SUPPORTING INFORMATION

Determining sequence-dependent DNA oligonucleotide hybridization and dehybridization mechanisms using coarse-grained molecular simulation, Markov state models, and infrared spectroscopy

Michael S. Jones,[†] Brennan Ashwood,[‡] Andrei Tokmakoff,[‡] and Andrew L. Ferguson[*,†]

†*Pritzker School of Molecular Engineering, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, United States*

‡*Department of Chemistry, Institute for Biophysical Dynamics, and James Franck Institute, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, United States*

E-mail: andrewferguson@uchicago.edu

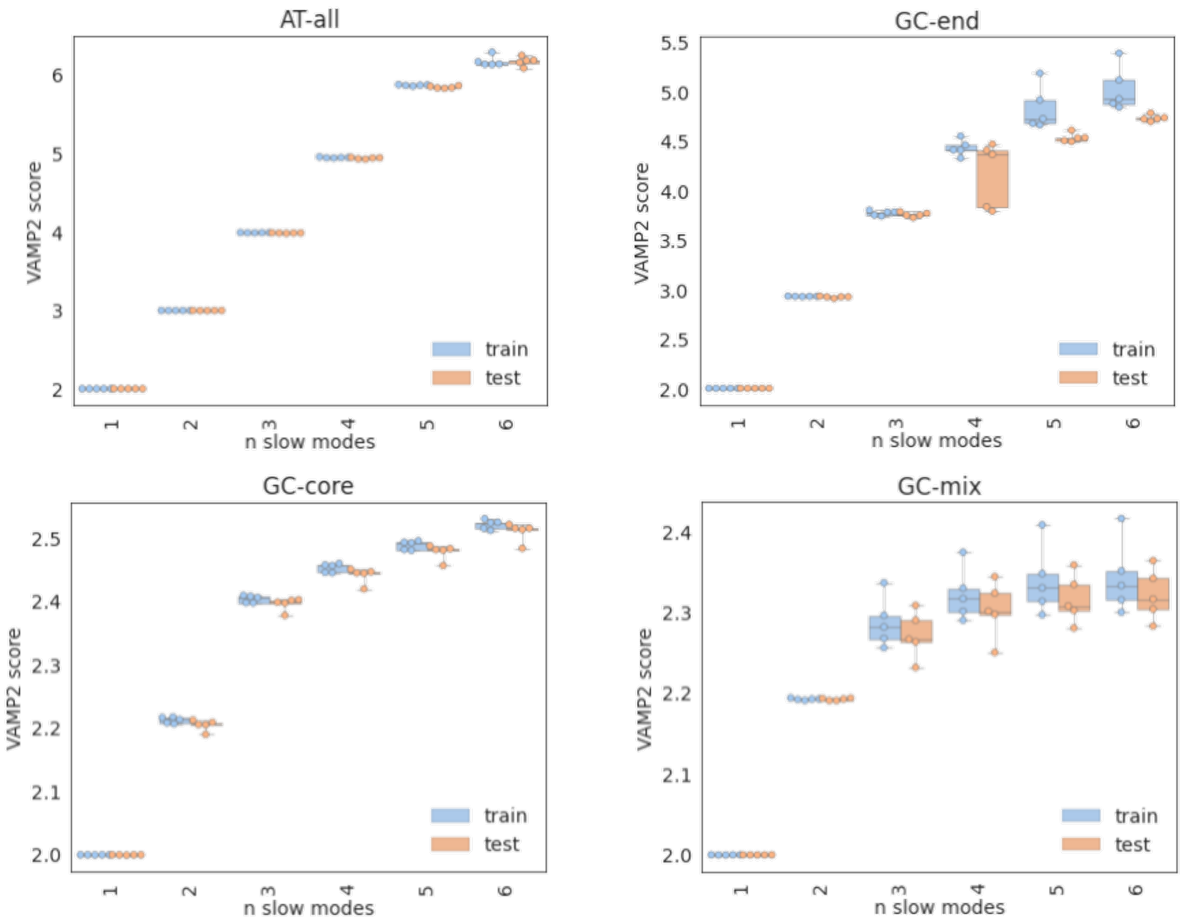


Figure S1: Five-fold cross-validation of the SRV VAMP-2 scores to select the optimal number of SRV coordinates for each sequence. A knee in the VAMP-2 plot was identified at the fifth, fourth, third, and second slow modes for AT-all, GC-end, GC-core, and GC-mix, respectively. An embedding of corresponding dimensionality was then used to cluster frames into discrete states. The absence of any significant separation in the training and testing VAMP-2 scores demonstrates that model is not overfitted.

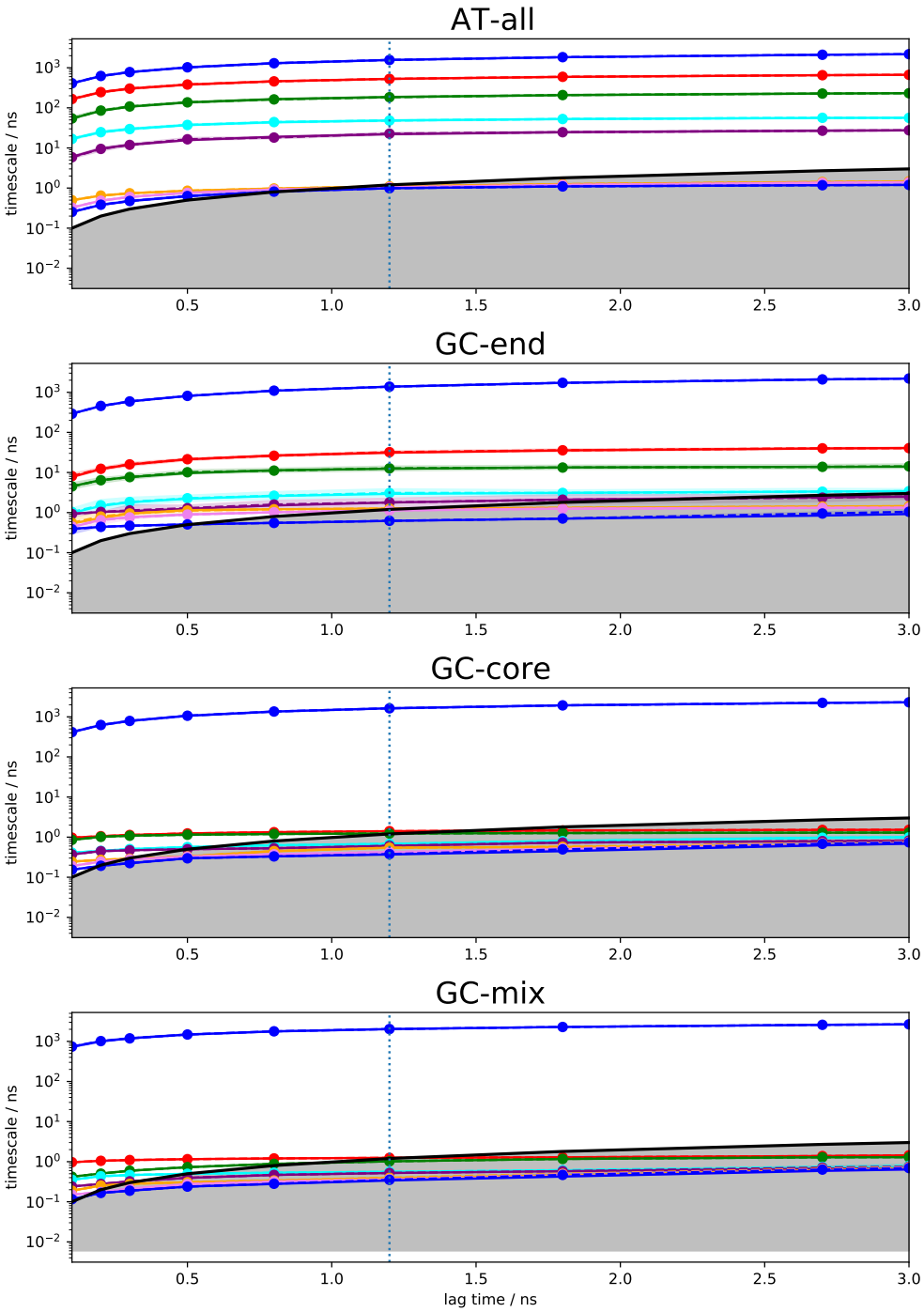


Figure S2: Convergence of the MSM implied time scales t_i as a function of lag time τ. Solid lines indicate maximum likelihood result while dashed lines show the Bayesian ensemble means. The implied time scales for all sequences converge at a lag time of $\tau = 1.2$ ns (vertical line). The black solid curve marks equality of the implied time scale and lag time and delimits the shaded region wherein the implied time scales are shorter than the lag time and cannot be resolved.

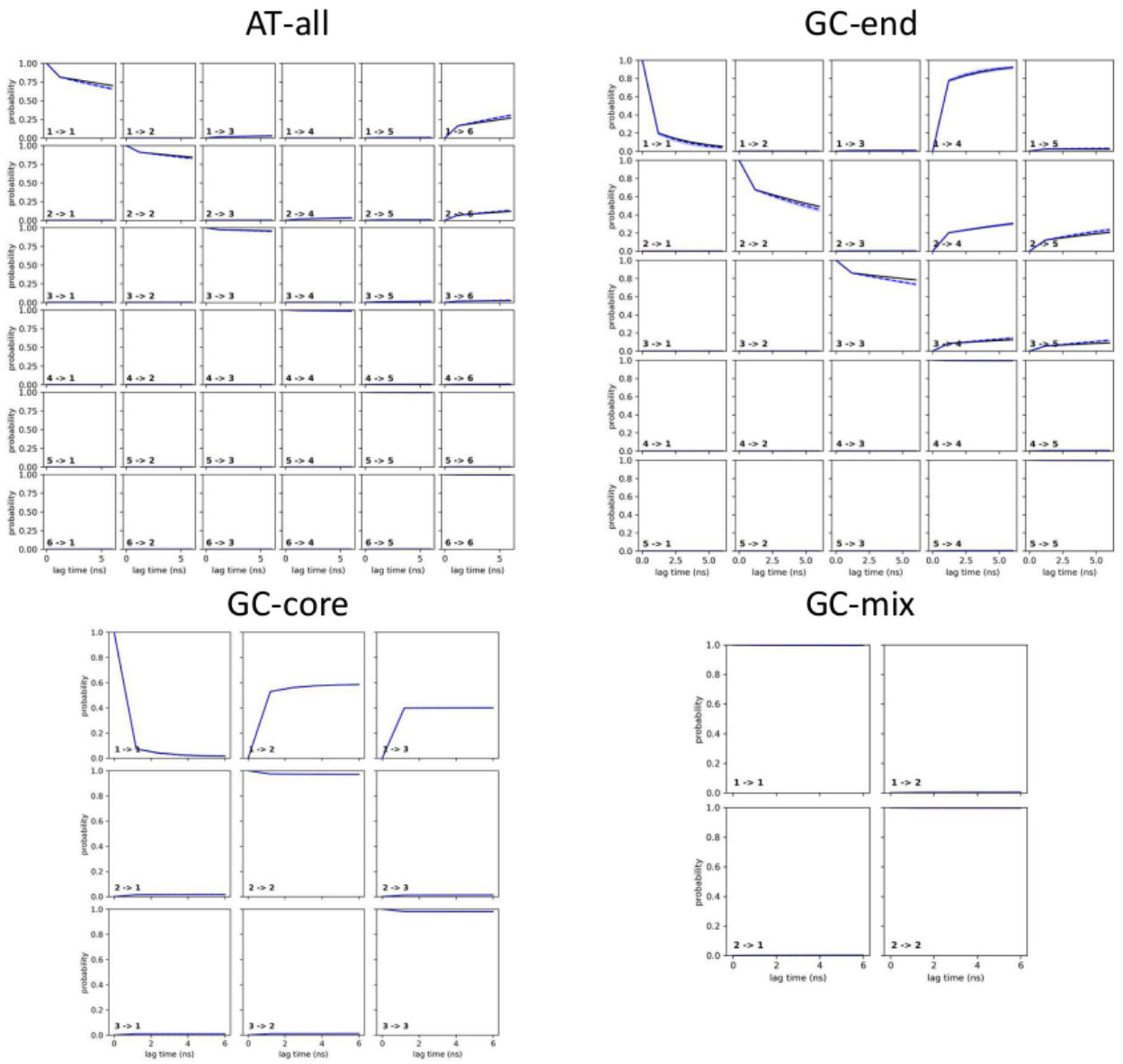


Figure S3: Chapman-Kolmogorov (CK) tests comparing the probabilities of remaining within each macrostates for each sequence as a function of lag time predicted by k applications of an MSM constructed at the $\tau = 1.2$ ns lag time $\mathbf{P}^k(\tau)$ (dashed blue line) versus those computed from an MSM constructed at that particular lag time $\mathbf{P}(k\tau)$ (solid black line). The good agreement between these two results provides numerical validation of the Markovian nature of the $\tau = 1.2$ ns lag time MSM.

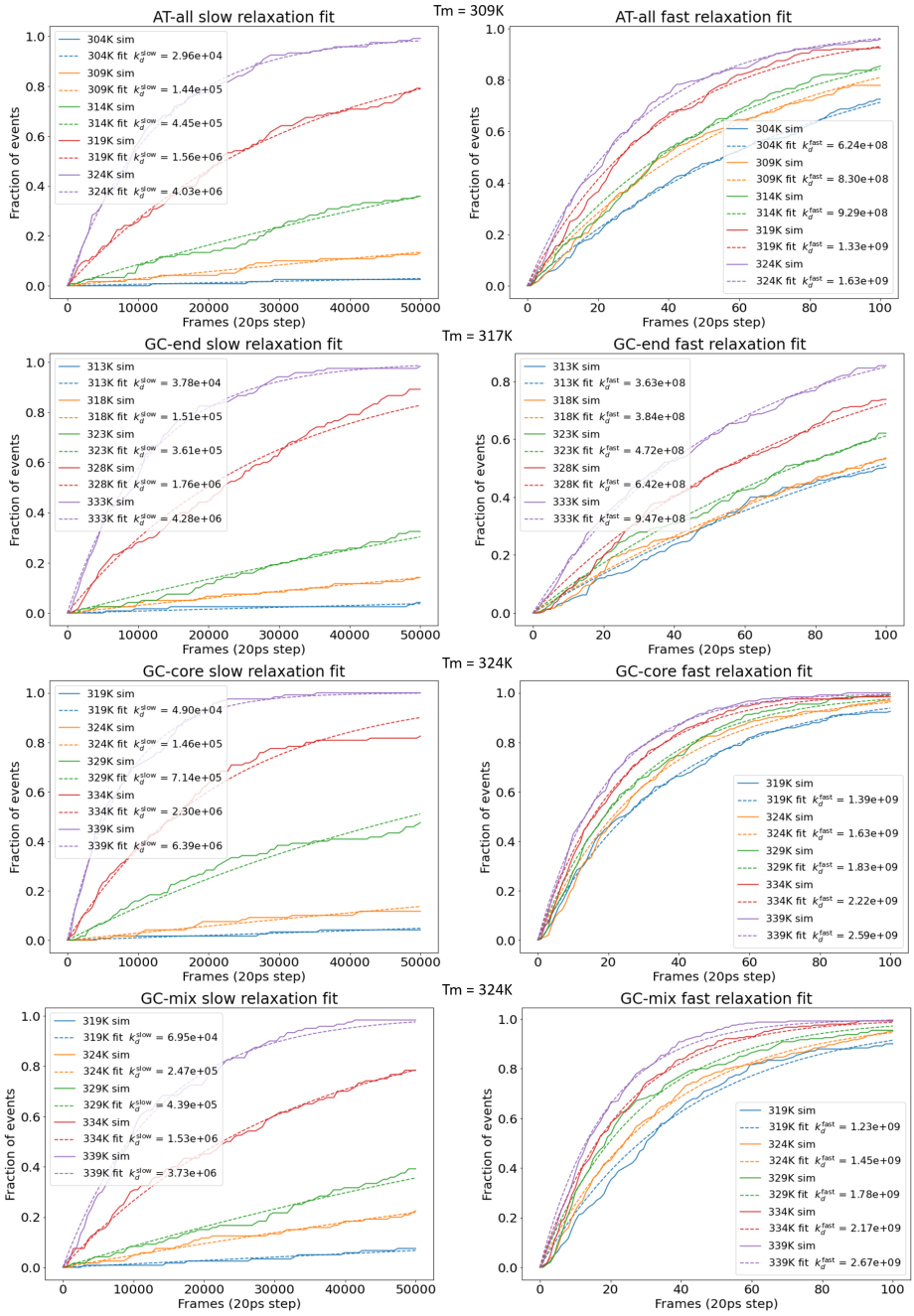
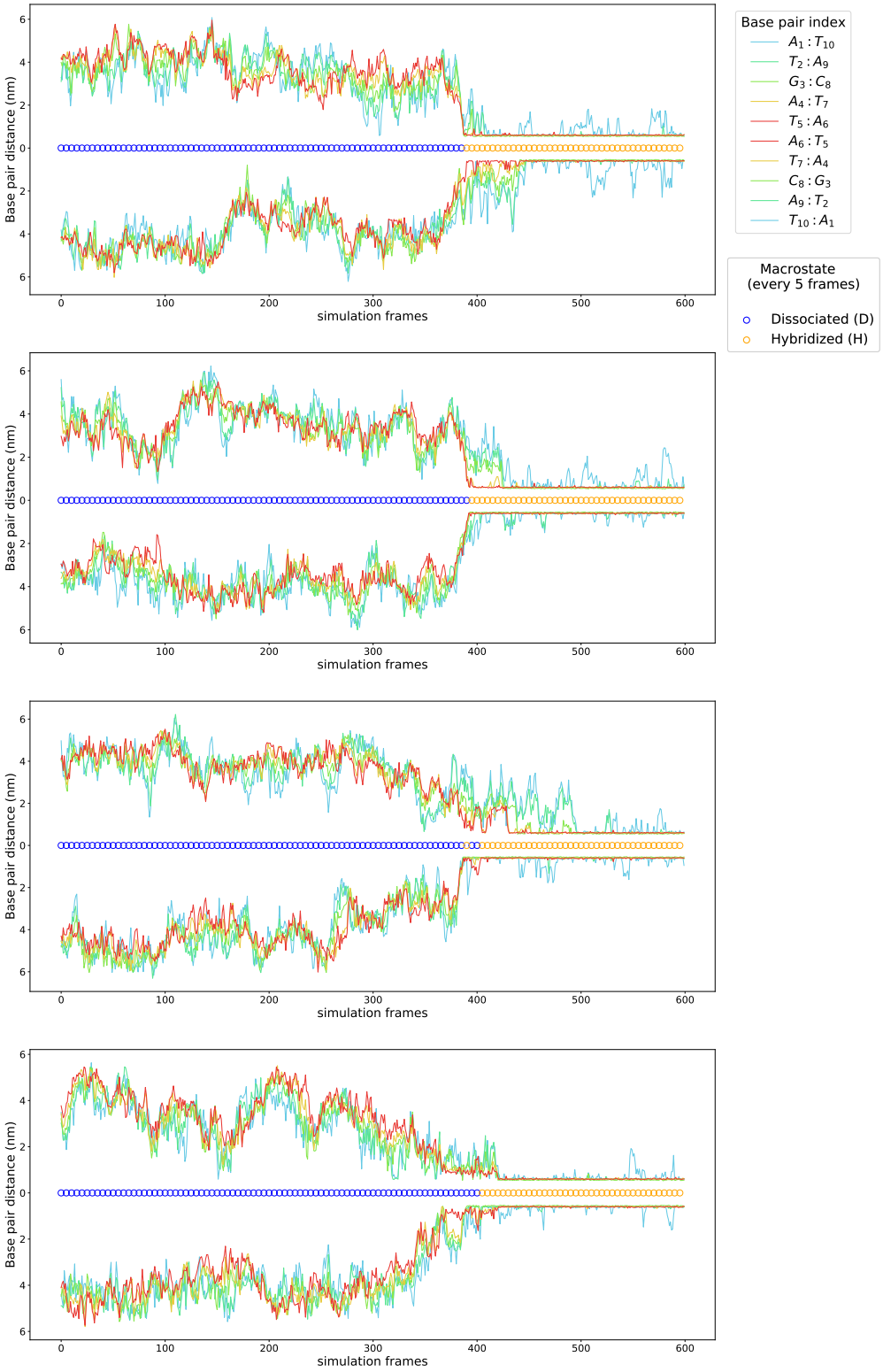


Figure S4: Exponential fits for both slow (dissociation) and fast (fraying) response for all four sequences during "computational T-jump" experiments. From 120 independent 1 μs simulations, we compiled the slow response data by recording the fraction of sequences with both central Watson-Crick base pairs intact as a function of time, and the fast response data as the fraction of sequences with both terminal Watson-Crick base pairs intact as a function of time. We define a Watson-Crick base pair to be intact if the two comlementary bases lie within a linear distance of 1.3 nm. We extracted our computational estimate of k_d^{fast} by fitting a decaying exponential to the fraction of bound A:T termini as a function of time $f_{\text{unfrayed}}(t) = \exp(-k_d^{\text{fast}} t)$. Similarly, we extracted our computational estimate of k_d^{slow} by fitting a decaying exponential to the fraction of hybridized sequences as a function of time $f_{\text{hybridized}}(t) = \exp(-k_d^{\text{slow}} t)$.


Base pair index
$A_1 : T_{10}$
$T_2 : A_9$
$G_3 : C_8$
$A_4 : T_7$
$T_5 : A_6$
$A_6 : T_5$
$T_7 : A_4$
$C_8 : G_3$
$A_9 : T_2$
$T_{10} : A_1$
Macrostate
(every 5 frames)
Dissociated (D)
Hybridized (H)
Base pair distance (nm)
simulation frames

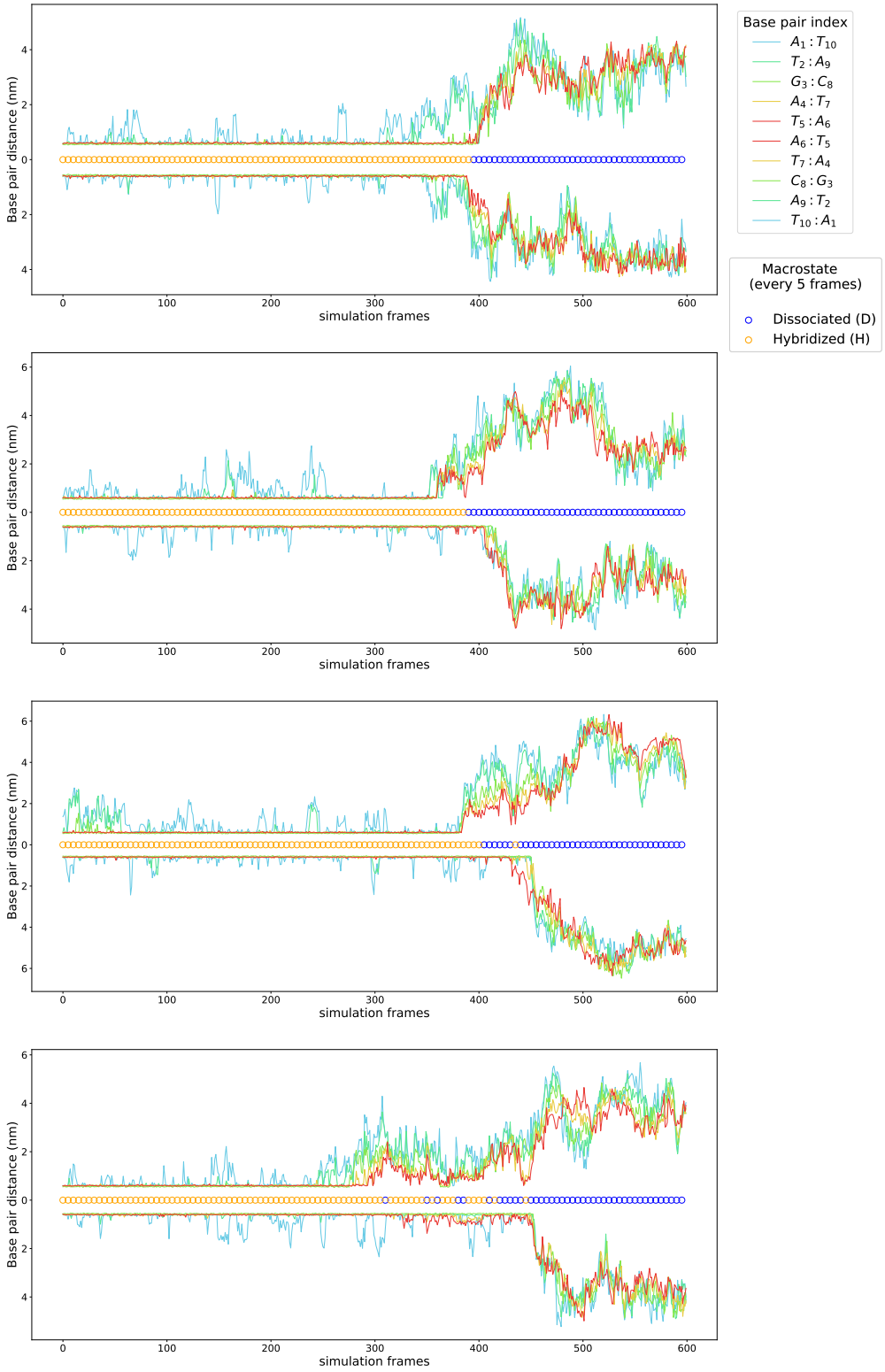

Base pair index
$A_1 : T_{10}$
$T_2 : A_9$
$G_3 : C_8$
$A_4 : T_7$
$T_5 : A_6$
$A_6 : T_5$
$T_7 : A_4$
$C_8 : G_3$
$A_9 : T_2$
$T_{10} : A_1$
Macrostate
(every 5 frames)
Dissociated (D)
Hybridized (H)
Base pair distance (nm)
simulation frames

Figure S5: GC-mix hybridizes by nucleation-zippering and dehybridizes by fraying-peeling. Tracking of the 10 intermolecular distances between native WC base pairs over the course of four additional hybridization and dissociation events. Symmetrically permutable distances (e.g., A_1:T_{10} and T_{10}:A_1) are reflected across the x-axis to avoid congestion in the plot. Circles superposed on the x-axis indicate the instantaneous MSM state assignment as dissociated D (blue) or hybridized H (orange). Hybridization tends to occur by a nucleation-zippering mechanism, wherein a native G:C WC pair and adjacent A:T pair or central A:T pairs first form prior to rapid formation of the complete duplex. Dehybridization tends to occur by a fraying-peeling mechanism wherein fraying of the two-base AT-tails on one or both sides of the duplex precedes dissociation of the central native base pairs and complete dissolution of the duplex.

Nearest neighbor model of duplex thermodynamics

DNA duplex hybridization thermodynamics are most commonly predicted using nearest-neighbor (NN) models where the stability of a given base pair is assumed to solely depend on the identity and orientation of its adjacent base pairs. We apply the NN model reported by SantaLucia[1] to estimate the stability of fully intact, shifted, and frayed duplex configurations observed in our Markov State Model.

The NN model assumes hybridization proceeds in a two-state all-or-nothing manner. The Gibbs free energy difference between the hybridized and dissociated states can be determined from calculation of NN enthalpy and entropy contributions,

$$\Delta G^{\circ}_{NN}(T) = \Delta H^{\circ}_{NN} - T\Delta S^{\circ}_{NN}. \tag{1}$$

ΔH°_{NN} and ΔS°_{NN} are assumed to be independent of temperature T, and are computed from the sum over all NN terms in a given configuration,

$$\Delta H^{\circ}_{NN} = \sum_{i}^{n_{NN}} \Delta H^{\circ}_{NN,i} + \sum_{j}^{n_{DE}} \Delta H^{\circ}_{DE,j} + \Delta H^{\circ}_{init}, \tag{2}$$

$$\Delta S^{\circ}_{NN} = \sum_{i}^{n_{NN}} \Delta S^{\circ}_{NN,i} + \sum_{j}^{n_{DE}} \Delta S^{\circ}_{DE,j} + \Delta S^{\circ}_{init} + \Delta S^{\circ}_{sym}. \tag{3}$$

$\Delta H^{\circ}_{NN,i}$ is the NN enthalpy for a given dinucleotide step within duplex DNA, $\Delta H^{\circ}_{DE,j}$ corresponds to the enthalpic contribution from a given dangling end (DE) base next to the duplex[2], and ΔH°_{init} is a duplex initiation term that accounts for terminal effects on duplex enthalpy. A schematic illustration of the NN and DE contributions for a fully hybridized duplex and an out-of-register shifted state is presented in Fig. S6. ΔS°_{NN} is constructed in an analogous manner from NN, dangling end, and initiation terms, plus a symmetry term, ΔS°_{sym}, that applies an entropic penalty to self-complementary sequences for maintaining C2 symmetry.

The SantaLucia NN parameters are defined for a solution condition at neutral pH with a sodium concentration of 1M. We use the empirical salt correction developed by Owczarzy[3] to determine the NN parameters at the sodium concentration used for simulations in this work. The Owczarzy model assumes the effect of counter ions on DNA hybridization thermodynamics to be purely entropic,

$$\Delta S^{\circ}_{NN}(c_{Na^+}) = S^{\circ}_{NN}(1M) + \Delta H^{\circ}_{NN}[(4.29 f_{GC} - 3.95) \times 10^{-5} \ln(c_{Na^+}) + 9.4 \times 10^{-6} (\ln(c_{Na^+}))^2], \tag{4}$$

where f_{GC} is the fractional G:C content of the duplex and c_{Na^+} is the concentration of sodium counter ions measured in M.

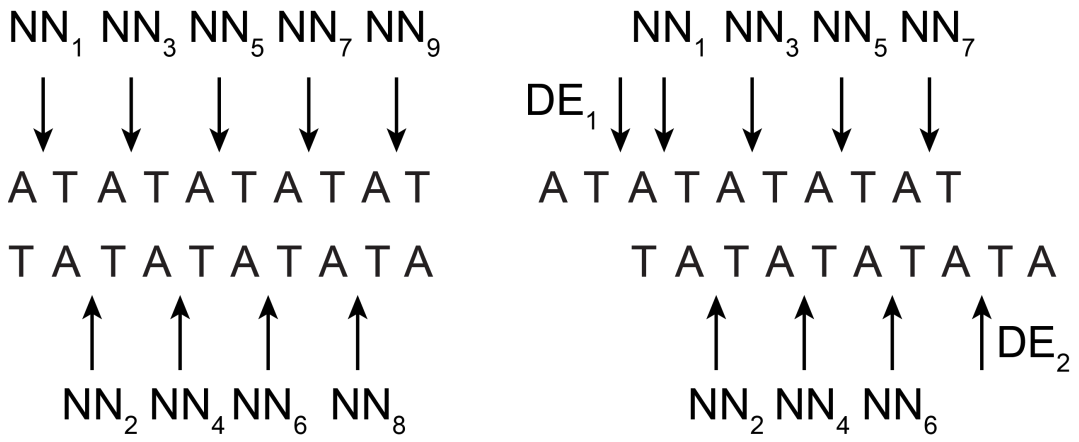


Figure S6: Schematic of nearest neighbor (NN) contributions for a fully hybridized AT-all duplex and an out-of-register duplex with dangling ends (DE).

References

(1) SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, *95*, 1460–1465.

(2) Santalucia, J.; Hicks, D. The thermodynamics of DNA structural motifs. *Annual Review of Biophysics and Biomolecular Structure* **2004**, *33*, 415–440.

(3) Owczarzy, R.; Moreira, B. G.; You, Y.; Behlke, M. A.; Wälder, J. A. Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry* **2008**, *47*, 5336–5353.