From fine-grain to coarse-grain modeling: Estimating kinetic parameters of DNA molecules

Jeremy Curuksu[1,2]

[1]Amazon.com LLC, 7 W34th Street, New York, 10001, NY, USA.
[2]Center for Data Science, New York University, 60 5th Ave, New York, 10011, NY, USA.

Contributing authors: curukj@amazon.com, jeremy.cur@nyu.edu;

Abstract

Coarse-grain models are essential to understand the biological function of DNA molecules because the length and time scales of the sequence-dependent physical properties of DNA are often beyond the reach of all-atom experimental and computational methods. Simulating coarse-grain models of DNA e.g., using Langevin dynamics, requires the parametrization of both a potential and kinetic energy functions. Many studies have shown that the flexibility (i.e., potential energy) of DNA molecules depends on its sequence. In contrast, little is known about the sequence-dependence of DNA mass parameters required to model its kinetic energy. In this paper, an algebraic expression is derived for the kinetic energy as a function of linear and angular velocities of each DNA base parameterized by its mass, center of mass, and rotational inertia tensor. The parameters of this function are then approximated from a set of fine-grain molecular dynamics simulations representing all combinations of the four DNA base pairs AT, TA, GC, and CG, in different sequence contexts. Compatibility conditions associated with the assumption of each base being modeled as a rigid body were verified to be good approximations. The kinetic parameters were found to be significantly different between the four G, C, A, and T bases, but to not be dependent on the sequence context. This suggests that the effective kinetic parameters of a DNA base may depend only on the base itself, not on its neighbors.

Keywords: DNA Models, Kinetic Energy, Molecular Simulations

1 Introduction

The mechanical and dynamical properties of DNA mediate important biological functions on length scales ranging from tens to several hundreds of base pairs (bp), such as protein/DNA interaction specificity [1], transcription regulation [2], and nucleosomes organization [3]. Modeling macromolecular complexes along a DNA sequence containing hundreds of base pairs using an all-atom description including the surrounding solvent is computationally too expensive. Simpler, more *coarse-grained* models are better suited in this context. But the parameterization of coarse-grained models is challenged by the lack of detailed experimental data.

Several methods have been developed to derive sequence-dependent, coarse-grained models of DNA molecules using sparse experimental data, or using more abundant (but more approximate) computer simulation data, of the atomic structure of DNA in its physiological conditions [4]. All-atom Molecular Dynamics (MD) simulations have successfully been used to parameterize the potential energy function of DNA elastic rod models [5, 6]. In particular, it has been shown that the two bases inside a DNA base pair undergo substantial fluctuations relative to each other and that significant improvement in modeling DNA as an elastic rod is obtained when using rigid bases as the building blocks instead of rigid base pairs [7].

The parameterization of DNA models using MD simulations offers the possibility to exploit the detailed structural representation of DNA molecules in solution, and to probe its dynamics i.e., its sequence dependent curvature and flexibility [8–10], which is not readily available from experiments. The MD method should also be well suited to the determination of kinetic energy parameters because these require dynamic information and sampling over both the configurational and momenta phase space, which is exactly what MD attempts to simulate.

In a rigid base DNA model [5, 7, 11], the kinetic energy function is parameterized by a total mass matrix which depends on an effective total mass and rotational inertia matrix for each chemical base as well as the position of its center of mass. These parameters can be estimated specifically for each of the four DNA bases G, C, A, and T, by matching the rigid base model to moments measured during molecular dynamics simulations including explicit water molecules and salt conditions representative of DNA *in vivo*. For example, this has been done successfully for the purine (A and G) and pyrimidine (T and C) bases, by simulating an homogeneous DNA fragment made of 16 alternating A and T bases within a box of explicit water molecules [5]. But no systematic estimation of DNA kinetic parameters for each base has been reported to date, which would require assessing the influence of all possible sequence contexts and of the simulation protocol.

In this paper, I derive the kinetic parameters for the four bases G, C, A, and T, in different DNA sequences simulated by molecular dynamics and show that the sequence context is negligeable. The representative set of six DNA sequences includes the four base pair steps CG, GC, TA, and AT, themselves flancked by different representative sequence contexts including homogeneous alternating pyrimidine (Py) / purine (Pu) and tracts of consecutive Pu/Pu/Pu and Py/Py/Py sequences. These different sequence contexts are known to induce very different conformational properties (as captured by the *potential energy* in a coarse-grained model) in term of bending and torsion of the double helix [11, 12]. The six sequences were also chosen to be representive of the systematic microsecond molecular dynamics investigations on tetranucleotide sequence effects in B-DNA [13]. In [13], it was shown that the statistical distribution of DNA conformations observed in the experimental Protein Data Bank (PDB, 14) is well approximated by a dataset of MD simulations including the six DNA sequences chosen here. For completeness, I also assess the effect of different solvent models, force fields, and MD protocols, and still I did not find significant changes in the kinetic parameter estimates. This paper suggests that the effective kinetic parameters of a DNA base may depend only on the base itself, not on its neighbors.

2 Methods

2.1 Mass matrix and finite difference approximation

The rigid base model for the three-dimensional, sequence dependent structure of DNA is specified by its kinetic and internal energy functions. A complete description of the model can be found in [5]. Each base X_a on the reference strand and on the complementary strand is considered rigid and represented by a reference point $\boldsymbol{r}^a$ and a right handed orthonormal frame $\{\boldsymbol{d}_1^a, \boldsymbol{d}_2^a, \boldsymbol{d}_3^a\}$ defined according to the Tsukuba convention [15]. $\boldsymbol{r}^a$ accounts for the position and $\{\boldsymbol{d}_1^a, \boldsymbol{d}_2^a, \boldsymbol{d}_3^a\}$ the orientation of the base X_a located at position a in the double helix ($a = 1, 2, ..., 2n$ where n is the number of base pairs). These are defined with respect to an arbitrary lab-fixed frame $\{e_i\}$ in term of component vectors r_i^a and rotation matrices $D_{ij}^a \in \mathbb{R}^{3\times 3}$ by:

$$r_i^a = e_i \cdot r^a, \quad D_{ij}^a = e_i \cdot d_j^a \tag{1}$$

The linear velocity v^a and angular velocity ω^a of each base X_a are defined as follows:

$$v^a = (D^a)^T \dot{r}^a \tag{2}$$

$$\omega^a = \mathrm{vec}[(D^a)^T \dot{D}^a] \tag{3}$$

where v^a is computed with respect to the reference point $\boldsymbol{r}^a$ that defines the position of X_a. The operation vec is defined for any matrix A as vec $(A) = (A_{32}, A_{13}, A_{21})$. When approximated using a discretized time step, v^a and ω^a can be defined as follows:

$$[v^a]^{(k)} = \left([D^a]^{(k)}\right)^T \frac{[r^a]^{(k+1)} - [r^a]^{(k-1)}}{t^{(k+1)} - t^{(k-1)}} \tag{4}$$

$$[\omega^a]^{(k)} = \mathrm{vec}\left[\mathrm{skew}\left(\left([D^a]^{(k)}\right)^T \frac{[D^a]^{(k+1)} - [D^a]^{(k-1)}}{t^{(k+1)} - t^{(k-1)}}\right)\right] \tag{5}$$

where $k = 1, 2, ..., N$ is the time step assuming a trajectory with a finite number of N time steps (e.g., obtained from a molecular computer simulation), and t^k is a measurement of the *actual* time associated with each recorded time frame k. The operation ”skew” is defined for any matrix A as $\text{skew}(A) = (AA^T)/2$. A skew symmetric projection is explicitly computed because the matrix $(D^a)^T \dot{D}^a$ is skew-symmetric in theory but its finite-difference approximation obtained from molecular computer simulations may not be [5].

The kinetic energy of each base X_a can be defined as:

$$\Phi^a = \tfrac{1}{2} m^a |v^a + \omega^a \times c^a|^2 + \tfrac{1}{2} \omega^a \cdot \Gamma^a \omega^a \tag{6}$$

where m^a is the total mass of the base, Γ^a is the symmetric rotational inertia tensor with respect to the mass center and c^a locates the mass center relative to $\boldsymbol{r}^a$. The first term in the right-hand-side of Eq.(6) represents the energy associated with the linear momentum of the center of mass of base X_a and the second term the energy associated with its angular momentum.

The total kinetic energy $\Phi(\mathbf{v})$ of a DNA molecule with n bp is obtained by summing over each base X_a on the two strands. Thus, it can be written algebraically as:

$$\Phi(\mathbf{v}) = \tfrac{1}{2}\, \mathbf{v} \cdot \mathbf{M}\, \mathbf{v} \tag{7}$$

where $\mathbf{v} = (v^1, \omega^1, v^2, \omega^2, ..., v^{2n}, \omega^{2n}) \in \mathbb{R}^{12n}$ is the vector of all momentum components and $\mathbf{M} \in \mathbb{R}^{12n \times 12n}$ is the total mass matrix which takes the following block diagonal expression:

$$\mathbf{M} = \begin{bmatrix} M^1 & 0_6 & \cdots & 0_6 \\ 0_6 & M^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_6 \\ 0_6 & \cdots & 0_6 & M^{2n} \end{bmatrix} \tag{8}$$

$\mathbf{M}$ is obtained by assembling the individual sub-blocks M^a associated with each base X_a. To be consistent with Eq.(6), each sub-block M^a is defined as:

$$M^a = \begin{bmatrix} m^a I & m^a [c^a \times]^T \\ m^a [c^a \times] & \Gamma^a + m^a [c^a \times][c^a \times]^T \end{bmatrix} \tag{9}$$

where $I \in \mathbb{R}^{3\times 3}$ is an identity matrix and $[c^a \times] \in \mathbb{R}^{3\times 3}$ denotes the skew-symmetric matrix:

$$[c^a \times] = \begin{bmatrix} 0 & -c^a_3 & c^a_2 \\ c^a_3 & 0 & -c^a_1 \\ -c^a_2 & c^a_1 & 0 \end{bmatrix} \tag{10}$$

To simplify notations in the rest of this development (see Eq.14), let us write each sub-block M^a as:

$$M^a = \begin{bmatrix} B_1^a & [B_2^a]^T \\ B_2^a & B_3^a \end{bmatrix} \tag{11}$$

To estimate the mass matrix $\mathbf{M}$ from trajectories produced by molecular dynamics simulations, we need to sample long enough so that we can assume ergodicity and replace the statistical mechanical averages with averages over the MD time frames (excluding DNA conformational frames containing one or more broken hydrogen bond(s) as explained in the next section). At thermal equilibrium, the average kinetic energy for each degree of freedom is $\frac{1}{2}k_BT$ due to equipartition of energy [16]. Distributed over the linear and angular momenta $\{v_x, v_y, v_z\}$ and $\{\omega_x, \omega_y, \omega_z\}$ of the $2n$ rigid bases of a coarse-grained DNA model, the following relations hold true:

$$\begin{aligned} \Phi(\mathbf{v}) &= \tfrac{1}{2}\, k_BT\,\mathbf{I} \\ \Leftrightarrow\ \tfrac{1}{2}\,\mathbf{M}\langle \mathbf{v}\otimes\mathbf{v}\rangle &= \tfrac{1}{2}\, k_BT\,\mathbf{I} \\ \Leftrightarrow\quad \langle \mathbf{v}\otimes\mathbf{v}\rangle &= \mathbf{M}^{-1}k_BT \\ \Leftrightarrow\quad \mathbf{M}^{-1} &= \frac{1}{k_BT}\langle \mathbf{v}\otimes\mathbf{v}\rangle \end{aligned} \tag{12}$$

where $\mathbf{I} \in \mathbb{R}^{12n\times 12n}$ is an identity matrix and $\otimes$ denotes the outer product. Eq.(12) corresponds to what is obtained when computing explicitly the statistical mechanical average expression for $\mathbf{v}\otimes\mathbf{v}$ [5].

Since MD trajectories consist of configuration variables at discrete times, a MD estimate of the inverse mass matrix M^{-1}, according to Eq.(12), is given by:

$$M^{-1}_{estimate} = \frac{1}{N \times k_BT}\sum_{k=1}^{N} \mathbf{v}^{(k)} \otimes \mathbf{v}^{(k)} \tag{13}$$

Given the above estimate of the mass matrix (after inversion), the kinetic parameter estimates for m^a, c^a and Γ^a can now be obtained using Eq.(9) and Eq.(11):

$$\begin{aligned} m^a &= \tfrac{1}{3}tr[B_1^a], \\ c^a &= \frac{1}{m^a}\,\mathrm{vec}(\mathrm{skew}(B_2^a)), \\ \Gamma^a &= B_3^a - m^a[c^a\times][c^a\times]^T \end{aligned} \tag{14}$$

The above development assumes that each base in a DNA molecule can be effectively modeled as a rigid body. This assumption leads to certain compatibility conditions apparent from Eq.(8) to Eq.(11) such as the matrix symmetry $M^T = M$ in Eq.(8), symmetries in the block-diagonal structure of M and in the fine structure of the diagonal blocks M^a defined in Eq.(9), and matrix positivity (positive eigenvalues). These conditions will be assessed and validated in the Results section to confirm that a rigid base model is a reasonable approximation when parameterizing the kinetic energy of a DNA molecule.

2.2 Molecular dynamics computer simulations

The kinetic parameters for the rigid base DNA model were derived from the all-atom, explicit solvent MD simulation of six representative 18-bp DNA sequences [13]:

s1. 5' d(GCTATATATATATATAGC) 3'

s2. 5' d(GCCGCGCGCGCGCGCGGC) 3'

s3. 5' d(GCGATCGATCGATCGAGC) 3'

s4. 5' d(GCCTAGCTAGCTAGCTGC) 3'

s5. 5' d(GCGCGGGCGGGCGGGCGC) 3'

s6. 5' d(GCATAAATAAATAAATGC) 3'

Only the central 10 bp (index number 5 to 14) were considered during the analysis of kinetic parameters to minimize errors due to end effect (i.e. larger fluctuation due to larger surface exposed to the solvent). This set of sequences contains the four DNA bases G, C, A, and T, in equal amount. The bp steps GC, CG, AT, and TA, are also evenly represented. And each of these bp steps is found in different, representative [13] sequence contexts of either alternating purines and pyrimidines, tracts of consecutive purines, or tracts of consecutive pyrimidines.

Many experimental and computational studies have indicated that G$_n$-tracts and A$_n$-tracts ($n \geq 3$) form peculiar double helical structures with locally increased rigidity [12, 17, 18]. In contrast, alternating sequences are frequently observed at positions in the genome where the transient melting or bending of DNA is needed to interact with proteins or initiate transcription/replication. For example, the TATA box motif leverages both rigid A$_n$-tracts and flexible alternating A/T flanking sequences, which has been extensively studied by simulation [9, 12] and experiment [4, 19].

The AMBER suite of programs [20] together with the parm-bsc0 force field [21] was used. For each of the six sequences, a DNA double helix was built using the fiber diffraction B-DNA coordinates avaliable in the NUCGEN module of AMBER. The LEAP module was used to add hydrogen atoms and 33 K+ neutralizing cations. The DNA molecule was then solvated with approximately 11,600 SPC-E water molecules [22] within a truncated octahedral box corresponding to a solvent layer > 10Å, and with 36 K+ and 36 Cl- ions corresponding to a physiological concentration of 150 mM. The effect of different solvent models, force fields, and MD protocols, was evaluated, by carriying out replicate simulations with an older AMBER force field called parm-94 [23], with the TIP3P water molecules [24], and with only K+ counterions (no extra K+ and Cl-). These replicate were carried out to evaluate how these conditions may affects the current estimation of DNA kinetic parameters.

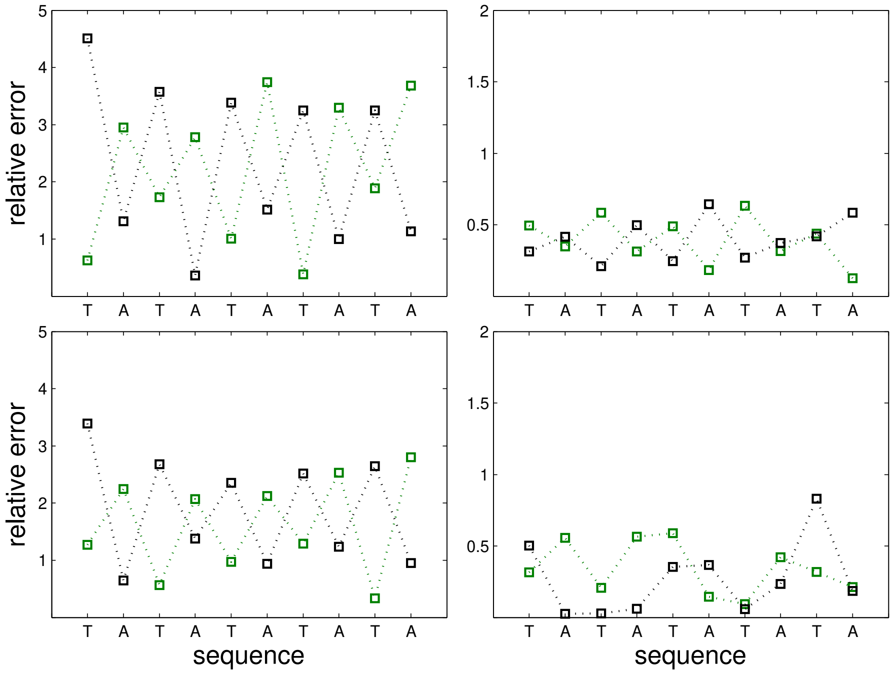

Fig. 1 Relative error (in %) in the diagonal sub-blocks B_1^a and B_2^a of the estimated mass matrix for each base along sequence s1. Upper-Left: Norm of the average difference between diagonal entries in B_1^a (for details see list I below) with respect to the norm of B_1^a. Upper-Right: Norm of the average off-diagonal entries in B_1^a (see II) with respect to the norm of B_1^a. Lower-Left: Norm of the average sum of the skew symmetric off-diagonal entries in B_2^a (see III) with respect to the norm of B_2^a. Lower-Right: Norm of the average diagonal entries in B_2^a (see IV) with respect to the norm of B_2^a.

I: $B_{1,11}^a - B_{1,22}^a$, $B_{1,22}^a - B_{1,33}^a$

II: $B_{1,12}^a$, $B_{1,13}^a$, $B_{1,23}^a$

III: $B_{2,12}^a + B_{2,21}^a$, $B_{2,13}^a + B_{2,31}^a$, $B_{2,23}^a + B_{2,32}^a$

IV: $B_{2,11}^a$, $B_{2,22}^a$, $B_{3,33}^a$

The starting state containing the DNA molecule 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. Only this final 1 ns was used to derive the DNA kinetic parameters and assumed to be at equilibrium where Eq.(12) holds. Simulations were performed at constant temperature (300K) and pressure (1 atm) applying periodic boundary conditions and the Particle-Mesh Ewald approach [25] with a 9-Å direct space sum cutoff.

The integration time step was 2 fs, using SHAKE [26] to freeze the vibrations of hydrogen bonds. The trajectory produced for each of the six sequences was 1 ns long and sampled every 2 fs (every time step). The conformations in these time frames were then used as the finite difference approximation to compute the linear velocities from Eq.(4) and angular velocities from Eq.(5), and ultimately the kinetic parameters from Eq.(14). More details on the general simulation protocol can be found in [13].

The conformations were analyzed using the program CURVES+ [27] which computes a reference point $\boldsymbol{r}^a$ and a right handed orthonormal frame $\{\boldsymbol{d}_1^a, \boldsymbol{d}_2^a, \boldsymbol{d}_3^a\}$ attached to each base as defined in Eq.(1). These reference points and frames were then used in Eq.(4) and Eq.(5) to compute the linear and angular velocities from the MD trajectories. To stay focused on properties of B-DNA where all bases on opposite strands are connected by hydrogen bonds to form Watson-Crick pairs, all conformations with at least one hydrogen bond broken anywhere in the oligomer were eliminated from the analysis. This is why Eq.(4) and Eq.(5) contain an explicit indexing of both k and t, these allow a consistent tracking and re-weighting over time of all estimates even for non-continuous sections of the trajectories. An hydrogen bond is considered broken if the distance between the donor and the acceptor is greater than 4 Å, as measured by the PTRAJ module of the AMBER suite of programs.

3 Results

3.1 Structure of the estimated mass matrices

The validity of the numerical simulations (e.g., assumption of ergodicity) and the rigidity assumption of the DNA bases was assessed by comparing the structure of the estimated *vs.* theoretical matrices defined from Eq.(8) to Eq.(11).

By definition of the outer product $\mathbf{v} \otimes \mathbf{v}$, every mass matrix estimated using Eq.(13) is symmetric with respect to the diagonal (i.e. $\mathbf{M} = \mathbf{M}^T$). But in addition, for any nonzero velocity field $\mathbf{v}$, the matrix $\mathbf{M}$ must be non negative. The smallest eigenvalue of the mass matrix estimated from the MD trajectory was positive for each of the six sequences s1 to s6 (Table 1), and thus all mass matrices estimated from these simulations are indeed positive definite.

The block diagonal assumption of the rigid base DNA model (Eq. 8) was evaluated by comparing the Euclidean norm of the off-diagonal portion of the total mass matrix with the norm of the entire matrix. As can be seen in Table 1, this ratio is less than 5% each of the six sequences s1 to s6 ($< 2\%$ for five of these sequences), so the velocity field extracted from the MD trajectories for these sequences appear consistent with the block diagonal assumption of the model.

Table 1 Smallest Eigenvalue of the mass matrix $\mathbf{M}$ estimated from MD simulations, and Euclidean norm of its off-diagonal blocks relative to the norm of the entire matrix

Sequence	min λ (10^{-16})	Norm ratio (on/off-diagonal)	Sequence	min λ (10^{-16})	Norm ratio (on/off-diagonal)
s1 (TATA)	1.02	0.014	s4 (AGCT)	0.96	0.042
s2 (CGCG)	1.03	0.015	s5 (GGGC)	1.02	0.013
s3 (TCGA)	1.03	0.015	s6 (AAAT)	1.01	0.014

Names given in parentheses follow a convention proposed in [28].

In the rigid base DNA model, some finer element symmetries exist in the particular matric structure of the diagonal block elements of $\mathbf{M}$ defined in Eq.(9), itself deduced from the expression chosen in this model for the kinetic energy i.e., Eq.(6). These symmetries were evaluated by analyzing the sparsity of sub-blocks B_1^a and B_2^a defined in Eq.(11) for each base X_a in each sequence.

Fig.1 shows the results for the sequence s1 (i.e., alternating T and A bases). The relative error in B_1^a in term of the Euclidean norm of the anisotropic part with respect to the norm of the entire sub-block is less than 5% for every base X_a in s1 (Fig.1, upper left). In case of the off-diagonal entries, which should all be zeros according to Eq.(9), the relative error is less than 2% (Fig.1, upper right). Thus, the sub-block B_1^a is nearly isotropic, in accordance with Eq.(9).

Similarly in the B_2^a sub-block, on average the norm of the skew symmetric off-diagonal entries is less than 4% of the norm of the entire sub-block (Fig.1, lower left). In case of the diagonal entries, which should all be zeros according to Eq.(9), the relative error is less than 2% (Fig.1, lower right). Thus, the sub-block B_2^a is nearly skew-symmetric, in accordance with Eq.(9).

Figures equivalent to Fig.1 for the five other sequences are available in supplementary material. The conclusions are the same: relative errors are systematically less than 5% for the non-zero elements and less than 2% for the near-to-zero elements of each sub-block B_1^a and B_2^a, for every base and every sequence. Thus, the eventual errors made in Eq.(14) when averaging the diagonal elements of B_1^a and using the skew-symmetric projection of B_2^a from the current MD trajectories is $< 5\%$.

The dependence of the estimated kinetic parameters on different simulation protocols was also evaluated (Fig.2). A significant difference was observed when using a time step larger than 2 fs in the finite difference approximations Eq.(4) and Eq.(5): the motion of the coordinate frames $\{\boldsymbol{d}_1^a, \boldsymbol{d}_2^a, \boldsymbol{d}_3^a\}$ that occurs within an interval of 2 fs is often infinitisemally small relative to the finite precision used by common programming languages, leading to instability ('divisions by zero') during the matrix inversion needed between Eq.(13) and Eq.(14) to estimate $\mathbf{M}$. To address this problem, larger timesteps $t^{(k+1)} - t^{(k-1)}$ were benchmarked going from 4 fs or 20 fs (Fig.2) in the two-sided finite difference approximations Eq.(4) and Eq.(5). In particular, the estimates of m^a in s1 increase when going from $t^{(k+1)} - t^{(k-1)} = 4$ fs to 8 fs to 12 fs, then remain steady when going to 16 fs and 20 fs. The steady estimates could also be recovered by manually removing each problematic element of the velocity field that led to divisions by zero, confirming the root cause of the instabilities. Thus, a finite difference time step of 20 fs was chosen for all estimates of kinetic parameter reported in this paper.

In contrast, as illustrated in Fig.2 for s1, using a different force field (parm-94, 23), or a different water model (TIP3P, 24), or a lengthy equilibration phase (100 ns) before to start the fine sampling used to extract the velocity field, or doubling the length of the fine sampling trajectory (2 ns), have no significant impact on the final estimates of m^a.

3.2 Estimation of kinetic parameters

Table 2 provides the final estimates of the kinetic parameters m^a, c^a, and Γ^a, for the four DNA bases G, C, A, and T.

The estimates of these kinetic parameters were analyzed and compared in the different sequence contexts of the sequences s1 to s6. For example, Fig.3 shows all

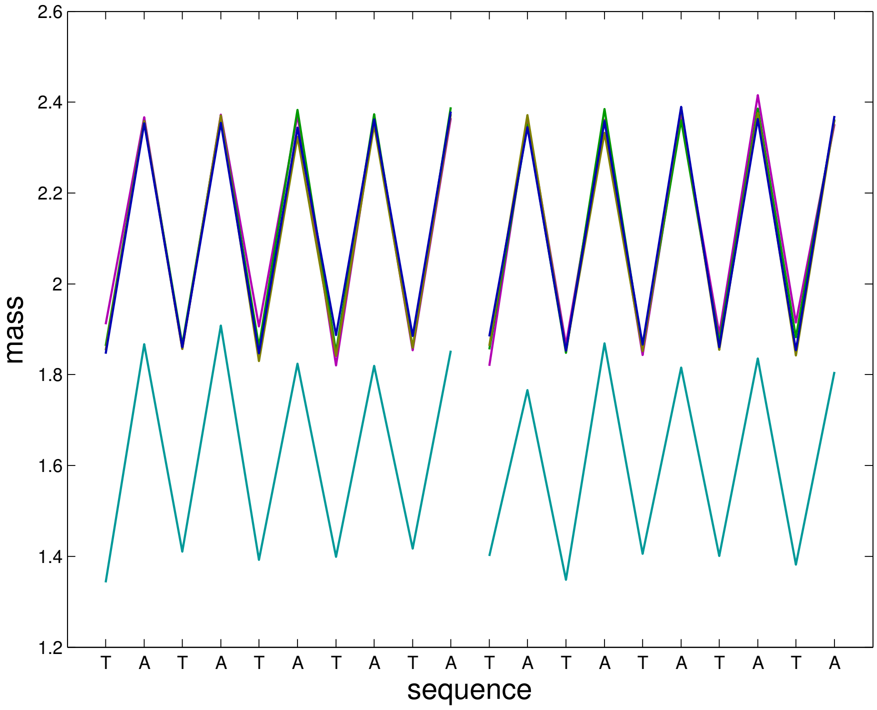

Fig. 2 Comparison of different MD simulation protocols and parameters on the estimation of the total mass (10^{-25} kg) for each base along sequence s1 (considering only the segment of ten central bp). First ten indexes on the x-axis correspond to strand 1 and last ten indexes correspond to strand 2, both read in the 5' to 3' direction. Blue line: $t^{(k+1)} - t^{(k-1)} = 20$ fs, cyan line: $t^{(k+1)} - t^{(k-1)} = 4$ fs, brown line: sampling during 2 ns, green line: pre-equilibration during 100 ns, violet line: parm-94 AMBER force field and TIP3P water model.

components of these kinetic parameters as defined in Eq.(9) for the sequence s1. And Fig.4 shows the average and standard deviation of m^a, c^a, and Γ^a, for each base and each sequence. In Fig.4, the standard deviation accounts for variations due to different locations of an identical base in a given sequence. As can be concluded from Fig.4, its value is always very small compared to the difference between purines (G,A) and pyrimidines (C,T), between the components c_x^a, c_y^a, and c_z^a, and between the components Γ_{xx}^a, Γ_{yy}^a, and Γ_{zz}^a.

Similarly, for a given DNA base, the difference observed between the different sequences (x-axis in Fig.4) appears negligeable. Thus, some sequence context effects are *not* observed for the DNA kinetic parameter estimated from the MD simulations of these six representative sequences. It is interesting to note that, as described in the introduction and the section 2.2, the three different motifs that characterize the chosen sequences, PyPu/PuPy, PyPy/PuPu, and PyPyPy/PuPuPu, are known to have significantly different effects on the stiffness and flexibility of DNA molecule under the same MD simulation protocol as used here [4, 9, 12, 13].

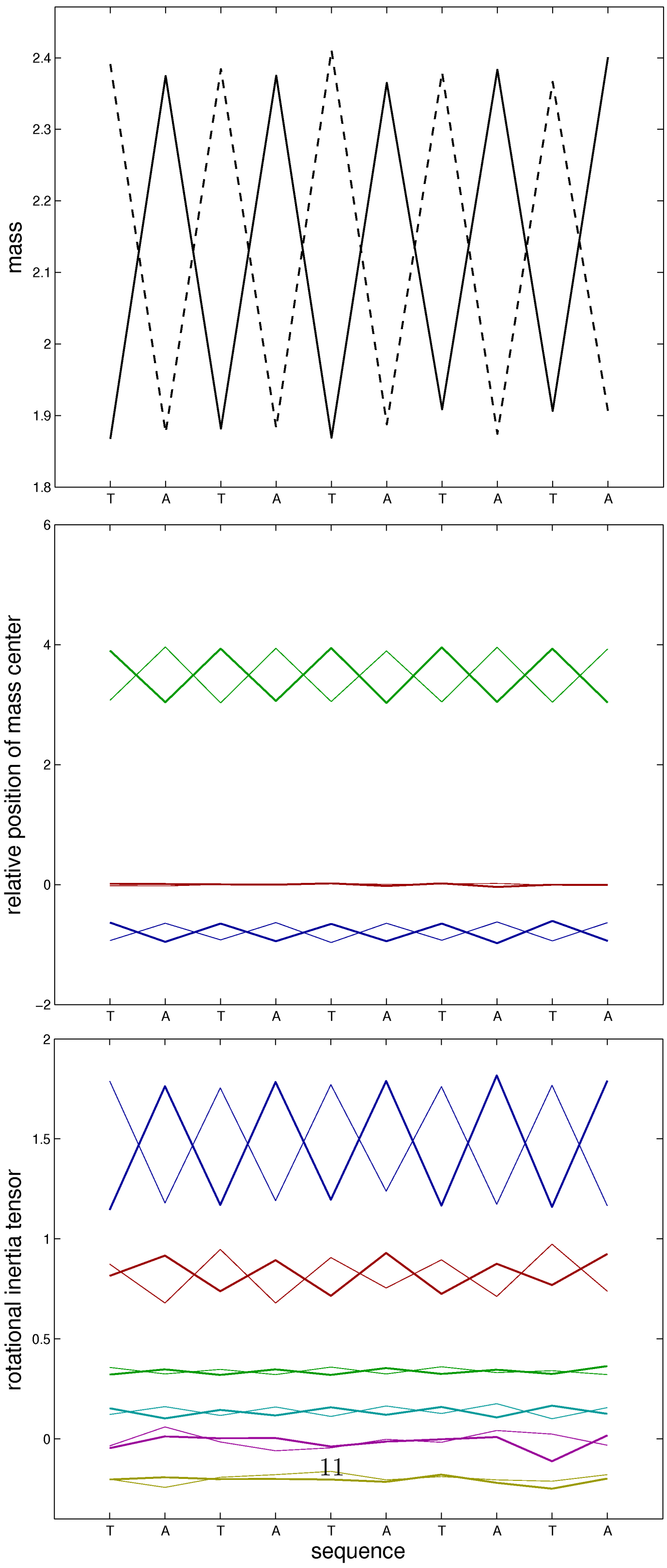

Fig. 3 Estimated kinetic parameters of each base along sequence s1 (ten central bp): Up: total mass m^a in 10^{-25} kg (full line: strand 1, dashed line: strand 2), Middle: mass center c^a in Å (blue line: c^a_x, green line: c^a_y, ochre line: c^a_z where the hair line is the complementary strand), Down: rotational inertia tensor Γ^a in 10^{-44} kg m^2 (blue line: Γ^a_{xx}, green line: Γ^a_{yy}, ochre line: Γ^a_{zz}, cyan line: Γ^a_{xy}, violet line: Γ^a_{xz}, gold line: Γ^a_{yz}, where the hair line is the complementary strand). Γ^a_{xx} values were all shifted by $+10^{-44}$ kg m^2 and Γ^a_{yz} by -2×10^{-45} kg m^2 for better visualization.

Table 2 Estimation of kinetic parameters m^a, c^a and principal values of Γ^a

	G	C	A	T
m (10^{-25} kg)	2.41 (0.04)	1.79 (0.02)	2.35 (0.03)	1.86 (0.02)
c_x (Å)	-0.97 (0.02)	-0.68 (0.02)	-0.94 (0.02)	-0.64 (0.03)
c_y (Å)	2.97 (0.02)	3.93 (0.02)	3.03 (0.03)	3.92 (0.02)
c_z (Å)	-0.01 (0.02)	0.0 (0.02)	0.0 (0.02)	-0.01 (0.02)
Γ_{xx} (*)	7.96 (0.41)	2.97 (0.42)	7.76 (0.34)	1.96 (0.30)
Γ_{yy} (*)	3.42 (0.11)	3.06 (0.06)	3.52 (0.10)	3.24 (0.06)
Γ_{zz} (*)	10.1 (0.41)	5.61 (0.43)	9.22 (0.38)	7.25 (0.33)

* in 10^{-45}kg m^2

Significant differences are observed between the estimated DNA kinetic parameters for purines (higher total mass) *vs.* pyrimidines (lower total mass), including differences between the detailed components of the mass center c^a_x, c^a_y, and c^a_z, and also the principal values of the symmetric rotational inertia tensor Γ^a_{xx}, Γ^a_{yy}, and Γ^a_{zz}. Compared to these differences between purines and pyrimidines, the differences observed between different purines (A *vs.* G and different pyrimidines (T *vs.* C) are negligeable.

When compared to the experimentally solved Arnott conformation (crosses in Fig.4) which is a reference model for B-DNA, the estimates obtained for the total mass m^a and the relative position of the mass center c^a are in very good agreement. For the principal values of the rotational inertia tensor Γ^a, some discrepancies are observed, but the fit appears reasonable given the much smaller magnitude for these parameters (all are multiples of 10^{-44}). In particular, the qualitative ranking of the three principal values is identical between the current simulations and the experimental data.

the systematic microsecond molecular dynamics investigations on tetranucleotide sequence effects in B-DNA [13]. In [13], it was shown that the statistical distribution of DNA conformations observed in the experimental Protein Data Bank (PDB, 14) is well approximated by a dataset of MD simulations including the six DNA sequences chosen here

The main motivation for this work, as explained in the introduction, is that experimental data is not available on the detailed atomistic and dynamic structure of DNA molecules, such as the kinetic variables of DNA in different sequence contexts. But a final evaluation of the current estimates was done by comparing a more exhaustive dataset of MD simulations (39 DNA seuqences as in 13, 28 including the 6 sequences used here) with the statistical ensemble made of all available DNA entries in the PDB [14]. A very good agreement was found between the two datasets in term of configurational variables such as helical parameter averages and stiffness matrices (in preparation). This supports the use of molecular dynamic simulations to also estimate the *kinetic* variables, as done here

Table 2 summarizes the results by providing the final estimates of the kinetic parameters averaged over all sequences studied for the four DNA bases G, C, A, and T. These values represent the best fit parameters inferred from the current MD simulations when ignoring sequence context effects (dinucleotides, trinucleotides, and

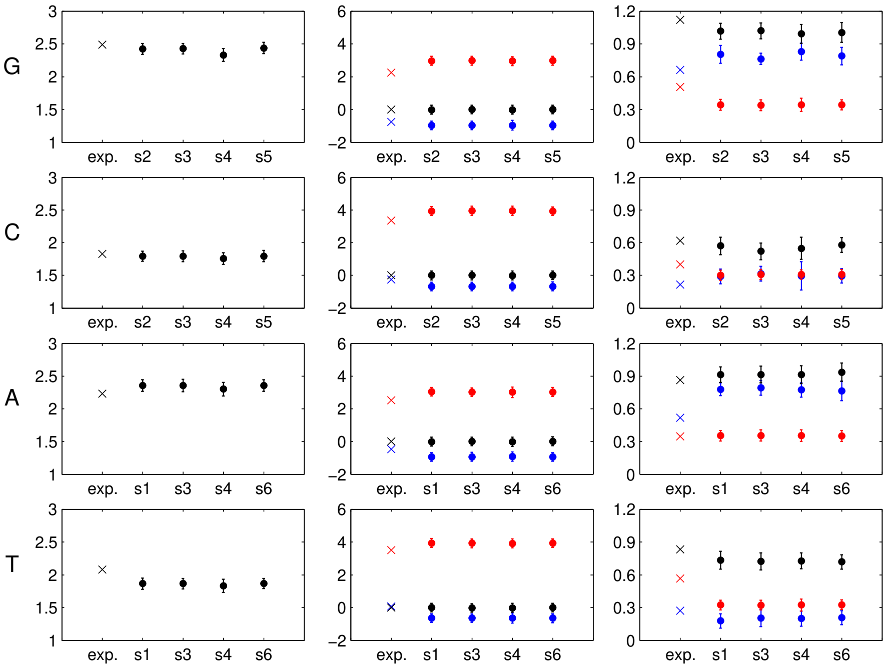

Fig. 4 Comparison of average kinetic parameters in each sequence for the base G (row 1), C (row 2), A (row 3) and T (row 4) over all its occurences in the segment of ten central bp of a given sequence. Left: total mass m^a in 10^{-25} kg, Middle: mass center c^a in Å (blue: c_x^a, red: c_y^a, black: c_z^a), Right: rotational inertia tensor Γ^a in 10^{-44} kg m^2 (blue: Γ_{xx}^a, red: Γ_{yy}^a, black: Γ_{zz}^a). The length of the error bar that accounts for the standard deviation is located outside the circles for better visualization. For each variable and each base, a cross indicates the reference value computed for the experimentally solved B-DNA Arnott conformation [5].

beyond). These parameters can be used as inputs to any rigid-base DNA model, and suggest that the coarse-grained kinetic energy of a DNA sequence, in contrast to its potential energy, depends on the nature of its bases and not on more global sequence contexts.

4 Conclusion

Fine-grain molecular dynamics simulations were used to estimate the kinetic energy parameters for a coarse-grain, rigid-base DNA model. Such estimates cannot yet be measured with existing experimental technologies. The closest published work [5] reported these parameters only based on the simulation of one sequence and one simulation protocol, and did not provide any information on the dependence of DNA kinetic parameters on their sequence context.

An algebraic expression was derived for the kinetic energy as a function of linear and angular velocities of each DNA base parameterized by its mass, center of mass, and rotational inertia tensor. The parameters of this function were then approximated from a set of fine-grain molecular dynamics simulations representing all combinations of the four DNA base pairs AT, TA, GC, and CG, in different sequence contexts.

A difference was observed between DNA kinetic parameters for purines (G,A) *vs.* pyrimidines (C,T), in particular in their total mass and rotational inertia tensor. But for a given DNA base, no difference in the kinetic parameters was observed between different sequence contexts, nor between different MD simulation protocols. A significant difference was observed due to instabilities arising from computing infinitisimally small quantities with a finite-precision computer. But using an integration time step above 14 fs in Eq.(4) and Eq.(5) eliminated this problem. A time step of 20 fs was used for the final estimates reported in Table 2.

The magnitudes of all current estimates are consistent with the limited but experimental data available i.e., the experimentally solved B-DNA structures. The assumptions of each base being modeled as a rigid body were verified to be good approximations. Overall, the relative error due to such assumptions was systematically less than 5%, and most often less than 2%. This was observed across all kinetic parameters, DNA sequences, and MD protocols.

The current estimates can thus be used to parameterize the kinetic energy function of coarse-grain DNA models with rigid-base resolution. Combined with sequence-dependent constitutive relations [29] that define the model's potential energy function, these results will help better model the dynamical structure of DNA on length and times scales most relevant in protein/DNA interaction, transcription regulation, and nucleosomes organization.

Acknowledgments. The author thanks Prof. John H. Maddocks (LCVMM, Department of Mathematics, EPFL, Switzerland) for his help with deriving the explicit algebraic expressions of the DNA model kinetic energy function, and his advices for deriving the kinetic parameter estimates from molecular dynamics simulations.

References

[1] Dickerson, R.E., Chiu, T.K.: Helix bending as a factor in protein/DNA recognition. Biopolymers **44**(4), 361–403 (1997)

[2] Rippe, K., Hippel, P.H., Langowski, J.: Action at a distance: DNA-looping and initiation of transcription. Trends in biochemical sciences **20**(12), 500–506 (1995)

[3] Morozov, A.V., Fortney, K., Gaykalova, D.A., Studitsky, V.M., Widom, J., Siggia, E.D.: Using DNA mechanics to predict in vitro nucleosome positions and formation energies. Nucleic Acids Research **37**(14), 4707–4722 (2009) https://doi.org/10.1093/nar/gkp475

[4] Da Rosa, G., Grille, L., Calzada, V., Ahmad, K., Arcon, J.P., Battistini, F., Bayarri, G., Bishop, T., Carloni, P., Cheatham Iii, T., et al.: Sequence-dependent structural properties of B-DNA: what have we learned in 40 years? Biophysical Reviews, 1–11 (2021) https://doi.org/10.1093/nar/gkp1061

[5] Lankaš, F., Gonzalez, O., Heffler, L.M., Stoll, G., Moakher, M., Maddocks, J.H.: On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. Physical Chemistry Chemical Physics **11**(45), 10565 (2009) https://doi.org/10.1039/b919565n

[6] De Bruin, L., Maddocks, J.H.: cgDNAweb: a web interface to the sequence-dependent coarse-grain model of double-stranded DNA. Nucleic Acids Research **46**(W1), 5–10 (2018) https://doi.org/10.1093/nar/gky351

[7] Gonzalez, O., Petkevičiūtė, D., Maddocks, J.: A sequence-dependent rigid-base model of DNA. The Journal of chemical physics **138**(5), 02–604 (2013) https://doi.org/10.1063/1.4789411

[8] Curuksu, J.: Adaptive conformational sampling based on replicas. Journal of mathematical biology **64**, 917–931 (2012) https://doi.org/10.1007/s00285-011-0432-6

[9] Pyne, A.L., Noy, A., Main, K.H., Velasco-Berrelleza, V., Piperakis, M.M., Mitchenall, L.A., Cugliandolo, F.M., Beton, J.G., Stevenson, C.E., Hoogenboom, B.W., *et al.*: Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides. Nature Communications **12**(1), 1053 (2021) https://doi.org/10.1038/s41467-021-21243-y

[10] Reymer, A., Zakrzewska, K., Lavery, R.: Sequence-dependent response of DNA to torsional stress: a potential biological regulation mechanism. Nucleic acids research **46**(4), 1684–1694 (2018) https://doi.org/10.1093/nar/gkx1270

[11] Sharma, R., Patelli, A.S., De Bruin, L., Maddocks, J.H.: cgNA+ web: A visual interface to the sequence-dependent statistical mechanics model of double-stranded nucleic acids. Journal of Molecular Biology, 167978 (2023) https://doi.org/10.1016/j.jmb.2023.167978

[12] Curuksu, J., Zacharias, M., Lavery, R., Zakrzewska, K.: Local and global effects of strong DNA bending induced during molecular dynamics simulations. Nucleic Acids Research **37**(11), 3766–3773 (2009) https://doi.org/10.1093/nar/gkp234

[13] Pasi, M., Maddocks, J.H., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dans, P.D., Jayaram, B., Lankas, F., Laughton, C., Mitchell, J., Osman, R., Orozco, M., Pérez, A., Petkevičiūtė, D., Spackova, N., Sponer, J., Zakrzewska, K., Lavery, R.: μabc: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Research **42**(19), 12272–12283 (2014) https://doi.org/10.1093/nar/gku855

[14] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer Jr, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.: The protein data bank: a computer-based archival file for macromolecular structures. Journal of

molecular biology **112**(3), 535–542 (1977)

[15] Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C., Berman, H.M.: A standard reference frame for the description of nucleic acid base-pair geometry. Journal of Molecular Biology **313**(1), 229–237 (2001) https://doi.org/10.1006/jmbi.2001.4987

[16] Huang, K.: Statistical Mechanics, pp. 136–138. John Wiley & Sons, Boston (1991)

[17] Hud, N.V., Plavec, J.: A unified model for the origin of DNA sequence-directed curvature. Biopolymers **69**(1), 144–158 (2003) https://doi.org/10.1002/bip.10364

[18] Strahs, D., Schlick, T.: A-tract bending: insights into experimental structures by computational model. Journal of Molecular Biology **301**(3), 643–663 (2000) https://doi.org/10.1006/jmbi.2000.3863

[19] Demurtas, D., Amzallag, A., Rawdon, E.J., Maddocks, J.H., Dubochet, J., Stasiak, A.: Bending modes of DNA directly addressed by cryo-electron microscopy of DNA minicircles. Nucleic Acids Research **37**(9), 2882–2893 (2009) https://doi.org/10.1093/nar/gkp137

[20] Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J.: The Amber biomolecular simulation programs. Journal of Computational Chemistry **26**(16), 1668–1688 (2005) https://doi.org/10.1002/jcc.20290

[21] Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T.E., Laughton, C.A., Orozco, M.: Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. Biophysical Journal **92**(11), 3817–3829 (2007) https://doi.org/10.1529/biophysj.106.097782

[22] Berendsen, H.J.C., Grigera, J.R., Straatsma, T.P.: The missing term in effective pair potentials. The Journal of Physical Chemistry **91**(24), 6269–6271 (1987) https://doi.org/10.1021/j100308a038

[23] Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., Kollman, P.A.: A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society **117**(19), 5179–5197 (1995) https://doi.org/10.1021/ja00124a002

[24] Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L.: Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics **79**(2), 926–935 (1983) https://doi.org/10.1063/1.445869

[25] Darden, T., York, D., Pedersen, L.: Particle mesh Ewald: an n × log(n) method for computing Ewald sums in large systems. The Journal of Chemical Physics **98**(12), 10089–10092 (1993) https://doi.org/10.1063/1.464397

[26] Ryckaert, J.-P., Ciccotti, G., Berendsen, H.J.: Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of computational physics **23**(3), 327–341 (1977)

[27] Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D., Zakrzewska, K.: Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Research **37**(17), 5917–5929 (2009) https://doi.org/10.1093/nar/gkp608

[28] Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dixit, S., Jayaram, B., Lankas, F., Laughton, C., Maddocks, J.H., Michon, A., Osman, R., Orozco, M., Perez, A., Singh, T., Spackova, N., Sponer, J.: A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. Nucleic Acids Research **38**(1), 299–313 (2009) https://doi.org/10.1093/nar/gkp834

[29] Petkevičiūtė, D., Pasi, M., Gonzalez, O., Maddocks, J.H.: cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. Nucleic Acids Research **42**(20), 153–153 (2014) https://doi.org/10.1093/nar/gku825