Electronic Supplementary Material (ESI) for Chemical Science.
This journal is © The Royal Society of Chemistry 2020

Supplementary Information for

Structure-Mechanics Statistical Learning Unravels the Linkage between Local Rigidity and Global Flexibility in Nucleic Acids

Yi-Tsao Chen[a], **Haw Yang**[b], **and Jhih-Wei Chu**[c]‡

[a] *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan 30068, ROC*
[b] *Department of Chemistry, Princeton University, Princeton, NJ 08544, USA*
[c] *Institute of Bioinformatics and Systems Biology, Department of Biological Science and Technology, and Institute of Molecular Medicine and Bioengineering, National Chiao Tung University, Hsinchu, Taiwan 30068, ROC*
‡ To whom correspondence should be addressed. Tel: +886 035 712121-56996; Email: jwchu@nctu.edu.tw

This PDF file includes:

Supporting Information Text

Details of all-atom MD simulations. Both systems of dsDNA and dsRNA are solvated in dodecahedron boxes of explicit water molecules with at least 10 Å between any nucleic atom and box edges. K^+ and Cl^- ions are added to achieve charge neutrality and ionic strength of 0.15 M. For the system of dsDNA, there are 60 K^+ and 30 Cl^-, and there are 56 K^+ and 26 Cl^- in the system of dsRNA. The resulting dsRNA system has 28555 atoms and the dsDNA system contains 32467 atoms. The cut-off radius for van dar Waals interactions and real-space particle-mesh Ewald terms of electrostatics [1] is 12 Å with a switching function effective at 10 Å. During the all-atom MD simulations, all bond lengths involving the hydrogen atom are constrained at the equilibrium values via LINCS [2]. After initial minimization and 12 ns equilibration period, the production run of 1 μs is conducted at constant temperature (310K) and pressure (1.013 bar) via the Langevin thermostat and the Parrinello-Rahman barostat [3]. A snapshot is saved every 100 ps for structural analysis and computing mechanical properties. To compute the order parameters Z_p [4] and χ-δ contour that indicate the A- and B-form contents in a nucleic acid structure, the 3DNA program [4] is employed.

Calculations of mechanical properties from all-atom MD simulations. The bending angle between tangent vectors along the chain is calculated to determine the persistence length. For each basepair i in dsDNA and dsRNA, the center position $\vec{r}_i$ is defined as the midpoint of the C6-C8 line [5]. Consecutive positions of $\vec{r}_i$ are thus a discrete representation of a worm-like chain. The vector for the i^{th} segment along the chain is $\vec{l}_i = \vec{r}_i - \vec{r}_{i-1}$. The angle θ_{ij} between $\vec{l}_i$ and $\vec{l}_j$ $(j > i)$ is a metric of bending deformation. For each configuration sampled in all-atom MD, the angle can be expressed as $\theta_{ij} = \overline{\theta}_{ij} + \delta\theta_{ij}$, where $\overline{\theta}_{ij}$ is the ensemble average and $\delta\theta_{ij}$ is the instantaneous angle fluctuation. Persistence length L_p is defined [6] as:

$$\langle \cos(\delta\theta_{ij}) \rangle = \exp\left(\frac{-(j-i)\bar{l}}{2L_p} \right). \quad (1)$$

Here, $\bar{l}$ is the averaged length between neighboring basepairs. For the 16-basepair dsDNA and dsRNA systems simulated here, the persistence lengths are calculated with $i = 4$ and $j = 13$. That is, the top and bottom three basepairs are excluded in calculating mechanical properties to avoid the fraying effects [7] that tend to occur at the ends. The values of persistence length calculated by using shorter base steps are shown in Figure S14.

To compute the local coordinate systems of bases and basepairs as well as to define the global helical axises for computing helical rises and twist angles between basepairs, the 3DNA program [4] is used. For the configurations sampled in all-atom simulations, the stretching deformation is measured via the contour length L, which is the sum of helical rises. The global twist Ω is the sum of helical twists between basepairs and is used to measure the twisting deformation. The two-by-two covariance matrix $\mathbf{C}$ of L and Ω is then calculated to determine the stretching modulus (η_{s}), twisting modulus (η_{s}), and twisting-stretching coupling (η_{ts}) [8]. The two-by-two modulus matrix $\mathbf{M}$ with η_{s} and η_{s} being the diagonal terms and η_{ts} being the off-diagonal term is inversely relate to $\mathbf{C}$ as $\mathbf{M} = L_0 k_{\text{B}} T \mathbf{C}^{-1}$. Here, L_0 is the total contour length averaged over the trajectory; k_{B} is the Boltzmann constant; and T is temperature. Calculations of L and Ω are conducted from $i = 4$ to 13 basepairs in our dsDNA and dsRNA systems. The values of stretching and twisting moduli by using shorter base pairs are shown in Figure S14.

Orthonormal expansion of order parameters for global deformations. The atomic coordinates sampled in a MD trajectory are aligned to obtain an averaged structure for computing the $3N \times 3N$ covariance matrix of positional fluctuations, where N is the number of heavy atoms. The quasi-harmonic analysis [9] considers the covariance of positional fluctuations in the Cartesian coordinate space and computes the eigenvalues λ_i and eigenvectors $\mathbf{e}_i$ via matrix diagonalization. The $3N$ QHA modes form a complete basis set in the Cartesian coordinate system, and six of which correspond to rigid-body translation and rotation with a zero eigenvalue. The rest $3N - 6$ vibrational modes can be used to express changes in molecular structures. The orthonormal basis can also be obtained by normal mode analysis [9] of the Hessian matrix of the heavy-atom elastic network model.

An order parameter Φ for measuring a deformation of nucleic acid can be expressed with vibrational modes in a linear approximation as:

$$\Phi(\mathbf{r}_j) = \Phi(\mathbf{R}) + \sum_{i=1}^{3N-6} c_{ji} \Phi'_i. \quad (2)$$

In this equation, $\mathbf{R}$ is the averaged coordinate of aligned structures and $\mathbf{r}_j$ is the j^{th} snapshot in an all-atom MD trajectory. The displacement with respect to the averaged structure represented using the orthonormal basis is calculated with $c_{ji} = (\mathbf{r}_j - \mathbf{R}) \cdot \mathbf{e}_i$ and the first derivative is $\Phi'_i = \nabla\Phi \cdot \mathbf{e}_i$. Here, $\nabla\Phi$ is the gradient vector of Φ in the Cartesian coordinate system. In this work, the explored order parameter included bending angle (θ), contour length (L), and global twist (Ω).

The variance of Φ in the trajectory data can be expressed by squaring eq 2 and averaging over the configurations to obtain contributions from each mode and mode couplings:

$$\sigma_\Phi^2 = \sum_{i=1}^{3N-6} \left(\Phi'_i\right)^2 \left\langle c_i^2 \right\rangle + 2 \sum_{i=1}^{3N-6} \sum_{j>i}^{3N-6} \left(\Phi'_i \Phi'_j\right) \left\langle c_i c_j \right\rangle. \quad (3)$$

Yi-Tsao Chen[a], Haw Yang[b], and Jhih-Wei Chu[c]‡

Ensemble averaged quantities are in angle brackets in the equation. Similarly, the covariance of two order parameters can be represented with vibrational modes. For twist-stretch coupling, the expression is:

$$\sigma^2_{\Omega L} = \sum_{i=1}^{3N-6} \left(\Omega'_i L'_i\right) \left\langle c_i^2 \right\rangle + 2 \sum_{i=1}^{3N-6} \sum_{j>i}^{3N-6} \left(\Omega'_i L'_j\right) \left\langle c_i c_j \right\rangle . \quad \textbf{(4)}$$

Computation of mode couplings terms from all-atom MD trajectories indicates that their values are orders of magnitudes smaller than the $\left\langle c_i^2 \right\rangle$ terms and hence are negligible. The gradient of an order parameter in the Cartesian coordinate space can be calculated analytically, if possible, to perform dot-product with vibrational modes for the Φ'_i terms. Alternatively, order parameter derivatives can be calculated numerically by taking the snapshots from the MD data, projecting along each mode, computing the order parameter from the projected coordinates, and calculating the variance of the order parameter along each mode. $|\Phi'_i|$ is then the square-root of the ratio of the order parameter variance to the valance of the projected displacement along each mode, and the sign of the derivative can be determined from the averaged values of the order parameter and its displacement along the mode. For bending angle θ, agreement of numerical and analytical calculations was observed. For contour length and global twist, the 3DNA procedure [4] involved numerical optimization, making analytical derivatives difficult to obtain. Therefore, the projection scheme desctibed above was used to numerically calculate the Φ'_i terms for L and Ω.

Heavy-atom elastic network model and its parametrization based on all-atom MD simulations. The probabilistic density function given by an elastic network model [10] of heavy atoms, haENM, at the temperature of MD simulation is employed to represent the mechanical coupling network in nucleic acids. All heavy atoms in dsDNA and dsRNA are incorporated to represent the structural topologies, and the potential energy function of haENM is:

$$V = \frac{1}{2} \sum_{m=1}^{M} k_m (b_m - b_m^0)^2. \quad \textbf{(5)}$$

Here, b_m^0 and k_m are the equilibrium bond lengths and spring constants of the m^{th} harmonic bond. The average bond lengths calculated from the trajectory data are taken as the equilibrium bond lengths of haENM. The vibrational partition function of haENM can be calculated via normal mode analysis (NMA) to determine the fluctuation of each spring, i.e., $\left\langle \delta b_m^2 \right\rangle_{\text{NMA}}$ [9]. The same quantity can also be calculated from the structures sampled in an all-atom MD trajectory, $\left\langle \delta b_m^2 \right\rangle_{\text{AA}}$. The harmonic spring constants are then calculated via an iterative scheme of fluctuation matching [11,12]:

$$k_m^{(n+1)} = k_m^{(n)} + \eta \left(\frac{1}{\langle \delta b_m^2 \rangle_{\text{NMA}}^{(n)}} - \frac{1}{\langle \delta b_m^2 \rangle_{\text{AA}}} \right). \quad \textbf{(6)}$$

The iterative step is (n) and η is a numerical learning factor. Bond length fluctuations $\left\langle \delta b_m^2 \right\rangle_{\text{AA}}$ were calculated from of the 1 μs trajectory data. A non-negative inequality constraint was imposed for each spring during fluctuation matching. Therefore, the list of non-zero k_m springs upon convergence define the mechanical coupling network in a nucleic acid structure.

Another key parameter in haENM is the cutoff distance R_c for including inter-atomic distances shorter than the cutoff in the list of springs. To determine R_c, a range of values (4-10 Å) is tested, and fluctuation-matching is applied to converge the list of spring constants for each R_c value. The variance of order parameter fluctuations at the temperature are then calculated using eq S3 and eq S4 to compare with the values calculated from all-atom MD simulations. Furthermore, mode by mode resemblance of the eigenvalues and eigenvectors of the fluctuation-matched haENM with those from quasi-harmonic analysis (QHA) of all-atom MD trajectories is also conducted for low-frequency modes 1-5 that contributed most to overall fluctuations.

The results are shown in Figure S12 and Figure S13. With the self-consistent iteration of fluctuation matching, consistent results are observed over the tested range of R_c values. The haENM is a simplified potential energy function and thus has limited capabilities in reproducing the quantitative values for all observables in all-atom MD simulations. The majored features for the flexibilities of global deformations, though, are robustly captured by the structural-mechanics with haENM, such as dsDNA is easier to bend and harder to stretch than dsRNA is, similarly flexible twist deformation, and the respective signs of twist-stretch couplings. Quantitative agreement can also be observed, albeit to different extents, in the flexibilities of different global deformations. Similar consistencies are also observed over the range of R_c values for comparing the eigenvalues and eigenvectors of vibrational modes calculated from haENM with those from all-atom MD. For haENM and all-atom MD, mode 1 similarities in both dsDNA and dsRNA exceeded 0.97 and the differences in the lowest eigenvalues are within 10 % for both biopolymers. Given similar performances in capturing the observables at the finer-grained scale, a smaller R_c is preferred as a few number of parameters are involved in the structural-mechanics statistical learning. As shown in Figure S12 and Figure S13, $R_c = 4.7$ Å is opted considering the performances over different observables. The insensitivity of haENM parameterization to the initial condition and the result of obtaining consistent behaviors with different R_c values highlight the robustness of our structural-mechanics statistical learning.

Profiling the compositions of local rigidities in structural changes. For a structural topology of N heavy atoms in haENM, the $3N \times 3N$ Hessian matrix $\mathbf{H}$ with spring constants converged after fluctuation matching can be represented via the $3N \times M$ Wilson's B matrix $\mathbf{B}$ [13] and the diagonal $M \times M$ matrix $\mathbf{K}$ that contains the M non-zero spring constants learned from the all-atom MD data as $\mathbf{H} = \mathbf{BKB}^T$. Furthermore, NMA of the haENM allows diagonalization of $\mathbf{H}$ into $3N \times 3N$ diagonal eigenvalue matrix $\mathbf{\Lambda}$ and $3N \times 3N$ orthogonal matrix $\mathbf{Q}$ of eigenvectors, i.e., $\mathbf{H} = \mathbf{Q\Lambda Q}^T$. Therefore, eigenvalues of the ENM Hessian can be expressed as $\mathbf{\Lambda} = \mathbf{Q}^T\mathbf{BKB}^T\mathbf{Q}$. The matrix components of $\mathbf{B}$ for the m^{th} spring are a vector containing partial derivatives of the inter-atomic distance with respect to the Cartesian coordinates of the two atoms a and b. Therefore, an eigenvalue λ_i of haENM comes from the combined effects of springs as:

$$\lambda_i = \sum_{m=1}^{M} k_m \left(\vec{s}_{ab}^{\,m} \cdot \vec{q}_{ab}^{\,i}\right)^2. \tag{7}$$

Here, $\vec{q}_{ab}^{\,i} = \vec{q}_b^{\,i} - \vec{q}_a^{\,i}$ comes from the vector components of eigenvector i that are associated with atoms a and b that spring m connects. The unit inter-atomic vector for spring m, $\vec{s}_{ab}^{\,m}$, arrives from the position vectors $\vec{r}_a$ and $\vec{r}_b$ of the two atoms:

$$\vec{s}_{ab}^{\,m} = \frac{\vec{r}_b - \vec{r}_a}{|\vec{r}_b - \vec{r}_a|}. \tag{8}$$

Along the particular mode in the $3N$-dimension, the level by which spring m that links atoms a and b is perturbed can thus be quantified as $\left(\vec{s}_{ab}^{\,m} \cdot \vec{q}_{ab}^{\,i}\right)^2$. Contributions from different spring groups to an eigenvalue discussed in the main text are calculated based on eq S7 by summing over the springs in each group.

Yi-Tsao Chen[a], Haw Yang[b], and Jhih-Wei Chu[c]‡

$\mathbf{e}_i^{\mathrm{LU}} \cdot \mathbf{e}_i^{\mathrm{haENM}}$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
dsDNA	0.834	0.834	0.959	0.142	0.133
dsRNA	0.940	0.963	0.688	0.672	0.226

Table S1. Comparison of locally uniform haENM with fluctuation-matched haENM to illustrate the resilience of vibrational modes to chemical details.

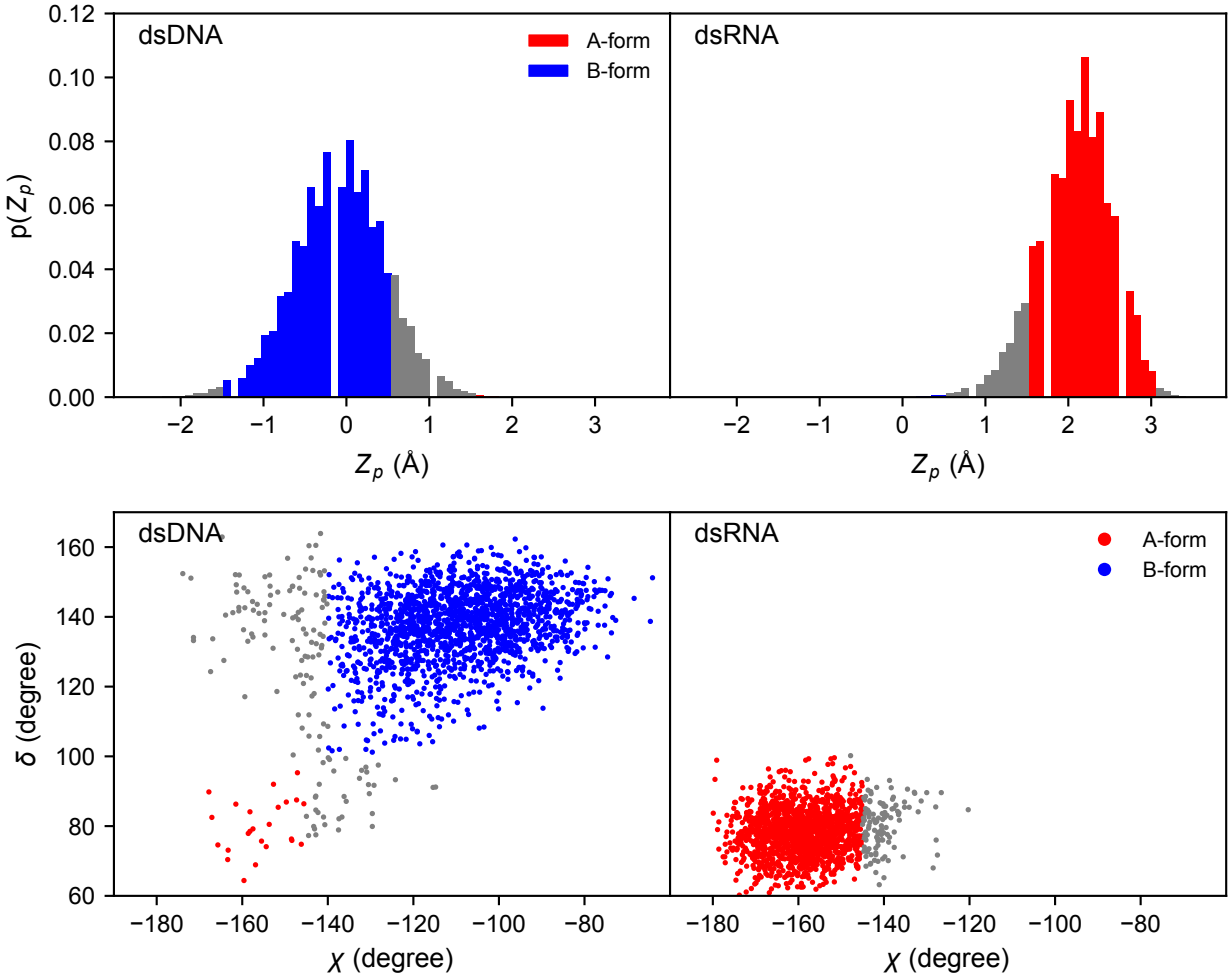

Fig. S1. The structural characterization of dsRNA and dsDNA in all-atom MD simulations. To avoid the fraying effect of the two ends along a nucleic acid strand, only the 10 central basepairs were used to compute the Z_p histograms and χ-δ contour[4,14]. The top panels show the histograms of Z_p in dsDNA and dsRNA simulations. The A-form structure is characterized by the value of Z_p in the range $1.5\,\text{Å} \leq\ Z_p \leq 3\,\text{Å}$, as indicated by the red region. The range of Z_p for the B-form structure is $-1.5\,\text{Å} \leq\ Z_p \leq 0.5\,\text{Å}$, as indicated by the blue region. In the bottom panels, the χ-δ contours for dsRNA and dsDNA simulations are shown. Red points are representative of A-form structures, and blue points are representative of B-form structures.

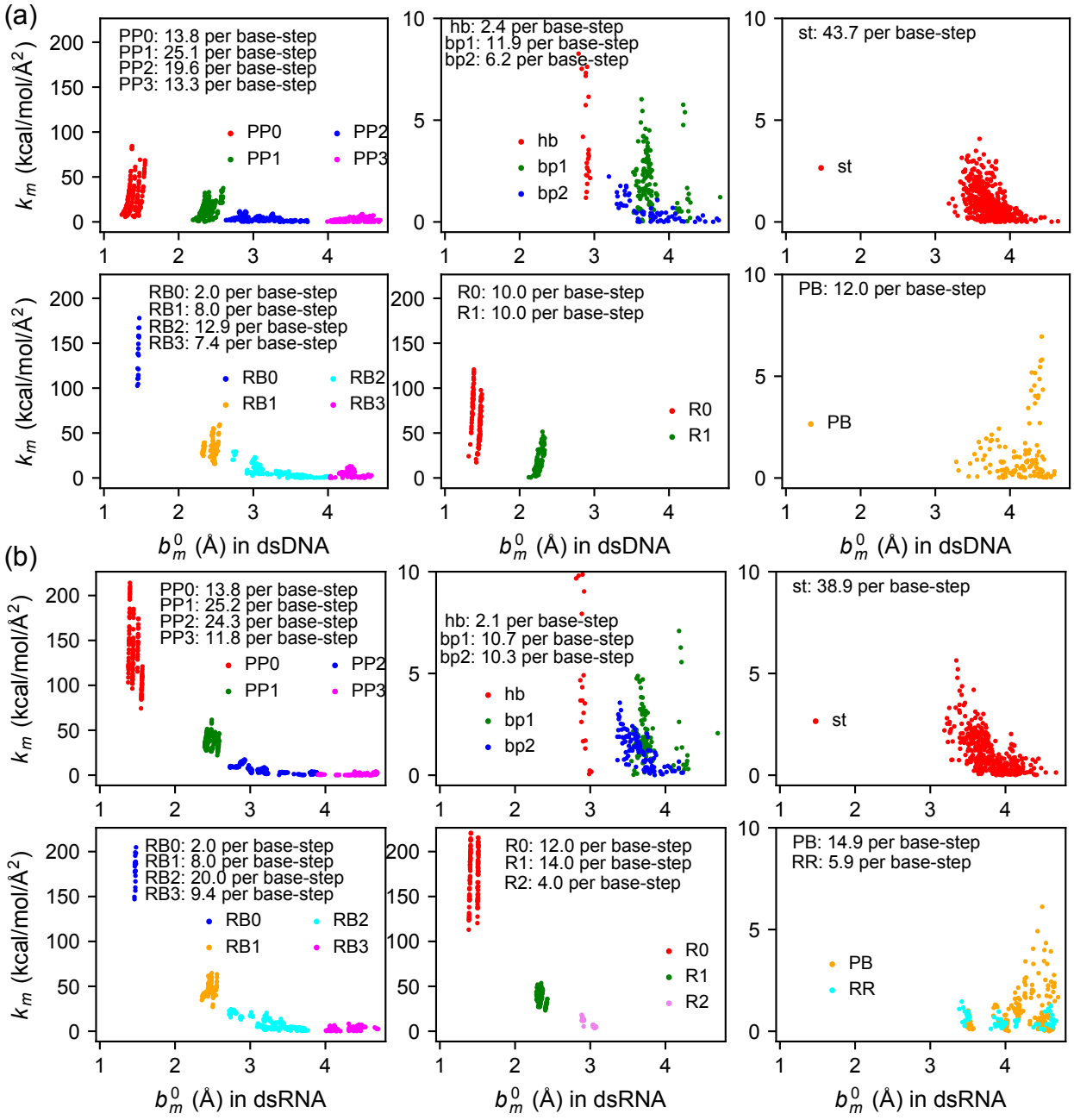

Fig. S2. The spring constant k_m in (kcal/mol/Å^2) versus equilibrium length b_m^0 in (Å) of a spring m in the haENM after fluctuation matching calculations for (a) dsDNA and (b) dsRNA. The sub-groups of springs are described as in Figure 2 of the main-text. The averaged number of springs per nucleotide for each sub-group is also labeled.

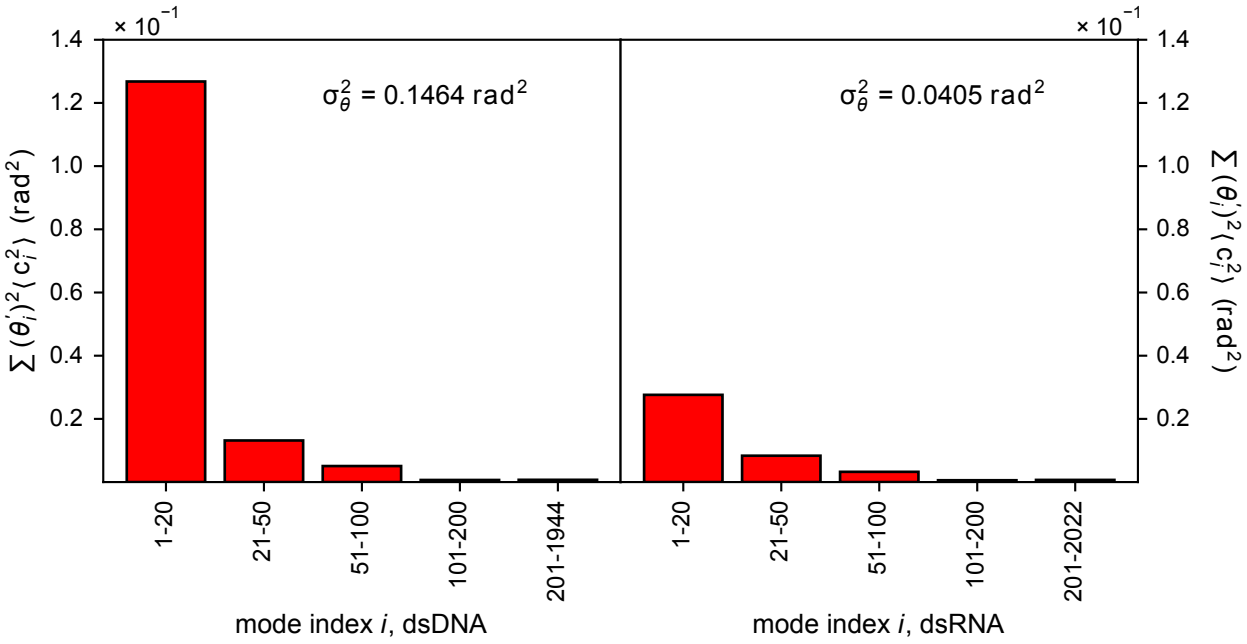

Fig. S3. Orthonormal expansion of bending-angle flexibility σ_θ^2 in the all-atom MD simulations of dsDNA and dsRNA for all of the vibrational modes.

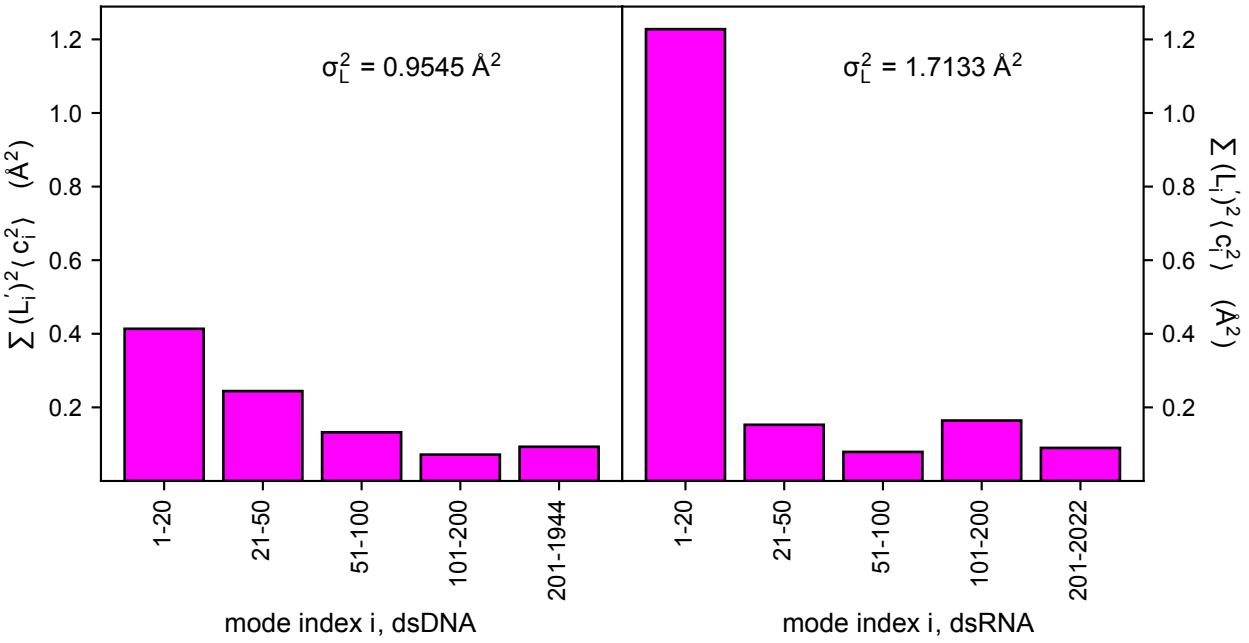

Fig. S4. Orthonormal expansion of contour-length flexibility σ_L^2 in the all-atom MD simulations of dsDNA and dsRNA for all of the vibrational modes.

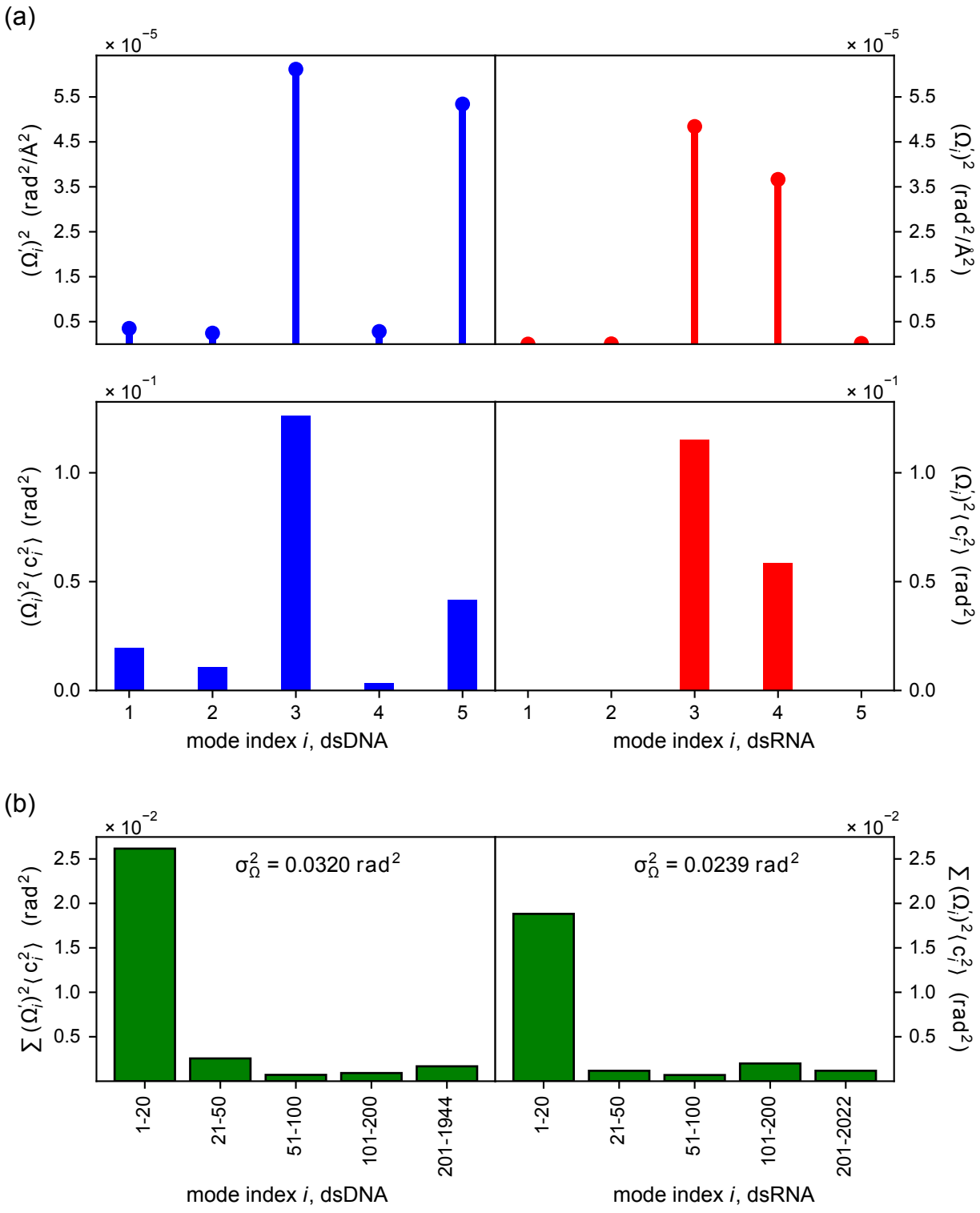

Fig. S5. Orthonormal expansion of twist-angle flexibility σ_Ω^2 in the all-atom MD simulations of dsDNA and dsRNA for (a) first five vibrational modes and (b) all of the vibrational modes.

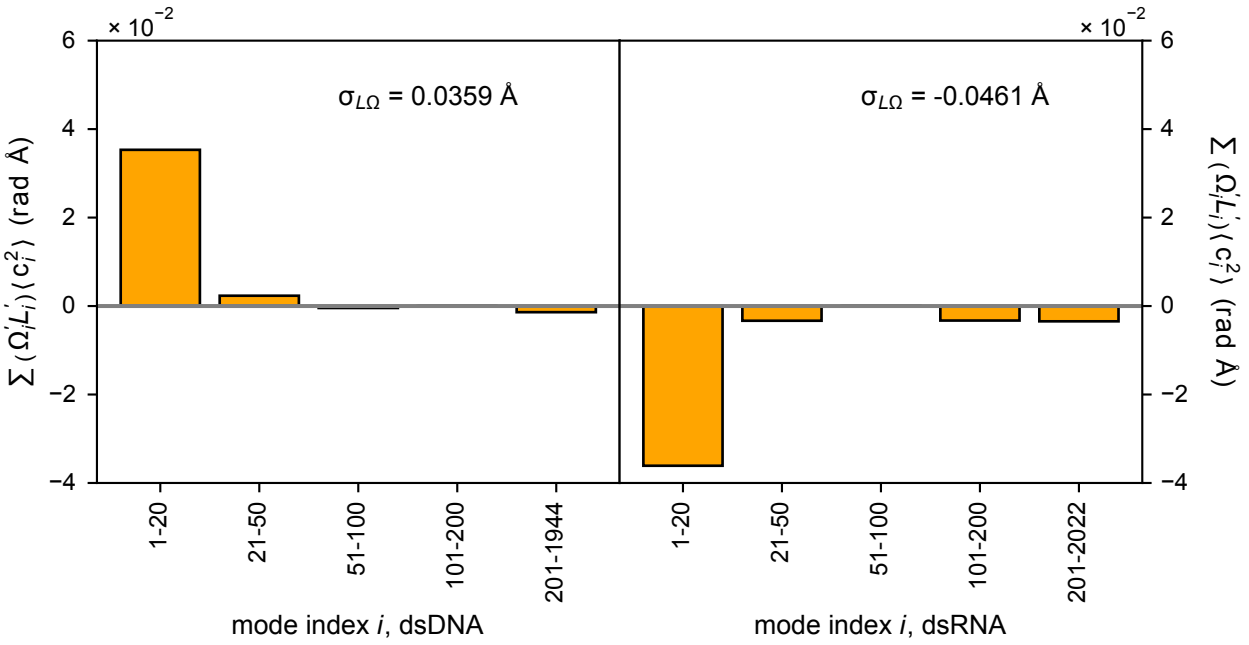

Fig. S6. Orthonormal expansion of twist-stretch coupling $\sigma_{\Omega L}$ in the all-atom MD simulations of dsDNA and dsRNA for all of the vibrational modes.

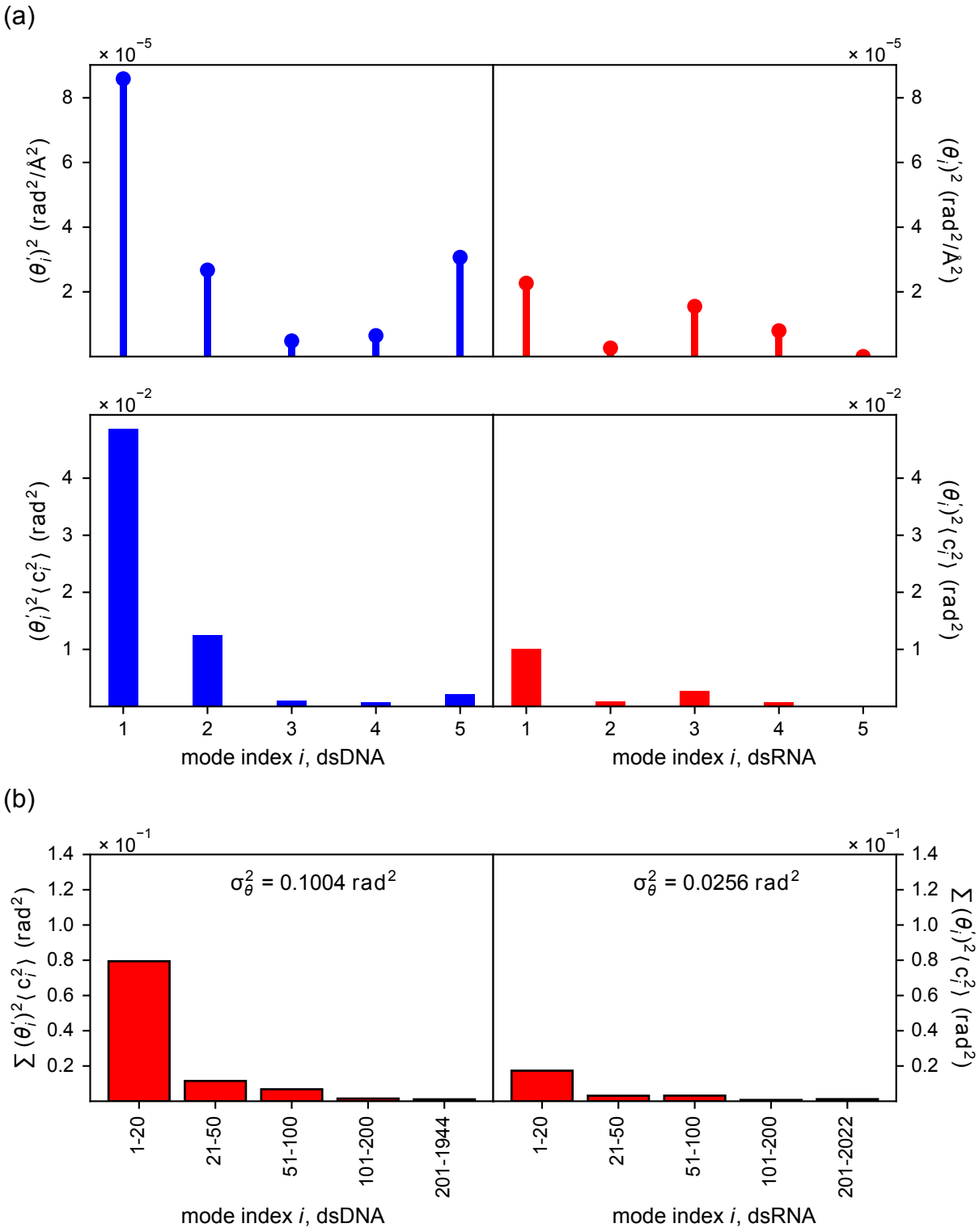

Fig. S7. Orthonormal expansion of bending-angle flexibility σ_θ^2 in the haENM of dsDNA and dsRNA for (a) first five vibrational modes and (b) all of the vibrational modes.

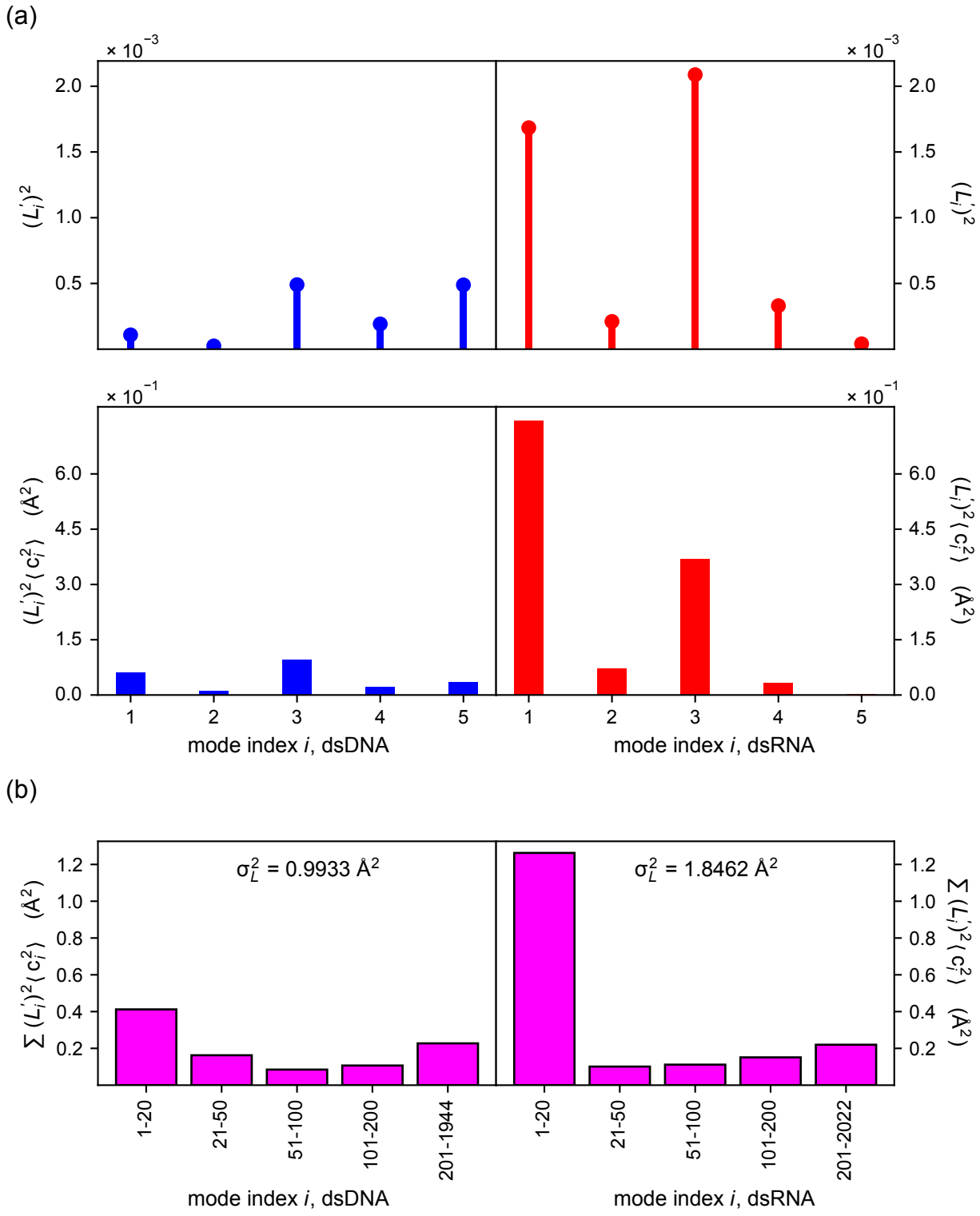

Fig. S8. Orthonormal expansion of contour-length flexibility σ_L^2 in the haENM of dsDNA and dsRNA for (a) first five vibrational modes and (b) all of the vibrational modes.

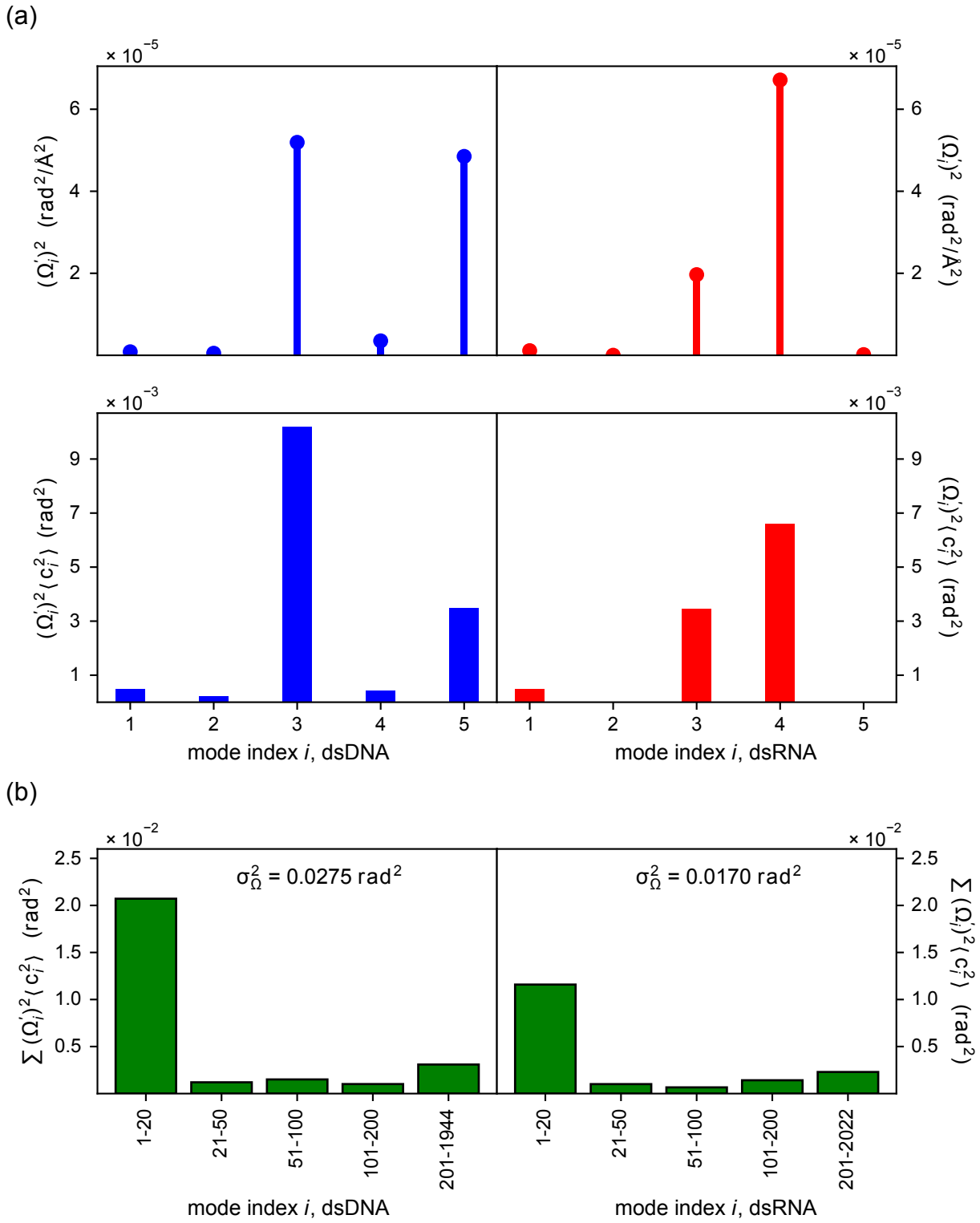

Fig. S9. Orthonormal expansion of twist-angle flexibility σ_Ω^2 in the haENM of dsDNA and dsRNA for (a) first five vibrational modes and (b) all of the vibrational modes.

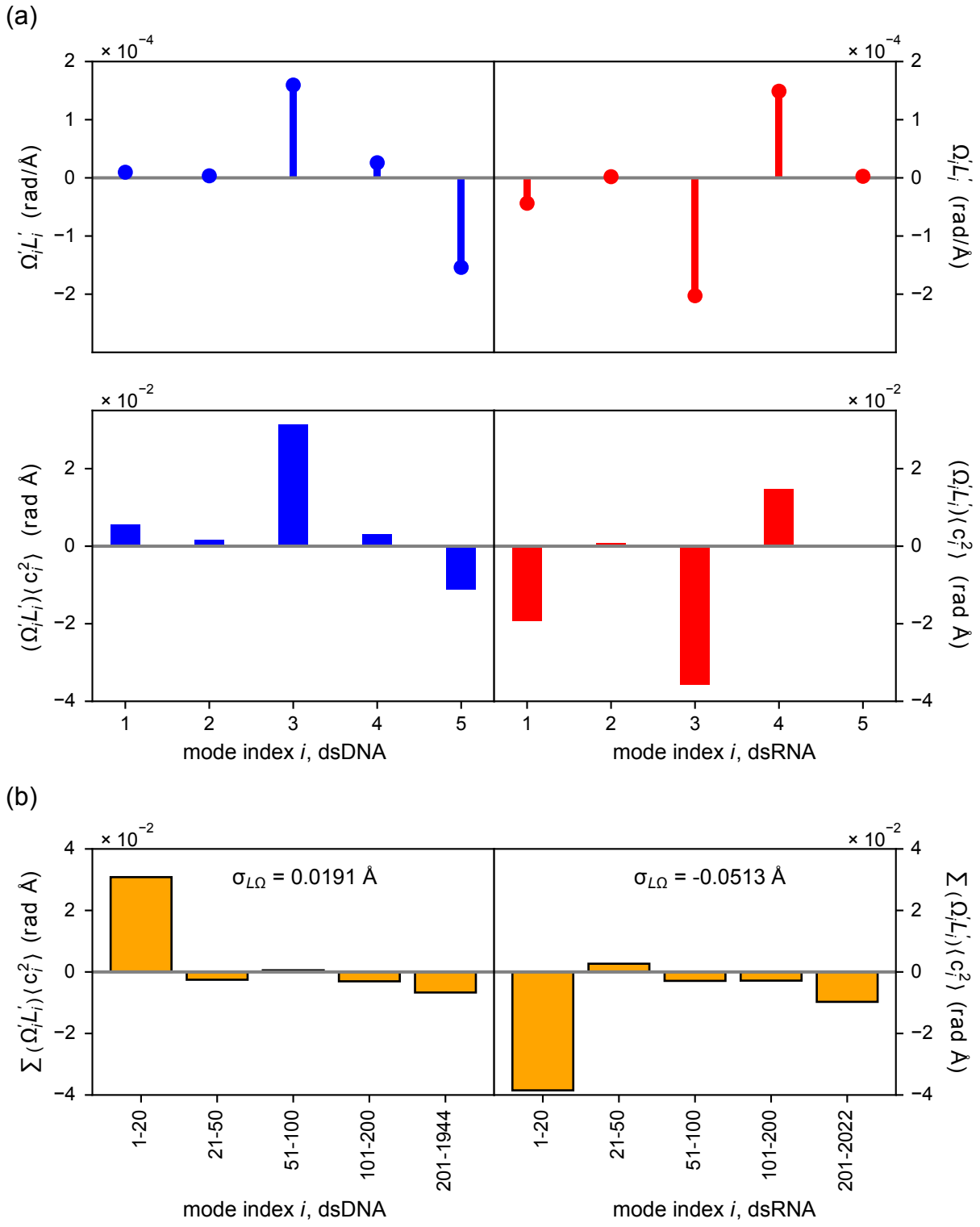

Fig. S10. Orthonormal expansion of twist-stretch coupling $\sigma_{\Omega L}$ in the haENM of dsDNA and dsRNA for (a) first five vibrational modes and (b) all of the vibrational modes.

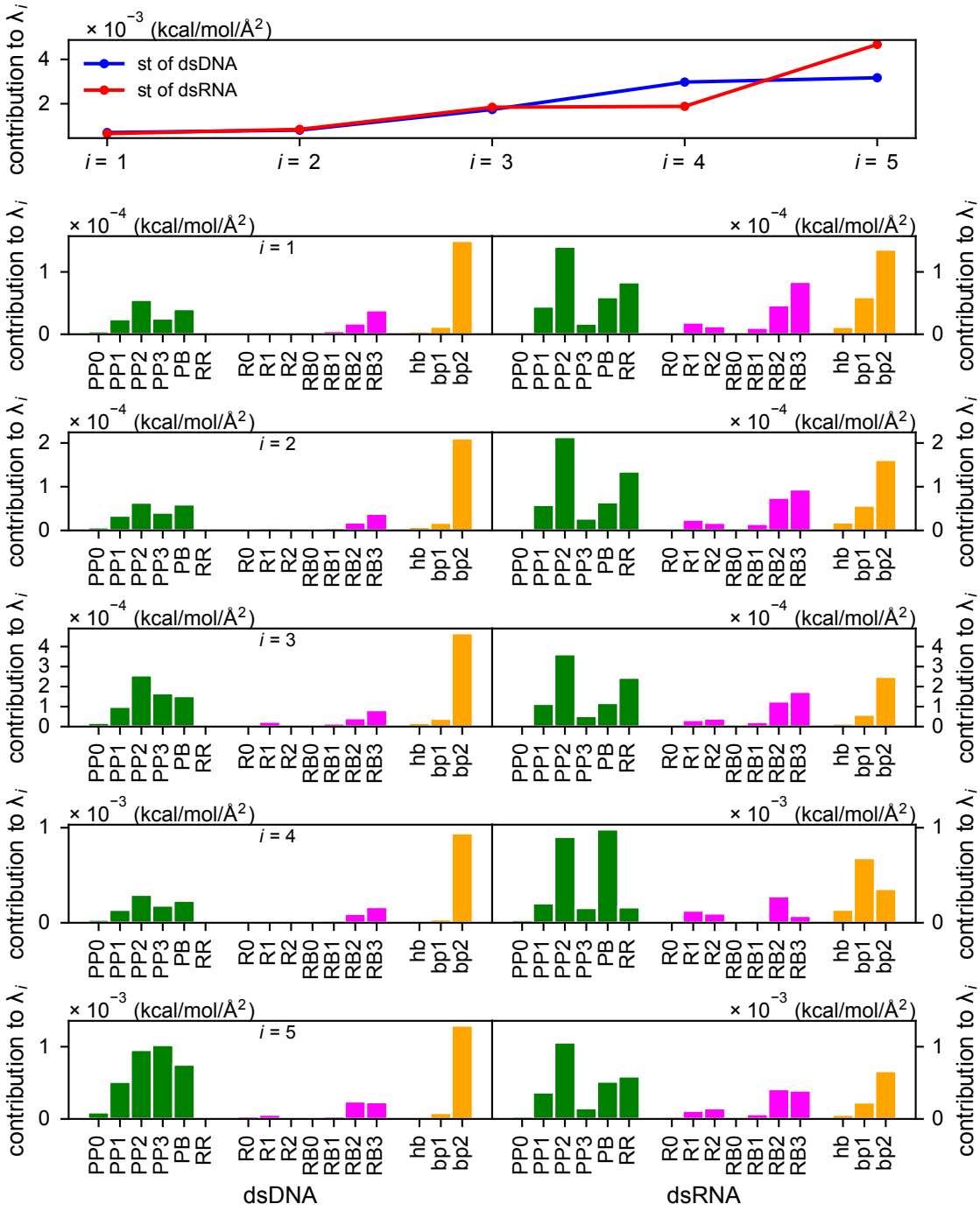

Fig. S11. Compositions of local rigidities in different sub-groups for the low-frequency modes of dsDNA and dsRNA. Top panel: The contribution from the st sub-group, i.e., base-stacking springs. Bottom panels: Contributions from springs in other sub-groups, one panel for each mode.

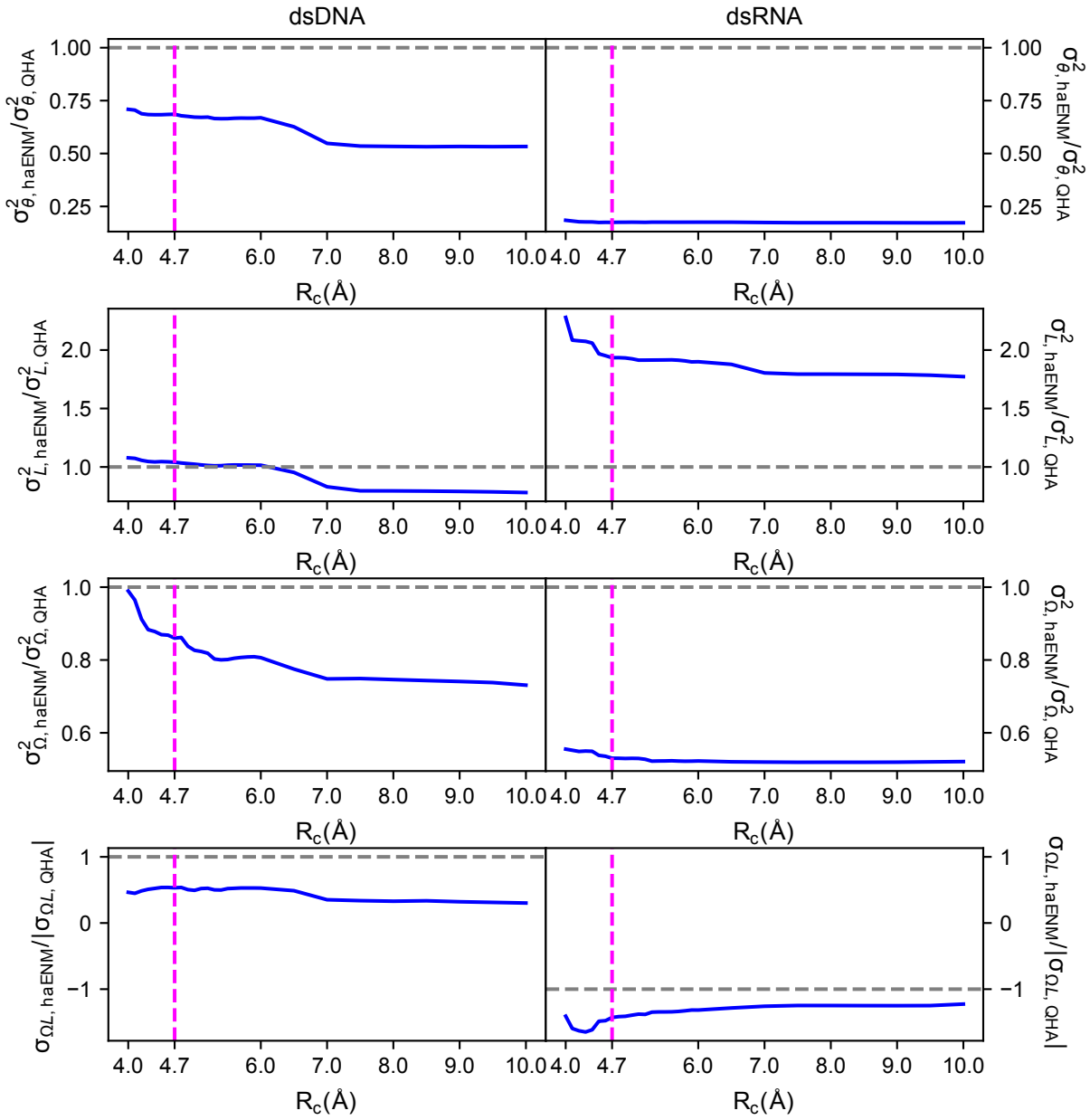

Fig. S12. Comparison of order parameter fluctuations of global shapes between the results from normal mode analysis (NMA) of fluctuation-matched haENM and those from quasi-harmonic analysis (QHA) of all-atom MD trajectories. The ratios of haENM flexibilities to those from QHA of different order parameters are shown for both dsDNA and dsRNA as a function of the cutoff radius $\mathrm{R_c}$ in (Å) for including an inter-atomic restraint in the haENM.

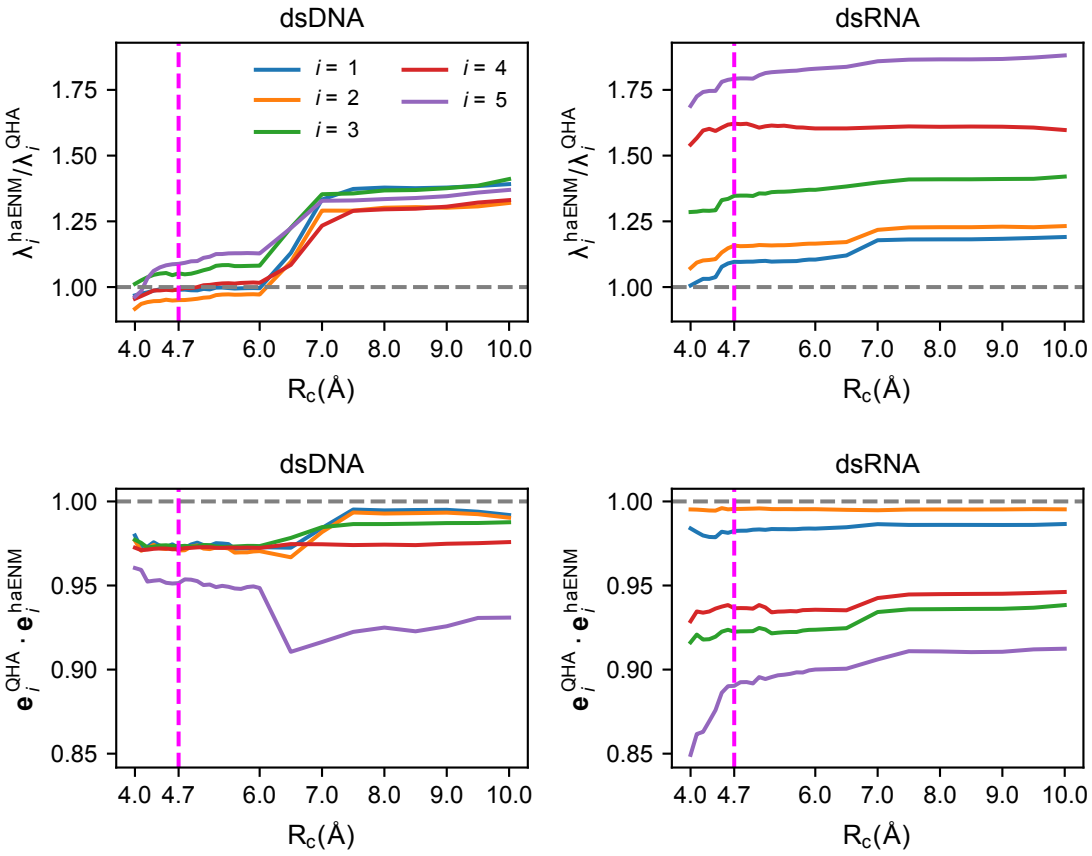

Fig. S13. Comparison of the eigenvalues and eigenvectors from NMA of fluctuation-matched haENM with those from QHA of all-atom MD trajectories. The ratios of haENM eigenvalues to those from QHA and dot-products of their eigenvectors are shown for both dsDNA and dsRNA as a function of the cutoff radius $\mathrm{R_c}$ in (Å) for including an inter-atomic restraint in the haENM.

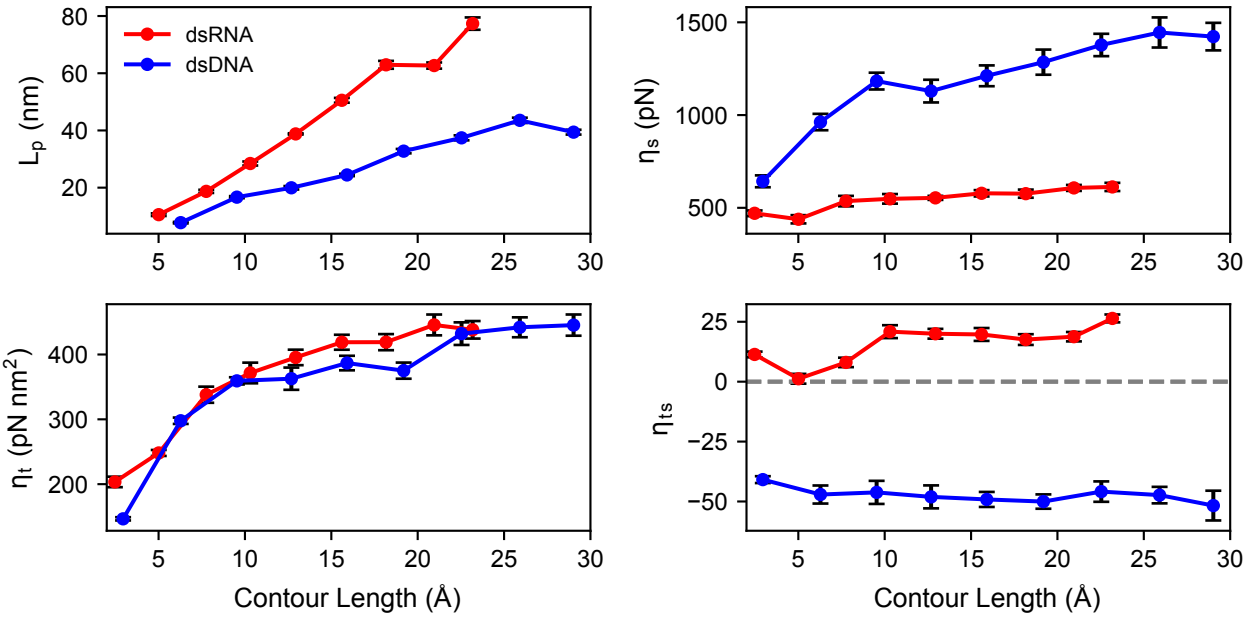

Fig. S14. The values of persistence length, moduli of stretching and twisting, and twist-stretch coupling calculated by using segments of different lengths. The persistence length (L_{p}) is calculated with $i = 4$ and $j = 5$,$\cdots$, 13 by using eq 1. For stretch modulus (η_{s}), twist modulus (η_{t}), and twisting-stretching coupling (η_{ts}), the 3DNA program[4] and the method described in the text are used for the calculation of one basepair ($i = 4$ to 5) to ten basepairs ($i = 4$ to 13). The corresponding contour length of dsDNA and dsRNA is based on the average structures of MD simulations.

References

1. T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089.
2. B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.
3. M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.
4. X. J. Lu and W. K. Olson, *Nucleic Acids Res.*, 2003, **31**, 5108–5121.
5. M. A. El Hassan and C. R. Calladine, *J. Mol. Biol.*, 1995, **251**, 648–664.
6. J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton*, Sinauer Associates, 2001.
7. Y. Y. Wu, L. Bao, X. Zhang and Z. J. Tan, *J. Chem. Phys.*, 2015, **142**, 125103.
8. K. Liebl, T. Drsata, F. Lankas, J. Lipfert and M. Zacharias, *Nucleic Acids Res.*, 2015, **43**, 10143–10156.
9. B. R. Brooks, D. Janežič and M. Karplus, *J. Comput. Chem.*, 1995, **16**, 1522–1542.
10. M. M. Tirion, *Phys. Rev. Lett.*, 1996, **77**, 1905–1908.
11. J. W. Chu and G. a. Voth, *Biophys. J.*, 2006, **90**, 1572–1582.
12. J. Silvestre-Ryan, Y. Lin and J. W. Chu, *PLoS Comput. Biol.*, 2011, **7**, e1002023.
13. E. B. Wilson, J. C. Decius and P. C. Cross, *Molecular Vibrations : The Theory of Infrared and Raman Vibrational Spectra*, New York : McGraw-Hill, 1955.
14. X. J. Lu, Z. Shakked and W. K. Olson, *J. Mol. Biol.*, 2000, **300**, 819–840.