On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations

F. Lankaš,†[a] O. Gonzalez,†[b] L. M. Heffler,‡[c] G. Stoll,§[c] M. Moakher[d] and J. H. Maddocks*[c]

Received 18th September 2009, Accepted 1st October 2009
First published as an Advance Article on the web 28th October 2009
DOI: 10.1039/b919565n

A method is described to extract a complete set of sequence-dependent material parameters for rigid base and basepair models of DNA in solution from atomistic molecular dynamics simulations. The method is properly consistent with equilibrium statistical mechanics, leads to effective shape, stiffness and mass parameters, and employs special procedures for treating spontaneous torsion angle flips and H-bond breaks, both of which can have a significant effect on the results. The method is accompanied by various analytical consistency checks that can be used to assess the equilibration of statistical averages, and different modeling assumptions pertaining to the rigidity of the bases and basepairs and the locality of the quadratic internal energy. The practicability of the approach is verified by estimating complete parameter sets for the 16-basepair palindromic oligomer $G(TA)_7C$ simulated in explicit water and counterions. Our results indicate that the method is capable of resolving sequence-dependent variations in each of the material parameters. Moreover, they show that the assumptions of rigidity and locality hold rather well for the base model, but not for the basepair model. For the latter, it is shown that the non-local nature of the internal energy can be understood in terms of a certain compatibility relation involving Schur complements.

1. Introduction

The sequence-dependent mechanical properties of DNA in solution, for example its effective shape, stiffness and mass, are critical for its packaging into the cell, recognition by other molecules, and conformational changes during biochemical processes. However, few methods are available which can probe all these properties at the base or basepair level. Various classic experimental methods can probe the structure of macromolecules in solution, for example electric dichroism and birefringence,[11] light scattering,[1] fluorescence polarization,[37] centrifugation[36] and various types of electrophoresis.[15] These methods in general yield only low-resolution information about shape and mass, and provide little if any information about stiffness. Other methods, such as X-ray diffraction,[29] probe the structure of macromolecules in crystalline states. These methods yield high-resolution information about shape, and given enough data, even stiffness,[31] but such data is static and may not be representative of molecules in solution. The nuclear magnetic resonance (NMR) of nucleic acids[38] has recently made a dramatic progress, related in part to advances in residual dipolar coupling techniques.[26] NMR methods can be applied to DNA in solution to obtain structural information among other things. However, the structures obtained may still depend on the force field used to refine the NMR data, and the information about structural fluctuations may be difficult to relate to a convenient set of internal molecular coordinates.

A different approach is to exploit the direct, detailed structural information available from a molecular dynamics (MD) simulation.[12] MD simulation probes the structure of a macromolecule through the numerical integration of the Newtonian laws of motion based on molecular forces computed from empirical potential functions. MD can generate an all-atom description of the dynamics of a macromolecule in solution, with solvent water and counterions included explicitly, and various structural characteristics of the macromolecule can be obtained from an analysis of the results. Although computationally intensive, MD has become widely used for the modeling of macromolecules in solution due in large part to the growing availability of increased computer power. Indeed, high-performance computing is emerging as an attractive complement to classic experimental methods. The field of MD applied to DNA has been under development for at least 20 years,[28] and the subject of force fields and simulation protocols has received considerable recent attention.[3–6,13,33,35] While any parameters found by MD can only be as good as the assumed atomistic potentials,

[a] *Center for Complex Molecular Systems and Biomolecules, Institute of Organic Chemistry and Biochemistry, Prague, Czech Republic*
[b] *Department of Mathematics, University of Texas at Austin, Austin, TX, USA*
[c] *Lab. for Computation and Visualization in Math. and Mech., Swiss Federal Institute of Technology, Lausanne, Switzerland*
[d] *Lab. for Mathematical and Numerical Modeling in Eng. Sci., National Engineering School at Tunis, Tunisia*
† These authors contributed equally to this work.
‡ Present address: School of Engineering, University of Newcastle, Australia.
§ Present address: Bioinformatics Laboratory, Curie Institute, Paris, France.

and are restricted by the time scale and system size that can be simulated, recent evidence suggests that MD simulation can capture many sequence-dependent effects on the structure, dynamics and mechanical properties of DNA.[4,5,13,33,32]

The sequence-dependent mechanical properties of DNA are determined by its internal and kinetic energy functions. The parameterization of the internal energy, which in general is assumed to be elastic and quadratic in specified internal coordinates, has been considered in various recent works. For example, the parameters for various different internal energy models have been inferred from crystal structure data[31] and from MD simulations.[20,21] In these works, as well as in related theoretical studies,[7,14] an assumption of locality is made. That is, the internal energy is assumed to be a sum of energies associated with either individual basepair steps (junction between consecutive base pairs)[31,20] or with the two bases in a pair.[21] The parameterization of the kinetic energy is more complicated and has been less well-studied. Indeed, the determination of kinetic energy parameters requires dynamic data, which is difficult to produce in physical experiments, but is naturally available in an MD simulation, and also requires a statistical mechanical description of the DNA model on its full phase space, not just the configuration space as is commonly considered.

In this article, we study two coarse-grained models of DNA, without any *a priori* assumptions of locality, and introduce an MD method for estimating the complete set of sequence-dependent internal and kinetic energy parameters for each. The models differ according to whether individual bases or basepairs are considered as independent, interacting rigid bodies. We consider models in which the internal elastic energy is a quadratic function of a set of internal coordinates describing the relative, three-dimensional rotation and displacement between bodies, and in which the kinetic energy is a quadratic function of the linear and angular velocities of each body as dictated by classical mechanical theory. We make no *a priori* assumptions on the internal energy other than its quadratic dependence on the internal coordinates; such an energy is often referred to as harmonic and is the only type considered here. We introduce the sequence-dependent shape, stiffness and mass parameters necessary to define each model, establish various results about their properties and derive statistical mechanical relations that connect the complete set of material parameters to the expected values of certain state functions.

For special sequences, we exploit the complementary nature of the two DNA strands and the objectivity of the internal and kinetic energies to derive various symmetry relations for the complete set of material parameters. Specifically, for palindromic sequences, we show that material parameters must be either symmetric or antisymmetric functions of position about the middle of the sequence. For general sequences, we exploit a certain factorability property of the canonical measure and show that the rigid base and basepair parameters must be compatible in an appropriate sense. For example, we show, under appropriate assumptions, that the elastic stiffness matrix of the basepair model is related to the stiffness matrix of the base model through a Schur complement. Such relations are of both theoretical and practical interest and can be exploited to obtain consistency checks on general parameter estimation methods. The statistical mechanical relations we derive are properly consistent with the canonical measure on the full phase space of the system and differ from the usual Gaussian-type relations by a Jacobian factor associated with the three-dimensional rotation group. While such factors are typically ignored, or equivalently assumed to be constant, we include them here.

We develop a method for estimating the complete set of material parameters for both the rigid base and basepair models from atomic-resolution MD simulation of a DNA oligomer. To obtain parameters consistent with the B-form DNA structural family, we propose special procedures for treating spontaneous torsion angle flips and H-bond breaks, both of which can have a significant effect on the results. We demonstrate the practicability of our proposed method by estimating material parameters for the 16-basepair palindromic oligomer $G(TA)_7C$. In particular, we use two different MD trajectories simulated with the inclusion of explicit water and counterions to estimate shape, stiffness and mass parameters for both the rigid base and basepair models. Various consistency checks indicate that the trajectories are sufficiently long so that the required statistical mechanical averages are estimated well. Our results indicate that the method is capable of resolving sequence-dependent variations in each of the material parameters.

We further use the estimated parameters to assess various modeling assumptions. Specifically, we study the assumption of rigidity of the bases and basepairs, and the property of locality of the internal elastic energy. Through an analysis of the predicted sparsity pattern for the generalized mass matrix of each model, we find that the simulated data is closely consistent with the assumption of rigid bases, but not rigid basepairs. Indeed, the estimated mass parameters for the bases are shown to compare favorably with various estimates based on canonical geometries. Through an analysis of the sparsity patterns of the generalized stiffness matrix, which *a priori* is not assumed to have any specific structure, we find that the simulated data is nearly consistent with a local internal energy for the rigid base model, but not for the rigid basepair model. Indeed, the estimated stiffness matrix for the basepair model is remarkably non-local. We show that this non-locality can be understood in terms of a certain compatibility relation.

It is well accepted that the mechanical properties of a DNA molecule depend on the sequence of bases that constitute it, and that this dependence is particularly pronounced at length scales ranging from tens to a few hundreds of basepairs. Such scales are prohibitively expensive for detailed atomistic-type models, and often involve important local features that are below the resolution of homogeneous chain- or rod-type models. In this respect, coarse-grained models in which bases or basepairs are modeled as rigid offer a promising approach to understand various structural features at these scales, such as sequence-dependent curvature and flexibility. Our results suggest that, at the scale of a few tens of basepairs, a model in which the bases are modeled as rigid, together with a local quadratic internal energy based on nearest neighbors as introduced here, is significantly more consistent with an atomic-resolution MD simulation than an analogous model in

which basepairs are modeled as rigid. Indeed, an analogous local rigid basepair model is visibly inconsistent with the simulated data. Moreover, the sequence-dependent variability in our results suggest that MD simulation could be used to estimate complete parameter sets for a local rigid base model of B-form DNA. Such parameter sets may be helpful in clarifying the relation between the base and basepair models, interpreting various experiments where sequence-dependent curvature and flexibility are believed to be important, and could potentially lead to useful predictions about biochemical processes such as transcription factor binding and nucleosome formation. On the other hand, our method is intrinsically limited to the assumption of quadratic, or harmonic energies, so that large-strain deformations of DNA, such as kinking that can arise in high-load regimes,[19] are beyond its scope.

The presentation is structured as follows. In sections 2 and 3 we outline our rigid base and basepair models of DNA, define the internal coordinates and velocities for each and establish various results about their material parameters. In section 4 we describe the MD protocol and special procedures used to simulate B-form DNA in explicit solvent and the methods used to compute the internal coordinates and velocities for each model. In section 5 we present results on the shape, stiffness and mass parameters for the oligomer $G(TA)_7C$ and examine various modeling assumptions pertaining to rigidity and locality. In section 6, we summarize our results and conclusions.

2. Rigid base model

Here we outline a rigid base model for the three-dimensional, sequence-dependent structure of DNA. We introduce the kinematic quantities necessary to define the configuration and velocity of each base, and the material parameters necessary to define the internal elastic energy and the kinetic energy of a given oligomer. We outline the statistical mechanical properties of the model and derive various symmetry relations that the material parameters must satisfy for special types of sequences.

2.1 Bases, configurations

We consider right-handed, double-helical DNA in which bases T, A, C and G are attached to two, oriented, anti-parallel backbone strands and form only the standard Watson–Crick pairs (A, T) and (C, G). Choosing one backbone strand as a reference, a DNA molecule consisting of n basepairs is identified with a sequence of bases $X_1X_2\cdots X_n$, listed in the $5'$ to $3'$ direction along the strand, where $X_a \in \{T, A, C, G\}$. The basepairs associated with this sequence are denoted by $(X, \bar{X})_1,(X, \bar{X})_2,\ldots,(X, \bar{X})_n$, where $\bar{X}$ is defined as the Watson–Crick complement of X in the sense that $\bar{A} = T$, $\bar{T} = A$, $\bar{C} = G$ and $\bar{G} = C$. The notation $(X, \bar{X})_a$ for a basepair indicates that base X is attached to the reference strand while $\bar{X}$ is attached to the opposite strand.

We adopt a model of DNA[9,10,30] in which each base is modeled as a rigid object. The configuration of an arbitrary base X_a is specified by giving the location of a reference point $\boldsymbol{r}^a$ fixed in the base, and the orientation of a right-handed, orthonormal frame $\{\boldsymbol{d}_i^a\}$ $(i = 1,2,3)$ attached to the base. The reference point and frame vectors are defined according to the Tsukuba convention.[30] The vector $\boldsymbol{d}_1^a$ points in the direction of the major groove along what would be the perpendicular bisector of the C1′–C1′ axis of an ideal basepair formed from X_a, whereas $\boldsymbol{d}_2^a$ points in the direction of the reference strand and is parallel to the C1′–C1′ axis. As a result, $\boldsymbol{d}_3^a = \boldsymbol{d}_1^a \times \boldsymbol{d}_2^a$ is perpendicular to the plane of X_a and normally points in the direction of X_{a+1}. The reference point $\boldsymbol{r}^a$ is located at the intersection of the perpendicular bisector of the C1′–C1′ axis with the axis defined by the pyrimidine C6 and the purine C8 atoms. Just as for X_a, the configuration of base $\bar{X}_a$ is specified by a reference point $\bar{\boldsymbol{r}}^a$ and frame $\{\bar{\boldsymbol{d}}_i^a\}$. The reference point and frame for $\bar{X}_a$ are defined in a manner exactly analogous to that for X_a using the same reference strand. As a result, when $(X, \bar{X})_a$ form an ideal basepair, the reference points and frames associated with each base coincide.

There are four possible basepairs $(X, \bar{X})_a$ corresponding to the choice $X_a \in \{T, A, C, G\}$. In a rigid base model, the positions of the non-hydrogen atoms in each base of each basepair with respect to the associated reference point and frame are considered to be constant. As a result, once the reference point and frame of each base are specified, so too are the positions of all the non-hydrogen atoms. Estimated values for these positions for each base in their ideal forms are tabulated in ref. 30. Thus the configuration of a DNA molecule consisting of n basepairs is completely defined by the reference points $\boldsymbol{r}^a$ and $\bar{\boldsymbol{r}}^a$ and the frames $\{\boldsymbol{d}_i^a\}$ and $\{\bar{\boldsymbol{d}}_i^a\}$ $(a = 1,\ldots,n)$. These points and frames are in turn uniquely defined by component vectors $r^a,\bar{r}^a \in \mathbb{R}^3$ and rotation matrices D^a, $\bar{D}^a \in \mathbb{R}^{3\times3}$, where $r_i^a = \boldsymbol{e}_i\cdot\boldsymbol{r}^a$, $\bar{r}_i^a = \boldsymbol{e}_i\cdot\bar{\boldsymbol{r}}^a$, $D_{ij}^a = \boldsymbol{e}_i\cdot\boldsymbol{d}_j^a$ and $\bar{D}_{ij}^a = \boldsymbol{e}_i\cdot\bar{\boldsymbol{d}}_j^a$. Here $\{\boldsymbol{e}_i\}$ denotes an arbitrary, lab-fixed frame. In terms of these components, we have

$$\begin{aligned} \boldsymbol{d}_j^a &= \sum_{i=1}^{3} D_{ij}^a \boldsymbol{e}_i,\ \boldsymbol{r}^a = \sum_{i=1}^{3} r_i^a \boldsymbol{e}_i, \\ \bar{\boldsymbol{d}}_j^a &= \sum_{i=1}^{3} \bar{D}_{ij}^a \boldsymbol{e}_i,\ \bar{\boldsymbol{r}}^a = \sum_{i=1}^{3} \bar{r}_i^a \boldsymbol{e}_i. \end{aligned} \tag{1}$$

2.2 Rotation, displacement coordinates

In a rigid base model, the three-dimensional shape of a DNA molecule is determined by the relative rotation and displacement between neighboring bases both across and along the two backbone strands. The relative rotation and displacement between X_a and $\bar{X}_a$ across the strands can be described in the general form

$$\boldsymbol{d}_j^a = \sum_{i=1}^{3} \Lambda_{ij}^a \bar{\boldsymbol{d}}_i^a,\ \boldsymbol{r}^a = \bar{\boldsymbol{r}}^a + \sum_{i=1}^{3} \xi_i^a \boldsymbol{g}_i^a, \tag{2}$$

where $\Lambda^a \in \mathbb{R}^{3\times3}$ is a rotation matrix which describes the orientation of frame $\{\boldsymbol{d}_i^a\}$ with respect to $\{\bar{\boldsymbol{d}}_i^a\}$, $\xi^a \in \mathbb{R}^3$ is a coordinate vector which describes the position of $\boldsymbol{r}^a$ with respect to $\bar{\boldsymbol{r}}^a$, and $\{\boldsymbol{g}_i^a\}$ is a right-handed, orthonormal frame in between the base frames $\{\boldsymbol{d}_i^a\}$ and $\{\bar{\boldsymbol{d}}_i^a\}$. The frame $\{\boldsymbol{g}_i^a\}$ will be referred to as the basepair frame associated with $(X, \bar{X})_a$ and will be defined in detail below. From (2) we deduce that the entries in Λ^a and ξ^a are given by

$$\Lambda_{ij}^a = \bar{\boldsymbol{d}}_i^a\cdot\boldsymbol{d}_j^a,\ \xi_i^a = \boldsymbol{g}_i^a\cdot(\boldsymbol{r}^a - \bar{\boldsymbol{r}}^a). \tag{3}$$

To describe the relative rotation and displacement between neighboring bases along the strands it is sufficient to consider a basepair frame $\{\boldsymbol{g}_i^a\}$ and reference point $\boldsymbol{q}^a$ associated with $(X, \bar{X})_a$, and a similar frame $\{\boldsymbol{g}_i^{a+1}\}$ and point $\boldsymbol{q}^{a+1}$ associated with $(X, \bar{X})_{a+1}$, as defined below. The relative rotation and displacement between $(X, \bar{X})_a$ and $(X, \bar{X})_{a+1}$ along the strands can then be described in the general form

$$\boldsymbol{g}_j^{a+1} = \sum_{i=1}^{3} L_{ij}^a \boldsymbol{g}_i^a,\ \boldsymbol{q}^{a+1} = \boldsymbol{q}^a + \sum_{i=1}^{3} \zeta_i^a \boldsymbol{h}_i^a, \tag{4}$$

where $L^a \in \mathbb{R}^{3\times 3}$ is a rotation matrix which describes the orientation of frame $\{\boldsymbol{g}_i^{a+1}\}$ with respect to $\{\boldsymbol{g}_i^a\}$, $\zeta^a \in \mathbb{R}^3$ is a coordinate vector which describes the position of $\boldsymbol{q}^{a+1}$ with respect to $\boldsymbol{q}^a$, and $\{\boldsymbol{h}_i^a\}$ is a right-handed, orthonormal frame in between the basepair frames $\{\boldsymbol{g}_i^a\}$ and $\{\boldsymbol{g}_i^{a+1}\}$. The frame $\{\boldsymbol{h}_i^a\}$ will be referred to as the junction frame associated with $(X, \bar{X})_a$ and $(X, \bar{X})_{a+1}$ and will be defined below. From (4) we deduce that the entries in L^a and ζ^a are given by

$$L_{ij}^a = \boldsymbol{g}_i^a \cdot \boldsymbol{g}_j^{a+1},\ \zeta_i^a = \boldsymbol{h}_i^a \cdot (\boldsymbol{q}^{a+1} - \boldsymbol{q}^a). \tag{5}$$

The rotation matrix Λ^a appearing in eqn (2) can be parameterized by a coordinate vector $\vartheta^a \in \mathbb{R}^3$ in a variety of ways. In this work, we parameterize rotation matrices using the Cayley (also referred to as Euler–Rodrigues or Gibbs) formula[18]

$$\Lambda^a = \mathrm{cay}[\vartheta^a] := I + \frac{4}{4 + |\vartheta^a|^2}([\vartheta^a\times] + \tfrac{1}{2}[\vartheta^a\times]^2), \tag{6}$$

where I is the identity matrix and $[\vartheta^a\times]$ denotes the skew-symmetric matrix

$$[\vartheta^a\times] = \begin{pmatrix} 0 & -\vartheta_3^a & \vartheta_2^a \\ \vartheta_3^a & 0 & -\vartheta_1^a \\ -\vartheta_2^a & \vartheta_1^a & 0 \end{pmatrix}. \tag{7}$$

The Cayley formula can be explicitly inverted as

$$\vartheta^a = \mathrm{cay}^{-1}[\Lambda^a] := \frac{2}{\mathrm{tr}[\Lambda^a] + 1}\mathrm{vec}[\Lambda^a - (\Lambda^a)^T], \tag{8}$$

where $\mathrm{tr}[\Lambda^a]$ and $(\Lambda^a)^T$ denote the trace and transpose of Λ^a, and, for an arbitrary skew-symmetric matrix A, we define $\mathrm{vec}[A] = (A_{32}, A_{13}, A_{21})$. Eqn (6) and (8) provide a one-to-one correspondence between rotation matrices Λ^a and coordinates ϑ^a provided that $\mathrm{tr}[\Lambda^a] \neq -1$. Matrices for which $\mathrm{tr}[\Lambda^a] = -1$ can be shown to correspond to a rotation through π-radians (180°), which are unlikely to occur between neighboring bases in our application to B-form DNA.

The Cayley parameterization of a rotation matrix has a straightforward geometrical interpretation. The matrix Λ^a in eqn (6) corresponds to a right-handed rotation about a unit vector $\boldsymbol{n}^a$ through an angle $\phi^a \in [0, \pi)$ where

$$\boldsymbol{n}^a = \frac{1}{|\vartheta^a|}\sum_{i=1}^{3} \vartheta_i^a \bar{\boldsymbol{d}}_i^a,\ \phi^a = 2\arctan\left(\frac{|\vartheta^a|}{2}\right). \tag{9}$$

From eqn $(9)_1$ we deduce that a simple rotation about the frame vector $\bar{\boldsymbol{d}}_1^a$ is obtained when $\vartheta^a = (\vartheta_1^a, 0, 0)$, where the angle of rotation is determined by $(9)_2$. Similar conclusions can be drawn for simple rotations about the other frame vectors. Further inspection of eqn $(9)_1$ and $(9)_2$ reveals that a rotation about an arbitrary unit vector $\boldsymbol{n}^a = \sum_{i=1}^{3}\mu_i^a \bar{\boldsymbol{d}}_i^a$ through an arbitrary angle $\sigma^a \in [0, \pi)$ is obtained when

$$\vartheta_i^a = 2\tan\left(\frac{\sigma^a}{2}\right)\mu_i^a. \tag{10}$$

The basepair frame $\{\boldsymbol{g}_i^a\}$ and reference point $\boldsymbol{q}^a$ associated with a pair $(X, \bar{X})_a$ can now be defined. The reference point is defined by $\boldsymbol{q}^a = \frac{1}{2}(\boldsymbol{r}^a + \bar{\boldsymbol{r}}^a)$. To define the frame, let Λ^a be the relative rotation matrix for frame $\{\boldsymbol{d}_i^a\}$ with respect to $\{\bar{\boldsymbol{d}}_i^a\}$. Then the coordinates ϑ^a, axis $\boldsymbol{n}^a$ and angle ϕ^a associated with this rotation are as given in eqn (8) and (9). The basepair frame is here defined by a relative rotation about the same axis $\boldsymbol{n}^a$, but through an angle of $\phi^a/2$. Using eqn (10) we obtain

$$\boldsymbol{g}_j^a = \sum_{i=1}^{3} \mathrm{cay}_{ij}[\widetilde{\vartheta}^a]\bar{\boldsymbol{d}}_i^a,\ \widetilde{\vartheta}_i^a = 2\tan\left(\frac{\phi^a}{4}\right)\frac{\vartheta_i^a}{|\vartheta^a|}. \tag{11}$$

Notice that, by construction, the basepair frame is midway between the base frames in the sense that the associated relative rotation matrix is the square root of the overall rotation matrix between the two bases.

Just as with Λ^a, the rotation matrix L^a in (4) can also be parameterized by a coordinate vector $\theta^a \in \mathbb{R}^3$. In particular, we use the Cayley parameterization $L^a = \mathrm{cay}[\theta^a]$, with explicit inverse $\theta^a = \mathrm{cay}^{-1}[L^a]$. Moreover, the junction frame $\{\boldsymbol{h}_i^a\}$ associated with adjacent pairs $(X, \bar{X})_a$ and $(X, \bar{X})_{a+1}$ can be defined in a manner exactly analogous to the basepair frame $\{\boldsymbol{g}_i^a\}$. That is, if L^a has coordinates θ^a, then the axis $\boldsymbol{m}^a$ and angle ψ^a associated with L^a are

$$\boldsymbol{m}^a = \frac{1}{|\theta^a|}\sum_{i=1}^{3} \theta_i^a \boldsymbol{g}_i^a,\ \psi^a = 2\arctan\left(\frac{|\theta^a|}{2}\right). \tag{12}$$

The junction frame is defined by a relative rotation about the same axis $\boldsymbol{m}^a$, but through an angle of $\psi^a/2$, which gives

$$\boldsymbol{h}_j^a = \sum_{i=1}^{3} \mathrm{cay}_{ij}[\widetilde{\theta}^a]\boldsymbol{g}_i^a,\ \widetilde{\theta}_i^a = 2\tan\left(\frac{\psi^a}{4}\right)\frac{\theta_i^a}{|\theta^a|}. \tag{13}$$

By construction, the junction frame is midway between the basepair frames in the sense that the associated relative rotation matrix is the square root of the overall rotation matrix between the two basepairs.

Thus the relative rotation and displacement between bases X_a and $\bar{X}_a$ across the strands is described by the coordinates $(\vartheta, \xi)^a$, whereas the relative rotation and displacement between the pairs $(X, \bar{X})_a$ and $(X, \bar{X})_{a+1}$ along the strands is described by the coordinates $(\theta, \zeta)^a$. The definitions of these coordinates can be shown to satisfy all the qualitative guidelines set forth in the Cambridge convention,[9] including the symmetry conditions associated with a change of reference strand. Accordingly, we refer to ϑ^a as buckle–propeller–opening, ξ^a as shear–stretch–stagger, θ^a as tilt–roll–twist and ζ^a as shift–slide–rise coordinates. Notice that ϑ^a and θ^a are not conventional angular coordinates as employed by many authors. Rather, they are abstract coordinates defined *via* the parameterization in eqn (6). These abstract coordinates can be put into correspondence with conventional angular ones, and are nearly identical in the case of small rotations when the angular ones are measured in radians.

The complete configuration of a DNA molecule can be specified by introducing coordinates θ^0 and ζ^0 for the first basepair frame $\{\boldsymbol{g}_i^1\}$ and reference point $\boldsymbol{q}^1$ with respect to the lab-fixed frame $\{\boldsymbol{e}_i\}$, that is,

$$\boldsymbol{g}_j^1 = \sum_{i=1}^{3} \mathrm{cay}_{ij}[\theta^0]\boldsymbol{e}_i,\ \boldsymbol{q}^1 = \sum_{i=1}^{3} \zeta_i^0 \boldsymbol{e}_i. \tag{14}$$

It can be shown that the coordinates $y^a = (\vartheta,\ \xi)^a \in \mathbb{R}^6$ $(a = 1,\ldots,n)$ and $z^a = (\theta, \zeta)^a \in \mathbb{R}^6$ $(a = 0,\ldots,n-1)$ completely define the configuration of a molecule with n basepairs as defined in section 2.1. Notice that z^0 are external coordinates that specify the spatial location of the molecule, whereas all others are internal coordinates which describe its shape.

2.3 Mass properties, kinematics

To each base X_a on the reference strand we ascribe a total mass m^a, a symmetric rotational inertia tensor $\boldsymbol{\Gamma}^a$ with respect to the mass center, and a vector $\boldsymbol{c}^a$ that locates the mass center relative to the base reference point $\boldsymbol{r}^a$, so that $\boldsymbol{\rho}^a = \boldsymbol{r}^a + \boldsymbol{c}^a$ is the position vector of the mass center. The kinematics of X_a are encapsulated in the relations

$$\dot{\boldsymbol{r}}^a = \boldsymbol{v}^a,\ \dot{\boldsymbol{d}}_i^a = \boldsymbol{\omega}^a \times \boldsymbol{d}_i^a, \tag{15}$$

where $\boldsymbol{v}^a$ is the velocity of the reference point, $\boldsymbol{\omega}^a$ is the angular velocity of the frame of base X_a, an over-dot denotes a derivative with respect to time and $\times$ denotes the standard vector product. Because $\boldsymbol{c}^a$ is fixed in the base, we have $\dot{\boldsymbol{c}}^a = \boldsymbol{\omega}^a \times \boldsymbol{c}^a$, and we find that the velocity of the center of mass of X_a is given by

$$\dot{\boldsymbol{\rho}}^a = \boldsymbol{v}^a + \boldsymbol{\omega}^a \times \boldsymbol{c}^a. \tag{16}$$

Let $\rho^a \in \mathbb{R}^3$ denote components in the lab-fixed frame, and let $c^a \in \mathbb{R}^3$, $v^a \in \mathbb{R}^3$, $\omega^a \in \mathbb{R}^3$ and $\Gamma^a \in \mathbb{R}^{3\times3}$ denote components in the associated base frame, so that $\rho_i^a = \boldsymbol{e}_i \cdot \boldsymbol{\rho}^a$, $c_i^a = \boldsymbol{d}_i^a \cdot \boldsymbol{c}^a$, $\Gamma_{ij}^a = \boldsymbol{d}_i^a \cdot \boldsymbol{\Gamma}^a \boldsymbol{d}_j^a$ and so on. Then from eqn (1), (15) and (16) we deduce the component relations

$$\dot{r}^a = D^a v^a,\ \dot{D}^a = D^a[\omega^a\times],\ \dot{\rho}^a = D^a[v^a + \omega^a \times c^a], \tag{17}$$

where $[\omega^a\times] \in \mathbb{R}^{3\times3}$ denotes the skew-symmetric matrix defined in eqn (7). As the notation suggests, this matrix has the property that $[\omega^a\times]g^a = \omega^a \times g^a$ for all component vectors g^a. Using the notation introduced in section 2.2, we have $\mathrm{vec}[\omega^a\times] = \omega^a$. Thus from eqn (17) we deduce the relations

$$v^a = (D^a)^T \dot{r}^a,\ \omega^a = \mathrm{vec}[(D^a)^T \dot{D}^a]. \tag{18}$$

Just as for the reference strand, to each base $\bar{X}_a$ on the opposite strand we ascribe a total mass $\bar{m}^a$, a symmetric rotational inertia tensor $\bar{\boldsymbol{\Gamma}}^a$ with respect to the mass center, and a vector $\bar{\boldsymbol{c}}^a$ that locates the mass center relative to the base reference point $\bar{\boldsymbol{r}}^a$, so that $\bar{\boldsymbol{\rho}}^a = \bar{\boldsymbol{r}}^a + \bar{\boldsymbol{c}}^a$ is the position vector of the mass center. The kinematics of $\bar{X}_a$ are encapsulated in the relations

$$\dot{\bar{\boldsymbol{r}}}^a = \bar{\boldsymbol{v}}^a, \dot{\bar{\boldsymbol{d}}}_i^a = \bar{\boldsymbol{\omega}}^a \times \bar{\boldsymbol{d}}_i^a, \tag{19}$$

where $\bar{\boldsymbol{v}}^a$ is the velocity of the reference point and $\bar{\boldsymbol{\omega}}^a$ is the angular velocity of the frame of base $\bar{X}_a$. As before, because $\bar{\boldsymbol{c}}^a$ is fixed in the base, we have $\dot{\bar{\boldsymbol{c}}}^a = \bar{\boldsymbol{\omega}}^a \times \bar{\boldsymbol{c}}^a$, and we find that the velocity of the center of mass of $\bar{X}_a$ is given by

$$\dot{\bar{\boldsymbol{\rho}}}^a = \bar{\boldsymbol{v}}^a + \bar{\boldsymbol{\omega}}^a \times \bar{\boldsymbol{c}}^a. \tag{20}$$

Let $\bar{\rho}^a \in \mathbb{R}^3$ denote components in the lab-fixed frame, and let $\bar{c}^a \in \mathbb{R}^3$, $\bar{v}^a \in \mathbb{R}^3$, $\bar{\omega}^a \in \mathbb{R}^3$ and $\bar{\Gamma}^a \in \mathbb{R}^{3\times3}$ denote components in the associated base frame, so that $\bar{\rho}_i^a = \boldsymbol{e}_i \cdot \bar{\boldsymbol{\rho}}^a$, $\bar{c}_i^a = \bar{\boldsymbol{d}}_i^a \cdot \bar{\boldsymbol{c}}^a$, $\bar{\Gamma}_{ij}^a = \bar{\boldsymbol{d}}_i^a \cdot \bar{\boldsymbol{\Gamma}}^a \bar{\boldsymbol{d}}_j^a$ and so on. Then from eqn (1), (19) and (20) we deduce as before

$$\dot{\bar{r}}^a = \bar{D}^a \bar{v}^a,\ \dot{\bar{D}}^a = \bar{D}^a[\bar{\omega}^a\times],\ \dot{\bar{\rho}}^a = \bar{D}^a[\bar{v}^a + \bar{\omega}^a \times \bar{c}^a], \tag{21}$$

from which we obtain

$$\bar{v}^a = (\bar{D}^a)^T \dot{\bar{r}}^a,\ \bar{\omega}^a = \mathrm{vec}[(\bar{D}^a)^T \dot{\bar{D}}^a]. \tag{22}$$

2.4 Change of reference strand

The sequence and configuration of a DNA molecule can be described in two different ways due to the freedom in choice of reference strand. Choosing one strand as a reference, the sequence and configuration is described by

$$\{X_a, \boldsymbol{r}^a, \bar{\boldsymbol{r}}^a, \{\boldsymbol{d}_i^a\}, \{\bar{\boldsymbol{d}}_i^a\}, y^a, z^{a-1}\}, \tag{23}$$

where the index $a = 1,\ldots,n$ increases in the 5′ to 3′ direction along this strand. Alternatively, by choosing the opposite strand as reference, the sequence and configuration is described by

$$\{X_{a_*}^*, \boldsymbol{r}_*^{a_*}, \bar{\boldsymbol{r}}_*^{a_*}, \{\boldsymbol{d}_{*i}^{a_*}\}, \{\bar{\boldsymbol{d}}_{*i}^{a_*}\}, y_*^{a_*}, z_*^{a_*-1}\}, \tag{24}$$

where the index $a_* = 1,\ldots,n$ increases in the 5′ to 3′ direction along this strand.

The above two descriptions are necessarily related. From the anti-parallel nature of the two strands we find that a and a_* denote the same basepair when $a_* = n - a + 1$. As a result, from the Watson–Crick pairing rules we find that $X_{n-a+1}^* = \bar{X}_a$ and $\bar{X}_{n-a+1}^* = X_a$, and from the convention for assigning reference points and frames to a base we deduce for all $a = 1,\ldots,n$ that

$$\begin{aligned}
\boldsymbol{r}_*^{n-a+1} &= \bar{\boldsymbol{r}}^a, \bar{\boldsymbol{r}}_*^{n-a+1} = \boldsymbol{r}^a,\\
\boldsymbol{d}_{*1}^{n-a+1} &= \bar{\boldsymbol{d}}_1^a, \bar{\boldsymbol{d}}_{*1}^{n-a+1} = \boldsymbol{d}_1^a,\\
\boldsymbol{d}_{*2}^{n-a+1} &= -\bar{\boldsymbol{d}}_2^a, \bar{\boldsymbol{d}}_{*2}^{n-a+1} = -\boldsymbol{d}_2^a,\\
\boldsymbol{d}_{*3}^{n-a+1} &= -\bar{\boldsymbol{d}}_3^a, \bar{\boldsymbol{d}}_{*3}^{n-a+1} = -\boldsymbol{d}_3^a.
\end{aligned} \tag{25}$$

Moreover, from the definitions of the relative rotation and displacement coordinates, together with the relation between the Cayley coordinates, axis and angle of a rotation matrix, we obtain

$$\begin{aligned}
y_*^{n-a+1} &= P y^a,\ (a = 1,\ldots,n),\\
z_*^{n-a} &= P z^a,\quad (a = 1,\ldots,n-1),
\end{aligned} \tag{26}$$

where $P = \mathrm{diag}(-1,1,1,-1,1,1) \in \mathbb{R}^{6\times6}$ is a constant, diagonal matrix with the property that $P = P^T = P^{-1}$. This property will be exploited throughout our developments. The transformation rule relating the external coordinates z_*^0 and z^0 is more complicated because it involves the relative rotation

and displacement coordinates along the entire length of a molecule. We omit this transformation since we will make no use of it.

Just as with the sequence and configuration variables, the mass properties and velocity variables can also be described in two different ways due to the freedom in choice of reference strand. Choosing one strand as a reference, these quantities are described by

$$\{m^a, \bar{m}^a, \boldsymbol{c}^a, \bar{\boldsymbol{c}}^a, \boldsymbol{\Gamma}^a, \bar{\boldsymbol{\Gamma}}^a, \boldsymbol{v}^a, \bar{\boldsymbol{v}}^a, \boldsymbol{\omega}^a, \bar{\boldsymbol{\omega}}^a\}, \tag{27}$$

where the index $a = 1,\ldots,n$ increases in the 5′ to 3′ direction along this strand. Alternatively, by choosing the opposite strand as reference, the mass properties and velocity variables are described by

$$\{m_*^{a_*}, \bar{m}_*^{a_*}, \boldsymbol{c}_*^{a_*}, \bar{\boldsymbol{c}}_*^{a_*}, \boldsymbol{\Gamma}_*^{a_*}, \bar{\boldsymbol{\Gamma}}_*^{a_*}, \boldsymbol{v}_*^{a_*}, \bar{\boldsymbol{v}}_*^{a_*}, \boldsymbol{\omega}_*^{a_*}, \bar{\boldsymbol{\omega}}_*^{a_*}\}, \tag{28}$$

where the index $a_* = 1,\ldots,n$ increases in the 5′ to 3′ direction along this strand.

As before, the above two descriptions are necessarily related due to the fact that a and a_* denote the same basepair when $a_* = n - a + 1$. From the transformation rules in eqn (25) for the configuration variables, and the fact that $X^*_{n-a+1} = \bar{X}_a$ and $\bar{X}^*_{n-a+1} = X_a$, we deduce for all $a = 1,\ldots,n$ that

$$\begin{aligned}
m_*^{n-a+1} &= \bar{m}^a,\ \bar{m}_*^{n-a+1} = m^a,\\
\boldsymbol{c}_*^{n-a+1} &= \bar{\boldsymbol{c}}^a,\ \bar{\boldsymbol{c}}_*^{n-a+1} = \boldsymbol{c}^a,\\
\boldsymbol{\Gamma}_*^{n-a+1} &= \bar{\boldsymbol{\Gamma}}^a,\ \bar{\boldsymbol{\Gamma}}_*^{n-a+1} = \boldsymbol{\Gamma}^a,\\
\boldsymbol{v}_*^{n-a+1} &= \bar{\boldsymbol{v}}^a,\ \bar{\boldsymbol{v}}_*^{n-a+1} = \boldsymbol{v}^a,\\
\boldsymbol{\omega}_*^{n-a+1} &= \bar{\boldsymbol{\omega}}^a,\ \bar{\boldsymbol{\omega}}_*^{n-a+1} = \boldsymbol{\omega}^a.
\end{aligned} \tag{29}$$

Let $c^a \in \mathbb{R}^3$, $\bar{c}_*^{a_*} \in \mathbb{R}^3$, $\Gamma^a \in \mathbb{R}^{3\times 3}$ and so on denote the components of the above vector and tensor quantities in the respective base frames, so that $c_i^a = \boldsymbol{d}_i^a \cdot \boldsymbol{c}^a$, $\bar{c}_{*i}^{a_*} = \boldsymbol{d}_{*i}^{a_*} \cdot \bar{\boldsymbol{c}}_*^{a_*}$, $\Gamma_{ij}^a = \boldsymbol{d}_i^a \cdot \boldsymbol{\Gamma}^a \boldsymbol{d}_j^a$ and so on. Then from eqn (29) and the transformation rules for the base frames in eqn (25) we deduce the component relations

$$\begin{aligned}
m_*^{n-a+1} &= \bar{m}^a,\ \bar{m}_*^{n-a+1} = m^a,\\
c_*^{n-a+1} &= \Psi \bar{c}^a,\ \bar{c}_*^{n-a+1} = \Psi c^a,\\
\Gamma_*^{n-a+1} &= \Psi \bar{\Gamma}^a \Psi,\ \bar{\Gamma}_*^{n-a+1} = \Psi \Gamma^a \Psi,\\
v_*^{n-a+1} &= \Psi \bar{v}^a,\ \bar{v}_*^{n-a+1} = \Psi v^a,\\
\omega_*^{n-a+1} &= \Psi \bar{\omega}^a,\ \bar{\omega}_*^{n-a+1} = \Psi \omega^a,
\end{aligned} \tag{30}$$

where $\Psi = \mathrm{diag}(1,-1,-1) \in \mathbb{R}^{3\times 3}$ is a constant, diagonal matrix with the property that $\Psi = \Psi^T = \Psi^{-1}$. This property will be exploited throughout. For later calculations, it will be convenient to introduce the velocity component vectors $\nu^a = (v^a, \omega^a) \in \mathbb{R}^6$ and $\bar{\nu}^a = (\bar{v}^a, \bar{\omega}^a) \in \mathbb{R}^6$ ($a = 1,\ldots,n$). From eqn (30) and the definition of Ψ we deduce that

$$\nu_*^{n-a+1} = -P\bar{\nu}^a, \bar{\nu}_*^{n-a+1} = -P\nu^a, \tag{31}$$

where P is the transformation matrix introduced above.

2.5 Internal elastic energy

For a molecule of n basepairs we consider an internal elastic energy function U of the general quadratic form

$$U(\mathrm{w}) = \tfrac{1}{2}(\mathrm{w} - \hat{\mathrm{w}})\cdot \mathrm{K}(\mathrm{w} - \hat{\mathrm{w}}), \tag{32}$$

where $\mathrm{w} = (y^1, z^1, y^2, z^2, \ldots, z^{n-1}, y^n) \in \mathbb{R}^{12n-6}$ is the vector of internal coordinates, $\mathrm{K} \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a symmetric, positive-definite matrix of stiffness parameters and $\hat{\mathrm{w}} \in \mathbb{R}^{12n-6}$ is a vector of shape parameters that represents the equilibrium value of w. The expression in (32) can be written in the equivalent form

$$U(\mathrm{w}) = \frac{1}{2}\sum_{\alpha,\beta=1}^{2n-1} (\mathrm{w}^\alpha - \hat{\mathrm{w}}^\alpha)\cdot \mathrm{K}^{\alpha\beta}(\mathrm{w}^\beta - \hat{\mathrm{w}}^\beta), \tag{33}$$

where $\mathrm{w}^\alpha \in \mathbb{R}^6$, $\hat{\mathrm{w}}^\alpha \in \mathbb{R}^6$ and $\mathrm{K}^{\alpha\beta} \in \mathbb{R}^{6\times 6}$ ($\alpha, \beta = 1,\ldots,2n-1$) denote the block entries of w, ŵ and K. Notice that $\mathrm{w}^{2a-1} = y^a$ for $a = 1,\ldots,n$ and $\mathrm{w}^{2a} = z^a$ for $a = 1,\ldots,n-1$. Thus the odd-numbered blocks in w correspond to the coordinates y^a and the even-numbered blocks correspond to the coordinates z^a.

We assume that the material parameters $\hat{\mathrm{w}}^\alpha$ and $\mathrm{K}^{\alpha\beta}$ are completely determined by the sequence $X_1\cdots X_n$ along the reference strand. Equivalently, we assume there exist functions $\mathbb{W}$ and $\mathbb{K}$ such that

$$\begin{aligned}
\hat{\mathrm{w}}^\alpha &= \mathbb{W}(X_1,\ldots,X_n,\alpha),\\
\mathrm{K}^{\alpha\beta} &= \mathbb{K}(X_1,\ldots,X_n,\alpha,\beta),
\end{aligned} \qquad \alpha,\beta = 1,\ldots,2n-1. \tag{34}$$

This rather mild assumption has some immediate consequences. For example, let $U(\mathrm{w})$ and $U_*(\mathrm{w}_*)$ denote the internal energies of a molecule computed using the two different choices of reference strand. Thus $U(\mathrm{w})$ is given by the expression in (33) with the parameters in (34), and $U_*(\mathrm{w}_*)$ is given by an exactly analogous expression with the parameters

$$\begin{aligned}
\hat{\mathrm{w}}_*^{\alpha_*} &= \mathbb{W}(X_1^*,\ldots,X_n^*,\alpha_*),\\
\mathrm{K}_*^{\alpha_*\beta_*} &= \mathbb{K}(X_1^*,\ldots,X_n^*,\alpha_*,\beta_*),
\end{aligned} \qquad \alpha_*,\beta_* = 1,\ldots,2n-1. \tag{35}$$

From the fact that $U(\mathrm{w})$ must equal $U_*(\mathrm{w}_*)$ for all possible configurations, together with the change of strand relations outlined in section 2.4, we deduce that the functions $\mathbb{W}$ and $\mathbb{K}$ must satisfy the relations

$$\begin{aligned}
\mathbb{W}(X_1,\ldots,X_n,\alpha) &= P\mathbb{W}(\bar{X}_n,\ldots,\bar{X}_1,2n-\alpha),\\
\mathbb{K}(X_1,\ldots,X_n,\alpha,\beta) &= P\mathbb{K}(\bar{X}_n,\ldots,\bar{X}_1,2n-\alpha,2n-\beta)P,
\end{aligned} \tag{36}$$

where P is the transformation matrix introduced in section 2.4. Thus the functions $\mathbb{W}$ and $\mathbb{K}$, and hence the internal energy parameters $\hat{\mathrm{w}}^\alpha$ and $\mathrm{K}^{\alpha\beta}$, cannot be completely arbitrary.

The general quadratic internal energy in eqn (33) allows couplings between bases along the entire length of a molecule. This assumption can be relaxed in various ways. For example, one may assume that couplings between distant bases along the chain are negligible and only consider interactions between neighboring bases. By a local or nearest-neighbor internal

energy for a rigid base model we mean an internal energy of the form

$$U(\mathrm{w}) = \sum_{a=1}^{n-1} U^a(y^a, z^a, y^{a+1}). \tag{37}$$

In this expression, each U^a can be interpreted as a local energy associated with the central junction between the four bases in the basepairs $(X, \bar{X})_a$ and $(X, \bar{X})_{a+1}$. Notice that the relative rotations and displacements between the four bases, both across and along the two backbone strands, are determined by the internal coordinates (y^a, z^a, y^{a+1}). In view of the ordering of the block entries of w, the quadratic energy in eqn (33) is of the local form eqn (37) when the block entries of K satisfy

$$\mathrm{K}^{\alpha\beta} = 0 \text{ when } \begin{cases} \alpha \text{ odd}, & |\beta - \alpha| > 2, \\ \alpha \text{ even}, & |\beta - \alpha| > 1. \end{cases} \tag{38}$$

2.6 Kinetic energy

To each base X_a on the reference strand we associate a kinetic energy Φ^a defined by its linear and angular velocity components as

$$\Phi^a = \tfrac{1}{2} m^a |v^a + \omega^a \times c^a|^2 + \tfrac{1}{2} \omega^a \cdot \Gamma^a \omega^a. \tag{39}$$

By expanding the first term, we find that this energy can be written in the convenient form

$$\Phi^a = \tfrac{1}{2} \nu^a \cdot M^a \nu^a, \tag{40}$$

where $\nu^a = (v^a, \omega^a) \in \mathbb{R}^6$ and $M^a \in \mathbb{R}^{6\times 6}$ with inverse $(M^a)^{-1} \in \mathbb{R}^{6\times 6}$ is a generalized mass matrix given in block form by

$$M^a = \begin{pmatrix} m^a I & m^a [c^a\times]^T \\ m^a [c^a\times] & \Gamma^a + m^a [c^a\times][c^a\times]^T \end{pmatrix},$$

$$(M^a)^{-1} = \begin{pmatrix} m_a^{-1} I + [c^a\times]\Gamma_a^{-1}[c^a\times]^T & [c^a\times]\Gamma_a^{-1} \\ \Gamma_a^{-1}[c^a\times]^T & \Gamma_a^{-1} \end{pmatrix}. \tag{41}$$

Here $[c^a\times] \in \mathbb{R}^{3\times 3}$ is the skew-symmetric matrix defined by the components $c^a \in \mathbb{R}^3$ as in eqn (7), $I \in \mathbb{R}^{3\times 3}$ is the identity matrix, $\Gamma_a \in \mathbb{R}^{3\times 3}$ is the matrix of rotational inertia components and $\Gamma_a^{-1} \in \mathbb{R}^{3\times 3}$ is its inverse.

Just as for the reference strand, to each base $\bar{X}_a$ on the opposite strand we associate a kinetic energy $\bar{\Phi}^a$ defined by its linear and angular velocity components as

$$\bar{\Phi}^a = \tfrac{1}{2} \bar{m}^a |\bar{v}^a + \bar{\omega}^a \times \bar{c}^a|^2 + \tfrac{1}{2} \bar{\omega}^a \cdot \bar{\Gamma}^a \bar{\omega}^a. \tag{42}$$

As before, by expanding the first term, we find that this energy can be written in the convenient form

$$\bar{\Phi}^a = \tfrac{1}{2} \bar{\nu}^a \cdot \bar{M}^a \bar{\nu}^a, \tag{43}$$

where $\bar{\nu}^a = (\bar{v}^a, \bar{\omega}^a) \in \mathbb{R}^6$ and $\bar{M}^a \in \mathbb{R}^{6\times 6}$ with inverse $(\bar{M}^a)^{-1} \in \mathbb{R}^{6\times 6}$ is a generalized mass matrix defined in a manner exactly analogous to eqn (41).

By summing over each base on each of the two strands we find that the total kinetic energy of a molecule with n basepairs can be written in the form

$$\Phi(\mathrm{v}) = \tfrac{1}{2} \mathrm{v} \cdot \mathrm{M} \mathrm{v}, \tag{44}$$

where $\mathrm{v} = (\nu^1, \bar{\nu}^1, \ldots, \nu^n, \bar{\nu}^n) \in \mathbb{R}^{12n}$ is a vector of velocity components and $\mathrm{M} = \mathrm{diag}[M^1, \bar{M}^1, \ldots, M^n, \bar{M}^n] \in \mathbb{R}^{12n\times 12n}$ is a block diagonal matrix of mass parameters. Notice that $\mathrm{v}^{2a-1} = \nu^a$ and $\mathrm{v}^{2a} = \bar{\nu}^a$. Thus the odd-numbered blocks $\mathrm{v}^\alpha \in \mathbb{R}^6$ and $\mathrm{M}^\alpha \in \mathbb{R}^{6\times 6}$ ($\alpha = 1,3,\ldots,2n-1$) correspond to velocity components and mass parameters of bases on the reference strand, whereas the even-numbered blocks $\mathrm{v}^\alpha \in \mathbb{R}^6$ and $\mathrm{M}^\alpha \in \mathbb{R}^{6\times 6}$ ($\alpha = 2, 4, \ldots, 2n$) correspond to velocity components and mass parameters of bases on the opposite strand.

We assume that the mass parameters M^α are completely determined by the sequence $X_1 \cdots X_n$ along the reference strand. Equivalently, we assume there exists a function $\mathbb{M}$ such that

$$\mathrm{M}^\alpha = \mathbb{M}(X_1, \ldots, X_n, \alpha), \ \alpha = 1, \ldots, 2n. \tag{45}$$

As before, this assumption has some immediate consequences. For example, let $\Phi(\mathrm{v})$ and $\Phi_*(\mathrm{v}_*)$ denote the kinetic energies of a molecule computed using the two different choices of reference strand. Thus $\Phi(\mathrm{v})$ is given by the expression in (44) with the parameters in (45), and $\Phi_*(\mathrm{v}_*)$ is given by an exactly analogous expression with the parameters

$$\mathrm{M}_*^{\alpha_*} = \mathbb{M}(X_1^*, \ldots, X_n^*, \alpha_*), \ \alpha_* = 1, \ldots, 2n. \tag{46}$$

From the fact that $\Phi(\mathrm{v})$ must equal $\Phi_*(\mathrm{v}_*)$ for all possible values of the velocity components, together with the change of strand relations outlined in section 2.4, we deduce that the function $\mathbb{M}$ must satisfy the relation

$$\mathbb{M}(X_1, \ldots, X_n, \alpha) = P\mathbb{M}(\bar{X}_n, \ldots, \bar{X}_1, 2n - \alpha + 1)P, \tag{47}$$

where P is the transformation matrix introduced in section 2.4.

2.7 Canonical measure

The equilibrium statistical properties of a rigid base model of DNA in contact with a heat bath are described by the standard canonical measure $\mathrm{d}\mu$. This measure is a function of the absolute heat bath temperature $T > 0$, the shape and stiffness parameters $\hat{\mathrm{w}}$ and K, and the mass parameters M. The explicit form of $\mathrm{d}\mu$ and the relative ease with which information can be extracted from it depend on the choice of variables used to describe the mechanical state of the model. Here we introduce a choice of variables which will simplify $\mathrm{d}\mu$ and prove convenient for estimating the parameters $\hat{\mathrm{w}}$, K and M.

The classic form of $\mathrm{d}\mu$ is obtained when model states are described in terms of canonical variables as defined in the theory of Hamiltonian systems. Standard canonical variables for a rigid base model as considered here take the form $(\xi, \zeta, \vartheta, \theta, \varphi_\xi, \varphi_\zeta, \varphi_\vartheta, \varphi_\theta)$, where $(\xi, \zeta, \vartheta, \theta) \in \mathbb{R}^{12n}$ are any independent coordinates for the base reference points and frames, and $(\varphi_\xi, \varphi_\zeta, \varphi_\vartheta, \varphi_\theta) \in \mathbb{R}^{12n}$ are their associated canonical momenta. If the kinetic energy Φ is expressed in terms of $(\xi, \zeta, \vartheta, \theta)$ and their time derivatives $(\dot{\xi}, \dot{\zeta}, \dot{\vartheta}, \dot{\theta})$, then the canonical momenta are defined by $\varphi_\xi = \partial\Phi/\partial\dot{\xi}$, $\varphi_\zeta = \partial\Phi/\partial\dot{\zeta}$, $\varphi_\vartheta = \partial\Phi/\partial\dot{\vartheta}$ and $\varphi_\theta = \partial\Phi/\partial\dot{\theta}$. In terms of standard canonical variables, the total mechanical energy or Hamiltonian function for the model is

$$H = U(\xi, \zeta, \vartheta, \theta) + \Phi(\xi, \zeta, \vartheta, \theta, \varphi_\xi, \varphi_\zeta, \varphi_\vartheta, \varphi_\theta), \tag{48}$$

and the measure $d\mu$ takes the usual form[17]

$$d\mu = (1/Z)e^{-H/k_B T}\, d\xi d\zeta d\vartheta d\theta d\varphi_\xi d\varphi_\zeta d\varphi_\vartheta d\varphi_\theta, \qquad (49)$$

where k_B is the Boltzmann constant and $Z > 0$ is a normalizing constant.

When $(\xi, \zeta, \vartheta, \theta)$ are the relative displacement and rotation coordinates defined in section 2.2, the form of the potential energy U is convenient, but the kinetic energy Φ expressed in the associated canonical momenta is configuration dependent. We find this form of the measure $d\mu$ to be inadequate for our purposes, because it provides little insight into the relation between moments of $d\mu$ and the parameters $\hat{w}$, K and M. One could instead choose momentum coordinates in which the form of the kinetic energy is simple, but in the associated canonical configuration coordinates, the potential energy would be complicated, and again the moments of the measure $d\mu$ would not be simply related to the parameters $\hat{w}$, K and M.

A more useful form for the measure $d\mu$ can be obtained by changing to the non-canonical variables $(\xi, \zeta, \vartheta, \theta, v, \bar{v}, \omega, \bar{\omega})$, where $(v, \bar{v}, \omega, \bar{\omega}) \in \mathbb{R}^{12n}$ denote the linear and angular velocity components introduced in section 2.3. In these variables, the Hamiltonian takes the simple, separable form

$$H = U(\xi, \zeta, \vartheta, \theta) + \Phi(v, \bar{v}, \omega, \bar{\omega}), \qquad (50)$$

and the measure $d\mu$ becomes

$$d\mu = (1/Z)e^{-H/k_B T} J\, d\xi d\zeta d\vartheta d\theta dv d\bar{v} d\omega d\bar{\omega}, \qquad (51)$$

where J is the Jacobian associated with the change of variables. A tedious application of the chain rule of multi-variable calculus shows that

$$J = \left[\prod_{a=0}^{n-1}\left(1+\frac{1}{4}|\theta^a|^2\right)^{-2}\right]\left[\prod_{a=1}^{n}\left(1+\frac{1}{4}|\vartheta^a|^2\right)^{-2}\right]. \qquad (52)$$

When expressed in the form (51) the measure $d\mu$ has the desirable feature that it is factorable into three independent measures $d\mu_{vel}$, $d\mu_{con}^{int}$ and $d\mu_{con}^{ext}$, where

$$\begin{aligned} d\mu_{vel} &= (1/Z_{vel})e^{-\Phi(v)/k_B T}\, dv, \\ d\mu_{con}^{int} &= (1/Z_{con}^{int})e^{-U(w)/k_B T} J'\, dw, \\ d\mu_{con}^{ext} &= (1/Z_{con}^{ext}) J^0\, dz^0. \end{aligned} \qquad (53)$$

Here v is the vector of all velocity components, w is the vector of all internal configuration coordinates, z^0 is the vector of external configuration coordinates, and J' and J^0 are reduced Jacobian factors given by

$$\begin{aligned} J' &= \left[\prod_{a=1}^{n-1}\left(1+\frac{1}{4}|\theta^a|^2\right)^{-2}\right]\left[\prod_{a=1}^{n}\left(1+\frac{1}{4}|\vartheta^a|^2\right)^{-2}\right], \\ J^0 &= \left(1+\frac{1}{4}|\theta^0|^2\right)^{-2}. \end{aligned} \qquad (54)$$

The measures $d\mu_{con}^{int}$ and $d\mu_{con}^{ext}$ necessarily involve Jacobian factors due to the non-Cartesian nature of the coordinates w and z^0. While such factors are typically ignored, or assumed to be constant, we include them here.

The statistical mechanical average of any state function $\phi = \phi(v, w, z^0)$ with respect to the measure $d\mu$ is given by

$$\langle\phi\rangle = \frac{\int \phi(v, w, z^0) e^{-H(v,w)/k_B T} J\, dv\, dw\, dz^0}{\int e^{-H(v,w)/k_B T} J\, dv\, dw\, dz^0}, \qquad (55)$$

where the integrations are performed over the domain $\mathbb{R}^{12n} \times \mathbb{R}^{12n-6} \times D^0$. Here $\mathbb{R}^{12n}$ is the domain for v, $\mathbb{R}^{12n-6}$ is the domain for w, and $D^0 \subset \mathbb{R}^6$ is a prescribed domain for z^0, which will play no role in our developments. Due to the factorability of $d\mu$, for any function $\psi = \psi(v)$ we find

$$\langle\psi\rangle = \frac{\int \psi(v) e^{-\Phi(v)/k_B T}\, dv}{\int e^{-\Phi(v)/k_B T}\, dv}. \qquad (56)$$

Moreover, for any function $\chi = \chi(w)$ we find

$$\frac{\langle\chi/J'\rangle}{\langle 1/J'\rangle} = \frac{\int \chi(w) e^{-U(w)/k_B T}\, dw}{\int e^{-U(w)/k_B T}\, dw}. \qquad (57)$$

2.8 Moment-parameter relations

Here we exploit the relations in (56) and (57) to derive explicit characterizations of the model parameters $\hat{w}$, K and M introduced in sections 2.5 and 2.6. In our developments below, we use the notation $w \otimes w$ and $v \otimes v$ to denote the usual outer or tensor product of the vectors w and v. Thus $[w \otimes w]_{pq} = w_p w_q$ and $[v \otimes v]_{pq} = v_p v_q$.

Notice that, although the potential energy $U(w)$ is quadratic, the measure $d\mu_{con}^{int}$ is non-Gaussian due to the presence of the Jacobian factor J'. As a result, the parameters $\hat{w}$ and K are not given by the usual moment relations for Gaussian measures. However, from (57) we see that the shape parameters $\hat{w}$ can be characterized as a ratio of expected values. In particular, substituting the vector-valued function $\chi(w) = w$ into (57) and carrying out the indicated integrations we obtain, by standard results for Gaussian integrals,[16]

$$\frac{\langle w/J'\rangle}{\langle 1/J'\rangle} = \hat{w}. \qquad (58)$$

Thus $\hat{w}$ is not equal to the expected value of w, but rather a ratio of expected values which are weighted by the Jacobian J'.

The stiffness matrix K can also be similarly characterized. Substituting the matrix-valued function $\chi = \Delta w \otimes \Delta w$ into (57), where $\Delta w = w - \hat{w}$, we get, again by standard results for Gaussian integrals,

$$\frac{\langle \Delta w \otimes \Delta w/J'\rangle}{\langle 1/J'\rangle} = k_B T K^{-1}. \qquad (59)$$

Thus, just as with $\hat{w}$, the matrix $k_B T K^{-1}$ is not equal to the expected value of $\Delta w \otimes \Delta w$, but rather a ratio of weighted expected values. By expanding the left-hand side of (59) and using (58) we deduce

$$\frac{\langle w \otimes w/J'\rangle}{\langle 1/J'\rangle} = k_B T K^{-1} + \hat{w} \otimes \hat{w}, \qquad (60)$$

which may be more convenient than (59) since the average $\langle w \otimes w/J'\rangle$ is independent of the parameters $\hat{w}$.

Due to intrinsic properties of the set of three-dimensional rotations, the Jacobian factor J' will be non-constant and

consequently the measure $d\mu_{con}^{int}$ will be non-Gaussian for arbitrary choices of internal coordinates. However, when the gradient of J′ is sufficiently small and $d\mu_{con}^{int}$ is sufficiently concentrated, it is reasonable to expect that variations in J′ can be neglected. By the Gaussian approximation of the internal configuration measure $d\mu_{con}^{int}$ we mean the measure obtained by assuming J′ to be constant. In this approximation, the relations in (58), (59) and (60) take the simplified forms

$$\langle \mathrm{w} \rangle = \hat{\mathrm{w}}, \langle \Delta\mathrm{w} \otimes \Delta\mathrm{w} \rangle = k_{\mathrm{B}}T\mathrm{K}^{-1}, \\ \langle \mathrm{w} \otimes \mathrm{w} \rangle = k_{\mathrm{B}}T\mathrm{K}^{-1} + \hat{\mathrm{w}} \otimes \hat{\mathrm{w}}. \quad (61)$$

The characterization of the mass matrix M is simpler due to the genuinely Gaussian form of the measure $d\mu_{vel}$. In particular, substituting the matrix-valued function $\psi(\mathrm{v}) = \mathrm{v} \otimes \mathrm{v}$ into eqn (56) and using standard results for Gaussian integrals we get

$$\langle \mathrm{v} \otimes \mathrm{v} \rangle = k_{\mathrm{B}}T\mathrm{M}^{-1}. \quad (62)$$

Thus the matrix $k_{\mathrm{B}}T\mathrm{M}^{-1}$ is equal to the expected value of v ⊗ v. Moreover, since $\mathrm{M} = \mathrm{diag}(M^1,\ldots,\bar{M}^n)$, we have $\mathrm{M}^{-1} = \mathrm{diag}([M^1]^{-1},\ldots,[\bar{M}^n]^{-1})$, where M^a and $\bar{M}^a$, along with their inverses, are explicit functions of the mass parameters of bases X_a and $\bar{X}_a$ as defined in eqn (41).

2.9 Material symmetries

Here we derive various implications of the internal and kinetic energy strand invariance relations (36) and (47) for the special case of a palindromic molecule. Thus we consider the case when the base sequence $\bar{X}_n \cdots \bar{X}_1$ is identical to $X_1 \cdots X_n$, which requires that n necessarily be even. This case will be considered in our numerical simulations described later. The results derived here follow directly from the assumed existence of the functions $\mathbb{W}$, $\mathbb{K}$ and $\mathbb{M}$. We stress that we do not assume or impose any structure on these functions such as locality or bandedness.

We first consider the internal energy or shape and stiffness parameters ŵ and K. Assuming $\bar{X}_n \cdots \bar{X}_1$ is identical to $X_1 \cdots X_n$, we deduce from (36) that

$$\hat{\mathrm{w}}^{\alpha} = P\hat{\mathrm{w}}^{2n-\alpha}, \mathrm{K}^{\alpha\beta} = P\mathrm{K}^{(2n-\alpha)(2n-\beta)}P, \alpha, \beta = 1,\ldots,2n-1, \quad (63)$$

where $\hat{\mathrm{w}}^{\alpha} = \mathbb{W}(X_1,\ldots,X_n,\alpha)$ and $\mathrm{K}^{\alpha\beta} = \mathbb{K}(X_1,\ldots,X_n,\alpha,\beta)$. Thus the parameters associated with different positions along the molecule must be related in a simple way through the matrix P.

From the first expression in eqn (63), and the definition of the entries in ŵ, we obtain $\hat{y}^1 = P\hat{y}^n$, $\hat{z}^1 = P\hat{z}^{n-1}$, $\hat{y}^2 = P\hat{y}^{n-1}$, $\hat{z}^2 = P\hat{z}^{n-2}$, and so on. From the expressions for the $\hat{y}$-parameters and the definition of P we deduce that the equilibrium values of propeller, opening, stretch and stagger must be symmetric about the middle of the molecule, whereas the equilibrium values of buckle and shear must be antisymmetric. Similarly, from the expressions for the $\hat{z}$-parameters we deduce that the equilibrium values of roll, twist, slide and rise must be symmetric about the middle of the molecule, whereas the equilibrium values of tilt and shift must be antisymmetric. Thus for a palindromic molecule the shape parameters must either be symmetric or antisymmetric functions of position about the middle of the molecule.

Further conclusions about the $\hat{z}$-parameters can be drawn based on the fact that n is even. In particular, setting $n = 2m$, and using the relations $\hat{z}^1 = P\hat{z}^{n-1}$, $\hat{z}^2 = P\hat{z}^{n-2}$, and so on, we get $\hat{z}^m = P\hat{z}^m$, which can be written in the equivalent form $(I - P)\hat{z}^m = 0$. From this result and the definition of P we deduce that the equilibrium values of tilt and shift at the middle junction m must vanish. Thus for a palindromic molecule there is a restriction on some of the equilibrium shape parameters at the precise middle of the molecule, which corresponds to a junction between basepairs.

From the second expression in (63), and for brevity considering only the diagonal blocks of K, we obtain $\mathrm{K}^{11} = P\mathrm{K}^{(2n-1)(2n-1)}P$, $\mathrm{K}^{22} = P\mathrm{K}^{(2n-2)(2n-2)}P$, and so on. Equivalently, when labeled according to the interactions they represent, we have $\mathrm{K}_{y^1y^1} = P\mathrm{K}_{y^ny^n}P$, $\mathrm{K}_{z^1z^1} = P\mathrm{K}_{z^{n-1}z^{n-1}}P$, and so on. Various conclusions similar to those outlined above can be drawn from these relations. For example, the diagonal stiffness associated with each of the twelve types of deformation variable must be symmetric about the middle of the molecule. Off-diagonal stiffnesses associated with different types of couplings can either be symmetric or antisymmetric. If we categorize the twelve types of deformation variables into two groups, odd (buckle, shear, tilt, shift) and even (all others), then the stiffness associated with an odd–odd or even–even coupling must be symmetric, whereas the stiffness associated with an odd–even coupling must be antisymmetric. Thus for a palindromic molecule the entries in the diagonal blocks of the stiffness matrix must either be symmetric or antisymmetric functions of position about the middle of the molecule.

Just as before, further conclusions about the K-parameters can be drawn based on the fact that n is even. In particular, setting $n = 2m$, we find that the diagonal block of K associated with junction m must satisfy $\mathrm{K}^{mm} = P\mathrm{K}^{mm}P$, or equivalently $\mathrm{K}^{mm} - P\mathrm{K}^{mm}P = 0$. From this result and the definition of P we deduce that the stiffness parameters corresponding to odd-even couplings, for example tilt–roll, tilt–twist, shift–roll, shift–twist and so on, must vanish. Thus for a palindromic molecule there is a restriction on the stiffness parameters in the diagonal block associated with the junction at the middle of the molecule.

We next consider the kinetic energy or mass parameters contained in the matrix M. Assuming $\bar{X}_n \cdots \bar{X}_1$ is identical to $X_1 \cdots X_n$, we deduce from eqn (47) that

$$\mathrm{M}^{\alpha} = P\mathrm{M}^{2n-\alpha+1}P, \alpha = 1,\ldots,2n, \quad (64)$$

where $\mathrm{M}^{\alpha} = \mathbb{M}(X_1,\ldots,X_n,\alpha)$. Thus the mass parameters associated with different positions along the molecule must be related in a simple way through the matrix P.

From the expression in (64), and the definition of the entries in M, we obtain $M^1 = P\bar{M}^nP$, $\bar{M}^1 = PM^nP$, $M^2 = P\bar{M}^{n-1}P$, and so on. These relations can be written in the general form

$$M^a = P\bar{M}^{n-a+1}P, \bar{M}^a = PM^{n-a+1}P, a = 1,\ldots,n. \quad (65)$$

By substituting $n - a + 1$ in place of a, and using the fact that $P^{-1} = P$, we find that the second relation is identical to the first. Thus the above relations are not independent.

Considering only the first relation, and using eqn (41) and the block representation $P = -\mathrm{diag}(\Psi, \Psi)$, where Ψ is the matrix introduced in section 2.4, we find

$$m^a = \bar{m}^{n-a+1},\ c^a = \Psi \bar{c}^{n-a+1},$$
$$\Gamma^a = \Psi \bar{\Gamma}^{n-a+1} \Psi,\ a = 1,\ldots,n. \qquad (66)$$

Thus for a palindromic molecule the mass parameters must have certain symmetry properties about the middle of the molecule as encapsulated in the above relations.

3. Rigid basepair model

Here we outline another model for the sequence-dependent structure of a DNA molecule. In contrast to the one considered in the previous section, the model here is coarser and based on basepairs rather than individual bases. We outline the theory for this model and derive compatibility relations that the material parameters of the two models must satisfy under appropriate assumptions. For brevity, we omit a discussion of material symmetries for special sequences as considered in the previous section, but note that analogous results can be derived exactly as before.

3.1 Basepairs, configurations

We consider a model of DNA in which each basepair is modeled as rigid. The configuration of an arbitrary basepair $(X, \bar{X})_a$ is specified by giving the location of a reference point $\boldsymbol{q}^a$ fixed in the basepair, and the orientation of a right-handed, orthonormal frame $\{\boldsymbol{g}_i^a\}$ $(i = 1,2,3)$ attached to the basepair. The reference point and frame vectors are defined as described in section 2. Thus in this model the configuration of a DNA molecule consisting of n basepairs is completely defined by the reference points $\boldsymbol{q}^a$ and frames $\{\boldsymbol{g}_i^a\}$ $(a = 1,\ldots,n)$. These points and frames are in turn uniquely defined by component vectors $q^a \in \mathbb{R}^3$ and rotation matrices $G^a \in \mathbb{R}^{3\times 3}$, where $q_i^a = \boldsymbol{e}_i \cdot \boldsymbol{q}^a$, $G_{ij}^a = \boldsymbol{e}_i \cdot \boldsymbol{g}_j^a$ and $\{\boldsymbol{e}_i\}$ denotes an arbitrary, lab-fixed frame. In terms of these components, we have

$$\boldsymbol{g}_j^a = \sum_{i=1}^{3} G_{ij}^a \boldsymbol{e}_i,\ \boldsymbol{q}^a = \sum_{i=1}^{3} q_i^a \boldsymbol{e}_i. \qquad (67)$$

3.2 Rotation, displacement coordinates

The three-dimensional shape of a DNA molecule within the rigid basepair model is determined entirely by the relative rotation and displacement between neighboring basepairs along the strands. The relative rotation and displacement between $(X, \bar{X})_a$ and $(X, \bar{X})_{a+1}$ along the strands is described by the relations

$$\boldsymbol{g}_j^{a+1} = \sum_{i=1}^{3} L_{ij}^a \boldsymbol{g}_i^a,\ \boldsymbol{q}^{a+1} = \boldsymbol{q}^a + \sum_{i=1}^{3} \zeta_i^a \boldsymbol{h}_i^a, \qquad (68)$$

where $L^a \in \mathbb{R}^{3\times 3}$ is the rotation matrix which describes the orientation of frame $\{\boldsymbol{g}_i^{a+1}\}$ with respect to $\{\boldsymbol{g}_i^a\}$, $\zeta^a \in \mathbb{R}^3$ is the coordinate vector which describes the position of $\boldsymbol{q}^{a+1}$ with respect to $\boldsymbol{q}^a$, and $\{\boldsymbol{h}_i^a\}$ is the junction frame midway between the basepair frames $\{\boldsymbol{g}_i^a\}$ and $\{\boldsymbol{g}_i^{a+1}\}$. The rotation matrix L^a is parameterized by a coordinate vector $\theta^a \in \mathbb{R}^3$ so that $L^a = \mathrm{cay}[\theta^a]$ and $\theta^a = \mathrm{cay}^{-1}[L^a]$, and from (68) we have

$$L_{ij}^a = \boldsymbol{g}_i^a \cdot \boldsymbol{g}_j^{a+1},\ \zeta_i^a = \boldsymbol{h}_i^a \cdot (\boldsymbol{q}^{a+1} - \boldsymbol{q}^a). \qquad (69)$$

As in the rigid base model, the complete configuration of a DNA molecule is specified by introducing coordinates θ^0 and ζ^0 for the first basepair frame $\{\boldsymbol{g}_i^1\}$ and reference point $\boldsymbol{q}^1$ with respect to the lab-fixed frame $\{\boldsymbol{e}_i\}$, that is,

$$\boldsymbol{g}_j^1 = \sum_{i=1}^{3} \mathrm{cay}_{ij}[\theta^0] \boldsymbol{e}_i,\ \boldsymbol{q}^1 = \sum_{i=1}^{3} \zeta_i^0 \boldsymbol{e}_i. \qquad (70)$$

Thus the coordinates $z^a = (\theta, \zeta)^a \in \mathbb{R}^6$ $(a = 0,\ldots,n-1)$ completely define the configuration of a molecule with n basepairs. Notice that z^0 are external coordinates that specify the spatial location of the molecule, whereas z^a $(a = 1,\ldots,n-1)$ are internal coordinates which describe its shape. We remark that $z^a = (\theta,\zeta)^a$ are tilt–roll–twist and shift–slide–rise variables defined exactly as before. Thus the internal coordinates for the rigid basepair model are a subset of those for the rigid base model introduced in section 2.2.

3.3 Mass properties, kinematics

To each basepair $(X, \bar{X})_a$ we ascribe a total mass m_{bp}^a, a symmetric rotational inertia tensor $\boldsymbol{\Gamma}_{\mathrm{bp}}^a$ with respect to the mass center, and a vector $\boldsymbol{c}_{\mathrm{bp}}^a$ that locates the mass center relative to the base reference point $\boldsymbol{q}^a$, so that $\boldsymbol{\rho}_{\mathrm{bp}}^a = \boldsymbol{q}^a + \boldsymbol{c}_{\mathrm{bp}}^a$ is the position vector of the mass center. The kinematics of $(X, \bar{X})_a$ are encapsulated in the relations

$$\dot{\boldsymbol{q}}^a = \boldsymbol{v}_{\mathrm{bp}}^a,\ \dot{\boldsymbol{g}}_i^a = \boldsymbol{\omega}_{\mathrm{bp}}^a \times \boldsymbol{g}_i^a, \qquad (71)$$

where $\boldsymbol{v}_{\mathrm{bp}}^a$ is the velocity of the basepair reference point and $\boldsymbol{\omega}_{\mathrm{bp}}^a$ is the angular velocity of the basepair frame. Because $\boldsymbol{c}_{\mathrm{bp}}^a$ is fixed in the basepair, we have $\dot{\boldsymbol{c}}_{\mathrm{bp}}^a = \boldsymbol{\omega}_{\mathrm{bp}}^a \times \boldsymbol{c}_{\mathrm{bp}}^a$, and we find that the velocity of the center of mass of $(X, \bar{X})_a$ is given by

$$\dot{\boldsymbol{\rho}}_{\mathrm{bp}}^a = \boldsymbol{v}_{\mathrm{bp}}^a + \boldsymbol{\omega}_{\mathrm{bp}}^a \times \boldsymbol{c}_{\mathrm{bp}}^a. \qquad (72)$$

Let $\rho_{\mathrm{bp}}^a \in \mathbb{R}^3$ denote components in the lab-fixed frame, and let $c_{\mathrm{bp}}^a \in \mathbb{R}^3$, $v_{\mathrm{bp}}^a \in \mathbb{R}^3$, $\omega_{\mathrm{bp}}^a \in \mathbb{R}^3$, $\Gamma_{\mathrm{bp}}^a \in \mathbb{R}^{3\times 3}$ denote components in the associated basepair frame, so that $(\rho_{\mathrm{bp}})_i^a = \boldsymbol{e}_i \cdot \boldsymbol{\rho}_{\mathrm{bp}}^a$, $(c_{\mathrm{bp}})_i^a = \boldsymbol{g}_i^a \cdot \boldsymbol{c}_{\mathrm{bp}}^a$, $(\Gamma_{\mathrm{bp}})_{ij}^a = \boldsymbol{g}_i^a \cdot \boldsymbol{\Gamma}_{\mathrm{bp}}^a \boldsymbol{g}_j^a$ and so on. Then from (67), (71) and (72) we deduce the component relations

$$\dot{q}^a = G^a v_{\mathrm{bp}}^a,\ \dot{G}^a = G^a[\omega_{\mathrm{bp}}^a\times],\ \dot{\rho}_{\mathrm{bp}}^a = G^a[v_{\mathrm{bp}}^a + \omega_{\mathrm{bp}}^a \times c_{\mathrm{bp}}^a]. \qquad (73)$$

Notice that, unless a physical basepair exactly satisfies the assumption of rigidity, the relations between the mass parameters and velocity variables introduced here and those for a rigid base model introduced in section 2.3 are not simple. Indeed, the rigid basepair quantities can be viewed as non-trivial averages of the rigid base quantities where the weighting depends on the relative placement and motion between the bases. Thus, in contrast to the internal coordinates discussed above, the rigid basepair mass parameters and velocity variables will in general not be subsets or simple combinations of those for the rigid base model.

3.4 Change of reference strand

Just as for the rigid base model, the sequence and configuration in a rigid basepair model can be described in two different ways due to the freedom in choice of reference strand. Using the same notation as in section 2.4, we deduce that the basepair reference points and frames from these two descriptions are related for all $a = 1,\ldots,n$ as

$$\begin{aligned} \boldsymbol{q}_*^{n-a+1} &= \boldsymbol{q}^a, \\ \boldsymbol{g}_{*1}^{n-a+1} &= \boldsymbol{g}_1^a, \\ \boldsymbol{g}_{*2}^{n-a+1} &= -\boldsymbol{g}_2^a, \\ \boldsymbol{g}_{*3}^{n-a+1} &= -\boldsymbol{g}_3^a. \end{aligned} \tag{74}$$

Moreover, the internal coordinates from these two descriptions must be related for all $a = 1,\ldots,n-1$ as

$$z_*^{n-a} = Pz^a. \tag{75}$$

As before, the transformation rule relating the external coordinates z_*^0 and z^0 is more complicated and is omitted.

The mass parameters and velocity variables can also be described in two different ways due to the freedom in choice of reference strand. For these we deduce for all $a = 1,\ldots,n$ that

$$\begin{aligned} &m_{\mathrm{bp}*}^{n-a+1} = m_{\mathrm{bp}}^a,\ \boldsymbol{c}_{\mathrm{bp}*}^{n-a+1} = \boldsymbol{c}_{\mathrm{bp}}^a,\ \boldsymbol{\Gamma}_{\mathrm{bp}*}^{n-a+1} = \boldsymbol{\Gamma}_{\mathrm{bp}}^a, \\ &\boldsymbol{v}_{\mathrm{bp}*}^{n-a+1} = \boldsymbol{v}_{\mathrm{bp}}^a, \quad \boldsymbol{\omega}_{\mathrm{bp}*}^{n-a+1} = \boldsymbol{\omega}_{\mathrm{bp}}^a. \end{aligned} \tag{76}$$

Let $c_{\mathrm{bp}}^a \in \mathbb{R}^3$, $c_{\mathrm{bp}*}^{a_*} \in \mathbb{R}^3$, $\Gamma_{\mathrm{bp}}^a \in \mathbb{R}^{3\times 3}$ and so on denote the components of the above vector and tensor quantities in the respective basepair frames, so that $(c_{\mathrm{bp}})_i^a = \boldsymbol{g}_i^a \cdot \boldsymbol{c}_{\mathrm{bp}}^a$, $(c_{\mathrm{bp}*})_i^{a_*} = \boldsymbol{g}_{*i}^{a_*} \cdot \boldsymbol{c}_{\mathrm{bp}*}^{a_*}$, $(\Gamma_{\mathrm{bp}})_{ij}^a = \boldsymbol{g}_i^a \cdot \boldsymbol{\Gamma}_{\mathrm{bp}}^a \boldsymbol{g}_j^a$ and so on. Then from (76) and the transformation rules for the base frames in (74) we deduce the component relations

$$\begin{aligned} &m_{\mathrm{bp}*}^{n-a+1} = m_{\mathrm{bp}}^a,\ c_{\mathrm{bp}*}^{n-a+1} = \Psi c_{\mathrm{bp}}^a,\ \Gamma_{\mathrm{bp}*}^{n-a+1} = \Psi \Gamma_{\mathrm{bp}}^a \Psi, \\ &v_{\mathrm{bp}*}^{n-a+1} = \Psi v_{\mathrm{bp}}^a,\ \omega_{\mathrm{bp}*}^{n-a+1} = \Psi \omega_{\mathrm{bp}}^a. \end{aligned} \tag{77}$$

As in section 2.4, it is convenient to introduce the velocity component vectors $\nu_{\mathrm{bp}}^a = (v_{\mathrm{bp}}^a, \omega_{\mathrm{bp}}^a) \in \mathbb{R}^6$ ($a = 1,\ldots,n$). From (77) and the definition of Ψ we deduce that

$$\nu_{\mathrm{bp}*}^{n-a+1} = -P\nu_{\mathrm{bp}}^a. \tag{78}$$

3.5 Internal elastic energy

For a molecule of n basepairs we consider an internal elastic energy function U_{bp} of the general quadratic form

$$U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}}) = \tfrac{1}{2}(\mathrm{w}_{\mathrm{bp}} - \hat{\mathrm{w}}_{\mathrm{bp}}) \cdot \mathrm{K}_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}} - \hat{\mathrm{w}}_{\mathrm{bp}}), \tag{79}$$

where $\mathrm{w}_{\mathrm{bp}} = (z^1,\ldots,z^{n-1}) \in \mathbb{R}^{6n-6}$ is a vector of internal coordinates, $\mathrm{K}_{\mathrm{bp}} \in \mathbb{R}^{(6n-6)\times(6n-6)}$ is a symmetric, positive-definite matrix of stiffness parameters and $\hat{\mathrm{w}}_{\mathrm{bp}} \in \mathbb{R}^{6n-6}$ is a vector of shape parameters that represents the equilibrium value of w_{bp}. The expression in (79) can be written in the equivalent form

$$U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}}) = \frac{1}{2}\sum_{\alpha,\beta=1}^{n-1} (\mathrm{w}_{\mathrm{bp}}^\alpha - \hat{\mathrm{w}}_{\mathrm{bp}}^\alpha) \cdot \mathrm{K}_{\mathrm{bp}}^{\alpha\beta}(\mathrm{w}_{\mathrm{bp}}^\beta - \hat{\mathrm{w}}_{\mathrm{bp}}^\beta), \tag{80}$$

where $\mathrm{w}_{\mathrm{bp}}^\alpha \in \mathbb{R}^6$, $\hat{\mathrm{w}}_{\mathrm{bp}}^\alpha \in \mathbb{R}^6$ and $\mathrm{K}_{\mathrm{bp}}^{\alpha\beta} \in \mathbb{R}^{6\times 6}$ ($\alpha,\beta = 1,\ldots,n-1$) denote the block entries of w_{bp}, $\hat{\mathrm{w}}_{\mathrm{bp}}$ and K_{bp}.

As before, we assume that the material parameters $\hat{\mathrm{w}}_{\mathrm{bp}}^\alpha$ and $\mathrm{K}_{\mathrm{bp}}^{\alpha\beta}$ are completely determined by the sequence $X_1\cdots X_n$ along the reference strand. Equivalently, we assume there exist functions $\mathbb{W}_{\mathrm{bp}}$ and $\mathbb{K}_{\mathrm{bp}}$ such that

$$\begin{aligned} \hat{\mathrm{w}}_{\mathrm{bp}}^\alpha &= \mathbb{W}_{\mathrm{bp}}(X_1,\ldots,X_n,\alpha), \\ \mathrm{K}_{\mathrm{bp}}^{\alpha\beta} &= \mathbb{K}_{\mathrm{bp}}(X_1,\ldots,X_n,\alpha,\beta), \end{aligned} \qquad \alpha,\beta = 1,\ldots,n-1. \tag{81}$$

From the condition that the internal energy be invariant to the choice of reference strand, together with the change of strand relations outlined in section 3.4, we deduce that the functions $\mathbb{W}_{\mathrm{bp}}$ and $\mathbb{K}_{\mathrm{bp}}$ must satisfy the relations

$$\begin{aligned} \mathbb{W}_{\mathrm{bp}}(X_1,\ldots,X_n,\alpha) &= P\mathbb{W}_{\mathrm{bp}}(\bar{X}_n,\ldots,\bar{X}_1,n-\alpha), \\ \mathbb{K}_{\mathrm{bp}}(X_1,\ldots,X_n,\alpha,\beta) &= P\mathbb{K}_{\mathrm{bp}}(\bar{X}_n,\ldots,\bar{X}_1,n-\alpha,n-\beta)P. \end{aligned} \tag{82}$$

The general quadratic internal energy in (80) allows couplings between basepairs along the entire length of a molecule. As before, this assumption can be relaxed to only allow couplings between neighboring basepairs. By a local or nearest-neighbor internal energy for a rigid basepair model we mean an internal energy of the form

$$U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}}) = \sum_{a=1}^{n-1} U_{\mathrm{bp}}^a(z^a). \tag{83}$$

Here each U_{bp}^a can be interpreted as a local energy associated with the junction between the basepairs $(X, \bar{X})_a$ and $(X, \bar{X})_{a+1}$. Notice that the relative rotation and displacement between these basepairs is determined entirely by the internal coordinate z^a. In view of the definition of w_{bp}, the quadratic energy in (80) is of the local form (83) when K_{bp} is block diagonal, that is

$$\mathrm{K}_{\mathrm{bp}}^{\alpha\beta} = 0 \text{ when } \beta \neq \alpha. \tag{84}$$

3.6 Kinetic energy

To each basepair $(X, \bar{X})_a$ we associate a kinetic energy Φ_{bp}^a defined by its linear and angular velocity components as

$$\Phi_{\mathrm{bp}}^a = \tfrac{1}{2} m_{\mathrm{bp}}^a |v_{\mathrm{bp}}^a + \omega_{\mathrm{bp}}^a \times c_{\mathrm{bp}}^a|^2 + \tfrac{1}{2}\omega_{\mathrm{bp}}^a \cdot \Gamma_{\mathrm{bp}}^a \omega_{\mathrm{bp}}^a. \tag{85}$$

By expanding the first term, we find that this energy can be written in the convenient form

$$\Phi_{\mathrm{bp}}^a = \tfrac{1}{2}\nu_{\mathrm{bp}}^a \cdot M_{\mathrm{bp}}^a \nu_{\mathrm{bp}}^a, \tag{86}$$

where $\nu_{\mathrm{bp}}^a = (v_{\mathrm{bp}}^a, \omega_{\mathrm{bp}}^a) \in \mathbb{R}^6$ and $M_{\mathrm{bp}}^a \in \mathbb{R}^{6\times 6}$ with inverse $(M_{\mathrm{bp}}^a)^{-1} \in \mathbb{R}^{6\times 6}$ is a generalized mass matrix given in block form by

$$M_{\mathrm{bp}}^a = \begin{pmatrix} m_{\mathrm{bp}}^a I & m_{\mathrm{bp}}^a [c_{\mathrm{bp}}^a \times]^T \\ m_{\mathrm{bp}}^a [c_{\mathrm{bp}}^a \times] & \Gamma_{\mathrm{bp}}^a + m_{\mathrm{bp}}^a [c_{\mathrm{bp}}^a \times][c_{\mathrm{bp}}^a \times]^T \end{pmatrix},$$

$$(M_{\mathrm{bp}}^a)^{-1} = \begin{pmatrix} (m_{\mathrm{bp}}^a)^{-1} I + [c_{\mathrm{bp}}^a \times](\Gamma_{\mathrm{bp}}^a)^{-1}[c_{\mathrm{bp}}^a \times]^T & [c_{\mathrm{bp}}^a \times](\Gamma_{\mathrm{bp}}^a)^{-1} \\ (\Gamma_{\mathrm{bp}}^a)^{-1}[c_{\mathrm{bp}}^a \times]^T & (\Gamma_{\mathrm{bp}}^a)^{-1} \end{pmatrix}. \tag{87}$$

By summing over each basepair, we find that the total kinetic energy of a molecule with n basepairs can be written in the form

$$\Phi_{\mathrm{bp}}(\mathrm{v}_{\mathrm{bp}}) = \tfrac{1}{2}\mathrm{v}_{\mathrm{bp}}\cdot\mathrm{M}_{\mathrm{bp}}\mathrm{v}_{\mathrm{bp}}, \tag{88}$$

where $\mathrm{v}_{\mathrm{bp}} = (\nu^1_{\mathrm{bp}},\ldots,\nu^n_{\mathrm{bp}}) \in \mathbb{R}^{6n}$ is a vector of velocity components and $\mathrm{M}_{\mathrm{bp}} = \mathrm{diag}[M^1_{\mathrm{bp}},\ldots,M^n_{\mathrm{bp}}] \in \mathbb{R}^{6n\times 6n}$ is a block diagonal matrix of mass parameters.

We again assume that the mass parameters $\mathrm{M}^{\alpha}_{\mathrm{bp}}$ are completely determined by the sequence $X_1\cdots X_n$ along the reference strand. Equivalently, we assume there exists a function $\mathbb{M}_{\mathrm{bp}}$ such that

$$\mathrm{M}^{\alpha}_{\mathrm{bp}} = \mathbb{M}_{\mathrm{bp}}(X_1,\ldots,X_n,\alpha),\ \alpha = 1,\ldots,n. \tag{89}$$

From the condition that the kinetic energy be invariant to the choice of reference strand, together with the change of strand relations outlined in section 3.4, we deduce that the function $\mathbb{M}_{\mathrm{bp}}$ must satisfy the relation

$$\mathbb{M}_{\mathrm{bp}}(X_1,\ldots,X_n,\alpha) = P\mathbb{M}_{\mathrm{bp}}(\bar{X}_n,\ldots,\bar{X}_1,n-\alpha+1)P. \tag{90}$$

3.7 Canonical measure

As in the rigid base model, a useful form for the canonical measure $\mathrm{d}\mu_{\mathrm{bp}}$ for the rigid basepair model can be obtained by employing the non-canonical variables $(\zeta,\theta,v_{\mathrm{bp}},\omega_{\mathrm{bp}})$, where $(v_{\mathrm{bp}},\omega_{\mathrm{bp}}) \in \mathbb{R}^{6n}$ denote the linear and angular elocity components introduced in section 3.3. In these variables, the Hamiltonian for the rigid basepair model takes the form

$$H_{\mathrm{bp}} = U_{\mathrm{bp}}(\zeta,\theta) + \Phi_{\mathrm{bp}}(v_{\mathrm{bp}},\omega_{\mathrm{bp}}), \tag{91}$$

and the measure $\mathrm{d}\mu_{\mathrm{bp}}$ becomes

$$\mathrm{d}\mu_{\mathrm{bp}} = (1/Z_{\mathrm{bp}})e^{-H_{\mathrm{bp}}/k_{\mathrm{B}}T}\mathrm{J}_{\mathrm{bp}}\,\mathrm{d}\zeta\,\mathrm{d}\theta\,\mathrm{d}v_{\mathrm{bp}}\,\mathrm{d}\omega_{\mathrm{bp}}, \tag{92}$$

where J_{bp} is the Jacobian associated with the change from canonical to non-canonical variables. An application of the chain rule similar to before yields

$$\mathrm{J}_{\mathrm{bp}} = \prod_{a=0}^{n-1}\left(1+\frac{1}{4}|\theta^a|^2\right)^{-2}. \tag{93}$$

In the form given in (92), the measure $\mathrm{d}\mu_{\mathrm{bp}}$ is factorable into three independent measures $\mathrm{d}\mu_{\mathrm{bp,vel}}$, $\mathrm{d}\mu^{\mathrm{int}}_{\mathrm{bp,con}}$ and $\mathrm{d}\mu^{\mathrm{ext}}_{\mathrm{bp,con}}$, where

$$\begin{aligned}
\mathrm{d}\mu_{\mathrm{bp,vel}} &= (1/Z_{\mathrm{bp,vel}})e^{-\Phi_{\mathrm{bp}}(\mathrm{v}_{\mathrm{bp}})/k_{\mathrm{B}}T}\,\mathrm{dv}_{\mathrm{bp}},\\
\mathrm{d}\mu^{\mathrm{int}}_{\mathrm{bp,con}} &= (1/Z^{\mathrm{int}}_{\mathrm{bp,con}})e^{-U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}})/k_{\mathrm{B}}T}\mathrm{J}'_{\mathrm{bp}}\,\mathrm{dw}_{\mathrm{bp}},\\
\mathrm{d}\mu^{\mathrm{ext}}_{\mathrm{bp,con}} &= (1/Z^{\mathrm{ext}}_{\mathrm{bp,con}})\mathrm{J}^0_{\mathrm{bp}}\,\mathrm{d}z^0.
\end{aligned} \tag{94}$$

Here v_{bp} is the vector of all velocity components, w_{bp} is the vector of all internal configuration coordinates, z^0 is the vector of external configuration coordinates, and $\mathrm{J}'_{\mathrm{bp}}$ and $\mathrm{J}^0_{\mathrm{bp}}$ are Jacobian factors given by

$$\mathrm{J}'_{\mathrm{bp}} = \prod_{a=1}^{n-1}\left(1+\frac{1}{4}|\theta^a|^2\right)^{-2},\ \mathrm{J}^0 = \left(1+\frac{1}{4}|\theta^0|^2\right)^{-2}. \tag{95}$$

The statistical mechanical average of any function $\phi = \phi(\mathrm{v}_{\mathrm{bp}},\mathrm{w}_{\mathrm{bp}},z^0)$ with respect to the measure $\mathrm{d}\mu_{\mathrm{bp}}$ is given by

$$\langle\phi\rangle_{\mathrm{bp}} = \frac{\int \phi(\mathrm{v}_{\mathrm{bp}},\mathrm{w}_{\mathrm{bp}},z^0)e^{-H_{\mathrm{bp}}(\mathrm{v}_{\mathrm{bp}},\mathrm{w}_{\mathrm{bp}})/k_{\mathrm{B}}T}\mathrm{J}_{\mathrm{bp}}\,\mathrm{dv}_{\mathrm{bp}}\,\mathrm{dw}_{\mathrm{bp}}\,\mathrm{d}z^0}{\int e^{-H_{\mathrm{bp}}(\mathrm{v}_{\mathrm{bp}},\mathrm{w}_{\mathrm{bp}})/k_{\mathrm{B}}T}\mathrm{J}_{\mathrm{bp}}\,\mathrm{dv}_{\mathrm{bp}}\,\mathrm{dw}_{\mathrm{bp}}\,\mathrm{d}z^0}, \tag{96}$$

where the integrations are performed over the domain $\mathbb{R}^{6n}\times\mathbb{R}^{6n-6}\times D^0$. Here R^{6n} is the domain for v_{bp}, $\mathbb{R}^{6n-6}$ is the domain for w_{bp}, and $D^0 \subset \mathbb{R}^6$ is a prescribed domain for z^0, which will play no role in our developments. Due to the factorability of $\mathrm{d}\mu_{\mathrm{bp}}$, for any function $\psi = \psi(\mathrm{v}_{\mathrm{bp}})$ we find

$$\langle\psi\rangle_{\mathrm{bp}} = \frac{\int \psi(\mathrm{v}_{\mathrm{bp}})e^{-\Phi_{\mathrm{bp}}(\mathrm{v}_{\mathrm{bp}})/k_{\mathrm{B}}T}\,\mathrm{dv}_{\mathrm{bp}}}{\int e^{-\Phi_{\mathrm{bp}}(\mathrm{v}_{\mathrm{bp}})/k_{\mathrm{B}}T}\,\mathrm{dv}_{\mathrm{bp}}}, \tag{97}$$

and for any function $\chi = \chi(\mathrm{w}_{\mathrm{bp}})$ we find

$$\frac{\langle\chi/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}}{\langle 1/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}} = \frac{\int \chi(\mathrm{w}_{\mathrm{bp}})e^{-U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}})/k_{\mathrm{B}}T}\,\mathrm{dw}_{\mathrm{bp}}}{\int e^{-U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}})/k_{\mathrm{B}}T}\,\mathrm{dw}_{\mathrm{bp}}}. \tag{98}$$

3.8 Moment–parameter relations

As before, the relations in (97) and (98) can be exploited to derive explicit characterizations of the rigid basepair model parameters $\hat{\mathrm{w}}_{\mathrm{bp}}$, K_{bp} and M_{bp}. Proceeding as in section 2.8, and using the notation $\Delta\mathrm{w}_{\mathrm{bp}} = \mathrm{w}_{\mathrm{bp}} - \hat{\mathrm{w}}_{\mathrm{bp}}$, we find

$$\frac{\langle\mathrm{w}_{\mathrm{bp}}/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}}{\langle 1/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}} = \hat{\mathrm{w}}_{\mathrm{bp}}, \tag{99}$$

and

$$\frac{\langle\Delta\mathrm{w}_{\mathrm{bp}}\otimes\Delta\mathrm{w}_{\mathrm{bp}}/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}}{\langle 1/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}} = k_{\mathrm{B}}T\mathrm{K}^{-1}_{\mathrm{bp}}, \tag{100}$$

or equivalently

$$\frac{\langle\mathrm{w}_{\mathrm{bp}}\otimes\mathrm{w}_{\mathrm{bp}}/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}}{\langle 1/\mathrm{J}'_{\mathrm{bp}}\rangle_{\mathrm{bp}}} = k_{\mathrm{B}}T\mathrm{K}^{-1}_{\mathrm{bp}} + \hat{\mathrm{w}}_{\mathrm{bp}}\otimes\hat{\mathrm{w}}_{\mathrm{bp}}, \tag{101}$$

and moreover

$$\langle\mathrm{v}_{\mathrm{bp}}\otimes\mathrm{v}_{\mathrm{bp}}\rangle_{\mathrm{bp}} = k_{\mathrm{B}}T\mathrm{M}^{-1}_{\mathrm{bp}}. \tag{102}$$

Just as for the rigid base model, the Jacobian factor J'_{bp} will be non-constant and consequently the measure $\mathrm{d}\mu^{\mathrm{int}}_{\mathrm{bp,con}}$ will be non-Gaussian for arbitrary choices of internal coordinates. However, when the gradient of J'_{bp} is sufficiently small and $\mathrm{d}\mu^{\mathrm{int}}_{\mathrm{bp,con}}$ is sufficiently concentrated, it is reasonable to expect that variations in J'_{bp} can be neglected. By the Gaussian approximation of the internal configuration measure $\mathrm{d}\mu^{\mathrm{int}}_{\mathrm{bp,con}}$ we mean the measure obtained by assuming J'_{bp} to be constant. In this approximation, the relations in (99), (100) and (101) take the simplified forms

$$\begin{aligned}
&\langle\mathrm{w}_{\mathrm{bp}}\rangle_{\mathrm{bp}} = \hat{\mathrm{w}}_{\mathrm{bp}},\ \langle\Delta\mathrm{w}_{\mathrm{bp}}\otimes\Delta\mathrm{w}_{\mathrm{bp}}\rangle_{\mathrm{bp}} = k_{\mathrm{B}}T\mathrm{K}^{-1}_{\mathrm{bp}},\\
&\langle\mathrm{w}_{\mathrm{bp}}\otimes\mathrm{w}_{\mathrm{bp}}\rangle_{\mathrm{bp}} = k_{\mathrm{B}}T\mathrm{K}^{-1}_{\mathrm{bp}} + \hat{\mathrm{w}}_{\mathrm{bp}}\otimes\hat{\mathrm{w}}_{\mathrm{bp}}.
\end{aligned} \tag{103}$$

3.9 Compatibility relations

Here we derive various compatibility relations that relate the rigid basepair parameters $\hat{\mathrm{w}}_{\mathrm{bp}}$, K_{bp} and M_{bp} to the rigid base parameters $\hat{\mathrm{w}}$, K and M. The results derived here depend only

on the quadratic nature of the internal and kinetic energies and on the assumed domain for the internal coordinates and velocities. In particular, they are independent of the base sequence $X_1 \cdots X_n$. For brevity, we focus attention on the shape and stiffness parameters and only briefly discuss the mass parameters.

From the definition of the internal coordinates, we observe that w_{bp} is related to w. Thus, in view of the moment–parameter relations in sections 2.8 and 3.8, it follows that $\hat{\mathrm{w}}_{\mathrm{bp}}$ and K_{bp} are related to $\hat{\mathrm{w}}$ and K. To make this relation explicit, it will be convenient to re-order the internal coordinate vector w so that $\mathrm{w} = (\mathrm{z},\mathrm{y})$, where $\mathrm{z} = (z^1,\ldots,z^{n-1}) \in \mathbb{R}^{6n-6}$ and $\mathrm{y} = (y^1,\ldots,y^n) \in \mathbb{R}^{6n}$, and the stiffness matrix K so that

$$\mathrm{K} = \begin{pmatrix} \mathrm{K}_{zz} & \mathrm{K}_{zy} \\ \mathrm{K}_{zy}^T & \mathrm{K}_{yy} \end{pmatrix}, \tag{104}$$

where $\mathrm{K}_{zz} \in \mathbb{R}^{(6n-6)\times(6n-6)}$, $\mathrm{K}_{yy} \in \mathbb{R}^{6n\times 6n}$ and $\mathrm{K}_{zy} \in \mathbb{R}^{(6n-6)\times 6n}$. Notice that, because the overall stiffness matrix K is assumed to be symmetric and positive-definite, so too must be the diagonal blocks K_{zz} and K_{yy}.

Several useful relations follow from the re-ordering introduced above. From the definitions of w_{bp} and w, we get $\mathrm{w}_{\mathrm{bp}} = \mathscr{P}\mathrm{w} = \mathrm{z}$, where $\mathscr{P} \in \mathbb{R}^{(6n-6)\times(12n-6)}$ is a simple projection matrix. Moreover, from the block-structure in (104), we obtain the identity

$$\tfrac{1}{2}(\mathrm{w}-\hat{\mathrm{w}})\cdot\mathrm{K}(\mathrm{w}-\hat{\mathrm{w}}) = \tfrac{1}{2}(\mathrm{z}-\hat{\mathrm{z}})\cdot\mathrm{K}_{zz}^{\mathrm{schur}}(\mathrm{z}-\hat{\mathrm{z}}) + \tfrac{1}{2}\mathrm{x}\cdot\mathrm{K}_{yy}\mathrm{x}, \tag{105}$$

where $\mathrm{K}_{zz}^{\mathrm{schur}} = \mathrm{K}_{zz} - \mathrm{K}_{zy}\mathrm{K}_{yy}^{-1}\mathrm{K}_{zy}^T$, $\mathrm{x} = \mathrm{K}_{yy}^{-1}\mathrm{K}_{zy}^T(\mathrm{z}-\hat{\mathrm{z}}) + \mathrm{y} - \hat{\mathrm{y}}$ and $(\hat{\mathrm{z}},\hat{\mathrm{y}}) = \hat{\mathrm{w}}$. The matrix $\mathrm{K}_{zz}^{\mathrm{schur}}$ is known as the Schur complement of K_{yy};[25] it has the property that $\mathrm{K}_{zz}^{\mathrm{schur}} = (\mathscr{P}\mathrm{K}^{-1}\mathscr{P}^T)^{-1}$. Also, for any fixed z, notice that x and y differ only by a fixed translation. Thus, if the domain for y is all of $\mathbb{R}^{6n}$, then so too is the domain for x. Furthermore, in view of (105) and (32), we have

$$U(\mathrm{w}) = U_1(\mathrm{z}) + U_2(\mathrm{x}), \tag{106}$$

where $U_1(\mathrm{z})$ and $U_2(\mathrm{x})$ correspond to the two terms on the right-hand side of (105).

Compatibility relations for the shape and stiffness parameters can now be derived. For an arbitrary function $\phi(z)$, compatibility between the rigid basepair and base models requires

$$\langle\phi\rangle_{\mathrm{bp}} = \langle\phi\rangle. \tag{107}$$

From the definitions of the averages in (96) and (55), we get, after cancelling factors independent of w_{bp} and w,

$$\frac{\int \phi(\mathrm{z}) e^{-U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}})/k_{\mathrm{B}}T}\mathbf{J}'_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}})\,\mathrm{dw}_{\mathrm{bp}}}{\int e^{-U_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}})/k_{\mathrm{B}}T}\mathbf{J}'_{\mathrm{bp}}(\mathrm{w}_{\mathrm{bp}})\,\mathrm{dw}_{\mathrm{bp}}} = \frac{\int \phi(\mathrm{z}) e^{-U(\mathrm{w})/k_{\mathrm{B}}T}\mathbf{J}'(\mathrm{w})\,\mathrm{dw}}{\int e^{-U(\mathrm{w})/k_{\mathrm{B}}T}\mathbf{J}'(\mathrm{w})\,\mathrm{dw}}. \tag{108}$$

We next assume that the two Jacobian factors are sufficiently weak functions of the coordinates so that the Gaussian approximations defined in sections 2.8 and 3.8 hold, or equivalently, so that each Jacobian may be treated as constant in each integral. Substituting for w_{bp} and w in terms of z and y, then using (106) and changing variable from y to x, and then cancelling common integrals over x, we get

$$\frac{\int \phi(\mathrm{z}) e^{-U_{\mathrm{bp}}(\mathrm{z})/k_{\mathrm{B}}T}\,\mathrm{dz}}{\int e^{-U_{\mathrm{bp}}(\mathrm{z})/k_{\mathrm{B}}T}\,\mathrm{dz}} = \frac{\int \phi(\mathrm{z}) e^{-U_1(\mathrm{z})/k_{\mathrm{B}}T}\,\mathrm{dz}}{\int e^{-U_1(\mathrm{z})/k_{\mathrm{B}}T}\,\mathrm{dz}}. \tag{109}$$

The fact that the above relation must hold for all functions ϕ implies that U_{bp} and U_1 must be equal up to a constant, which without loss of generality may be taken as zero since the energies themselves are defined only up to a constant. Thus, under the Gaussian approximation, compatibility between the rigid basepair and base models implies

$$\hat{\mathrm{w}}_{\mathrm{bp}} = \mathscr{P}\hat{\mathrm{w}},\ \mathrm{K}_{\mathrm{bp}} = (\mathscr{P}\mathrm{K}^{-1}\mathscr{P}^T)^{-1}. \tag{110}$$

Similar calculations can be used to derive a relation between the parameters M_{bp} and M. To begin, notice that (z,z^0) and $(\mathrm{y},\mathrm{z},z^0)$ are complete sets of configuration coordinates for the rigid basepair and base models. From this we find that the velocity components v_{bp} and v can be expressed as invertible linear functions of $(\dot{\mathrm{z}},\dot{z}^0)$ and $(\dot{\mathrm{y}},\dot{\mathrm{z}},\dot{z}^0)$, respectively. If we let $\mathrm{v}_{\mathrm{int}}$ be a representation of $\dot{\mathrm{y}}$ in any convenient basis, then from the above remarks we deduce the linear relation $\mathrm{v} = \mathscr{A}(\mathrm{v}_{\mathrm{bp}},\mathrm{v}_{\mathrm{int}})$, where the coefficient matrix $\mathscr{A} \in \mathbb{R}^{12n\times 12n}$ in general depends on $(\mathrm{y},\mathrm{z},z^0)$. Inverting this relation, we find $\mathrm{v}_{\mathrm{bp}} = \mathscr{B}\mathrm{v}$, where $\mathscr{B} \in \mathbb{R}^{6n\times 12n}$ is an appropriate matrix, which also depends on $(\mathrm{y},\mathrm{z},z^0)$. The expressions $\mathrm{v} = \mathscr{A}(\mathrm{v}_{\mathrm{bp}},\mathrm{v}_{\mathrm{int}})$ and $\mathrm{v}_{\mathrm{bp}} = \mathscr{B}\mathrm{v}$ are analogous to the expressions $\mathrm{w} = (\mathrm{w}_{\mathrm{bp}},\mathrm{y})$ and $\mathrm{w}_{\mathrm{bp}} = \mathscr{P}\mathrm{w}$ employed above. Thus, proceeding as before it is possible to derive a relation between M_{bp} and M, but in this case the relation is highly implicit due to the configuration dependence of the matrix $\mathscr{A}$. This relation is omitted since no use will be made of it.

4. Methods

Here we outline the molecular dynamics method that was used to estimate the rigid base and basepair parameters for the 16-basepair palindromic oligomer $\mathrm{G(TA)_7C}$ in explicit solvent. We discuss special procedures for simulating the B-form structure over relatively long time periods, and for extracting the configuration and velocity data and material parameters for both models.

4.1 Simulation protocol

The 16-basepair DNA oligomer was simulated using atomic resolution, explicit solvent molecular dynamics. The AMBER suite of programs together with the *parm94* force field[8] was used. The DNA duplex was built using the fiber diffraction B-DNA coordinates as implemented in the *nucgen* module of AMBER, hydrogen atoms were added using the *leap* module. The structure then underwent *in vacuo* energy minimization in which heavy atoms were restrained to their initial positions. The resulting structure will be referred to as canonical B-DNA. It was used as the starting structure for the subsequent simulation, and for the calculation of mass parameters of DNA bases subsequently referred to as canonical mass parameters. These were used for comparison with the mass parameters obtained from the MD simulation.

The *leap* module was employed to add 30 K^+ neutralizing cations, and to hydrate the system with TIP3P water molecules in an octahedral periodic box, with a minimum distance of 10 Å between the wall of the box and the closest atom of DNA or an ion. This results in a system of *ca.* 28 000 atoms in total. The *ptraj* module was used to swap the positions of ions and those of randomly chosen water molecules in such a way that no ion came less that 3 Å away from another ion, nor less than 5 Å away from the DNA. The system was equilibrated by a series of energy minimizations and short MD runs with DNA atoms attached to their initial positions by restraints that were gradually released, followed by 1 ns of unrestrained MD. The production MD simulation was then started. It was performed in the *NpT* ensemble with temperature 300 K and pressure 1 atm. Temperature and pressure were regulated using a Berendsen thermostat and barostat with coupling constants of 5 ps, which was more than twice the estimated value of the relaxation time of the atomic velocities, and the integration time step was 2 fs with *SHAKE* applied to hydrogens only, which should have a minimal impact on the dynamics of interest. The particle mesh Ewald method was used to treat long-range electrostatic interactions beyond a cutoff of 9 Å. Details of the simulation protocol can be found in ref. 2.

Two MD trajectories were produced, both started from the same set of initial conditions at the end of the equilibration run. One trajectory was 1 ns long, sampled every 2 fs (every time step) and used to calculate velocities and mass parameters. This trajectory will be referred to as the fine-sampled one. The second trajectory was 180 ns long, sampled every 1 ps and used to calculate elastic energy parameters. This trajectory will be referred to as the coarse-sampled one. Both trajectories were stripped of water and ions and divided into individual snapshots saved in *pdb* format, using the *ptraj* module. The coordinates of the 180 ns trajectory were centered and imaged into the primary box, and rms-fitted to the first snapshot. However, no centering, imaging or rms fitting was performed for the 1 ns trajectory.

The snapshots were analyzed using the program *Curves*,[23] modified to compute and report a reference point and an orthonormal frame attached to each base as defined by the Tsukuba convention.[30] The points and frames were reported by *Curves* as component vectors and rotation matrices in a fixed coordinate frame. From these we computed the corresponding component vectors and rotation matrices for each basepair, and then the full set of internal coordinates and velocities for both the base and basepair models. *Curves* was also used to compute backbone torsional angles which were used to monitor the simulation as described next.

It was recently discovered[2] that the *parm94* force field overstabilizes particular non-canonical ''flipped'' states of the DNA backbone. These states get increasingly populated as the simulation proceeds and their high occupancy is followed, in a linear DNA fragment, by a distortion of the helical geometry into a configuration in which helical twist is close to zero. In order to avoid such a behavior and keep the simulated DNA within the B-DNA family, we coupled our simulation to an information bias procedure (which may be called ''Maxwell daemon molecular dynamics'', or MDMD): whenever any backbone fragment flipped into a non-canonical state, the simulation was stopped and restarted from a time point *ca.* 100 ps before the flip. The flip then never happened again at the same time and location, and a flip-free trajectory is produced without restraining the system in any way. The restart had to be done every 5 ns on average. Based on previous experience,[2] we defined a flipped backbone fragment as one in which the torsion angle γ is in t instead of its canonical $g+$ state (around $180 \pm 30^\circ$ instead of $60 \pm 30^\circ$). In a later stage of preparation of this article, an updated force field called *parmbsc0* was published,[35] which largely eliminates the problem of the backbone flips.

We also observed that basepairs transiently broke and re-formed during our MD simulation. Since we were interested only in the properties of B-DNA, with bases on opposite strands connected by hydrogen bonds (H-bonds) to form Watson–Crick pairs, we eliminated from our analysis all snapshots with at least one H-bond broken anywhere in the oligomer. We consider an H-bond broken if the distance between donor and acceptor is greater than 4 Å. This criterion was suggested by Lu and Olson to identify basepairs in structures of nucleic acids.[27] We used the *ptraj* module of AMBER to measure the donor–acceptor distances.

4.2 Parameter estimation

Let N be the total number of snapshots in a flip-free trajectory, which we label consecutively by $k = 1,\ldots,N$. For each k, the *Curves* program was used to determine the rigid base configuration variables $[r^a]^{(k)}$, $[\bar{r}^a]^{(k)}$ and $[D^a]^{(k)}$, $[\bar{D}^a]^{(k)}$, and then the rigid basepair configuration variables $[q^a]^{(k)}$ and $[G^a]^{(k)}$. From these, we calculated the internal coordinates $\mathrm{w}^{(k)}$ and $\mathrm{w}_{\mathrm{bp}}^{(k)}$ and the velocity components $\mathrm{v}^{(k)}$ and $\mathrm{v}_{\mathrm{bp}}^{(k)}$. To estimate parameters, we assumed ergodicity and replaced the statistical mechanical averages appearing in sections 2.8 and 3.8 with averages over the snapshots, excluding those with one or more broken H-bond as described below. Throughout the remainder of our developments we use a subscript ''E'' to indicate the value of a parameter that was estimated in this way. We omit this subscript when referring to its exact or theoretical value.

The shape and stiffness parameters for the rigid base model were estimated by replacing the statistical mechanical averages with snapshot or time series averages. Considering only snapshots with no broken H-bond anywhere in the oligomer, we defined an estimate for the shape vector $\hat{\mathrm{w}}$ by

$$\hat{\mathrm{w}}_E = \frac{\sum_{k\in\mathscr{I}} \mathrm{w}^{(k)}/[\mathrm{J}']^{(k)}}{\sum_{k\in\mathscr{I}} 1/[\mathrm{J}']^{(k)}}, \qquad (111)$$

and an estimate for the stiffness matrix K by

$$k_{\mathrm{B}}T[\mathrm{K}_E]^{-1} = \frac{\sum_{k\in\mathscr{I}} \Delta_E\mathrm{w}^{(k)} \otimes \Delta_E\mathrm{w}^{(k)}/[\mathrm{J}']^{(k)}}{\sum_{k\in\mathscr{I}} 1/[\mathrm{J}']^{(k)}}, \qquad (112)$$

where $\Delta_E\mathrm{w} = \mathrm{w} - \hat{\mathrm{w}}_E$, J' is the Jacobian factor defined in $(54)_1$, T is the simulated temperature and $\mathscr{I}$ is the set of indices corresponding to snapshots with no broken H-bond. Parameters for the rigid basepair model were defined similarly.

In order to estimate the mass parameters for the rigid base model it was necessary to first compute the linear and angular velocity components (v^a,ω^a) and $(\bar{v}^a,\bar{\omega}^a)$. Since the available data consisted only of configuration variables at discrete times,

we employed a finite difference approximation for the velocities. Consistent with (18), the linear and angular velocities (v^a,ω^a) at snapshot k were defined as

$$[v^a]^{(k)} = ([D^a]^{(k)})^T \frac{[r^a]^{(k+1)} - [r^a]^{(k-1)}}{t^{(k+1)} - t^{(k-1)}},$$

$$[\omega^a]^{(k)} = \mathrm{vec}\left[\mathrm{skew}\left(([D^a]^{(k)})^T \frac{[D^a]^{(k+1)} - [D^a]^{(k-1)}}{t^{(k+1)} - t^{(k-1)}}\right)\right], \tag{113}$$

where $t^{(k)}$ is the time associated with snapshot k and for any matrix A we define $\mathrm{skew}(A) = (A - A^T)/2$. Whereas the matrix $(D^a)^T\dot{D}^a$ is always skew-symmetric, its finite-difference approximation is in general not. For this reason, we employed the skew-symmetric projection in $(113)_2$. For $k = 1$ and $k = N$ the above expressions were replaced with simple one-sided differences, and similar definitions consistent with (22) were used to compute the velocities $(\bar{v}^a,\bar{\omega}^a)$. From these components and those above, we formed the global velocity vector $\mathrm{v}^{(k)}$ at each snapshot k.

Just as with the shape and stiffness parameters, the mass parameters for the rigid base model were estimated by replacing the statistical mechanical average with a time series average. We included snapshot k in the average only if each of the snapshots $k - 1$, k and $k + 1$ had no broken H-bond anywhere in the oligomer, which ensured that the velocity vector $\mathrm{v}^{(k)}$ obtained from the finite-difference approximation (113) was representative of DNA with only Watson–Crick pairs. The snapshots with $k = 1$ and $k = N$ were checked analogously, consistent with the one-sided differences employed. Thus we defined an estimate for the global mass matrix M by

$$k_\mathrm{B}T[\mathrm{M}_E]^{-1} = \frac{1}{I'}\sum_{k\in\mathscr{I}'} \mathrm{v}^{(k)} \otimes \mathrm{v}^{(k)}, \tag{114}$$

where $\mathscr{I}'$ is the set of admissible snapshots as described above and I' is the number of snapshots in $\mathscr{I}'$.

Similar to the exact mass matrix M, the estimate M_E is in general symmetric and positive-definite. However, in contrast to M, the estimate M_E is in general not block-diagonal. To estimate the mass parameters of the individual bases, we considered only the diagonal blocks of M_E and denoted them by M_E^a and $\bar{M}_E^a$ in direct analogy with the diagonal blocks M^a and $\bar{M}^a$ of M. Using symmetry, we partitioned each matrix M_E^a as

$$M_E^a = \begin{pmatrix} B_1^a & [B_2^a]^T \\ B_2^a & B_3^a \end{pmatrix}, \tag{115}$$

where B_1^a, B_2^a and B_3^a are 3×3 matrices. Each matrix $\bar{M}_E^a$ was partitioned similarly. Consistent with the form of the exact mass matrix M^a in (41), we defined the mass parameter estimates m_E^a, c_E^a and Γ_E^a by

$$m_E^a = \frac{1}{3}\mathrm{tr}[B_1^a],\ c_E^a = \frac{1}{m_E^a}\mathrm{vec}(\mathrm{skew}(B_2^a)),$$
$$\Gamma_E^a = B_3^a - m_E^a[c_E^a\times][c_E^a\times]^T. \tag{116}$$

The estimates $\bar{m}_E^a$, $\bar{c}_E^a$ and $\bar{\Gamma}_E^a$ were defined similarly, and the parameters for the rigid basepair model were defined in an exactly analogous way.

5. Results

Here we describe the parameter estimation results obtained with the fine- and coarse-sampled molecular dynamics trajectories described in section 4.1. We begin with results on spontaneous H-bond breaking and its effects on internal coordinates, and then outline various results concerning the mass, shape and stiffness parameters. We use the estimated parameters to assess various modeling assumptions pertaining to the rigidity of the bases and basepairs, and the property of locality of the internal energy. When convenient, we present results in dimensionless or reduced form using the energy scale $k_\mathrm{B}T$, length scale 5.2 Å and time scale 1 ps. These scales are motivated by the parameters for canonical B-form DNA (the length scale is equal to the ratio of rise to twist), and are chosen purely for convenience.

In our analysis, we performed tests to assess the convergence of the statistical averages outlined in section 4.2. Specifically, we chose several non-overlapping time windows within each trajectory, computed the averages for each window, and compared them to those computed for the whole trajectory. Our results showed that each of the two trajectories was sufficiently long to estimate the relevant averages well. We also studied the influence of the Jacobian factor in these statistical averages and found that, for our choice of internal coordinates, the influence was rather small. Various averages computed with and without the Jacobian differed by less than 3%. We nevertheless included the Jacobian for completeness. The details of these results are omitted for brevity.

5.1 Base pairing

In our simulations, we observed that no basepair stays intact throughout the entire course of a trajectory. Rather, at least one H-bond in each basepair is temporarily broken according to our criteria described in section 4. Here we outline various results pertaining to the breaking of H-bonds and its effect on internal coordinates. For brevity, we present results only for the coarse-sampled trajectory. Results for the fine-sampled trajectory were similar.

Fig. 1 shows the fraction of the total snapshots for which a given H-bond was broken. We see that the fraction of broken states for each basepair was less than 1%, and was higher towards the oligomer ends than in the interior. Moreover, the H-bonds on the major groove side were more prone to breaking than those on the minor groove side. Indeed, the fraction of broken states for H-bonds on the minor groove side for each of basepairs 4 to 15 was less than 0.005%, which corresponds to less than 10 snapshots out of the entire 180 000 snapshot trajectory. The fraction of snapshots which had no broken H-bond anywhere in the oligomer was observed to be 93.4%, which corresponds to more than 168 000 snapshots.

Fig. 2 shows the maximum lifetime of a broken state for the individual H-bonds. We see that lifetimes were in general longer at the ends of the oligomer than in the interior, and longer on the major groove side than on the minor. The lifetimes at the ends were mostly well over 50 ps, whereas those in the interior were mostly under 10 ps. For basepairs 4 to 15, the lifetimes on the minor groove side hardly exceeded 5 ps. A notable exception occurred at basepair 2. There the

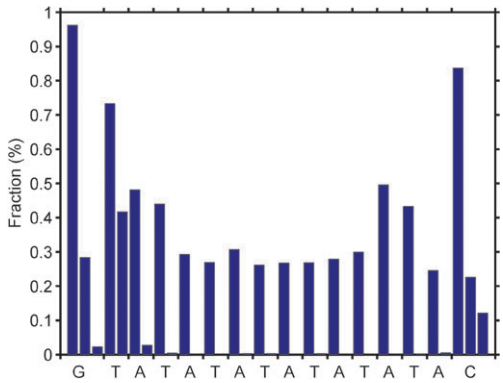

Fig. 1 Fraction of snapshots in the coarse-sampled trajectory in which individual H-bonds were broken. The H-bonds are shown for consecutive basepairs as defined by the base sequence of the reference strand. The three bars for each of the terminal (G,C) and (C,G) basepairs correspond to the major, middle and minor groove H-bonds (left-to-right). The two bars for each of the interior (A,T) or (T,A) basepairs correspond to the major and minor groove H-bonds (left-to-right). The fractions for H-bonds on the minor groove side for basepairs 4 to 15 are small ($<0.005\%$) and barely visible on this scale.

lifetime was 753 ps on the major groove side and 743 ps on the minor groove side. Using this data, we investigated possible cooperativity in H-bond breaking. For each basepair, we first identified the broken H-bond with the longest lifetime, which was usually the major groove H-bond, and recorded the time interval in which the break took place. We then considered the donor–acceptor distance for all the H-bonds in the adjacent basepairs over this interval. From this procedure we found no simultaneous breaking of H-bonds in adjacent basepairs, with only two exceptions: basepairs 15 and 16 had a broken H-bond which overlapped for about 6 ps, and basepairs 6 and 7 had a broken H-bond which overlapped for about 2 ps. Thus H-bond breaking in our simulation appeared to be largely uncooperative.

Fig. 3 shows the distribution of the internal coordinate Shear for basepair 2. The distribution is shown for raw data corresponding to all snapshots in the trajectory, and filtered data corresponding to snapshots containing no broken

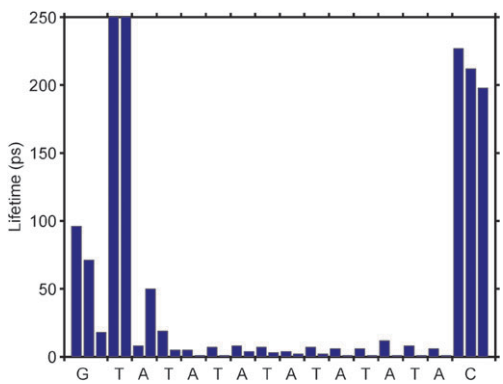

Fig. 2 Maximum lifetimes of broken H-bonds in the coarse-sampled trajectory. The ordering of the H-bonds follows the same convention as in Fig. 1. The lifetimes for both the major groove (753 ps) and minor groove (743 ps) H-bonds for basepair 2 fall outside the range of the Figure. The lifetimes for the broken minor groove H-bonds for basepairs 4 to 15 are mostly small (<5 ps) and barely visible on this scale.

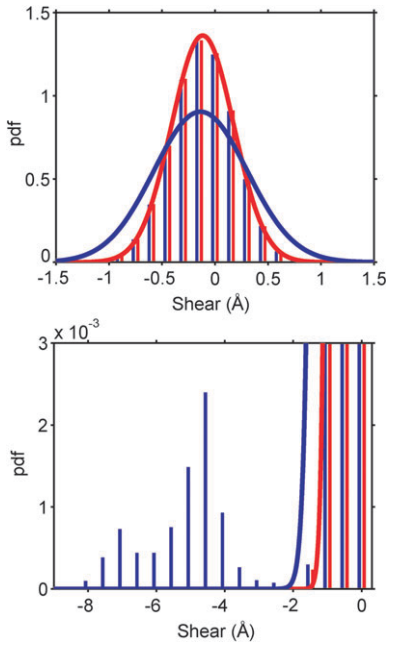

Fig. 3 Normalized distributions (pdf's) of the internal coordinate Shear for basepair 2 for the coarse-sampled trajectory. The distribution is shown for two sets of snapshot data: raw data (blue), which includes all the snapshots from the trajectory, and filtered data (red), which includes only those snapshots containing no broken H-bond anywhere in the oligomer. The histograms denote the actual data, whereas the solid curves denote Gaussian fits with the same mean and variance as the data. Top panel: illustration of mean and variance of raw and filtered data (histogram bin size 0.15 Å). Bottom panel: illustration of tail in raw data (histogram bin size 0.5 Å).

H-bonds anywhere in the oligomer. Both sets of data exhibit a bell-shaped profile with similar means: -0.14 Å raw, -0.11 Å filtered. However, the standard deviations differ by roughly 50%: 0.44 Å raw, 0.29 Å filtered. This difference leads to a dramatic difference in the Gaussian fits as can be seen in the top panel of the figure, and can be attributed to an extended tail in the raw data as shown in the bottom panel. By comparison, no such tail appears in the filtered data. Similar phenomenon at varying degrees of intensity occurred in other internal coordinates and at other basepair locations. Throughout the remainder of our developments we shall consider only the filtered data since it may provide a better representation of the properties of B-form DNA than the raw data.

5.2 Mass parameters

Here we outline results on the estimation of mass parameters for the rigid base and basepair models from the fine-sampled trajectory described in section 4. We illustrate various consistency checks based on the theory in sections 2 and 3. These checks provide a means of assessing the quality of the numerical simulations and the rigidity assumptions in the base and basepair models.

Fig. 4 shows a portion of the estimated mass matrix $\mathbf{M}_E$ for the rigid base model. The 6×6 diagonal blocks in Fig. 4 correspond to the mass matrices M_E^a and $\bar{M}_E^a$ for the eight individual bases in basepairs $a = 7,\ldots,10$. To assess the

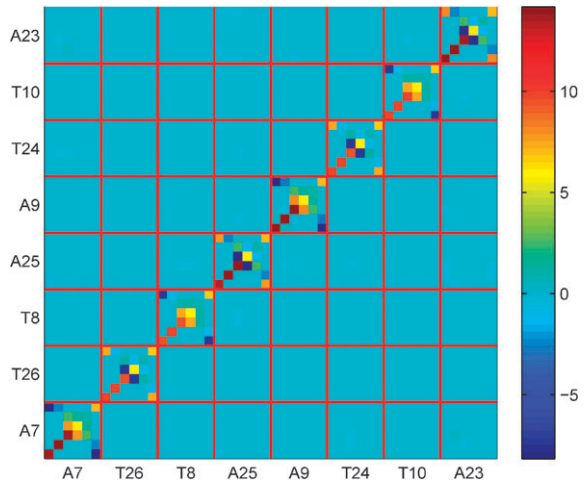

Fig. 4 A portion of the estimated mass matrix M_E in reduced form for the rigid base model computed with the fine-sampled trajectory. The portion shown corresponds to the 6 × 6 block entries $\mathrm{M}_E^{\alpha,\beta}$ for $\alpha,\beta = 7,\ldots,14$, which are marked by the grid lines. The diagonal blocks correspond to the mass matrices M_E^a and $\bar{M}_E^a$ for the eight individual bases in basepairs $a = 7,\ldots,10$. The ordering of the matrix entries is such that the first row and column begin at the lower-left corner. Thus the matrix diagonal proceeds from lower-left to upper-right. For clarity, the block entries are labeled according to the base they represent, with numbers between 1 and n denoting bases on the reference strand, and numbers between $n + 1$ and $2n$ denoting the complementary bases on the opposite strand, with numbers increasing from the 5′ to 3′ direction on each strand.

simulations and the rigidity assumption on the bases, we compared the structure of the estimated matrices with the structure of the exact matrices defined in section 2.6. First, we found that the estimated matrix M_E was nearly block-diagonal in accordance with the exact matrix M. Indeed, the Euclidean norm of the off-diagonal portion of the matrix was only 1.5% of the norm of the entire matrix. Second, we found that each of the matrices M_E^a and $\bar{M}_E^a$ had a structure in accordance with M^a and $\bar{M}^a$ as encapsulated in eqn (41). Specifically, using the notation from (115), each sub-block B_1^a and $\bar{B}_1^a$ was nearly isotropic (scalar multiple of the identity), and each sub-block B_2^a and $\bar{B}_2^a$ was nearly skew-symmetric, as can be seen in Fig. 5. Indeed, for each sub-block B_1^a and $\bar{B}_1^a$, we found that the Euclidean norm of the anisotropic part was less than 3% of

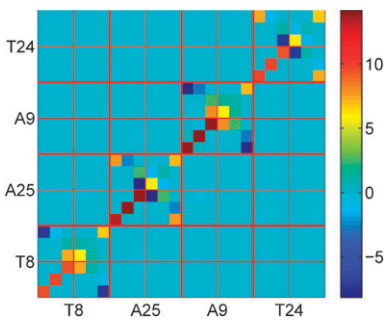

Fig. 5 A zoom-in of Fig. 4. The thick grid lines denote 6 × 6 blocks whereas thin grid lines denote 3 × 3 sub-blocks. The 6 × 6 diagonal blocks correspond to the mass matrices M_E^a and $\bar{M}_E^a$ for the four individual bases in basepairs $a = 8, 9$. The structure of the 3 × 3 sub-blocks of these diagonal blocks is consistent with the rigid base model.

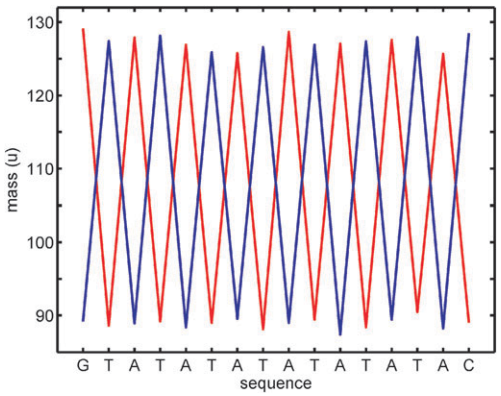

Fig. 6 Estimated mass parameters m_E^a and $\bar{m}_E^a$ *versus* sequence for the rigid base model computed with the fine-sampled trajectory. The parameters m_E^a are represented by the vertices of the red curve and $\bar{m}_E^a$ by the vertices of the blue curve. Masses are expressed in atomic mass units (u). The base sequence of the reference strand is indicated on the horizontal axis. The visible symmetry of the data is consistent with the palindromic symmetry relations for a rigid base model.

the norm of the sub-block. Moreover, for each sub-block B_2^a and $\bar{B}_2^a$, we found that the Euclidean norm of the symmetric part was less than 3.5% of the norm of the sub-block.

Fig. 6 shows results for the estimated mass parameters m_E^a and $\bar{m}_E^a$ *versus* sequence for the rigid base model. As another check on the simulation, we tested the estimated values against the palindromic symmetry relation $m^a = \bar{m}^{n-a+1}$ from section 2.9, which states that the mass of the reference-strand base in basepair 1 should be equal to the mass of the complementary-strand base in basepair n, and so on. The data in the Figure show that this relation is nearly satisfied for all basepairs. Indeed, we found that the maximum value of the relative difference $|\Delta m_E^a|/m_E^a$ was 1.8%, where $\Delta m_E^a = m_E^a - \bar{m}_E^{n-a+1}$. We further checked the symmetry relations $c^a = \Psi\bar{c}^{n-a+1}$ and $\Gamma^a = \Psi\bar{\Gamma}^{n-a+1}\Psi$ for the center of mass and rotational inertia components. For the center of mass components, we found that the maximum difference $|\Delta c_E^a|$ (Euclidean norm) was 0.04 Å, which is small compared to inter-atomic distances (about 1 Å), or even to the precision of atomic Cartesian coordinates in X-ray structures (roughly 0.1 Å). For the rotational inertia components, we compared principal values and axes rather than the matrices themselves. Denoting the principal values of Γ_E^a by $\lambda_{1,E}^a \le \lambda_{2,E}^a \le \lambda_{3,E}^a$, and similarly for $\Psi\bar{\Gamma}_E^{n-a+1}\Psi$ (notice that the principal values of this matrix are identical to those for $\bar{\Gamma}_E^{n-a+1}$), we found that the maximum value of the relative difference $|\Delta\lambda_{3,E}^a|/\lambda_{3,E}^a$ was 10%, $|\Delta\lambda_{2,E}^a|/\lambda_{2,E}^a$ was 14% and $|\Delta\lambda_{1,E}^a|/\lambda_{1,E}^a$ was 34%. We remark that this last value occurred at an isolated basepair ($a = 4$) and that $|\Delta\lambda_{1,E}^a|/\lambda_{1,E}^a$ did not exceed 16% for the other basepairs ($a \neq 4$). For all a, the matrices Γ_E^a and $\Psi\bar{\Gamma}_E^{n-a+1}\Psi$ were found to have three distinct principal values $\lambda_{i,E}^a$ and $\bar{\lambda}_{i,E}^{n-a+1}$, and the relative rotation angles between corresponding principal axes triads were less than 12 degrees. Thus the palindromic symmetry relations were all approximately satisfied. Notice that the satisfaction of these relations does not preclude context effects. That is, the relations can be satisfied even when a base at one location in the oligomer exhibits parameters different from those of an identical base at another location.

Table 1 Selected mass parameters for the bases A, G, C and T in the rigid base model. The data in normal font are estimated values obtained from the fine-sampled trajectory. For bases A and T, the average over the oligomer is given, with standard deviation shown in parentheses. For bases G and C, the two values from each oligomer end are given, ordered in the 5′ to 3′ direction of the reference strand. The data in bold font refer to bases in their canonical geometries as defined in section 4. The data in italic font refer to purine or pyrimidine rings in their canonical geometries (bases with exocyclic heavy atoms and hydrogens removed)

	A	G	C	T
	127 (1)	129, 128	89, 89	89 (1)
m (u)	**134**	**150**	**110**	**125**
	116	*116*	*76*	*76*
	−0.93 (0.02)	−0.93, −0.95	−0.60, −0.59	−0.56 (0.01)
c_1 (Å)	**−0.46**	**−0.76**	**−0.25**	**0.08**
	−0.75	*−0.75*	*−0.22*	*−0.22*
	3.00 (0.02)	2.99, 3.02	3.92, 3.94	3.93 (0.01)
c_2 (Å)	**2.52**	**2.25**	**3.36**	**3.50**
	2.74	*2.74*	*3.70*	*3.70*
	0.00 (0.01)	0.02, 0.02	0.02, 0.01	0.00 (0.02)
c_3 (Å)	**0.00**	**0.00**	**0.00**	**0.00**
	0.00	*0.00*	*0.00*	*0.00*
	2.51 (0.08)	2.70, 2.51	1.04, 1.00	1.1 (0.12)
λ_1 (10^{-45} kg m^2)	**3.46**	**4.22**	**2.11**	**2.56**
	1.90	*1.90*	*1.18*	*1.18*
	5.05 (0.23)	5.15, 4.92	2.18, 2.27	2.50 (0.11)
λ_2 (10^{-45} kg m^2)	**5.18**	**7.44**	**4.06**	**5.81**
	4.36	*4.36*	*1.23*	*1.23*
	8.57 (0.29)	8.60, 8.57	4.54, 4.23	4.43 (0.21)
λ_3 (10^{-45} kg m^2)	**8.64**	**11.66**	**6.17**	**8.31**
	6.26	*6.26*	*2.41*	*2.41*

Table 1 shows selected mass parameters for the bases A, G, C and T in the rigid base model obtained by different means. Shown are estimated values obtained from the fine-sampled trajectory, values obtained for the bases in their canonical geometries as defined in section 4, and values obtained for only the purine or pyrimidine rings in their canonical geometries. In all cases, the estimated mass m_E is consistently smaller than that of the canonical base, but larger than that of the mere canonical ring. The estimated center of mass coordinates $c_{1,E}$ and $c_{2,E}$ are all systematically shifted compared to the canonical values, whereas the estimated coordinates $c_{3,E}$ are all nearly identical to the canonical values. The estimated principal values $\lambda_{i,E}$ of the rotational inertia matrix are nearly all smaller than the canonical base values and larger than the canonical ring values. The only exceptions occur for the smallest principal values of C and T. Here the estimated values fall below those of the canonical bases and rings. The differences between the estimated and canonical values of parameters may be due to the flexibility of the bases. Indeed, the exocyclic groups are attached to the rings in a somewhat flexible manner, leading to bases that are not entirely rigid. Thus it is natural to expect that estimated values based on a dynamical simulation will be different from canonical values based on a single, fixed configuration. These differences may in fact provide a measure of the real flexibility of a base.

Fig. 7 shows a portion of the estimated mass matrix $(\mathrm{M_{bp}})_E$ for the rigid basepair model. The 6 × 6 diagonal blocks in Fig. 7 correspond to the mass matrices $(M_{\mathrm{bp}})^a_E$ for the eight basepairs $a = 5,\ldots,12$. To assess the simulations and rigidity assumption on the basepairs, we compared the structure of the estimated matrices with the structure of the exact matrices defined in section 3.6. First, we found that the estimated matrix $(\mathrm{M_{bp}})_E$ was nearly block-diagonal in accordance with the exact matrix $\mathrm{M_{bp}}$. Indeed, the Euclidean norm of the off-diagonal portion of the matrix was only 3.6% of the norm of the entire matrix. Second, we found that some of the matrices $(M_{\mathrm{bp}})^a_E$ had a structure that was inconsistent with that of M^a_{bp} as encapsulated in eqn (87). Specifically, using notation analogous to (115), some sub-blocks $(B_{\mathrm{bp}})^a_1$ were not close to isotropic, and some sub-blocks $(B_{\mathrm{bp}})^a_2$ were not close to skew-symmetric, as can be seen in Fig. 8. Indeed, for some

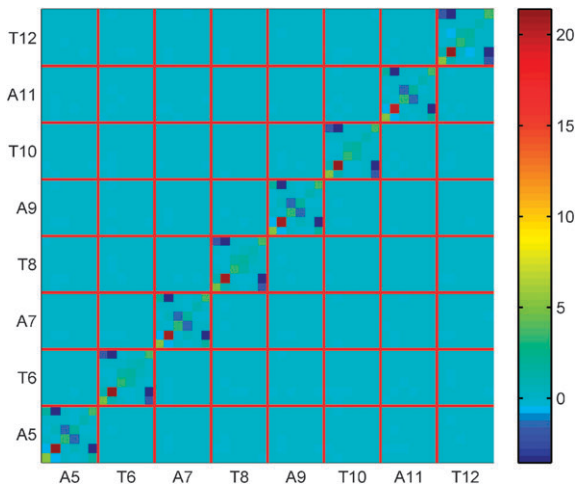

Fig. 7 A portion of the estimated mass matrix $(\mathrm{M_{bp}})_E$ in reduced form for the rigid basepair model computed with the fine-sampled trajectory. The portion shown corresponds to the 6×6 block entries $(\mathrm{M_{bp}})^{\alpha\beta}_E$ for $\alpha,\beta = 5,\ldots,12$, which are marked by the grid lines. The diagonal blocks correspond to the mass matrices $(M_{\mathrm{bp}})^a_E$ for the eight basepairs $a = 5,\ldots,12$. The ordering of the matrix entries is such that the first row and column begin at the lower-left corner. Thus the matrix diagonal proceeds from lower-left to upper-right. For clarity, the block entries are labeled according to the base and position on the reference strand.

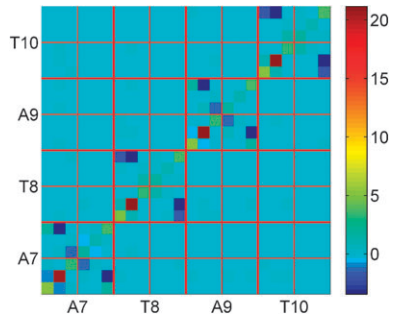

Fig. 8 A zoom-in of Fig. 7. The thick grid lines denote 6 × 6 blocks whereas thin grid lines denote 3 × 3 sub-blocks. The 6 × 6 diagonal blocks correspond to the mass matrices $(M_{\mathrm{bp}})^a_E$ for the four basepairs $a = 7,\ldots,10$. The structure of some of the 3 × 3 sub-blocks of these diagonal blocks is inconsistent with the rigid basepair model.

sub-blocks $(B_{\mathrm{bp}})^a_1$, we found that the Euclidean norm of the anisotropic part was greater than 62% of the norm of the sub-block. Moreover, for some sub-blocks $(B_{\mathrm{bp}})^a_2$, we found that the Euclidean norm of the symmetric part was greater than 52% of the norm of the sub-block. Thus, in contrast to the base model, the rigidity assumption in the basepair model is not supported by the data. This is not surprising, since a mere glance at the visualized trajectory reveals, in some cases, substantial fluctuations of the two bases in a pair relative to each other, indicating that basepairs do not generally fluctuate as rigid entities.

Although the rigidity assumption in the basepair model may be overly idealistic, it might still provide a useful compromise between simplicity and accuracy in applications where a low-resolution model of DNA is acceptable. In this respect, we remark that the mass parameters estimated here, as well as the shape and stiffness parameters discussed below, should be interpreted simply as best-fit parameters obtained by matching the model to moments of the simulated data.

5.3 Shape, stiffness parameters

Here we outline results on the estimation of shape and stiffness parameters for the rigid base and basepair models from the coarse-sampled trajectory described in section 4. As before, we illustrate various consistency checks to assess the quality of the numerical simulations and various assumptions in the models. By the intra-basepair coordinates we mean the internal coordinates $y^a = (\vartheta,\xi)^a$, where ϑ^a are buckle–propeller–opening and ξ^a are shear–stretch–stagger, and by the inter-basepair coordinates we mean the internal coordinates $z^a = (\theta,\zeta)^a$, where θ^a are tilt–roll–twist and ζ^a are shift–slide–rise coordinates, as defined in sections 2 and 3. We recall that ϑ^a and θ^a are non-dimensional Cayley coordinates defined *via* the parameterization in (6).

Fig. 9 shows the estimated shape parameters $\hat{y}^a_E$ and $\hat{z}^a_E$ *versus* sequence for the rigid base model. As a check on the simulation, we tested the estimated values against the palindromic symmetry relation $(63)_1$ from section 2.9, which requires that all the shape parameters be symmetric functions of position about the middle of the oligomer, except buckle, shear, tilt and shift, which should be antisymmetric. The data in the Figure show that these conditions are nearly satisfied. Indeed, we found that these conditions were satisfied to within

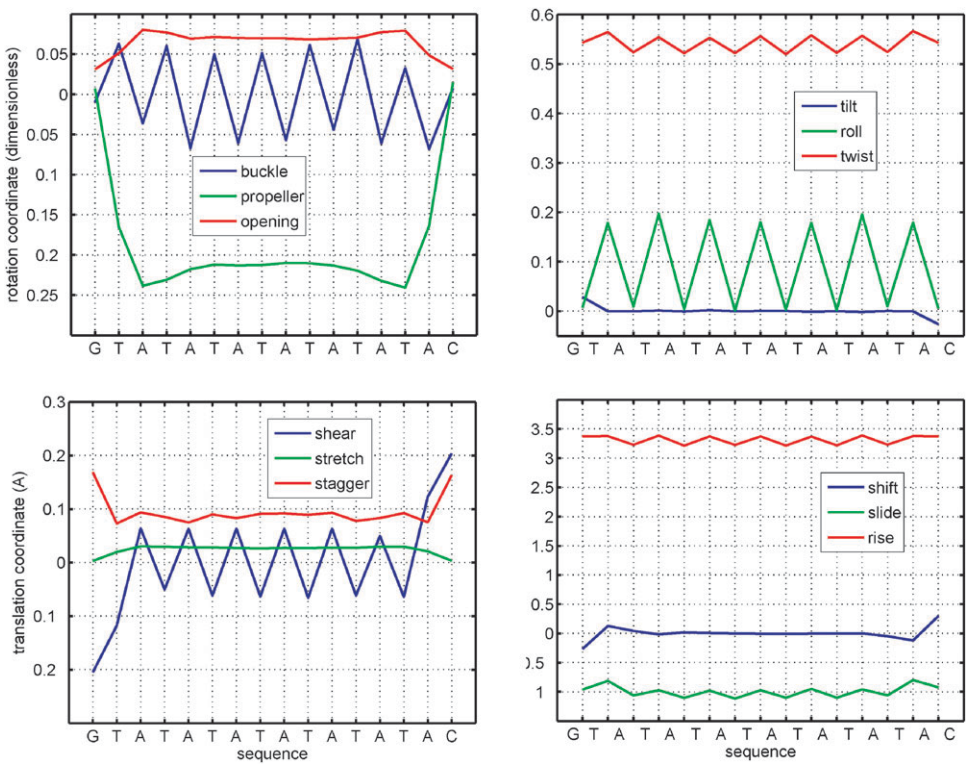

Fig. 9 Estimated shape parameters $\hat{y}^a_E$ and $\hat{z}^a_E$ *versus* sequence for the rigid base model computed with the coarse-sampled trajectory. The base sequence of the reference strand is indicated on the horizontal axis, and the parameter values are interpolated by piecewise-linear curves. The visible symmetry of the data is consistent with the palindromic symmetry relations for the rigid base model.

a maximum absolute error of 0.008 for buckle–propeller–opening and tilt–roll–twist, and 0.04 Å for shear–stretch–stagger and shift–slide–rise, both of which are small compared to the scales in the figure. The values of the parameters fall within expected ranges for the B-DNA structural family. Some of the parameter curves, for example buckle, roll and shear, are remarkably periodic and reflect strong differences in values for the TA and AT dimer steps. In contrast, some of the curves, for example tilt, stretch and shift, are rather flat and reflect only weak differences. Other curves, for example propeller and shear, show pronounced behavior near the two ends of the oligomer. It is unknown whether this behavior is due to specific sequence effects or simply the influence of the free ends.

Fig. 10 shows a portion of the estimated covariance matrix K_E^{-1} and stiffness matrix K_E for the rigid base model. The portion shown corresponds to the internal coordinates (y^a,z^a) for the individual bases in basepairs $a = 6,\ldots,11$. To assess the simulations, we checked the block entries of the estimated

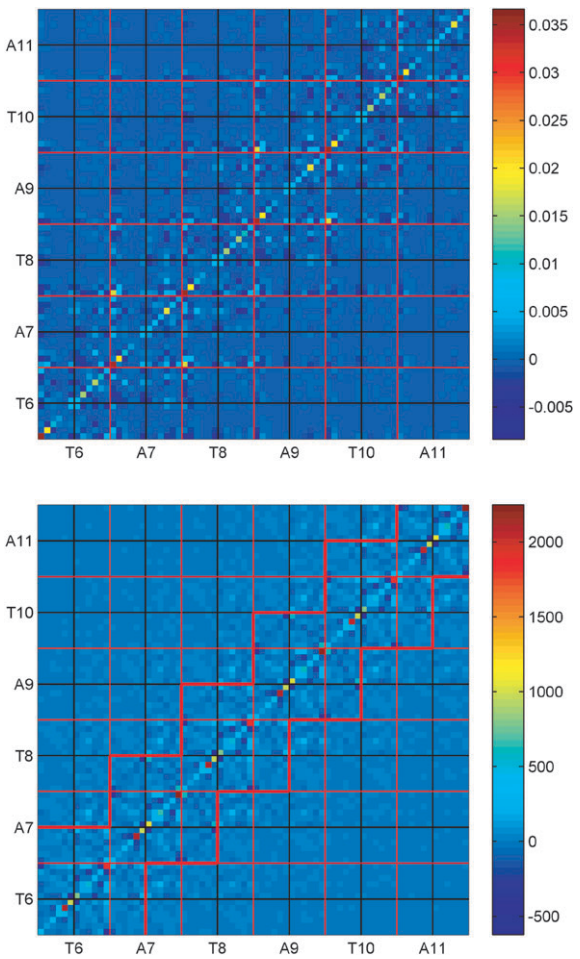

Fig. 10 A portion of the estimated covariance matrix K_E^{-1} (top) and stiffness matrix K_E (bottom) in reduced form for the rigid base model computed with the coarse-sampled trajectory. Horizontal and vertical bands marked by thin red lines contain entries corresponding to (y^a,z^a) $(a = 6,\ldots,11)$. Entries corresponding to y^a and z^a are further separated from each other by thin black lines. The diagonal region of the stiffness matrix marked by the thick red lines denotes the structure associated with a local internal energy model.

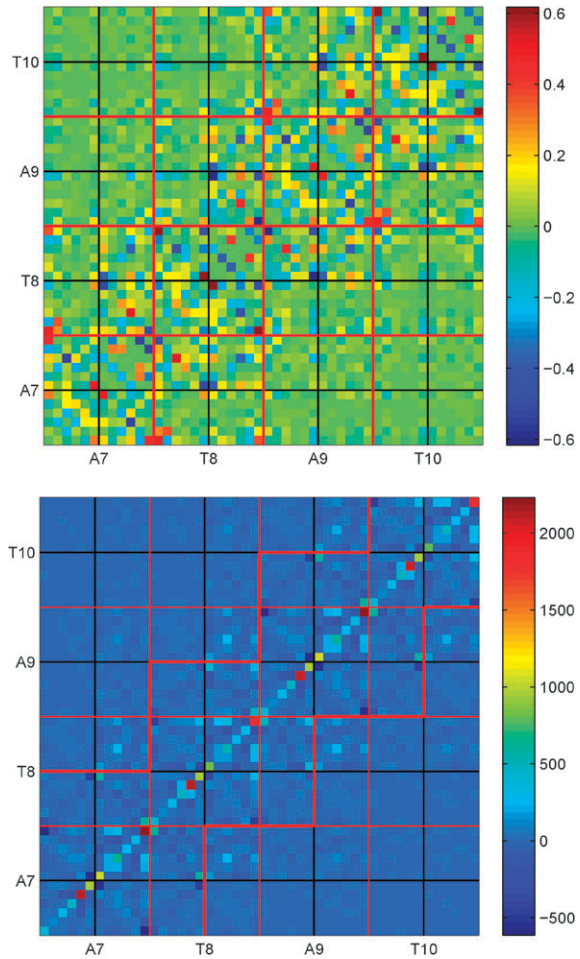

Fig. 11 A zoom-in of the two matrices in Fig. 10. Coordinate orderings and line markings are the same as before. For clarity, the covariance matrix has been normalized to produce a correlation matrix with unit diagonal entries, which have been omitted. The estimated stiffness matrix has a structure that is nearly consistent with a local internal energy model, in which all entries outside of the diagonal portion marked by the thick red lines are zero.

stiffness matrix against the palindromic symmetry relation in $(63)_2$ and found that the relations were all satisfied to within a maximum absolute error of approximately 50 in the Euclidean norm, which is small compared to the dimensionless scale in the bottom panel of the Figure. A zoom-in of both matrices is shown in Fig. 11, where the zoomed-in version of the covariance matrix has been normalized into a correlation matrix. The data shows that the correlation patterns in the 12×12 diagonal blocks marked by red lines, and indeed each of their 6×6 sub-blocks marked by black lines, are noticeably periodic. Distinct differences in the correlation patterns of the intra-basepair coordinates y^a are visible between the (A,T) and (T,A) base pairs, and in the inter-basepair coordinates z^a between the AT and TA dimer steps. The observed patterns are in partial agreement with those computed from crystallographic databases,[31,34] but a detailed comparison is difficult due to differences in the definitions of the internal coordinates among other things. The data from the stiffness matrix shows that the largest entries are concentrated in a region near the

diagonal, which implies that interactions between proximal bases make a dominant contribution to the internal energy. Indeed, most of the largest entries fall within the region marked by the thick red lines, which corresponds to a local model as described in section 2.5, and remarkably, this portion of the matrix is itself positive-definite. Some notable entries that fall outside this region correspond to twist–twist, tilt–tilt and twist–rise couplings between adjacent dimer steps.

Fig. 12 shows diagonal entries of the estimated stiffness matrix *versus* sequence for the rigid base model. As before, we tested the estimated values against the palindromic symmetry relation $(63)_2$, which requires that all diagonal entries be symmetric functions of position about the middle of the oligomer. The data in the Figure show that these conditions are nearly satisfied. Indeed, we found that these conditions were satisfied to within a maximum relative error of 6% for the slide–slide entry, 5.5% for twist–twist, 3.5% for roll–roll and 2.5% for all the other diagonal entries in the stiffness matrix. As with the shape parameters, some of the diagonal stiffness curves, for example, tilt–tilt, roll–roll, twist–twist and rise–rise, are remarkably periodic and reflect strong differences in values for the TA and AT dimer steps. In contrast, the remaining curves are rather flat and reflect only weak differences. Most of the curves show pronounced behavior near the two ends of the oligomer. As before, it is unknown whether this behavior is due to specific sequence effects or simply the influence of the free ends.

Fig. 13 shows a portion of the estimated stiffness matrix $(\mathrm{K_{bp}})_E$ and a certain block-diagonal or local approximation $(\mathrm{K'_{bp}})_E$ for the basepair model, both obtained using a Gaussian approximation as described in section 3.8. Whereas the matrix $\mathrm{K_{bp}}$ is defined by the covariance of the full internal coordinate vector $\mathrm{w_{bp}} = (z^1,\ldots,z^{n-1})$ as recorded in eqn (103), each diagonal block of the matrix $\mathrm{K'_{bp}}$ is defined by the covariance of each individual coordinate vector z^a $(a = 1,\ldots,n-1)$. The matrix $\mathrm{K'_{bp}}$ is referred to as local since it is consistent with a local internal energy model as defined in section 3.5. Matrices of the form $\mathrm{K'_{bp}}$ have been investigated in various molecular dynamics and crystal database analyses.[20,31,34] Indeed, any covariance analysis in which individual dimer steps are treated as independent yields a block diagonal matrix $\mathrm{K'_{bp}}$ as defined here. As can be seen from the Figure, the estimated basepair stiffness matrix $(\mathrm{K_{bp}})_E$ and the block-diagonal approximation $(\mathrm{K'_{bp}})_E$ are significantly different; the Euclidean norm of the difference is 66% of the norm of $(\mathrm{K_{bp}})_E$. Thus, while the estimated stiffness matrix for the rigid base model is nearly local (Fig. 10 and 11, bottom), the estimated stiffness matrix for the basepair model is noticeably non-local (Fig. 13, top). This non-locality can be understood as a consequence of the compatibility relations derived in section 3.9. In the Gaussian approximation, the rigid basepair stiffness matrix $\mathrm{K_{bp}}$ is related to the rigid base stiffness matrix K through a Schur complement as encapsulated in eqn (110). From this we deduce, by properties of the matrix inverse, that a nearly local form for K does not imply the same for $\mathrm{K_{bp}}$.

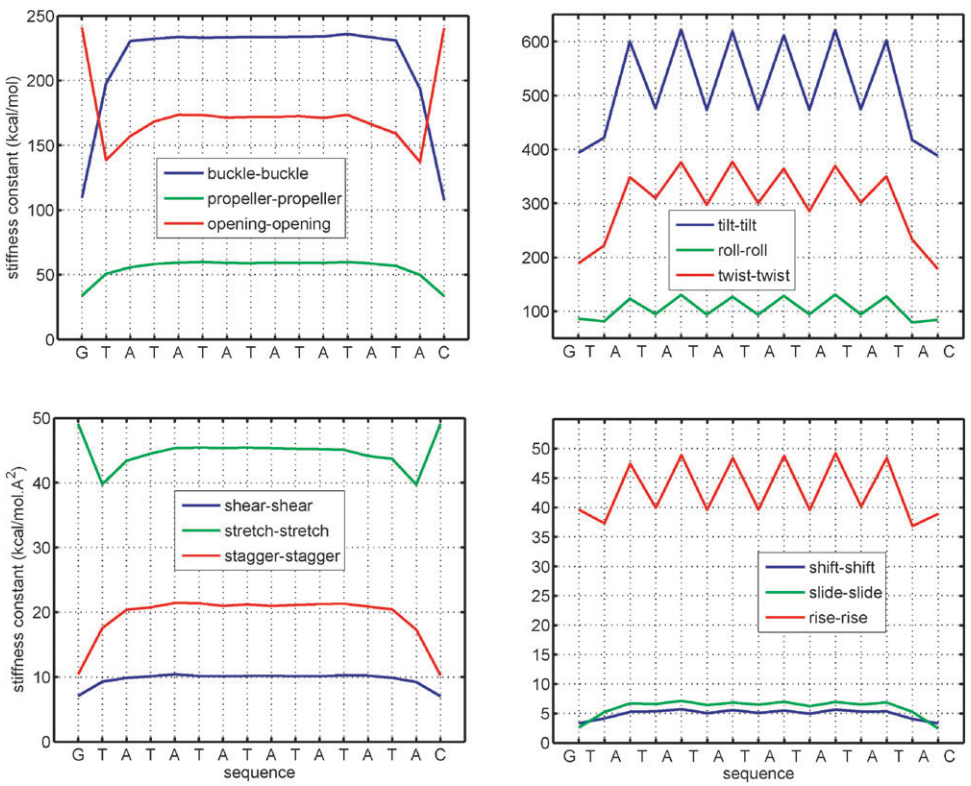

Fig. 12 Diagonal entries of the estimated stiffness matrix K_E *versus* sequence for the rigid base model computed with the coarse-sampled trajectory. The base sequence of the reference strand is indicated on the horizontal axis. The diagonal entries are labeled according to the interactions they represent and are interpolated by piecewise-linear curves. The visible symmetry of the data is consistent with the palindromic symmetry relations for the rigid base model.

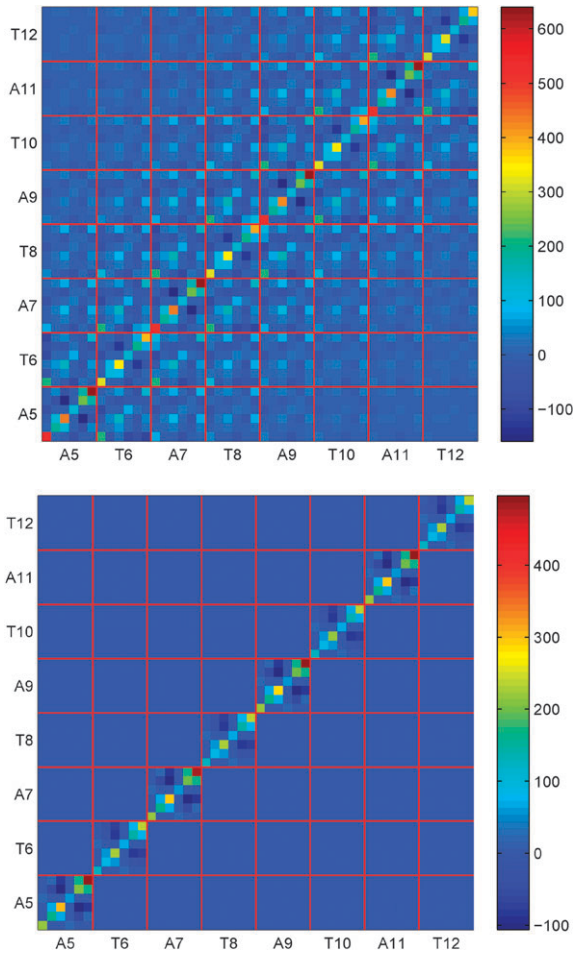

Fig. 13 A portion of the estimated stiffness matrix $(K_{bp})_E$ and a block-diagonal approximation $(K'_{bp})_E$ in reduced form for the basepair model computed with the coarse-sampled trajectory. Horizontal and vertical bands marked by red lines contain entries corresponding to the internal coordinates z^a ($a = 5,\ldots,12$). Each diagonal block of $(K'_{bp})_E$ was computed from the covariance of each individual coordinate vector z^a. The estimated matrix $(K_{bp})_E$ is significantly different from the block-diagonal approximation $(K'_{bp})_E$, which indicates that the simulated data is inconsistent with a local internal energy model.

6. Summary and conclusions

A method has been developed to estimate a complete set of sequence-dependent shape, stiffness and mass parameters for rigid base and basepair models of DNA in solution. The method is based on atomic-resolution trajectories obtained by explicit-solvent MD simulations, uses a coarse-graining procedure which is properly consistent with equilibrium statistical mechanics on the full phase space of the models, employs special procedures for keeping the simulated DNA within the B-DNA family by avoiding non-canonical backbone flips, and furthermore uses a data filter to eliminate simulated structures with broken intra-basepair hydrogen bonds.

The sequence-dependent models we consider are specified by their kinetic and internal energy functions. The kinetic energy function is parameterized by the effective mass, center-of-mass coordinates and rotational inertia matrices of the bases or basepairs. The internal energy function is assumed to be elastic and quadratic in a natural set of internal coordinates satisfying the Cambridge convention and is parameterized by the equilibrium (minimum energy) shape parameters and the stiffness matrix associated with base or basepair interactions. In contrast to previous studies in the literature, we make no assumption about the locality of the internal elastic energy and include the appropriate Jacobian factors associated with the non-Cartesian coordinates which describe the relative, three-dimensional rotations between bases or basepairs.

The method is accompanied by various analytical consistency checks that can be used to assess the equilibration of statistical averages, and different modeling assumptions pertaining to the rigidity of the bases or basepairs and the locality of the quadratic internal energy. For general sequences, it was shown that the rigidity and locality assumptions imply certain theoretical sparsity patterns for the global mass and stiffness matrices for each model, and moreover, in the Gaussian approximation, the shape and stiffness parameters of the two models must necessarily satisfy certain compatibility relations involving a Schur complement. For special sequences, such as the palindromic sequence considered here, it was furthermore shown using objectivity arguments that the sequence dependence of the mass, shape and stiffness parameters must necessarily satisfy various symmetry requirements. Specifically, we showed that material parameters must be either symmetric or antisymmetric functions of position about the middle of the sequence.

The practicability of our method was demonstrated by applying it to estimate a complete parameter set for the 16-basepair oligomer $G(TA)_7C$ simulated in explicit water and counterions. Two different trajectories were considered: a fine-sampled 1 ns trajectory for estimating mass parameters, and a coarse-sampled 180 ns trajectory for estimating shape and stiffness parameters. The analytical consistency checks, together with detailed convergence tests based on the comparison of values from non-overlapping time windows, suggested that the trajectories were sufficiently long for the relevant averages to equilibrate. Our results indicate that sequence-dependent variations in the material parameters can be resolved rather well. Moreover, they show that the assumptions of rigidity and locality hold rather well for the base model, but not for the basepair model. Whereas the non-rigid nature of basepairs is intuitively and mechanically clear, we showed that the non-local nature of the internal energy in the basepair model can be understood in terms of a compatibility relation involving a Schur complement. We stress that the locality of the rigid base internal energy is a result implied by the simulated data; it was not assumed *a priori*.

The quadratic assumption on the internal elastic energy is not invariant under general changes of coordinates. Consequently, any results on the details of such an energy, for example the local or non-local structure of the associated stiffness matrix, will in general be coordinate dependent. In contrast, the quadratic form of the kinetic energy expressed in terms of physical velocity components as done here and the

associated physical mass matrix are invariant. Our choice to use internal coordinates based on a Cayley parameterization of three-dimensional rotations was made primarily for mathematical convenience. Indeed, the Cayley parameterization has a straightforward geometrical interpretation and leads to elegant expressions for the mid-rotation frames among other things. To what extent the material symmetry and stiffness matrix locality results described here would also hold for other choices of internal coordinates such as those employed in *3DNA*,[27] *Curves*,[23] *Curves+*[22] and other structural analysis programs is an open question that will be pursued in future work. We remark that the Cayley parameterization of rotations adopted here, in which the rotation vector can take any value in $\mathbb{R}^3$ with its norm approaching infinity as the angle of rotation approaches π radians or 180 degrees, is closely related to the rotation parameterization adopted in *Curves+*, with the only difference being that, in the *Curves+* convention, the norm of the rotation vector is used to encode the angle of the rotation expressed in degrees, so that it lies in the ball of radius 180.

The purpose of this article was to examine coarse-grained models of DNA, including methods for extracting coarse-grained parameters from fine-grained MD simulations. Indeed, our results show that, through an appropriate analysis of atomic-resolution simulations, various coarse-grained modeling assumptions pertaining to rigidity and locality can be assessed. It was not our objective to test the validity or otherwise of any specific MD force field or protocol. This study was initiated and completed using the atomistic force field *parm94*.[8] However, the same methodology and analysis could be applied to simulations based on newer force fields such as *parmbsc0*,[35] where some of the filtering steps we adopt may no longer be necessary. We believe that our conclusions pertaining to rigidity and locality in the rigid base and basepair models would be unaffected. Recently, simulations of 39 different DNA oligomers containing all possible sequence tetramers using *parmbsc0* have become available.[24] The analysis of those simulations using the methods developed here will be pursued in future work.

A novel feature of our approach is the ability to estimate effective mass parameters for both the rigid base and basepair models. For the model system studied here, these parameters were deduced from a fine-sampled 1 ns trajectory. Despite its shortness as compared to the coarse-sampled 180 ns trajectory used for the shape and stiffness parameters, we believe this trajectory was sufficiently long to extract converged estimates of the mass parameters. This belief is supported not only by the convergence tests described above, but also by analytical consistency checks based on the theoretical structure of the mass matrices. Indeed, for the rigid base model, for which rigidity is clearly a realistic assumption, the theoretically predicted sparsity pattern of the mass matrix was realized with rather high accuracy in the estimates obtained from the 1 ns trajectory.

The estimates of shape, stiffness and mass parameters obtained from time series generated by one MD force field and set of protocols as compared to others will of course vary. The parameters may well also depend on the sequence context, ions and their concentration, and temperature. The study of these effects is of specific interest in developing a better understanding of the coarse-grained consequences of fine-grained details.

Acknowledgements

The following support is gratefully acknowledged. FL, the J. E. Purkyně Fellowship and grant no. Z40550506 from the Academy of Sciences of the Czech Republic, project no. LC512 from the Ministry of Education, Youth and Sports (MŠMT) of the Czech Republic, and the Swiss National Science Foundation. OG, the US National Science Foundation. LMH, GS, MM and JHM, the Swiss National Science Foundation.

References

1 B. Berne and R. Pecora, *Dynamic Light Scattering: With Applications to Chemistry, Biology and Physics*, Wiley, New York, 1976.
2 D. Beveridge, G. Barreiro, K. Byun, D. Case, T. Cheatham III, S. Dixit, E. Giudice, F. Lankas, R. Lavery, J. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. Thayer, P. Varnai and M. Young, Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps, *Biophys. J.*, 2004, **87**, 3799–3813.
3 D. Beveridge and K. McConnell, Nucleic acids: Theory and computer simulation, Y2K, *Curr. Opin. Struct. Biol.*, 2000, **10**, 182–196.
4 T. Cheatham III, Simulation and modeling of nucleic acid structure, dynamics and interactions, *Curr. Opin. Struct. Biol.*, 2004, **14**(3), 360–367.
5 T. Cheatham III and D. Case, *Using Amber to simulate DNA and RNA*, in *Computational Studies of RNA and DNA*, ed. J. Sponer and F. Lankas, Springer, 2006, pp. 45–71.
6 T. Cheatham III and P. Kollman, Molecular dynamics simulation of nucleic acids, *Annu. Rev. Phys. Chem.*, 2000, **51**, 435–471.
7 B. Coleman, W. Olson and D. Swigon, Theory of sequence-dependent DNA elasticity, *J. Chem. Phys.*, 2003, **118**(15), 7127–7140.
8 W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz Jr, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell and P. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.*, 1995, **117**, 5179–5197.
9 R. Dickerson, M. Bansal, C. Calladine, S. Diekmann, W. Hunter, O. Kennard, R. Lavery, H. Nelson, W. Olson, W. Saenger, Z. Shakked, H. Sklenar, D. Soumpasis, C.-S. Tung, E. von Kitzing, A. Wang and V. Zhurkin, Definitions and nomenclature of nucleic acid structure parameters, *J. Mol. Biol.*, 1989, **205**, 787–791.
10 M. El Hassan and C. Calladine, The assessment of the geometry of dinucleotide steps in double-helical DNA: A new local calculation scheme, *J. Mol. Biol.*, 1995, **251**, 648–664.
11 E. Fredericq and C. Houssier, *Electric Dichroism and Electric Birefringence*, Clarendon, Oxford, 1973.
12 D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, London, 2002.
13 E. Giudice and R. Lavery, Simulations of nucleic acids and their complexes, *Acc. Chem. Res.*, 2002, **35**, 350–357.
14 O. Gonzalez and J. Maddocks, Extracting parameters for base-pair level models of DNA from molecular dynamics simulations, *Theor. Chem. Acc.*, 2001, **106**, 76–82.
15 D. Hawcroft, *Electrophoresis*, Oxford University Press, Oxford, 1997.
16 R. Hogg and A. Craig, *Introduction to Mathematical Statistics*, Macmillan, New York, 3rd edn, 1970.
17 K. Huang, *Statistical Mechanics*, Wiley, New York, 2nd edn, 1987.
18 P. Hughes, *Spacecraft Attitude Dynamics*, Wiley, Boston, 1986.
19 F. Lankas, R. Lavery and J. Maddocks, Kinking occurs during molecular dynamics simulations of small DNA minicircles, *Structure*, 2006, **14**, 1527–1534.

20 F. Lankas, J. Sponer, J. Langowski and T. Cheatham III, DNA basepair step deformability inferred from molecular dynamics simulations, *Biophys. J.*, 2003, **85**, 2872–2883.
21 F. Lankas, J. Sponer, J. Langowski and T. Cheatham III, DNA deformability at the base pair level, *J. Am. Chem. Soc.*, 2004, **126**, 4124–4125.
22 R. Lavery, M. Moakher, J. Maddocks, D. Petkeviciute and K. Zakrzewska, Conformational analysis of nucleic acids revisited: *Curves+*, *Nucleic Acids Res.*, 2009, **37**, 5917–5929.
23 R. Lavery and H. Sklenar, Defining the structure of irregular nucleic acids: Conventions and principles, *J. Biomol. Struct. Dyn.*, 1989, **6**(4), 655–667.
24 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.*, 2009, DOI: 10.1093/nar/gkp834, Advance Access published on 22 October.
25 D. Lay, *Linear Algebra and its Applications*, Addison-Wesley, 2nd edn, 2000.
26 R. Lipsitz and N. Tjandra, Residual dipolar couplings in NMR structure analysis, *Annu. Rev. Biophys. Biomol. Struct.*, 2004, **33**, 387–413.
27 X.-J. Lu and W. Olson, 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, *Nucleic Acids Res.*, 2003, **31**(17), 5108–5121.
28 J. McCammon and S. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, 1987.
29 A. McPherson, *Introduction to Macromolecular Crystallography*, Wiley-Liss, 2003.
30 W. Olson, M. Bansal, S. Burley, R. Dickerson, M. Gerstein, S. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.-S. Tung, E. Westhof, C. Wolberger and H. Berman, A standard reference frame for the description of nucleic acid base-pair geometry, *J. Mol. Biol.*, 2001, **313**, 229–237.
31 W. Olson, A. Gorin, X.-J. Lu, L. Hock and V. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 11163–11168.
32 M. Orozco, A. Noy and A. Perez, Recent advances in the study of nucleic acid flexibility by molecular dynamics, *Curr. Opin. Struct. Biol.*, 2008, **18**(2), 185–193.
33 M. Orozco, A. Perez, A. Noy and F. Luque, Theoretical methods for the simulation of nucleic acids, *Chem. Soc. Rev.*, 2003, **32**, 350–364.
34 A. Perez, F. Lankas, F. Luque and M. Orozco, Towards a molecular dynamics consensus view of B-DNA flexibility, *Nucleic Acids Res.*, 2008, **36**(7), 2379–2394.
35 A. Perez, I. Marchan, D. Svozil, J. Sponer, T. Cheatham III, C. Laughton and M. Orozco, Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers, *Biophys. J.*, 2007, **92**, 3817–3829.
36 *Modern Analytical Ultracentrifugation*, ed. T. Schuster and T. Laue, Birkhauser, Boston, 1994.
37 B. Valeur, *Molecular Fluorescence: Principles and Applications*, Wiley-VCH, Weinheim, 2002.
38 L. Zidek, R. Stefl and V. Sklenar, NMR methodology for the study of nucleic acids, *Curr. Opin. Struct. Biol.*, 2001, **11**(3), 275–281.

