Theor Chem Acc (2001) 106:76–82
DOI 10.1007/s002140100256

Theoretical
Chemistry Accounts

Regular article

Extracting parameters for base-pair level models of DNA from molecular dynamics simulations

Oscar Gonzalez[1], **John H. Maddocks**[2]

[1] Department of Mathematics, The University of Texas, Austin, TX 78712, USA
[2] Département de Mathématiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Received: 12 July 2000 / Accepted: 5 January 2001 / Published online: 3 May 2001

Abstract. A method is described to extract a complete set of sequence-dependent energy parameters for a rigid base-pair model of DNA from molecular dynamics (MD) simulations. The method is properly consistent with equilibrium statistical mechanics and leads to effective inertia parameters for the base-pair units as well as stacking and stiffness parameters for the base-pair junctions. We give explicit formulas that yield a complete set of base-pair model parameters in terms of equilibrium averages that can be estimated from a time series generated in an MD simulation. The expressions to be averaged depend strongly both on the choice of coordinates used to describe rigid-body orientations and on the choice of strain measures at each junction.

1 Introduction

The deformations of double-helical DNA in solvent have been studied by modeling DNA in many different ways, for example, as an atomistic system, as a system of interacting rigid bases or base pairs, or as a continuous elastic rod [1, 2, 3]. These models resolve DNA deformations to different levels of detail, possess different practical limitations, and collectively provide a means to explore deformations over a wide range of scales. A basic issue for each model is the determination of the necessary parameters, or constitutive relations, for the kinetic and potential energies. For a family of models possessing a natural hierarchy in resolution, it is reasonable to expect that parameters for lower-resolution models can be determined from those of higher-resolution models. One example of this would be the extraction of base-pair level parameters from atomistic level parameters via the analysis of molecular dynamics (MD) trajectories. Here we describe the necessary statistical mechanics computations to carry out this program and examine the various choices and assumptions that underly the approach.

A rigid base-pair model of DNA, with a quadratic potential energy at each junction, is completely defined by specifying a set, $\mathbb{K}$, of stacking and stiffness parameters for each of the base-pair junctions, along with a set, $\mathbb{M}$, of inertia parameters for each of the base-pair units. In a homogeneous model, each base pair would be assigned the same $\mathbb{M}$, and each junction the same $\mathbb{K}$. In a non-homogeneous model, each base pair and junction would be assigned independent parameter sets depending on sequence; for example, $\mathbb{M}$ could be monomer-based and $\mathbb{K}$ dimer-based. In this case there would be two independent sets $\mathbb{M}$, one for each of the independent base pairs identified (on one strand) by A or C, and ten independent sets $\mathbb{K}$, one for each of the independent base-pair dimers AT, AG, AC, AA, CG, CC, CA, GC, GA, TA. The parameters $\mathbb{M}$ could also be trimer-based and $\mathbb{K}$ tetramer-based and so on. Indeed, recent studies [4, 5] suggest that junction parameters such as $\mathbb{K}$ depend on the sequence beyond the surrounding dimer. The stacking and stiffness parameters $\mathbb{K}$, as well as the degree to which a quadratic potential energy itself is valid, depend upon the specific choice of variables used to describe the relative positions and orientations of the rigid base pairs. The inertia parameters $\mathbb{M}$, by their nature, are independent of this choice.

Stacking and stiffness parameters have previously been determined for various models by a number of different means [6, 7, 8, 9]. For example, dimer-based stacking parameters were estimated in Refs. [6, 7] by curve-fitting procedures applied to gel electrophoresis data. In Ref. [8], dimer-based stacking and stiffness parameters were estimated from protein–DNA crystal structures. Other levels of sequence-dependence have also been considered. For example, trimer-based stiffness parameters for a simplified base-pair model were estimated in Ref. [9] from enzyme binding data. The techniques used in these studies were adapted to exper-

Correspondence to: J. H. Maddocks
Contribution to the Symposium Proceedings of Computational Biophysics 2000

imental data that provided only indirect or incomplete information on the desired parameters. In contrast, our objective is to develop a technique for the quantitative determination of parameters that exploits the direct, detailed structural information available from an atomistic simulation of DNA.

Compared to their stacking and stiffness counterparts, inertia parameters have received relatively little attention. Their determination requires dynamic data that capture not only environmental and hydration effects, but also sequence-dependence. While difficult to produce by physical experiments, such data are produced by computational experiments on an atomistic model, for example, MD simulations. The extraction of rigid-body inertia parameters from dynamic data is itself an intrinsically difficult task. The main problem is how to explicitly relate the desired parameters to computational observables. As we show, such relations can be obtained from the full phase-space distribution function of the rigid-body system formulated in terms of noncanonical momentum variables. We identify a particular choice of variables that leads to a factorization of the phase-space distribution, which we then exploit to derive explicit expressions for inertia parameters.

We develop a method by which a complete set of sequence-dependent kinetic-energy and potential-energy parameters for a rigid base-pair model may be determined from an atomistic simulation. The method is similar to the inverse harmonic analysis in Ref. [8], but differs in four main respects: it is properly consistent with the canonical distribution of equilibrium statistical mechanics; it exploits full phase-space data rather than only configurational data; it yields effective inertia parameters for the base-pair units in addition to stacking and stiffness parameters; and it allows for the control of environmental conditions to the extent possible in an MD simulation. Our method may be viewed as an adaptation to the case of rigid bodies of the methodology in Ref. [10] for atomistic systems, but we do not assume that the appropriate metric Jacobian factors are constant as was done there.

We give explicit formulas for the base-pair model parameters in terms of equilibrium averages in the statistical mechanics sense. The expressions to be averaged depend strongly both on the choice of coordinates used to describe rigid-body orientations and on the choice of strain measures at each junction. We implicitly assume that sufficiently long MD trajectories are available so that the required equilibrium averages can be estimated well. The parameters found by our method can only be as good as the atomistic potentials employed in the MD simulations. However, recent evidence [11] suggests that these potentials can reliably capture sequence-dependent structural effects in double-helical DNA.

The presentation is structured as follows. In Sect. 2 we introduce the parameter sets $\mathbb{M}$ and $\mathbb{K}$ that completely define our rigid base-pair model of DNA, and in Sect. 3 we outline the equilibrium statistical mechanics of this model. In Sect. 4 we identify state functions whose equilibrium average yield the base-pair parameter sets, and in Sect. 5 we make explicit, for two choices of junction variables, the metric Jacobian factors that appear in the expressions to be averaged. In Sect. 6 we summarize our proposed method for the determination of the parameter sets $\mathbb{M}$ and $\mathbb{K}$, and we discuss various choices, assumptions, and consistency checks associated with the overall approach.

2 Rigid base-pair model

We consider a description of double-helical DNA in which each base pair is modeled as a rigid body. A chain or segment of $n+1$ base pairs corresponds to a list of rigid bodies indexed by $a = 0, \ldots, n$, where each body is defined by a vector $\mathbf{r}^{(a)}$ and an orthonormal frame $\{\mathbf{d}_i^{(a)}\}$ $(i = 1, 2, 3)$. The vector $\mathbf{r}^{(a)}$ describes the position of a body reference point, while the frame $\{\mathbf{d}_i^{(a)}\}$ is fixed in the body and describes its orientation relative to a frame $\{\mathbf{e}_i\}$ fixed in the lab. The configuration of a chain is thus completely defined by the $n+1$ coordinate vectors $r^a \in \mathbb{R}^3$ and the $n+1$ rotation matrices $Q^{(a)} \in \mathbb{R}^{3\times 3}$, where

$$r_i^{(a)} = \mathbf{r}^{(a)} \cdot \mathbf{e}_i \quad \text{and} \quad Q_{ik}^{(a)} = \mathbf{e}_i \cdot \mathbf{d}_k^{(a)}$$

for $i, k = 1, 2, 3$.

The potential energy for a chain is defined in terms of strain or junction variables $u^{(a)} \in \mathbb{A} \subset \mathbb{R}^3$ and $v^{(a)} \in \mathbb{R}^3$. The variables $u^{(a)}$ might be taken as tilt-roll-twist angles and $v^{(a)}$ as shift-slide-rise displacements as described in the Cambridge convention [12], which may be implemented in various ways [2, 13, 14]. However, other choices for the junction variables are possible, and certain choices may be more compatible with the assumption of a quadratic potential energy than others. Thus, to maintain some level of generality, we assume only that $u^{(a)}$ are independent coordinates, with domain $\mathbb{A}$, for the relative rotation matrices $\Lambda^{(a)}$ defined as

$$\Lambda^{(0)} = Q^{(0)},$$

$$\Lambda^{(a)} = \left[Q^{(a-1)}\right]^T Q^{(a)}, \quad (a = 1, \ldots, n)$$

and $v^{(a)}$ are relative displacements defined as

$$v^{(0)} = r^{(0)},$$

$$v^{(a)} = G(Q^{(a)}, Q^{(a-1)})[r^{(a)} - r^{(a-1)}], \quad (a = 1, \ldots, n)$$

where $G(Q^{(a)}, Q^{(a-1)}) \in \mathbb{R}^{3\times 3}$ is a specified function.

The relative rotation matrix $\Lambda^{(a)}$ describes the orientation of frame $\{\mathbf{d}_i^{(a)}\}$ in frame $\{\mathbf{d}_i^{(a-1)}\}$ according to

$$\mathbf{d}_j^{(a)} = \sum_{i=1}^{3} \Lambda_{ij}^{(a)} \mathbf{d}_i^{(a-1)}, \quad (j = 1, 2, 3)$$

and $G(Q^{(a)}, Q^{(a-1)})$ represents a general reference frame (possibly nonorthogonal) in which to measure the relative displacement vector $\mathbf{r}^{(a)} - \mathbf{r}^{(a-1)}$. The function G should have the property $G(QR, QT) = G(R, T)Q^T$ for all rotation matrices Q, R and T. This property ensures that the variables $v^{(1)}, \ldots, v^{(n)}$ are invariant under rigid displacements of the chain and thus depend only on the chain shape. The variables $u^{(1)}, \ldots, u^{(n)}$ automatically enjoy this property as a result of their definition.

We consider a chain potential energy, U, of the form

$$U(v,u) = \sum_{a=0}^{n} U_a(v^{(a)}, u^{(a)}) \ ,$$

where U_a is a general quadratic energy associated with the junction between bodies a and $a-1$. In particular, we have

$$U_a(v^{(a)}, u^{(a)}) = \frac{1}{2} \begin{Bmatrix} v^{(a)} - \widehat{v}^{(a)} \\ u^{(a)} - \widehat{u}^{(a)} \end{Bmatrix} \cdot \mathsf{K}^{(a)} \begin{Bmatrix} v^{(a)} - \widehat{v}^{(a)} \\ u^{(a)} - \widehat{u}^{(a)} \end{Bmatrix} \ ,$$

where $\mathsf{K}^{(a)} \in \mathbb{R}^{6\times 6}$ is a symmetric matrix of stiffness parameters, and $\widehat{v}^{(a)} \in \mathbb{R}^3$ and $\widehat{u}^{(a)} \in \mathbb{R}^3$ are parameters representing the equilibrium values of $v^{(a)}$ and $u^{(a)}$. Here and throughout we use the notation $v = (v^{(0)}, \ldots, v^{(n)})$, $u = (u^{(0)}, \ldots, u^{(n)})$, and so on. Whether the quadratic potential energy U is viewed as the true energy, or simply as an approximation of it, the parameters $\mathsf{K}^{(a)}$, $\widehat{v}^{(a)}$, and $\widehat{u}^{(a)}$ depend upon the specific choice of junction variables used to describe the relative positions and orientations of the rigid base pairs. The degree to which the quadratic assumption itself is valid also depends on this choice.

The kinetic energy of a chain depends on the inertia properties of each constituent rigid body. To each body we ascribe a total mass $m^{(a)}$, a symmetric rotational inertia tensor $\mathbf{\Gamma}^{(a)}$ with respect to the mass center, and a vector $\mathbf{c}^{(a)}$ that locates the mass center relative to the body reference point, so that $\mathbf{r}^{(a)} + \mathbf{c}^{(a)}$ is the position vector of the mass center. In terms of these quantities, we find that the total linear momentum, $\mathbf{p}^{(a)}$, and angular momentum, $\boldsymbol{\pi}^{(a)}$, about the body reference point are given by the vector relations

$$\mathbf{p}^{(a)} = m^{(a)}(\dot{\mathbf{r}}^{(a)} + \boldsymbol{\sigma}^{(a)} \times \mathbf{c}^{(a)}),$$
$$\boldsymbol{\pi}^{(a)} = \mathbf{c}^{(a)} \times \mathbf{p}^{(a)} + \mathbf{\Gamma}^{(a)}\boldsymbol{\sigma}^{(a)} \ ,$$

where $\boldsymbol{\sigma}^{(a)}$ is the angular velocity vector for body a. In terms of the components with respect to the frame $\{\mathbf{d}_i^{(a)}\}$, i.e., $\Gamma_{ij}^{(a)} = \mathbf{d}_i^{(a)} \cdot \mathbf{\Gamma}^{(a)}\mathbf{d}_j^{(a)}$, $c_i^{(a)} = \mathbf{c}^{(a)} \cdot \mathbf{d}_i^{(a)}$ and so on, the kinetic energy, Φ, of a chain becomes

$$\Phi(p,\pi) = \sum_{a=0}^{n} \Phi_a(p^{(a)}, \pi^{(a)}) \ ,$$

where Φ_a is the kinetic energy of body a. In particular, we have

$$\Phi_a(p^{(a)}, \pi^{(a)}) = \frac{1}{2} \begin{Bmatrix} p^{(a)} \\ \pi^{(a)} \end{Bmatrix} \cdot [\mathsf{M}^{(a)}]^{-1} \begin{Bmatrix} p^{(a)} \\ \pi^{(a)} \end{Bmatrix} \ ,$$

where, omitting the superscript for clarity, the generalized mass matrix $\mathsf{M} \in \mathbb{R}^{6\times 6}$ and its inverse are given in block form by

$$\mathsf{M} = \begin{pmatrix} mI & m[c\times]^T \\ m[c\times] & \Gamma + m[c\times][c\times]^T \end{pmatrix}$$
$$\mathsf{M}^{-1} = \begin{pmatrix} m^{-1}I + [c\times]\Gamma^{-1}[c\times]^T & [c\times]\Gamma^{-1} \\ \Gamma^{-1}[c\times]^T & \Gamma^{-1} \end{pmatrix} .$$

Here $[c\times] \in \mathbb{R}^{3\times 3}$ is the skew-symmetric matrix defined by the coordinate vector c as

$$[c\times] = \begin{pmatrix} 0 & -c_3 & c_2 \\ c_3 & 0 & -c_1 \\ -c_2 & c_1 & 0 \end{pmatrix}$$

and $I \in \mathbb{R}^{3\times 3}$ is the identity matrix.

A rigid base-pair model, based on a quadratic potential energy at each junction, is thus completely defined by specifying a kinetic energy parameter set $\mathbb{M} = \{m, c, \Gamma\}$ for each of the base pairs and a potential-energy parameter set $\mathbb{K} = \{\widehat{v}, \widehat{u}, \mathsf{K}\}$ for each of the junctions. There are ten independent parameters in the set $\mathbb{M}$ and 27 independent parameters in the set $\mathbb{K}$. The number of independent sets $\mathbb{M}$ and $\mathbb{K}$ depends on how sequence-dependence is modeled.

3 Equilibrium statistical mechanics

The equilibrium statistical properties of a rigid base-pair chain in contact with a heat bath are described by the standard canonical measure $\mathrm{d}\mu$ as in Eq. (1). This measure is a function of the absolute temperature $\Theta > 0$ of the heat bath, the inertia parameter sets $\mathbb{M}$ and the stacking and stiffness parameter sets $\mathbb{K}$. The explicit form of $\mathrm{d}\mu$, and the relative ease with which information can be extracted from it, depends on the choice of configuration and momentum variables used to define the mechanical state of the chain.

The classic form of $\mathrm{d}\mu$ is obtained when chain states are described in terms of canonical variables as defined in the theory of Hamiltonian systems. Standard canonical variables for a chain of $N = n+1$ rigid base-pairs are $(v,\zeta,u,\xi) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N} \times \mathbb{A}^N \times \mathbb{R}^{3N}$, where (v,u) may be any strain or junction variables and (ζ,ξ) are their associated canonical momenta. In particular, if the chain kinetic energy Φ is expressed in terms of (v,u) and their time derivatives $(\dot{v},\dot{u})$, then the canonical momenta are defined by $\zeta = \partial\Phi/\partial\dot{v}$ and $\xi = \partial\Phi/\partial\dot{u}$. These relations yield expressions for $(\dot{v},\dot{u})$ in terms of (v,ζ,u,ξ). In terms of these variables, the total mechanical energy or Hamiltonian function for the chain is

$$H(v,\zeta,u,\xi) = U(v,u) + \Phi(v,\zeta,u,\xi)$$

and the measure $\mathrm{d}\mu$ takes the usual form [16]

$$\mathrm{d}\mu = \frac{1}{Z} \mathrm{e}^{-H(v,\zeta,u,\xi)/k_\mathrm{B}\Theta} \, \mathrm{d}v \, \mathrm{d}\zeta \, \mathrm{d}u \, \mathrm{d}\xi \ , \qquad (1)$$

where k_B is the Boltzmann constant and $Z > 0$ is a normalizing constant. When (v,u) are chosen as junction variables, the form of the potential energy $U(v,u)$ is convenient, but then the kinetic energy $\Phi(v,\zeta,u,\xi)$ expressed in the associated canonical variables is configuration-dependent. This leads to a form of the measure $\mathrm{d}\mu$ that is inadequate for our purposes, because it provides little insight into the relation between moments of $\mathrm{d}\mu$ and the parameters in $\mathbb{M}$ and $\mathbb{K}$. 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 $\mathrm{d}\mu$ would not be simply related to the parameters in $\mathbb{M}$ and $\mathbb{K}$.

A more useful form for the measure $\mathrm{d}\mu$ can be obtained by changing from canonical variables (v,ζ,u,ξ) to the noncanonical variables (v,p,u,π). With this choice, the Hamiltonian takes the simple, separable form

$$H(v,p,u,\pi) = U(v,u) + \Phi(p,\pi) \ ,$$

while the measure $\mathrm{d}\mu$ becomes

$$\mathrm{d}\mu = \frac{1}{Z}\mathrm{e}^{-H(v,p,u,\pi)/k_{\mathrm{B}}\Theta}\prod_{a=0}^{n} J_a^u \,\mathrm{d}v\,\mathrm{d}p\,\mathrm{d}u\,\mathrm{d}\pi \ . \tag{2}$$

Here the additional terms $J_a^u = J(u^{(a)})$ are Jacobian factors associated with the change of variables. These factors are defined by

$$J_a^u = \begin{cases} \det S_0^u, & a = 0 \\ \det S_a^u / \det G_a^u, & a = 1,\ldots,n \end{cases} \ , \tag{3}$$

where $G_a^u, S_a^u \in \mathbb{R}^{3\times 3}$ are the matrix functions

$$\begin{aligned} G_a^u &= G(I, [\Lambda^{(a)}]^T), \\ [S_a^u]_{mj} &= \frac{1}{2}\sum_{i,k,l=1}^{3} \varepsilon_{imk}\Lambda_{li}^{(a)}\frac{\partial \Lambda_{lk}^{(a)}}{\partial u_j^{(a)}} \ , \end{aligned} \tag{4}$$

and $\varepsilon_{imk} = \mathbf{e}_i \cdot (\mathbf{e}_m \times \mathbf{e}_k)$ is the permutation symbol. The precise form of the Jacobians J_a^u depends on the choice of junction variables. Explicit examples for some standard choices are given later.

When expressed in the variables (v,p,u,π) the measure $\mathrm{d}\mu$ has the desirable feature that it is factorable into a configuration measure, $\mathrm{d}\mu_{\mathrm{con}}$, and a momentum measure, $\mathrm{d}\mu_{\mathrm{mom}}$:

$$\begin{aligned} \mathrm{d}\mu_{\mathrm{con}} &= \frac{1}{Z_{\mathrm{con}}}\mathrm{e}^{-U(v,u)/k_{\mathrm{B}}\Theta}\prod_{a=0}^{n} J_a^u \,\mathrm{d}v\,\mathrm{d}u, \\ \mathrm{d}\mu_{\mathrm{mom}} &= \frac{1}{Z_{\mathrm{mom}}}\mathrm{e}^{-\Phi(p,\pi)/k_{\mathrm{B}}\Theta}\,\mathrm{d}p\,\mathrm{d}\pi \ . \end{aligned}$$

Whether derived from Eq. (1) or Eq. (2), $\mathrm{d}\mu_{\mathrm{con}}$ necessarily involves Jacobian factors J_a^u because of the non-Cartesian nature of the coordinates u. While these factors are typically ignored [8], or assumed to be constant [10], we include them here.

Since U and Φ are the sum of local energies U_a and Φ_a, we find that $\mathrm{d}\mu_{\mathrm{con}}$ and $\mathrm{d}\mu_{\mathrm{mom}}$ can each be factored further into a product of entirely localized measures, namely,

$$\begin{aligned} \mathrm{d}\mu_{\mathrm{con}} &= \prod_{a=0}^{n}\frac{1}{Z_{\mathrm{con}}^{(a)}}\mathrm{e}^{-U_a/k_{\mathrm{B}}\Theta}J_a^u \,\mathrm{d}v^{(a)}\,\mathrm{d}u^{(a)}, \\ \mathrm{d}\mu_{\mathrm{mom}} &= \prod_{a=0}^{n}\frac{1}{Z_{\mathrm{mom}}^{(a)}}\mathrm{e}^{-\Phi_a/k_{\mathrm{B}}\Theta}\,\mathrm{d}p^{(a)}\,\mathrm{d}\pi^{(a)} \ . \end{aligned}$$

This localization is a special property of the class of variables (v,p,u,π) introduced in Sect. 2. For example, in addition to the specific separable form of the Hamiltonian, the Jacobian arising in the change of variables must also be factorable into localized terms $J_a^u = J(u^{(a)})$.

The statistical mechanical average of any chain state function ϕ with respect to the measure $\mathrm{d}\mu$ is given by

$$\langle\phi\rangle = \frac{\int \phi\,\mathrm{e}^{-H/k_{\mathrm{B}}\Theta}\prod_{a=0}^{n} J_a^u \,\mathrm{d}v\,\mathrm{d}p\,\mathrm{d}u\,\mathrm{d}\pi}{\int \mathrm{e}^{-H/k_{\mathrm{B}}\Theta}\prod_{a=0}^{n} J_a^u \,\mathrm{d}v\,\mathrm{d}p\,\mathrm{d}u\,\mathrm{d}\pi} \ ,$$

where all integrations are performed over $\mathbb{R}^{3N}\times\mathbb{R}^{3N}\times\mathbb{A}^N\times\mathbb{R}^{3N}$ unless mentioned otherwise. Due to the factorability of $\mathrm{d}\mu$ into localized measures we find

$$\langle\psi_a\rangle = \frac{\int_{\mathbf{R}^3\times\mathbf{R}^3}\psi_a\,\mathrm{e}^{-\Phi_a/k_{\mathrm{B}}\Theta}\,\mathrm{d}p^{(a)}\,\mathrm{d}\pi^{(a)}}{\int_{\mathbf{R}^3\times\mathbf{R}^3}\mathrm{e}^{-\Phi_a/k_{\mathrm{B}}\Theta}\,\mathrm{d}p^{(a)}\,\mathrm{d}\pi^{(a)}} \tag{5}$$

for any function $\psi_a = \psi(p^{(a)},\pi^{(a)})$, and

$$\frac{\langle\vartheta_a/J_a^u\rangle}{\langle 1/J_a^u\rangle} = \frac{\int_{\mathbf{R}^3\times\mathbb{A}}\vartheta_a\,\mathrm{e}^{-U_a/k_{\mathrm{B}}\Theta}\,\mathrm{d}v^{(a)}\,\mathrm{d}u^{(a)}}{\int_{\mathbf{R}^3\times\mathbb{A}}\mathrm{e}^{-U_a/k_{\mathrm{B}}\Theta}\,\mathrm{d}v^{(a)}\,\mathrm{d}u^{(a)}} \tag{6}$$

for any function $\vartheta_a = \vartheta(v^{(a)},u^{(a)})$. Notice that $\langle\psi_a\rangle$ is equivalent to a Boltzmann average defined by Φ_a over the local noncanonical momentum variables $(p^{(a)},\pi^{(a)})$. In contrast, it is the ratio $\langle\vartheta_a/J_a^u\rangle/\langle 1/J_a^u\rangle$, and not $\langle\vartheta_a\rangle$, that is equivalent to a Boltzmann average defined by U_a over the local junction variables $(v^{(a)},u^{(a)})$.

4 Extraction relations

Here we exploit the facts that in the variables (v,p,u,π) the equilibrium measure is factorable and that the kinetic and potential energies are of Gaussian form. We identify state functions ψ_a and ϑ_a that depend only on chain kinematics and whose equilibrium averages yield a complete set of base-pair model parameters.

4.1 Inertia parameters

A kinematic state function related to local inertia parameters may be defined as follows. Let $\boldsymbol{\sigma}^{(a)}$ denote the angular velocity of base-pair frame $\{\mathbf{d}_i^{(a)}\}$, namely,

$$\boldsymbol{\sigma}^{(a)} = \frac{1}{2}\sum_{i,j,k=1}^{3}\varepsilon_{ijk}[\dot{\mathbf{d}}_j^{(a)}\cdot\mathbf{d}_k^{(a)}]\,\mathbf{d}_i^{(a)} \ ,$$

let $\sigma^{(a)}\in\mathbb{R}^3$ denote the coordinate vector of components $\sigma_i^{(a)} = \boldsymbol{\sigma}^{(a)}\cdot\mathbf{d}_i^{(a)}$, and let $v^{(a)}\in\mathbb{R}^6$ be the state function defined by

$$v^{(a)} = \begin{Bmatrix} [Q^{(a)}]^T\dot{r}^{(a)} \\ \sigma^{(a)} \end{Bmatrix} \ .$$

The function $v^{(a)}$ can be recognized as the components, in the local frame $\{\mathbf{d}_i^{(a)}\}$, of the base-pair linear and angular velocities $\dot{\mathbf{r}}^{(a)}$ and $\boldsymbol{\sigma}^{(a)}$. These velocity components are connected to the local momentum variables and inertia parameters through the relation

$$v^{(a)} = [\mathsf{M}^{(a)}]^{-1}\begin{Bmatrix} p^{(a)} \\ \pi^{(a)} \end{Bmatrix} \ .$$

Using this function in Eq. (5) and carrying out the indicated integrations [17] we obtain

$$\langle v^{(a)}\otimes v^{(a)}\rangle = k_{\mathrm{B}}\Theta[\mathsf{M}^{(a)}]^{-1}, \quad (0\le a\le n) \ . \tag{7}$$

Here we use the notation $v^{(a)}\otimes v^{(a)}$ to denote the 6×6 matrix with components $[v^{(a)}\otimes v^{(a)}]_{\mu\lambda} = v_\mu^{(a)}v_\lambda^{(a)}$, where $\mu,\lambda = 1,\ldots,6$.

4.2 Stacking and stiffness parameters

A kinematic state function related to local stacking and stiffness parameters is provided by $w^{(a)} = (v^{(a)}, u^{(a)}) \in \mathbb{R}^6$. If we suppose that the coordinate domain $\mathbb{A}$ for the variables $u^{(a)}$ is all of $\mathbb{R}^3$, then Eq. (6) yields the exact expressions

$$\frac{\langle w^{(a)}/J_a^u \rangle}{\langle 1/J_a^u \rangle} = \widehat{w}^{(a)}, \quad (1 \le a \le n) \tag{8}$$

and

$$\frac{\langle \Delta w^{(a)} \otimes \Delta w^{(a)}/J_a^u \rangle}{\langle 1/J_a^u \rangle} = k_{\mathrm{B}}\Theta[\mathsf{K}^{(a)}]^{-1}, \quad (1 \le a \le n) \ , \tag{9}$$

where $\Delta w^{(a)} = w^{(a)} - \widehat{w}^{(a)}$. By expanding the left-hand side of Eq. (9) and using Eq. (8) we deduce

$$\frac{\langle w^{(a)} \otimes w^{(a)}/J_a^u \rangle}{\langle 1/J_a^u \rangle} = k_{\mathrm{B}}\Theta[\mathsf{K}^{(a)}]^{-1} + \widehat{w}^{(a)} \otimes \widehat{w}^{(a)} \ , \tag{10}$$

which may be more convenient than Eq. (9) since the average $\langle w^{(a)} \otimes w^{(a)}/J_a^u \rangle$ is independent of the parameters $\widehat{w}^{(a)}$.

5 Junction variables and Jacobians

The Jacobian factors J_a^u appearing in the statistical mechanical averages depend upon the specific variables $u^{(a)}$ used to describe a relative rotation matrix $\Lambda^{(a)}$ and on the choice of the function G used to define the relative displacement variables $v^{(a)}$. There are many independent options for both, and these lead to nontrivial Jacobians in general.

5.1 Cambridge convention variables

The shift-slide-rise and tilt-roll-twist variables described in the Cambridge convention [12] can be interpreted and implemented in a variety of ways [2, 13, 14]. Each different implementation will generally lead to a different Jacobian factor. To illustrate the necessary computations, we suppose that the shift-slide-rise displacements $v^{(a)} \in \mathbb{R}^3$ are the components in the local frame $\{\mathbf{d}_i^{(a-1)}\}$ of the relative displacement vector $\mathbf{r}^{(a)} - \mathbf{r}^{(a-1)}$, namely,

$$v^{(a)} = [Q^{(a-1)}]^T[r^{(a)} - r^{(a-1)}] \ .$$

Here v_1 is the shift, v_2 is the slide and v_3 is the rise. From this definition we deduce $G(Q^{(a)}, Q^{(a-1)}) = [Q^{(a-1)}]^T$, thus

$$G_a^u = G(I, [\Lambda^{(a)}]^T) = \Lambda^{(a)} \quad \text{and} \quad \det G_a^u = 1$$

since $\Lambda^{(a)}$ is a rotation matrix. Furthermore, we suppose that the tilt-roll-twist parameterization of the relative rotation matrix $\Lambda^{(a)}$ is defined by

$$\mathbb{A} = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \times (-\pi, \pi) \times (-\pi, \pi) \ ,$$

$$\Lambda^{(a)} = \begin{pmatrix} c_2c_3 - s_1s_2s_3 & -c_1s_3 & s_2c_3 + s_1c_2s_3 \\ c_2s_3 + s_1s_2c_3 & c_1c_3 & s_2s_3 - s_1c_2c_3 \\ -c_1s_2 & s_1 & c_1c_2 \end{pmatrix} \ ,$$

where $c_i = \cos(u_i^{(a)})$, $s_i = \sin(u_i^{(a)})$. Here u_1 is the tilt, u_2 is the roll, and u_3 is the twist. From Eq. (4) we deduce

$$S_a^u = \begin{pmatrix} c_2 & 0 & -c_1s_2 \\ 0 & 1 & s_1 \\ s_2 & 0 & c_1c_2 \end{pmatrix} \ ,$$

which implies

$$\det S_a^u = \cos(u_1^{(a)}) \ .$$

Insertion of the expressions for $\det G_a^u$ and $\det S_a^u$ into Eq. (3) leads to the Jacobian factors

$$J_a^u = \cos(u_1^{(a)}), \quad a = 1, \ldots, n \ .$$

5.2 Discretized continuum variables

Another choice of junction variables is motivated by the continuum theory of elastic rods [15]. Here, relative displacements between base-pairs are naturally measured by

$$v^{(a)} = \tfrac{1}{2}[Q^{(a)} + Q^{(a-1)}]^T[r^{(a)} - r^{(a-1)}] \ ,$$

from which we deduce

$$G(Q^{(a)}, Q^{(a-1)}) = \tfrac{1}{2}[Q^{(a)} + Q^{(a-1)}]^T$$

and

$$G_a^u = G(I, [\Lambda^{(a)}]^T) = \tfrac{1}{2}(I + \Lambda^{(a)}) \ .$$

The parameterization for the relative rotation matrix $\Lambda^{(a)}$ is defined by

$$\mathbb{A} = \mathbb{R}^3 \ ,$$

$$\Lambda^{(a)} = (I + \tfrac{1}{2}[u\times]^{(a)})(I - \tfrac{1}{2}[u\times]^{(a)})^{-1} \ ,$$

from which we deduce

$$S_a^u = (I - \tfrac{1}{2}[u\times]^{(a)}) \Big/ \left[1 + \tfrac{1}{4}|u^{(a)}|^2\right] \ .$$

The determinants of G_a^u and S_a^u are

$$\det G_a^u = \left[1 + \tfrac{1}{4}|u^{(a)}|^2\right]^{-1} \ ,$$

$$\det S_a^u = \left[1 + \tfrac{1}{4}|u^{(a)}|^2\right]^{-2} \ ,$$

which leads to the Jacobian factors

$$J_a^u = [1 + \tfrac{1}{4}|u^{(a)}|^2]^{-1}, \quad a = 1, \ldots, n \ .$$

This choice of junction variables has a straightforward physical interpretation. The variables $v^{(a)}$ are the components of the relative displacement vector in the average, nonorthogonal frame defined by $\frac{1}{2}[Q^{(a)} + Q^{(a-1)}]$. The variables $u^{(a)}$ define the rotation axis and the total rotation angle of $\Lambda^{(a)}$ in the local frame $\{\mathbf{d}_i^{(a-1)}\}$. In particular, the rotation axis is parallel to the vector with components $u_i^{(a)}$, and the total rotation angle $\theta^{(a)} \in [0, \pi)$ about this (oriented) axis satisfies $\cos\theta^{(a)} = [1 - \frac{1}{4}|u^{(a)}|^2]/[1 + \frac{1}{4}|u^{(a)}|^2]$. The variables $u^{(a)}$ may be determined from the relative rotation matrix $\Lambda^{(a)}$ according to the formula

$$u_i^{(a)} = \frac{2}{1+\mathrm{tr}[\Lambda^{(a)}]} \sum_{j,k=1}^{3} \varepsilon_{jik} \Lambda_{jk}^{(a)} \quad ,$$

where $\mathrm{tr}[\Lambda^{(a)}] = \sum_{i=1}^{3} \Lambda_{ii}^{(a)}$ and ε_{jik} is the permutation symbol introduced in Sect. 3.

These discretized continuum variables possess a simple symmetry property with respect to complementary strands. To see this, first notice that base pairs may be labeled by $a = 0, \ldots, n$ along one strand of double-helical DNA in the 5′-to-3′ direction or by $b = 0, \ldots, n$ along the complementary strand in the 5′-to-3′ direction. If $\{\mathbf{r}, \mathbf{d}_i\}^{(a-1)}$ and $\{\mathbf{r}, \mathbf{d}_i\}^{(a)}$ are the local frames for a dimer step along the first strand and $\{\mathbf{r}, \mathbf{d}_i\}^{(b-1)}$ and $\{\mathbf{r}, \mathbf{d}_i\}^{(b)}$ are the frames for this step along the complementary strand, then $\mathbf{r}^{(b)} = \mathbf{r}^{(a-1)}$, $\mathbf{d}_1^{(b)} = \mathbf{d}_1^{(a-1)}$, and $\mathbf{d}_i^{(b)} = -\mathbf{d}_i^{(a-1)}$ for $i = 2, 3$ in accordance with the Cambridge convention. Similar relations hold for the local frames with the complementary labels $b-1$ and a. From the definition of the discretized continuum variables we deduce

$$(v_1, v_2, v_3)^{(b)} = (-v_1, v_2, v_3)^{(a)},$$
$$(u_1, u_2, u_3)^{(b)} = (-u_1, u_2, u_3)^{(a)} \quad .$$

Thus, variables computed along one strand and those computed along its complement differ only by a sign in the first component. While there is always a functional relationship between junction variables computed along complementary strands, this relationship is particularly simple for these discretized continuum variables. (For a discussion of symmetry conditions in other variables see Refs. [2, 13], and references therein.)

6 Proposed extraction method

Our proposed method to extract rigid base-pair parameters from MD simulation data involves six main considerations:

1. Identification of rigid base pairs. A rule is needed for assigning a reference point $r^{(a)}$ and a frame $Q^{(a)}$ to each base pair in the atomistic MD model. There are many possible ways to construct such a rule; one standard example is provided in the program Curves [18, 19]. Such a rule is a necessary, but independent, part of the extraction method. The choice of rule should be consistent with the assumption that each base pair is a kinematically independent rigid body as outlined in Sect. 2. For example, the rule should be based only on individual base pairs, and not on any averages with its nearest neighbors.

2. Choice of junction variables. A choice must be made regarding the local junction variables $w^{(a)} = (v^{(a)}, u^{(a)})$ that measure the relative displacement and rotation between adjacent base pairs. As discussed in Sect. 2, there are many choices for these variables and certain choices may well be more compatible than others with the assumption of a quadratic potential energy in those variables. Each different choice gives rise to a different set of Jacobian factors, J_a^u, as defined in Sect. 3. These factors appear in the extraction relations presented in Sect. 4, and are explicitly computed in Sect. 5 for two different choices of junction variables.

3. Estimation and testing of equilibrium averages. The central step of the extraction method is the estimation of the equilibrium averages

$$\begin{aligned} &\langle 1/J_a^u \rangle, \quad \langle w^{(a)}/J_a^u \rangle, \\ &\langle w^{(a)} \otimes w^{(a)}/J_a^u \rangle, \quad \langle v^{(a)} \otimes v^{(a)} \rangle \quad , \end{aligned} \tag{11}$$

from a time series of rigid-body configurations extracted from an MD simulation. Since the arguments of these averages are all functions of the chain variables r, Q, $\dot{r}$, and $\dot{Q}$, the averages can indeed be estimated from time series data on r and Q. Notice that $\dot{r}$ and $\dot{Q}$ themselves can be estimated from the data via an appropriate finite-difference approximation in time. The practical success or failure of our extraction method will be dependent on whether sufficiently long MD trajectories are available so that the required equilibrium averages can be estimated well. Of course, statistical tests should be applied whenever possible to determine whether or not time averages have reached equilibrium values.

4. Extraction of parameters. Given numerical estimates for the equilibrium averages Eq. (11), the rigid base-pair model parameters may be determined from Eqs. (7), (8), and (10). Since the form of $[\mathsf{M}^{(a)}]^{-1}$ is known explicitly, the inertia parameters $m^{(a)}$, $c^{(a)}$, and $\Gamma^{(a)}$ can be determined directly from the estimate of $\langle v^{(a)} \otimes v^{(a)} \rangle$. Similarly, the stacking parameters $\widehat{v}^{(a)}$ and $\widehat{u}^{(a)}$ can be determined directly from the estimates of $\langle 1/J_a^u \rangle$ and $\langle w^{(a)}/J_a^u \rangle$. The determination of the stiffness parameters $\mathsf{K}^{(a)}$ from the estimate of $\langle w^{(a)} \otimes w^{(a)}/J_a^u \rangle$ requires a matrix inversion.

5. Consistency checks on model. Consistency checks are available to assess the validity of a rigid base-pair model with a quadratic potential energy. The degree to which each base-pair in the MD model behaves as a rigid body can be assessed from the moments of the variable $v^{(a)}$. All moments of this variable should be consistent with equilibrium averages computed from Eq. (5). For example, all first- and third-order moments should vanish, and the matrix of second-order moments, which is proportional to $[\mathsf{M}^{(a)}]^{-1}$, should have only ten independent elements. Assumptions on the potential energy can also be checked. For example, third- and higher-order moments of the variable $\Delta w^{(a)}$, scaled by J_a^u as in Eq. (6), provide a check on the validity of the assumed quadratic dependence of the potential energy in the chosen coordinates $w^{(a)}$.

6. Interpretation of results. For an MD simulation of a given DNA oligomer consisting of $n+1$ base pairs, our extraction method will yield n sets of stacking and stiffness parameters $\mathbb{K}$, one for each base-pair junction, and $n+1$ sets of inertia parameters $\mathbb{M}$, one for each base-pair unit. All these sets will likely have different numerical values regardless of the oligomer sequence composition. The extra steps of data processing neces-

sary to reduce an arbitrary number of parameter sets to a fixed number or table of sequence-dependent sets are not addressed in this article. Nevertheless, we remark that the degree to which the parameter sets $\mathbb{K}$ and $\mathbb{M}$ depend on sequence is a modeling assumption. For example, it is reasonable to assume that the parameters $\mathbb{K}$ for a particular base-pair junction depend on the surrounding dimer (two nearest neighbors) or the surrounding tetramer (four nearest neighbors), and so on. Similarly, it is reasonable to assume that the parameters $\mathbb{M}$ for a particular base-pair unit depend on the monomer itself or on the trimer of nearest neighbors (i.e., the unit itself and two nearest neighbors), and so on. For a given level of sequence-dependence, the sets $\mathbb{K}$ and $\mathbb{M}$ should depend on the local sequence composition and not the global location within an oligomer. This observation could be exploited to derive consistency checks on assumed models of sequence dependence.

Acknowledgements. It is a pleasure to thank R. Lavery, other members of the IBPC, Paris, and members of the Maddocks group for insightful discussions leading to this work. O. G was partially supported by the US National Science Foundation.

References

1. Schlick T (1995) Curr Opin Struct Biol 5: 245–262
2. Olson WK (1996) Curr Opin Struct Biol 6: 242–256
3. Manning RS, Maddocks JH, Kahn JD (1996) J Chem Phys 105: 5626–5646
4. Packer MJ, Dauncey MP, Hunter CA (2000) J Mol Biol 295: 71–83
5. Packer MJ, Dauncey MP, Hunter CA (2000) J Mol Biol 295: 85–103
6. De Santis P, Palleschi A, Savino M, Scipioni A (1990) Biochemistry 29: 9269–9273
7. Bolshoy A, McNamara P, Harrington RE, Trifonov EN (1991) Proc Natl Acad Sci USA 88: 2312–2316
8. Olson WK, Gorin AG, Lu X-J, Hock IM, Zhurkin VB (1998) Proc Natl Acad Sci USA 95: 11163–11168
9. Munteanu MG, Vlahovičcek K, Parthasarathy S, Simon I, Pongor S (1998) TIBS 23: 341–347
10. Karplus M, Kushick JN (1981) Macromolecules 14: 325–332
11. Beveridge DL, McConnell KJ (2000) Curr Opin Struct Biol 10: 182–196
12. Dickerson RE (1989) J Biomol Struct Dyn 6: 627–634
13. Lu X-J, Olson WK (1999) J Mol Biol 285: 1563–1575
14. Lavery R, Zakrzewska K (1999) In: Neidle S (ed) Oxford handbook of nucleic acid structure. Oxford University Press, New York, pp 39–76
15. Dichmann DJ, Li Y, Maddocks JH (1996) In: Mesirov JP, Schulten K, Sumners D (eds) Mathematical approaches to biomolecular structure and dynamics. Springer, Berlin Heidelberg, New York, pp 71–113
16. Huang K (1987) Statistical mechanics, 2nd edn. Wiley, New York
17. Hogg RV, Craig AT (1970) Introduction to mathematical statistics, 3rd edn. Macmillan, New York
18. Lavery R, Sklenar H (1988) J Biomol Struct Dyn 6: 63–91
19. Lavery R, Sklenar H (1989) J Biomol Struct Dyn 6: 655–667