Supplementary material for "*cgDNA*: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA" by D. Petkevičiūtė, M. Pasi, O. Gonzalez and J.H. Maddocks

Nucleic Acids Research, 2014
doi: 10.1093/nar/gku825

S1 Extended description of the rigid base model (Supplement to Section 2.1: Model)

As described in the main article, we consider the coarse-grain model of DNA (originally presented in [S1]) in which each base on each strand is modeled as an independent rigid body. We here present a complete but non-mathematical description of this model. Choosing one strand as a reference, a DNA molecule consisting of n base pairs is identified with a sequence of bases $\mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$, listed in the $5'$ to $3'$ direction along the strand, where $\mathtt{X}_a \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$. The base pairs associated with this sequence are denoted by $(\mathtt{X},\overline{\mathtt{X}})_1, (\mathtt{X},\overline{\mathtt{X}})_2, \ldots, (\mathtt{X},\overline{\mathtt{X}})_n$, where $\overline{\mathtt{X}}$ is defined as the Watson-Crick complement of $\mathtt{X}$ as illustrated in Figure S1. In the model, the backbones themselves are not considered explicitly; only the position and orientation of each base is determined by the conformational variables, as shown in Figure S2.

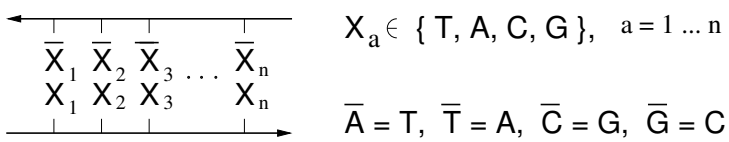


Figure S1: *Identification of DNA bases.* $\mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$ *denote bases on the reference strand, listed in the* $5'$ *to* $3'$ *direction, and* $\overline{\mathtt{X}}_1\overline{\mathtt{X}}_2\cdots\overline{\mathtt{X}}_n$ *are the corresponding bases on the complementary strand. Each base has five nearest neighbours, for example,* $\mathtt{X}_2$ *has nearest neighbours* $\mathtt{X}_1, \mathtt{X}_3, \overline{\mathtt{X}}_1, \overline{\mathtt{X}}_2, \overline{\mathtt{X}}_3$.

The three-dimensional configuration of a DNA molecule is determined by the relative rotation and displacement between neighboring bases both across and along the two strands. To this end, we use the *Curves+* [S2] implementation of the Tsukuba convention [S3] to assign a reference frame to each base $\mathtt{X}_a$ and each complementary base $\overline{\mathtt{X}}_a$. As illustrated in Figure S3, we then define a reference frame for each base pair $(\mathtt{X},\overline{\mathtt{X}})_a$ via an appropriate average of the two base frames, and also define a reference frame for the junction between each pair of base pairs $(\mathtt{X},\overline{\mathtt{X}})_a$ and $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$ via an appropriate average of the two base-pair frames [S4]. The relative rotation and displacement between the bases $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$ across the strands (see Figures S3.a and S4.b) is then described in the associated base-pair frame by an intra-base-pair coordinate vector y^a with six entries comprising three rotation coordinates (Buckle-Propeller-Opening) and three displacement coordinates (Shear-Stretch-Stagger). Similarly, the relative rotation and displacement between the base pairs $(\mathtt{X},\overline{\mathtt{X}})_a$ and $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$ along the strands (see Figures S3.b and S4.c) is described in the associated junction frame by an inter-base-pair coordinate vector z^a with six entries comprising three rotation coordinates (Tilt-Roll-Twist) and three displacement coordinates (Shift-Slide-Rise). For a molecule of n base pairs, there are therefore a total of n intra-base-pair coordinate vectors y^a ($a = 1, \ldots, n$) and a total of $n-1$ inter-base-pair coordinate vectors z^a ($a = 1, \ldots, n-1$), cf. Figure S4.a. The collection

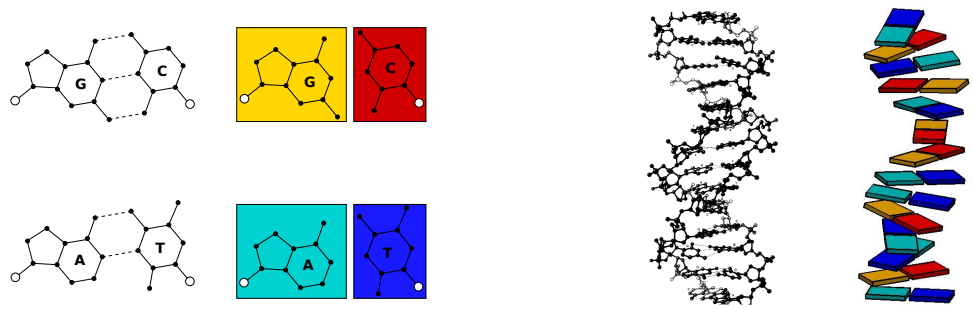


Figure S2: *In our coarse grain, rigid base, model of DNA a configuration of a molecule is described by the relative positions and orientations of all of its bases, the atoms of which imply a rigid body position and orientation. The atoms forming the sugar-phosphate backbones are not explicitly included in the conformational variables.*

of all intra- and inter-base-pair coordinates is denoted by $\mathsf{w} = (y^1, z^1, y^2, z^2, \ldots, z^{n-1}, y^n)$, which has a total of $N = 12n - 6$ entries.

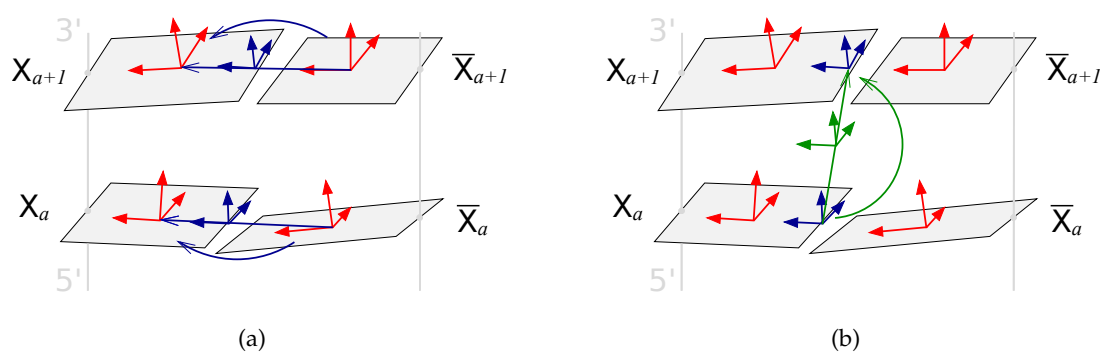
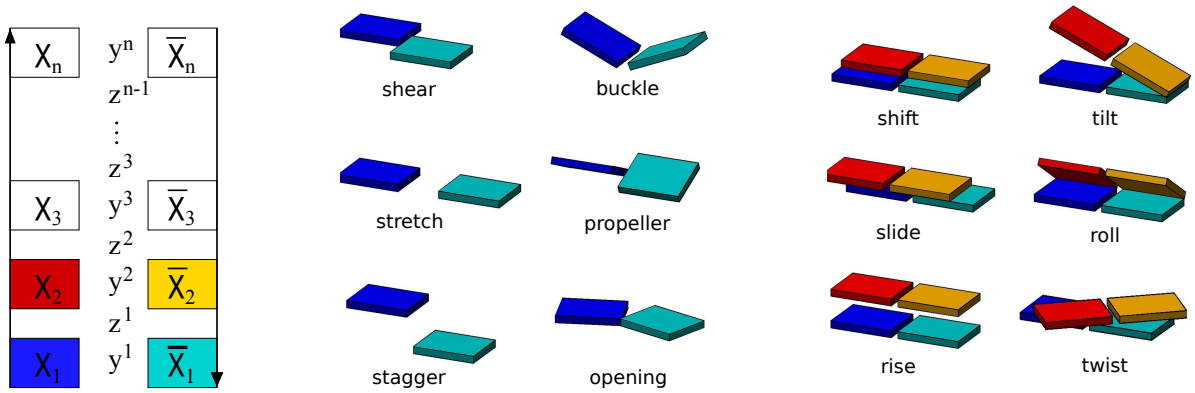


Figure S3: *A schematic illustration of the base (red), base-pair (blue), and junction (green) frames for an isolated base-pair step (or equivalently dinucleotide or junction). Note that the Tsukuba convention requires that the choice of the orientation of the base normal be approximately parallel to the 5′-3′ backbone direction on the Watson strand along which the sequence* S *is read (here on the left). Each of the seven illustrated frames is made up of a right-handed orthonormal triad* $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$ *where, by convention, it is always the third vector* $\mathbf{d}_3$ *of each triad that is approximately perpendicular to the base planes and parallel to the 5′-3′ Watson backbone, the second vectors* $\mathbf{d}_2$ *are oriented from the Crick strand toward the Watson strand, and consequently the* $\mathbf{d}_1$ *vectors point into the major groove.*

To a DNA molecule with n base pairs and sequence $\mathsf{S} = \mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$ we associate a free energy of the form

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}[\mathsf{w} - \widehat{\mathsf{w}}(\mathsf{S})] \cdot \mathsf{K}(\mathsf{S})[\mathsf{w} - \widehat{\mathsf{w}}(\mathsf{S})] + \widehat{\mathsf{U}}(\mathsf{S}), \tag{S1}$$

where $\mathsf{K}(\mathsf{S})$ is a symmetric, positive-definite matrix of size $N \times N$ of stiffness parameters, $\widehat{\mathsf{w}}(\mathsf{S})$ is a vector of size N of coordinates that define the ground or minimum energy state, and $\widehat{\mathsf{U}}(\mathsf{S})$ is a constant, depending on an arbitrary choice of reference zero for the energy, which will play no role in our current developments. The ground-state coordinate vector $\widehat{\mathsf{w}}(\mathsf{S})$ and stiffness matrix $\mathsf{K}(\mathsf{S})$ are completely determined by the sequence S through a construction rule based on local interaction energies [S1, S5]. Specifically, we consider two such energies: a mononucleotide, or intra-base-pair, interaction energy U_m for each base pair along the molecule, and a dinucleotide interaction energy U_d for each base-pair step. As illustrated below, each interaction energy U_m

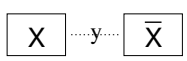


(a) (b) intra-base-pair coordinates, y^a (c) inter-base-pair coordinates, z^a

Figure S4: *A DNA conformation coordinate vector (on the left) comprises sextuplets of intra-base-pair y^a and inter-base-pair z^a coordinates of rigid base DNA, as shown schematically on the right. Note that, as shown in Figure S3, the intra coordinates are components in the associated base-pair frame (blue in Figure S3) which describe the transformation from the Crick-base frame to the Watson-base frame, while the inter coordinates are components in the associated junction frame (green in Figure S3) which describe the transformation between adjacent base-pair frames in the Watson 5′-3′ direction. With the convention that the Watson-strand bases are drawn on the left throughout, the four deformations illustrated in the top rows of panels b) and c) all correspond to positive coordinates with respect to the appropriate $\mathbf{d}_1$ axis, the four in the middle rows to positive coordinates with respect to the appropriate $\mathbf{d}_2$ axis, and the four in the bottom rows to positive coordinates with respect to the appropriate $\mathbf{d}_3$ axis.*

is a shifted quadratic function of the mononucleotide coordinate vector w_m, and is defined by the mononucleotide parameters $\widehat{\mathsf{w}}_m^{\mathtt{X}}$ and $\mathsf{K}_m^{\mathtt{X}}$ which depend on the mononucleotide sequence $\mathtt{X}$. Similarly, each interaction energy U_d is a shifted quadratic function of the dinucleotide coordinate vector w_d, and is defined by the dinucleotide parameters $\widehat{\mathsf{w}}_d^{\mathtt{XY}}$ and $\mathsf{K}_d^{\mathtt{XY}}$ which depend on the dinucleotide sequence $\mathtt{XY}$. Notice that the configuration of a mononucleotide is determined by the associated intra-base-pair coordinate vector y, whereas the configuration of a dinucleotide is determined by the associated up-stream intra-base-pair coordinate vector y_-, the inter-base-pair coordinate vector z, and the down-stream intra-base-pair coordinate vector y_+.

mononucleotide interaction energy

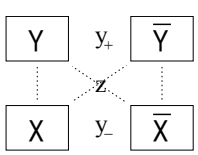


$\mathsf{w}_m = y \in \mathbb{R}^6$

$$\mathsf{U}_m = \tfrac{1}{2}(\mathsf{w}_m - \widehat{\mathsf{w}}_m^{\mathtt{X}}) \cdot \mathsf{K}_m^{\mathtt{X}}(\mathsf{w}_m - \widehat{\mathsf{w}}_m^{\mathtt{X}})$$

$$\widehat{\mathsf{w}}_m^{\mathtt{X}} \in \mathbb{R}^6, \quad \mathsf{K}_m^{\mathtt{X}} \in \mathbb{R}^{6\times 6}.$$

dinucleotide interaction energy



$\mathsf{w}_d = (y_-, z, y_+) \in \mathbb{R}^{18}$

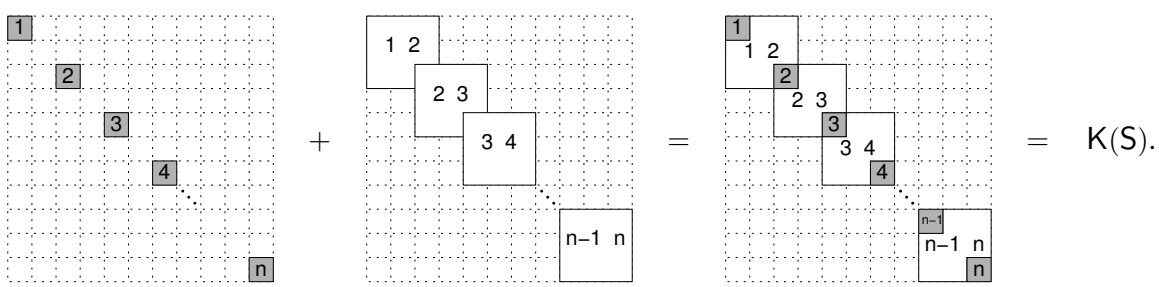
$$\mathsf{U}_d = \tfrac{1}{2}(\mathsf{w}_d - \widehat{\mathsf{w}}_d^{\mathtt{XY}}) \cdot \mathsf{K}_d^{\mathtt{XY}}(\mathsf{w}_d - \widehat{\mathsf{w}}_d^{\mathtt{XY}})$$

$$\widehat{\mathsf{w}}_d^{\mathtt{XY}} \in \mathbb{R}^{18}, \quad \mathsf{K}_d^{\mathtt{XY}} \in \mathbb{R}^{18\times 18}.$$

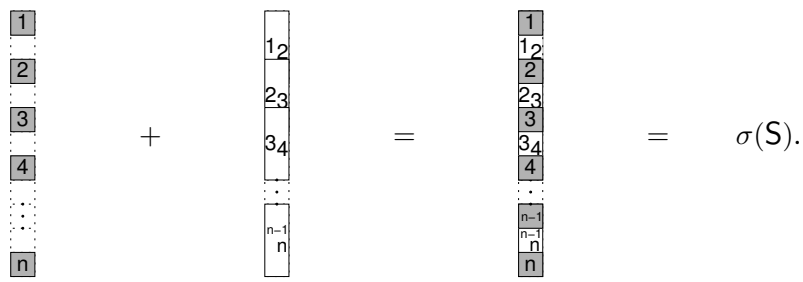
By summing the local mononucleotide and dinucleotide interaction energies along a molecule,

we obtain the total energy $\mathsf{U}(\mathsf{w})$ in Equation (S1). The ground-state stiffness matrix $\mathsf{K}(\mathsf{S})$ and coordinate vector $\widehat{\mathsf{w}}(\mathsf{S})$ are hence determined by the local parameters corresponding to the mononucleotide and dinucleotide content of the sequence S. Notice that the energy of a molecule with an arbitrary sequence can be constructed from a relatively small set of model parameters. Specifically, introducing the quantities $\sigma_m^{\mathtt{X}} := \mathsf{K}_m^{\mathtt{X}}\widehat{\mathsf{w}}_m^{\mathtt{X}}$ and $\sigma_d^{\mathtt{XY}} := \mathsf{K}_d^{\mathtt{XY}}\widehat{\mathsf{w}}_d^{\mathtt{XY}}$ for convenience, a complete model parameter set consists of the mononucleotide parameters $\{\mathsf{K}_m^{\mathtt{X}}, \sigma_m^{\mathtt{X}}\}$ for the 4 possible choices of $\mathtt{X}$ (only 2 of which are independent), and the dinucleotide parameters $\{\mathsf{K}_d^{\mathtt{XY}}, \sigma_d^{\mathtt{XY}}\}$ for the 16 possible choices of $\mathtt{XY}$ (only 10 of which are independent). As discussed in the main article, a complete model parameter set for DNA in solvent under standard environmental conditions has been estimated using an extensive database [S6] of molecular dynamics simulations, and this set can be updated as more data becomes available, or estimation techniques are improved.

The ground-state stiffness matrix $\mathsf{K}(\mathsf{S})$ is determined by the sequence $\mathsf{S} = \mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$ through a matrix assembly step [S1, S5]. Specifically, as illustrated below, the mononucleotide matrices $\mathsf{K}_m^{\mathtt{X}}$ of size 6×6 and the dinucleotide matrices $\mathsf{K}_d^{\mathtt{XY}}$ of size 18×18 are assembled and summed as shown on the left, and this produces the matrix $\mathsf{K}(\mathsf{S})$ of size $N \times N$ as shown on the right. The solid grey squares denote the mononucleotide matrices, the solid white squares denote the dinucleotide matrices, the grid lines denote blocks of size 6×6, and entries in the double and triple overlaps are summed in the obvious way. Moreover, on the left-hand side, each single number in a grey square denotes a dependence on the mononucleotide $\mathtt{X}_a$, while each pair of numbers in a white square denotes a dependence on the dinucleotide $\mathtt{X}_a\mathtt{X}_{a+1}$. On the right-hand side, notice that the blocks with triple overlaps exhibit an effective dependence on the trinucleotide $\mathtt{X}_{a-1}\mathtt{X}_a\mathtt{X}_{a+1}$ corresponding to the overlap of two adjacent dinucleotides and the implied central mononucleotide.



The ground-state coordinate vector $\widehat{\mathsf{w}}(\mathsf{S})$ is determined by the sequence $\mathsf{S} = \mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$ through a combined matrix inversion and assembly step [S1, S5]. Specifically, we have $\widehat{\mathsf{w}}(\mathsf{S}) = [\mathsf{K}(\mathsf{S})]^{-1}\sigma(\mathsf{S})$, where $\mathsf{K}(\mathsf{S})$ is the matrix outlined above and $\sigma(\mathsf{S})$ is a vector that is assembled in an analogous fashion to the matrix $\mathsf{K}(\mathsf{S})$. As illustrated below, the mononucleotide vectors $\sigma_m^{\mathtt{X}}$ of size 6×1 and the dinucleotide vectors $\sigma_d^{\mathtt{XY}}$ of size 18×1 are assembled and summed as shown on the left, and this produces the vector $\sigma(\mathsf{S})$ of size $N \times 1$ as shown on the right. The solid grey squares denote the mononucleotide vectors, the solid white squares denote the dinucleotide vectors, the grid lines denote blocks of size 6×1, and entries in the double and triple overlaps are summed in the obvious way. The numbers in the solid grey and white squares denote a dependence on the mononucleotide $\mathtt{X}_a$ and dinucleotide $\mathtt{X}_a\mathtt{X}_{a+1}$ in the same way as before.



The relation $\widehat{\mathsf{w}}(\mathsf{S}) = [\mathsf{K}(\mathsf{S})]^{-1}\sigma(\mathsf{S})$ occurs naturally in the model when summing the local mononucleotide and dinucleotide interaction energies to obtain the total energy. In particular it arises when a matrix version of completing-the-square is performed to obtain the specific form in Equation (S1). This relation, which is a unique feature of the model considered herein, implies that the ground-state configuration of a molecule will in general depend nonlocally on its sequence. This follows from the observation that, while local changes in the sequence S give rise to only local changes in the entries of $\mathsf{K}(\mathsf{S})$ and $\sigma(\mathsf{S})$, the entries of $\widehat{\mathsf{w}}(\mathsf{S})$ will in general change nonlocally due to the matrix inversion. The changes in $\widehat{\mathsf{w}}(\mathsf{S})$ will typically be largest at the sites of the local changes in S, and decay with increasing distance from these sites.

The function $\mathsf{U}(\mathsf{w})$ in Equation (S1) is to be interpreted as the free energy of a DNA molecule in solvent under standard environmental conditions. Accordingly, the configurational statistics of the molecule are described by an associated Boltzmann probability density function on the set of coordinate vectors w, namely

$$\rho(\mathsf{w}) = \frac{1}{Z_J} e^{-\mathsf{U}(\mathsf{w})/kT}\ J(\mathsf{w}), \tag{S2}$$

where k is the Boltzmann constant, T is the solvent temperature, Z_J is a normalizing constant, and $J(\mathsf{w})$ is a Jacobian factor that arises due to the non-Cartesian nature of the rotational coordinates; more details and an explicit expression for the Jacobian can be found in [S4]. Notice that, although the free energy $\mathsf{U}(\mathsf{w})$ is quadratic, the density is non-Gaussian due to the appearance of $J(\mathsf{w})$. However, there is increasing evidence that the effect of the Jacobian is rather small in the types of applications considered here, namely the modeling of structural variations within relatively stiff, B-form DNA, so that the Jacobian can be approximated as being constant when the density is expressed in *Curves+* coordinates [S4, S1, S5]. In this case, the density takes the standard Gaussian form

$$\rho(\mathsf{w}) = \frac{1}{Z} e^{-\mathsf{U}(\mathsf{w})/kT}. \tag{S3}$$

This density, defined by the free energy introduced above, can be combined with an efficient sampling method to obtain a powerful tool for studying sequence-dependent structural variations in B-form DNA. Specifically, although the DNA backbones are not direct observables, there is sufficient structural resolution in the model to capture variations in bending, twisting, stretching and shearing of the base pairs along the contour of the double helix, as well as variations in the deformation between the bases within each base pair across the double helix. Hence, among other things, variations in the spacing of the major and minor grooves of the double helix can be predicted and studied.

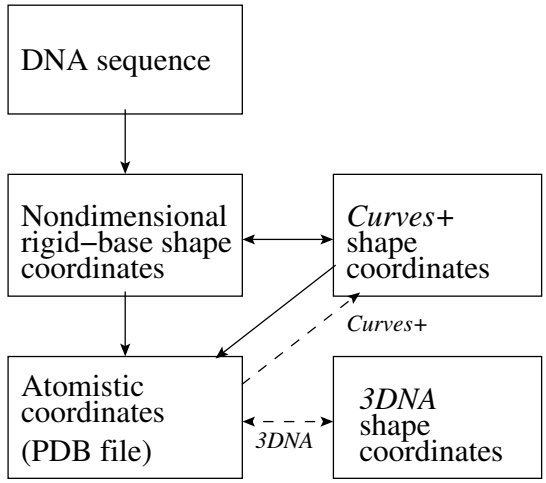


Figure S5: *A diagram of data flow from a DNA sequence to non-dimensional, Curves+, 3DNA and atomistic (PDB) coordinates. Solid pointers correspond to functions available in the package cgDNA and dashed pointers show functions available in the packages Curves+ and 3DNA.*

S2 Extended description of the *cgDNA* package (Supplement to Section 2.2: Software)

As described in the main article, the *cgDNA* package (http://lcvmwww.epfl.ch/cgDNA) is a suite of Matlab programs for implementing the rigid base model. The heart of the package is a parameter set file that contains a complete model parameter set, namely the mononucleotide parameters $\{\mathsf{K}_m^{\mathtt{X}}, \sigma_m^{\mathtt{X}}\}$ for two independent, or distinct, choices of $\mathtt{X}$, and the dinucleotide parameters $\{\mathsf{K}_d^{\mathtt{XY}}, \sigma_d^{\mathtt{XY}}\}$ for 10 independent choices of $\mathtt{XY}$; parameters for the remaining choices of $\mathtt{X}$ and $\mathtt{XY}$ are obtained from these using the symmetries that are fully described in [S1]. The basic input to the package is the sequence S along a DNA molecule, and the basic output is the vector $\widehat{\mathsf{w}}(\mathsf{S})$ of ground-state coordinates, and the matrix $\mathsf{K}(\mathsf{S})$ of ground-state stiffness coefficients, which are computed according to the matrix assembly and inversion steps described above. The core computations in *cgDNA* are performed using a non-dimensional form of the *Curves+* coordinates [S1, S5]. However, to make the results accessible to the widest possible audience, the intra- and inter-base-pair coordinates contained in $\widehat{\mathsf{w}}(\mathsf{S})$ can be obtained in both the standard (dimensional) *Curves+* [S2] helical parameters or as an atomistic representation of the ground state in PDB format. Via the PDB format, helical DNA parameters in other conventions such as *3DNA* [S7] can be obtained as indicated below. In the diagram in Figure S5, solid pointers correspond to functions available in the package *cgDNA*, whereas dashed pointers correspond to functions available in the packages *Curves+* and *3DNA*.

Although *cgDNA* can report the ground-state conformation vector $\widehat{\mathsf{w}}(\mathsf{S})$ using different coordinate conventions, it reports the ground-state stiffness matrix $\mathsf{K}(\mathsf{S})$ for only the non-dimensional version of the *Curves+* convention. The primary reason for this is that the free energy $\mathsf{U}(\mathsf{w})$ does not remain quadratic under a general change of coordinates. Specifically, if w denotes the intra- and inter-base-pair coordinates defined according to the *Curves+* convention, and w' denotes the same coordinates defined according to any other convention, then the relation between coordinate values has the general form $\mathsf{w} = f(\mathsf{w}')$ for some explicitly known but potentially quite complicated function f. When this relation is substituted into $\mathsf{U}(\mathsf{w})$, the resulting composite function $\mathsf{U}(f(\mathsf{w}'))$

is in general no longer quadratic. Although a quadratic approximation of this composite function could always be made, and a corresponding stiffness matrix for the associated coordinates could be constructed, we here opt for simplicity and report only a non-dimensional stiffness matrix in the *Curves+* coordinates.

cgDNA can also predict the free energy difference between two configurations of a molecule of B-form DNA of a given sequence in standard environmental conditions. The sequence and the two configurations can be input directly, or can be read from a file. For example, a sequence along with a configuration can be read from a .lis format file, which can be obtained by running *Curves+* on a .pdb file.

A listing of all the script files of *cgDNA* and a description of their functionality, input and output can be found in the README file, which is part of the package. Each script file also contains its own documentation, which can be accessed directly using the Matlab or Octave *help* command.

S3 Additional comparisons of ground-state coordinates (Supplement to Section 3: RESULTS)

Figures S6 – S9 are expanded versions of Figures 1–4 in the main text, which now show the ground-state values of all the intra- and inter-base-pair coordinates at each position along the molecule for each of the sequences in Table 1 of the main article. The Figures are completely analogous, namely they use the *Curves+* definitions, the base sequence along the reference strand of the molecule is shown on the horizontal axis, the intra-base-pair coordinate values are indicated at each base pair, and inter-base-pair coordinate values indicated at each junction. As pointed out in the main article, for each of the molecules, the ground-state values of the coordinates predicted using *cgDNA* (solid curves) are rather close to the results obtained from MD simulation (dotted curves) throughout the interior and end regions of the molecules. (MD estimates at the ends were not available for the sequence *DD* in Figure S6.) This further supports our assertion that *cgDNA* can effectively serve as a substitute for MD simulation in studies of ground-state structures. We emphasize that none of the sequences shown in the figures were part of the training set that was used to estimate the *cgDNA* model parameter set. Hence the results shown here provide an illustration of the predictive capability of *cgDNA* for arbitrary sequences.

The *cgDNA* and MD results are also in reasonable agreement with the experimental data. As pointed out in the main text, all of the sequences we consider have a palindromic symmetry, which implies that their equilibrium or ground-state coordinate values should either be even or odd functions of position (depending on the coordinate type) about the center of the molecule. However, the experimental data are not generally consistent with this property. Thus, to enhance our comparisons, for each NMR or X-ray estimate in Figures S6 – S9, we also include a symmetrized value which is more consistent with the symmetric nature of each molecule as discussed in the main text. The figures show that the *cgDNA* and MD results are in reasonable agreement with the NMR and X-ray data, both original and symmetrized values. There are, however, some exceptions, for example Stretch in Figure S6, Propeller and Roll in Figure S8, and Propeller and Slide in Figure S9. Due to the fact that the parametrisation of the rigid base model underlying the *cgDNA* package was trained using a database of MD simulations, the level of agreement between the *cgDNA* predictions and the experimental data can naturally be no better than the fit which can be expected from MD, but in fact they are also no worse, while the computations involved in making *cgDNA* predictions are several orders of magnitude less intensive.

References

[S1] O. Gonzalez, D. Petkevičiūtė, and J.H. Maddocks. A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.*, 138(5), 2013.

[S2] R. Lavery, M. Moakher, J.H. Maddocks, D. Petkeviciute, and K. Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, 37:5917–5929, 2009.

[S3] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X. J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C. S. Tung, E. Westhof, C. Wolberger, and H. M. Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *J Mol Biol*, 313(1):229–37, 10 2001.

[S4] F. Lankaš, O. Gonzalez, L.M. Heffler, G. Stoll, M. Moakher, and J.H. Maddocks. On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Physical Chemistry Chemical Physics*, 11:10565–10588, 2009.

[S5] D. Petkevičiūtė. *A DNA Coarse-Grain Rigid Base Model and Parameter Estimation from Molecular Dynamics Simulations*. PhD thesis, EPFL, 2012.

[S6] R. Lavery, K. Zakrzewska, D. Beveridge, T. Bishop, D. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, 38(1):299–313, 2010.

[S7] X.-J. Lu and W.K. Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, 31(17):5108–5121, 2003.

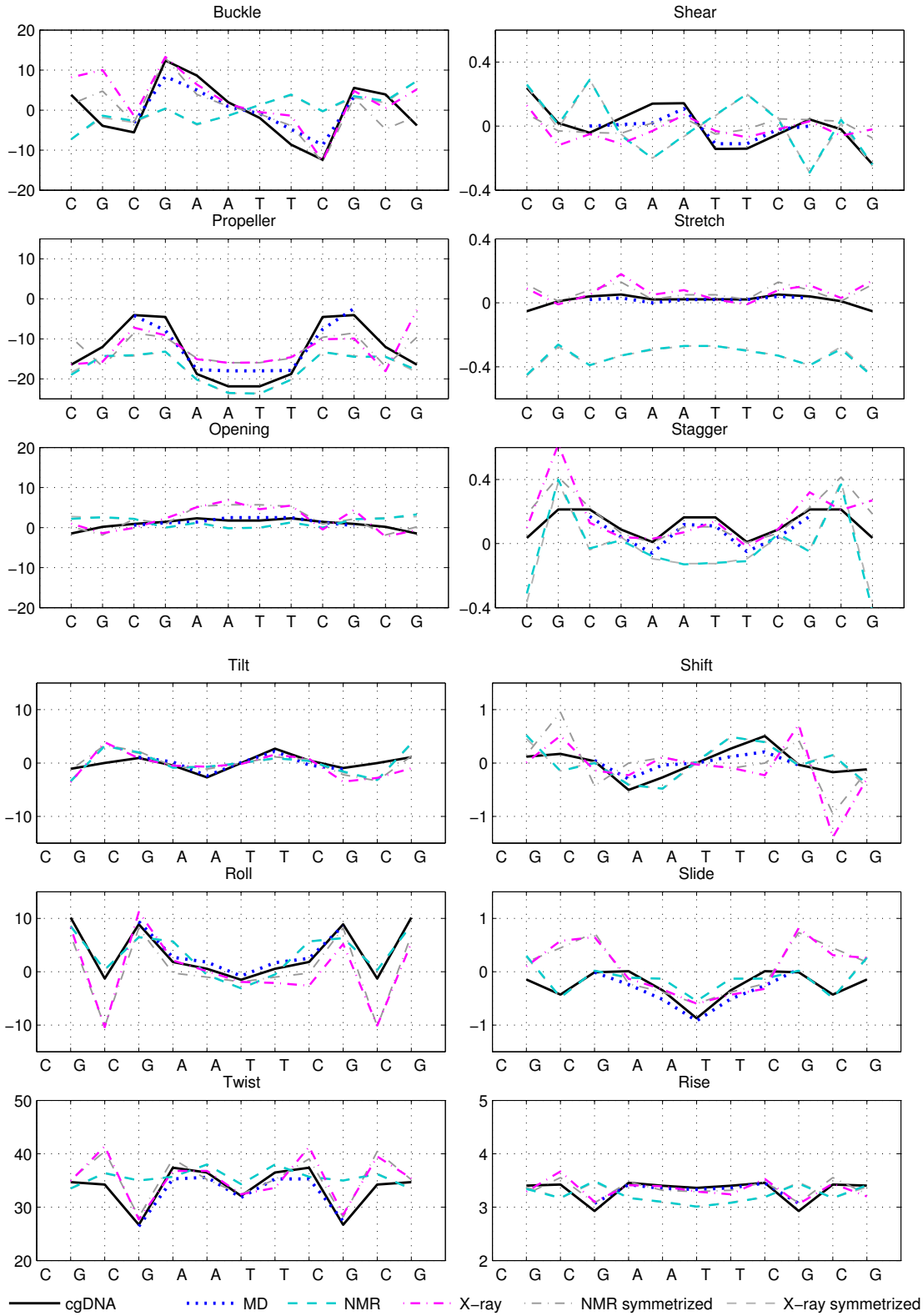


Figure S6: *Ground-state coordinate values for the sequence DD (Dickerson-Drew dodecamer). MD results for end base pairs are not available for this sequence. In Figures S6 to S9, rotations (left column) are in degrees, and displacements (right column) are in Å. Sequence position is indicated on the horizontal axis and coordinate values are interpolated by piecewise linear curves.*

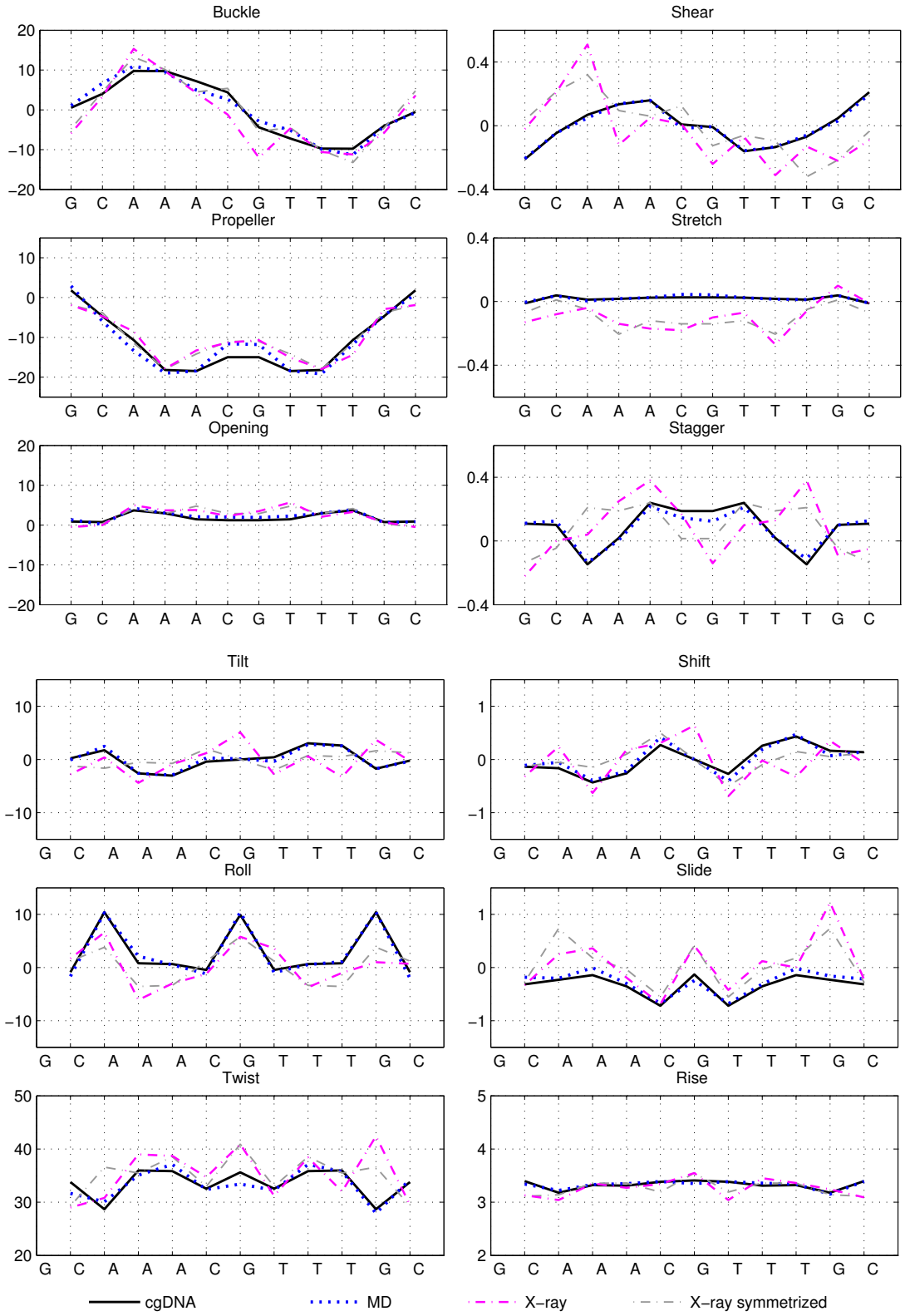


Figure S7: *Ground-state coordinate values for the sequence A3CGT3. See also caption of Figure S6.*

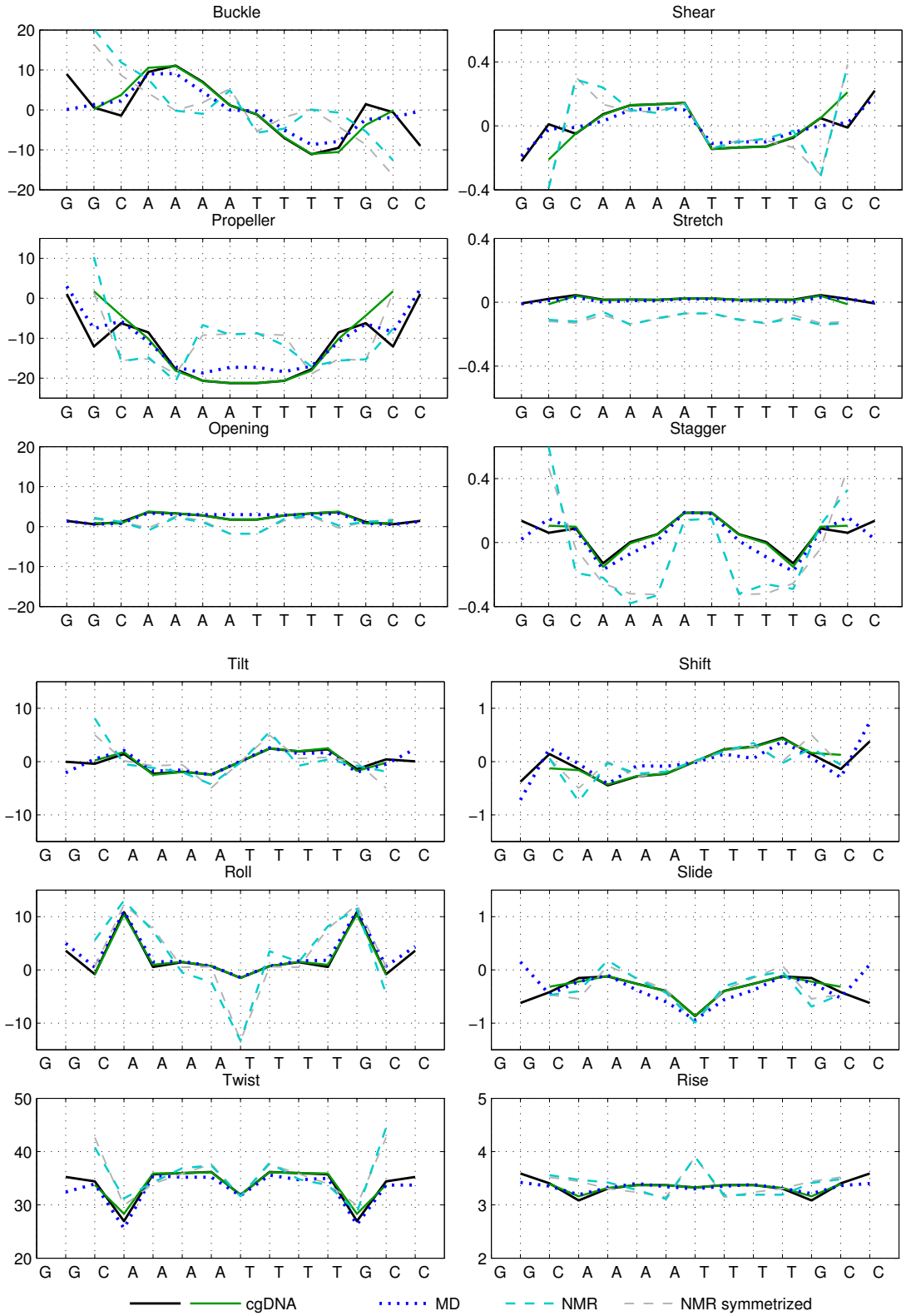


Figure S8: *Ground-state coordinate values for the sequences A4T4 and A4T4_mod. See also caption of Figure S6. NMR results are for the sequence A4T4 whereas MD results are for the sequence A4T4_mod. cgDNA predictions are given for both A4T4 (green) and A4T4_mod (black).*

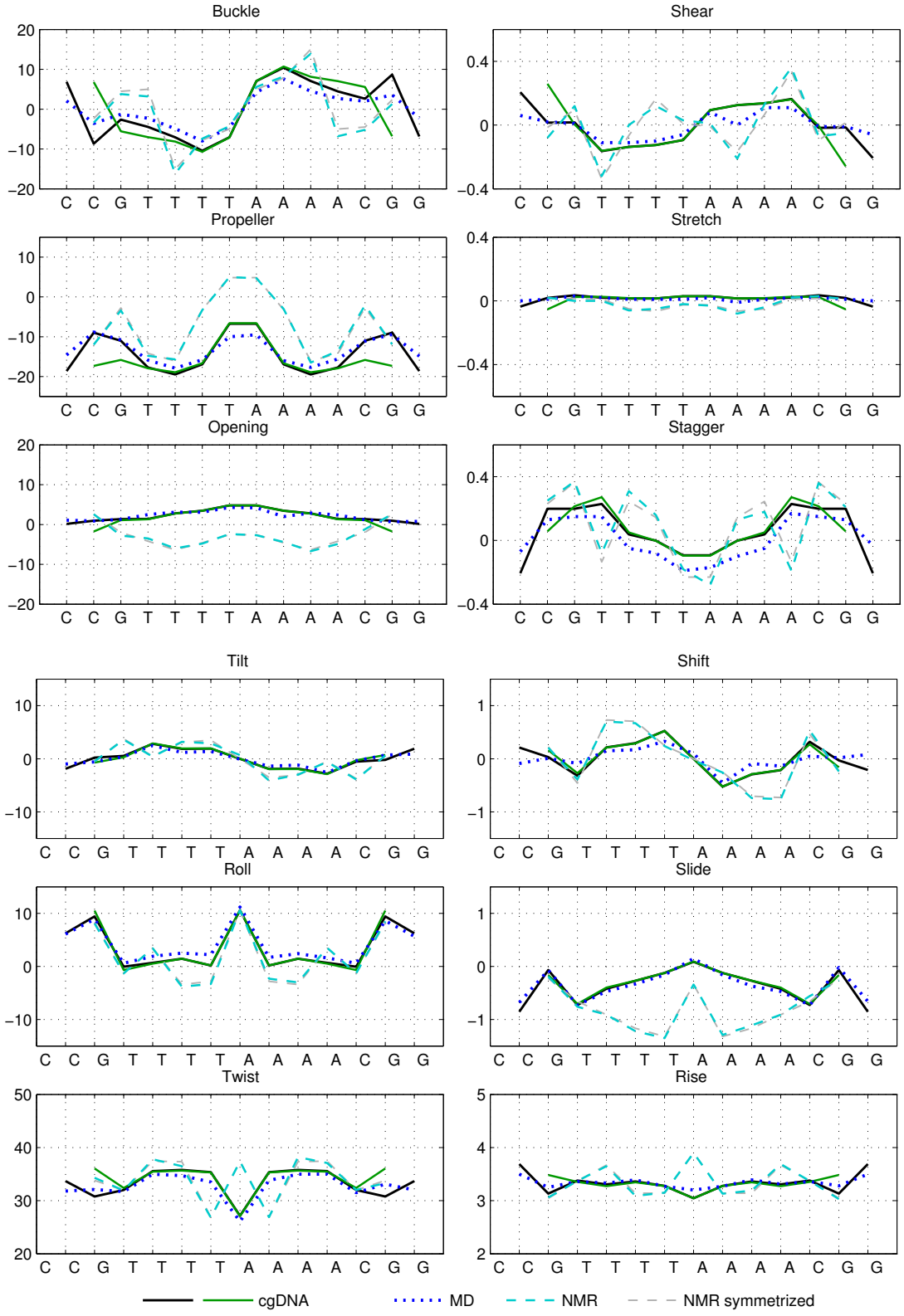


Figure S9: *Ground-state coordinate values for the sequences T4A4 and T4A4_mod. See also captions of Figures S6 and S8.*