Molecular Kinetics: Generation and Validation. J. Chem. Phys. 134:174105.

[17] Reichl L. E. (2016) A Modern Course in Statistical Physics, 4 edition (Weinheim: Wiley-VCH).

[18] van Kampen N. G. (2007) Stochastic Processes in Physics and Chemistry, 3rd edition (Amsterdam ; Boston: North Holland).

[19] Scherer M. K., Trendelkamp-Schroer B., Paul F., Pérez-Hernández G., Hoffmann M., Plattner N., Wehmeyer C., Prinz J. H., Noé F. (2015) PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. J. Chem. Theo. Comp. 11:5525–42