MULTISCALE MODEL. SIMUL.
Vol. 2, No. 4, pp. 527–553

MULTISCALE METHOD FOR SIMULATING PROTEIN-DNA COMPLEXES[*]

ELIZABETH VILLA[†], ALEXANDER BALAEFF[‡], L. MAHADEVAN[§], AND KLAUS SCHULTEN[¶]

Abstract. We present a multiresolution approach to modeling complexes between protein and DNA that contain looped or coiled DNA. The approach combines a coarse-grained model of the DNA loop, based on the classical theory of elasticity, with an atom level model of proteins and protein-DNA interfaces based on molecular dynamics. The coarse-grained DNA description is controlled through the atom level protein description and vice versa. The feasibility of the resulting multiscale modeling approach is demonstrated for a protein-DNA complex in which a protein called the *E. coli lac* repressor forces DNA into a 76 base pair loop. The required simulation involves 230,000 atoms, a number that would triple if both protein and DNA loops were described at the atomic level.

Key words. multiscale, coarse-grained, protein-DNA interaction, gene control, elastic rod model of DNA, molecular dynamics

AMS subject classifications. 9208, 92C05, 92C10

DOI. 10.1137/040604789

1. Introduction.

Computational studies are ideally suited for investigations of structure, function, and dynamics of biological molecules at the atomic level [23]. The dramatic growth of computational power makes it feasible to simulate larger systems for longer timescales: a decade ago the limit was typically simulations of ∼30,000 atoms for ∼200 ps (picoseconds); today simulations cover hundreds of thousands of atoms for many nanoseconds [63]. Still, due to their size, biomolecular systems of relevance are often beyond the reach of computational methods like molecular dynamics (MD). Examples of such systems [2] are the ribosome ($2 \cdot 10^6$ atoms), a molecular machine, which reads messenger RNA and synthesizes proteins [38]; the nucleosome ($6 \cdot 10^5$ atoms), which packs DNA into a compact structure by winding it up [56]; the ATPase ($5 \cdot 10^5$ atoms), which reversibly converts a membrane potential into chemical energy [16]; virus capsid ($\sim 10^6$ atoms), an icosahedral coat of typically 240 proteins that encloses viral DNA for release from infected cells and infection of new cells [4]; and the bacterial flagellum ($\sim 10^7$ atoms), a large aggregate of proteins that propels and reorients swimming bacteria [13]. The size limitations call for a multiscale approach, in which descriptions of biomolecules are simplified using coarse-grained models, preferably models capable of furnishing idealized full atomic level detail when needed. Such coarse-grained models, when applied to all or some of the simulated volume, can reduce the atom count through MD simulations with rather

[*] Received by the editors March 4, 2004; accepted for publication April 11, 2004; published electronically September 2, 2004. This work was supported by the National Institutes of Health (PHS 2 P41 RR05969). Computer time was provided by National Resource Allocations Committee grant MCA93S028.

http://www.siam.org/journals/mms/2-4/60478.html

[†] Center for Biophysics and Computational Biology and Beckman Institute for Advanced Science and Technology, University of Illinois, 405 N. Mathews Ave., Urbana, IL 61802 (villa@ks.uiuc.edu).

[‡] IBM T.J. Watson Research Center, 1101 Kitchawan Road, Route 134, Yorktown Heights, NY 10598 (balaeff@us.ibm.com).

[§] Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 (lm@deas.harvard.edu).

[¶] Physics Department and Beckman Institute for Advanced Science and Technology, University of Illinois, 405 N. Mathews Ave., Urbana, IL 61802 (kschulte@ks.uiuc.edu).

few effective atoms or replace MD with other mathematical descriptions, e.g., continuum theory. The resulting models become computationally more tractable, partially due to reduced atom count, but mainly due to smoothing built into the coarse-grained models that accelerates dynamic processes by orders of magnitude (see, e.g., [1]). The effectiveness of those methodologies was discussed and demonstrated in [53, 5, 35].

We introduce one such multiscale method for the investigation of protein-DNA complexes where the DNA is looped or coiled. Looped or coiled configurations of DNA arise often in the regulation of gene expression [42]. Respective protein-DNA complexes include the nucleosome and regulatory proteins such as the *gal* and *lac* repressors. The *lac* repressor is a protein of ~22,000 atoms; a simulation in a suitable solvent environment involves 230,000 atoms, as detailed below. If the loop that the *lac* repressor induces in the bound DNA is included in the system and solvated, the size of the system triples. The expected slow dynamics of the DNA loop due to strong coupling of loop and solvent motions would be a further hindrance to computational descriptions.

Obviously, protein-DNA complexes of ~700,000 atoms with slow DNA dynamics suggest themselves for a multiscale approach. A way to accomplish this approach is by using simplified models of DNA [51, 41]. Due to its one "long" dimension and two "short" dimensions, DNA can be approximated as a thin rod [8] by means of the theory of elasticity [65, 58, 47]. Such an approach replaces a full-atom MD simulation with a continuum (elasticity) theory treatment with vastly accelerated DNA loop dynamics; in fact, loop relaxation can be assumed to be instantaneous if an equilibrium loop model is adopted.

In the present paper we outline a multiscale methodology that links a full-atom MD description applied to a protein with short DNA segments bound to it and a coarse-grained (continuum) elastic rod model applied to the DNA between these segments. The multiscale methodology can be used to address fundamental questions posed by the structure and function of protein-DNA complexes [10]. We apply this method to the *lac* repressor-DNA complex.

The *lac* repressor protein, shown in Figure 1.1, is the most widely known regulatory protein and has helped to establish the paradigm of gene control through protein-DNA interaction [45]. The protein functions as a negative switch which clamps DNA and induces a loop in a key DNA segment, which contains the promoter for a set of genes, *lacZ*, *lacY*, and *lacA*, that code for proteins involved in lactose uptake and metabolism [43, 44]. In the absence of lactose, the protein binds with high specificity to two 21 bp (base pair) DNA segments called "operators" [34], folding the DNA between them into a loop [42] (cf. Figure 1.1) and inhibiting the expression of *lacZ*, *lacY*, and *lacA*. The formation of the loop has been shown to be critical for full repression [46]. When lactose is present, the repressor dissociates from the DNA, allowing the transcription of the genes [45]. Studying the dynamics of the *lac* repressor-DNA complex is important for a basic understanding of the mechanisms of gene control [52].

Despite extensive studies [45, 55, 44] the mechanical properties of the complex and its *in vivo* configuration remain unknown. The crystal structure of the protein clamped to two short pieces of DNA but without the DNA loop is available [34], as shown in Figure 1.2. However, the *lac* repressor is not bound to such disjointed segments of DNA in the real cell. Rather, it interacts with continuous DNA subject to forces that result from the formation of the DNA loop. Presently, the geometry of the loop must be inferred from modeling [8, 9]. The following questions arise: What is the structure of the DNA loop? How is the structure of the protein optimized for the role of controlling the DNA? What forces does the DNA exert on the protein?

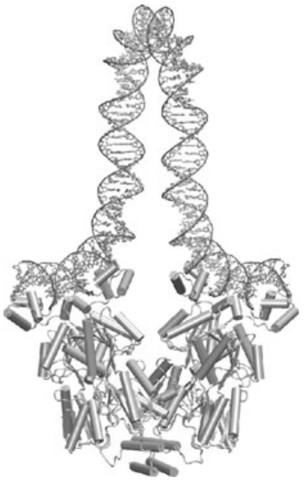

Fig. 1.1. *The* lac *repressor protein with full-atom DNA loop. The all-atom structure of the loop is constructed from the elastic rod description, as explained in section* 2.2.4 *and Appendix* B.

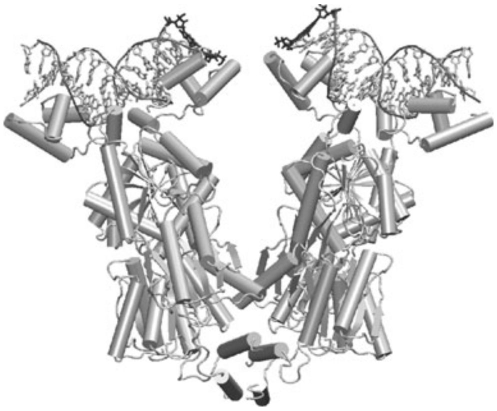

Fig. 1.2. *The* lac *repressor protein. Structure of the* lac *repressor binding two segments of DNA, modeled from crystal and nuclear magnetic resonance structures. The terminal base pairs (tbp's) of the bound pieces of DNA are highlighted in black.*

Are these forces strong enough to alter the configuration of the protein? If so, what is the *in vivo* structure of the protein-DNA complex?

In the present work we show by means of a short simulation that the suggested multiscale methodology can describe the dynamics of the *lac* repressor-DNA complex. A detailed investigation of the equilibrium and dynamical properties of the complex will be presented in a forthcoming publication [64].

2. Multiscale approach to protein-DNA simulations. The multiscale methodology combines two levels of description: the DNA loop is described by means of elasticity theory in a continuum representation, and the protein is described by means of MD at atomic level resolution. Protein MD simulations position the DNA loop end points; these points then serve as crucial boundary conditions to the elasticity theory in determining the shape of the corresponding loop. Elasticity theory in turn determines the forces arising at the end points of the new loop; the forces enter into the protein MD simulation closing the cycle. As a result, the elasticity theory description of the DNA loop and the MD calculation of the protein are intertwined through the exchange of boundary conditions and forces. In this section we present first the protein MD description, then the DNA elasticity theory description, and, finally, the linkage between the two descriptions that defines the multiscale methodology.

Before we introduce these descriptions, we note that the MD description and DNA elasticity theory description are implemented in two different coordinate systems, shown in Figure 2.1. The MD simulation occurs in the "laboratory coordinate system" (LCS) that we will denote by unprimed coordinates, e.g., $\mathbf{r}(s), \hat{\mathbf{d}}_i$. The elasticity theory description occurs in the "rod coordinate system" (RCS) that we will denote by primed coordinates, e.g., $\mathbf{r}'(s), \hat{\mathbf{d}}'_i$.

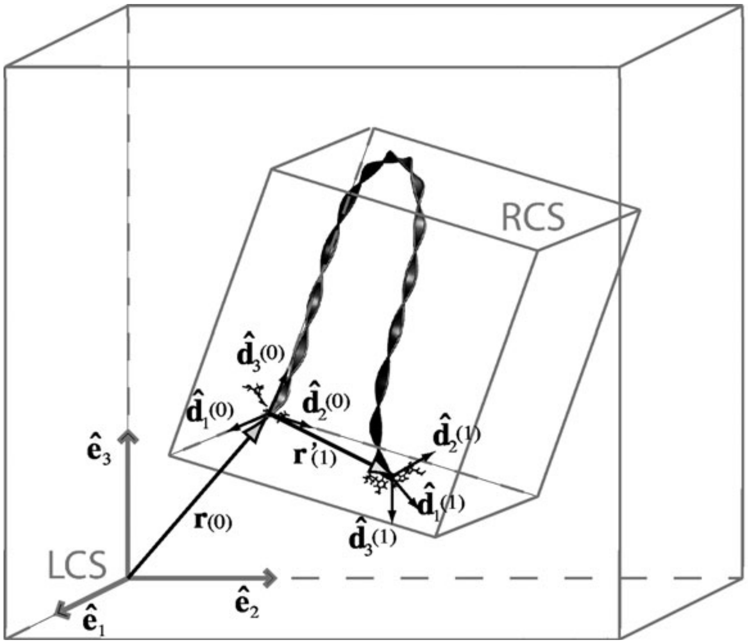

FIG. 2.1. *Lab coordinate system (LCS) versus rod coordinate system (RCS). The LCS, shown in the larger box, is the frame of reference describing the MD simulation, defined by the director basis* $\{\hat{\mathbf{e_1}}, \hat{\mathbf{e_2}}, \hat{\mathbf{e_3}}\}$. *The RCS has its origin at the location of the terminal base pair (tbp) at* $\mathbf{r}(s = 0)$.

2.1. Molecular dynamics. MD is a computational method that calculates the time-dependent behavior of a molecular system at the atomic level [23]. Today, it represents one of the principal tools in the theoretical study of biological molecules. The method permits one to include the natural environment of a biomolecular system by explicitly including water and ions in the simulation [33].

The simulations provide detailed information on the fluctuations and conformational changes of biopolymers, e.g., proteins and nucleic acids. The MD method is also used in the refinement of structures obtained from x-ray crystallography and NMR (nuclear magnetic resonance) spectroscopy. This section introduces the conventional use of MD. We also explain the treatment given inside the MD simulation to the DNA segments bound to the protein that define the loop termini.

2.1.1. Conventional molecular dynamics. This method is based on classical mechanics and propagates the positions $\mathbf{r}_i$ and velocities $\mathbf{v}_i$ of a set of interacting atoms $i = 1, 2, \dots, N$ in time t by integrating the Newtonian equations of motion, with each atom being represented as a point of mass m_i. A potential energy function $U(\mathbf{R})$ describes the interaction of all particles in the system in terms of the atomic positions, described in terms of $3N$ Cartesian coordinates collected in the vector $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$.

The use of Cartesian coordinates rather than internal coordinates leads to efficient integration of the equations of motion due to a simple form of the inertia terms; cf. (2.2)–(2.4) and [59].

The potential energy function used in MD simulations is designed to provide a compromise between accuracy and computational efficiency. The contributions to the potential energy can be classified as bonded and nonbonded interactions; the former describe interactions between atoms linked by covalent bonds, including U_{bond}, which describes high frequency vibrations along covalent chemical bonds; U_{angle}, which describes bending motions between two adjacent bonds; $U_{dihedral}$, which describes torsional motion around a bond; and $U_{improper}$, which describes the planar orientation of one atom relative to three others. The nonbonded terms describe interactions between atoms which are not covalently bonded or atoms separated by three or more covalent bonds, and include U_{vdW}, the pairwise van der Waals energy, and U_{elec}, the pairwise Coulomb energy between charged atoms. The form of the empirical potential energy function is stated in internal coordinates

$$\begin{aligned}
U(\mathbf{R}) = {} & \underbrace{\sum_{bonds,\alpha} k_\alpha^{bond}(r_\alpha - r_{0\alpha})^2}_{U_{bond}} + \underbrace{\sum_{angles,\beta} k_\beta^{angle}(\theta_\beta - \theta_{0\beta})^2}_{U_{angle}} \\
& + \underbrace{\sum_{dihedrals,\gamma} k_\gamma^{dihed}([1 + \cos(n_\gamma \psi_\gamma + \delta_\gamma)])}_{U_{dihedral}} + \underbrace{\sum_{impropers,\delta} k_\delta^{impr}(\phi_\delta - \phi_{0\delta})^2}_{U_{improper}} \\
& + \underbrace{\sum_i \sum_{i \neq j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]}_{U_{vdW}} + \underbrace{\sum_i \sum_{i \neq j} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{elec}},
\end{aligned} \tag{2.1}$$

where r_α is the bond length, $r_{0\alpha}$ the associated equilibrium bond length, and k_α^{bond} the respective bond spring constant; θ_β is the angle between two bonds, $\theta_{0\beta}$ the associated equilibrium bond angle, and k_β^{angle} the respective spring constant for the angle; ψ_γ is the angle of rotation around a bond, k_γ^{dihed} the rotational spring constant, proportional to the energy barrier for rotation, and n_γ the number of maxima (or minima) in one full rotation with δ_γ the angular offset; ϕ_δ is the improper torsion angle, $\phi_{0\delta}$ the associated equilibrium value, and k_δ^{impr} the associated angle spring constant. The van der Waals interaction between atoms i and j is modeled using the Lenard-Jones 6-12 potential, with ϵ_{ij} being the depths of the functions contributing to U_{vdW}, the minima being located at $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j| = \sigma_{ij}$. The electrostatic contribution U_{elect} accounts for interaction between atomic partial charges q_i and q_j, r_{ij} is the separation between them, i.e., $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, ϵ_0 is the permittivity of free space, and ϵ is the relative dielectric constant of the medium in which the charges are placed.

The parameters for the force field defined through (2.1), e.g., charges q_i, are obtained by calibration of experimental results and quantum mechanical simulations of small model compounds. The energy parameters used in the present study are those of the CHARMM22 force field [37].

The numerical method used by most MD programs to integrate the equations of motion is based on the velocity Verlet algorithm [3]. This method provides a direct solution of the Newtonian equations of motion, where the atomic positions $\mathbf{u}$, i.e., $\mathbf{u} = \mathbf{r_i}$, velocities $\mathbf{v}$, i.e., $\mathbf{v} = \dot{\mathbf{r_i}}$, and accelerations $\mathbf{a}$, i.e., $\mathbf{a} = \ddot{\mathbf{r_i}}$, are obtained by discretizing time in units of δt:

$$m\,\mathbf{a}(t) \;=\; -\nabla U\left(\mathbf{u}(t)\right) + \mathbf{F}_{ext}, \tag{2.2}$$

$$\mathbf{v}(t+\delta t) \;=\; \mathbf{v}(t) + \frac{1}{2}\left[\mathbf{a}(t+\delta t) + \mathbf{a}(t)\right]\delta t, \tag{2.3}$$

$$\mathbf{u}(t+\delta t) \;=\; \mathbf{u}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2. \tag{2.4}$$

Here $U\left(\mathbf{u}(t)\right)$ is the potential energy function defined in (2.1), where only the dependence on one atom's position $\mathbf{u}$ is explicitly stated. $\mathbf{F}_{ext}$ is any external force applied to the respective atom inside the MD calculation. Through $U(\mathbf{R})$, the motions of atoms are coupled to each other. Excellent introductions to MD can be found in [3, 23, 33].

The choice of discretization δt, here $\delta t = 1\,\mathrm{fs}$ (femtoseconds), is made in order to properly represent the fastest motion arising in biopolymers, namely, the vibration of hydrogen atoms along their covalent bonds. This time step is many orders of magnitude below the timescales of relevant processes in biological cells. The limitations on computer power today permit simulations of $\sim 10^7$ integration steps, i.e., in the multinanosecond range. This timescale accessible to MD is too short by a factor of 10^3–10^6 for many important biological processes. This limitation is a key reason for introducing the multiscale method for the protein-DNA complex; the dynamics of the loop could not be described on the timescale available to the MD method, whereas elasticity theory effectively covers larger timescales, as explained in sections 2.2 and 2.3.

The presented multiscale implementation uses the MD program NAMD2 [32], with applied external forces obtained from the elasticity theory description of the DNA loop (Figure 1.1). The simulations are described in further detail below.

2.1.2. MD treatment of terminal base pairs. The multiscale method suggested here can be applied only when the termini of the DNA loop are resolved in a structure of a protein-DNA complex, such that a full-atom MD simulation can include both the protein and the bound DNA segments as illustrated in Figure 1.2. We call the last base pair defining the beginning and ending of the loop the "terminal base pair" (tbp) of the DNA loop. There are two tbp's, one for each bound DNA segment, as highlighted in Figure 1.2. In the present application, the structure of the *lac* repressor with DNA bound segments, each with 19 bp, is fortunately known (Figure 1.2). The tbp's are subject to special treatment inside the MD simulation, illustrated in Figure 2.2: they are constrained to preserve their planar structure and are subject to external forces determined by the elasticity theory description of the DNA loop. Most important, the positions and orientations of the tbp's define the boundary conditions for the latter description.

Preserving the tbp plane. Harmonic constraints are introduced in the tbp's in order to preserve their planar structure. In general, the Watson–Crick hydrogen bonds at the ends of a DNA segment are easily broken, since there is a strong competition between water and the exposed base pair atoms for hydrogen bonding in the actual

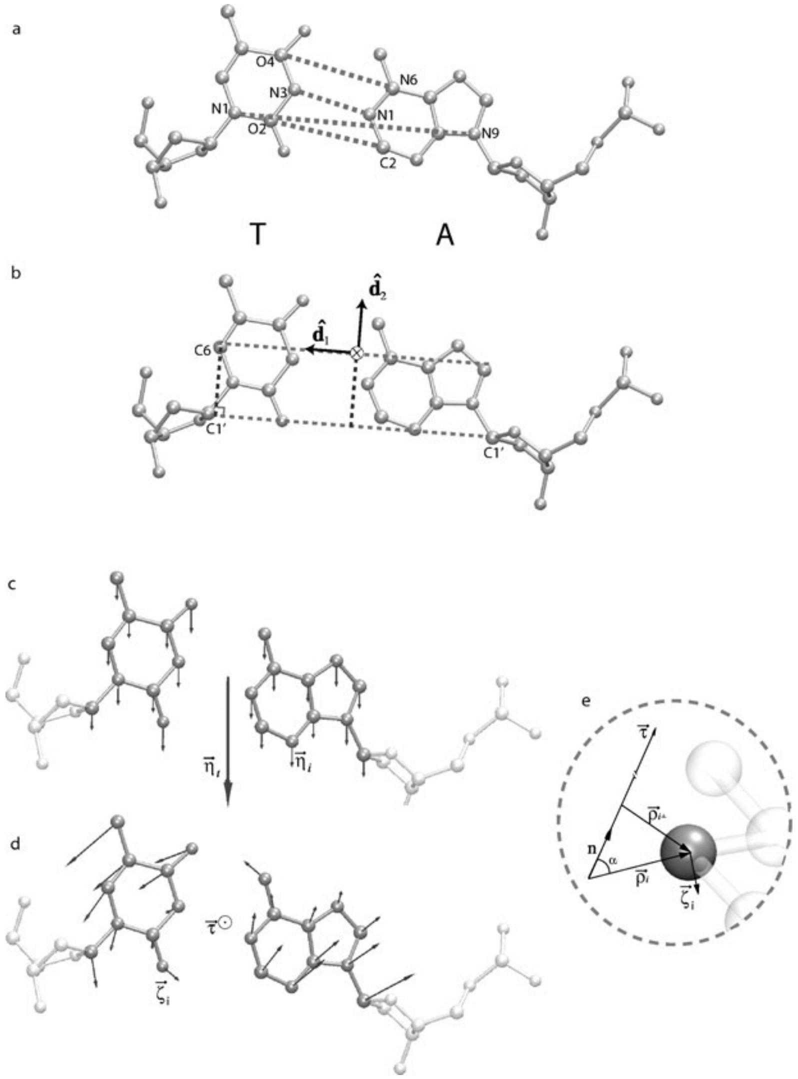

FIG. 2.2. *Tbp's.* (a) *Constraints imposed on the tbp's to preserve the Watson–Crick structure for the case of an adenine-thymine (A-T) base pair. Springs were placed between the atoms $N3$-$N1$, $O4$-$N6$, $O2$-$C2$, and $N1$-$N9$, where the first atom name refers to the atom belonging to thymine (T), and the second to the atom in adenine (A), and assigned a spring constant of $k = 100\,kcal/mol{\cdot}\AA^2$ and equilibrium distances of $2.9\,\AA$, $2.88\,\AA$, $3.71\,\AA$, and $8.97\,\AA$, respectively (distances taken from an ideal Watson–Crick base pair).* (b) *Coordinate frame associated with a DNA base pair* [48]. (c) *Application of the strain obtained from the elastic rod theory to the tbp's during the MD simulation. The total force is applied to the center of mass of the nonhydrogen atoms in the bases.* (d) *Application of the torque τ obtained from the elastic rod theory to the tbp during the MD simulation. The total torque is applied to the nonhydrogen atoms in the base, shown in dark gray, excluding the atoms in the sugars and phosphates, shown in light gray. Each atom is subject to a torque generating force η_i as depicted in* (e). (e) *Detail of the application of the total torque, shown for a single atom.*

MD simulation. However, it is desirable to mimic DNA continuity at the ends as if the DNA that forms the loop were present. One can do this by attaching springs to pairs of atoms in order to constrain their relative distances and preserve the planar configuration of a Watson–Crick base pair. The tbp's in the application presented in this work are A-T pairs. Springs were placed between the atoms N3-N1, O4-N6, O2-C2, and N1-N9, where the first atom name refers to the atom belonging to thymine (T), and the second to the atom in adenine (A), as shown in Figure 2.2(a). The constraints for the distances N3-N1, O4-N6, O2-C2, and N1-N9 were assigned a spring constant $k = 100\,\text{kcal/mol}\cdot\text{Å}^2$ (cf. $k \sim 300\,\text{kcal/mol}\cdot\text{Å}^2$ for C-C covalent bonds) and equilibrium distances of 2.9 Å, 2.88 Å, 3.71 Å, and 8.97 Å, respectively, which correspond to their distances in the ideal Watson–Crick base pair. A similar treatment can be adopted for a G-C base pair.

Local coordinate system at a tbp. Each tbp determines the boundary conditions of the elastic rod model of DNA by defining a coordinate frame, also referred to as the director basis, $\{\hat{\mathbf{d}}_1(s), \hat{\mathbf{d}}_2(s), \hat{\mathbf{d}}_3(s)\}$, at the location of the base pair, with origin $\mathbf{r}(s)$, according to a general convention [48] introduced below, illustrated in Figure 2.2(b). s is a coordinate along the DNA loop, defined in section 2.2 below. The tbp defining the beginning of the loop is denoted by a local frame of reference with origin at

$$\mathbf{r}(0) \tag{2.5}$$

and director basis

$$\{\hat{\mathbf{d}}_1(0), \hat{\mathbf{d}}_2(0), \hat{\mathbf{d}}_3(0)\}. \tag{2.6}$$

Likewise, the tbp defining the end of the loop is denoted by a local frame of reference with origin at

$$\mathbf{r}(1) \tag{2.7}$$

and director basis

$$\{\hat{\mathbf{d}}_1(1), \hat{\mathbf{d}}_2(1), \hat{\mathbf{d}}_3(1)\}. \tag{2.8}$$

The local frame of reference $\{\hat{\mathbf{d}}_1, \hat{\mathbf{d}}_2, \hat{\mathbf{d}}_3\}$ and origin $\mathbf{r}(s)$ at a base pair are obtained as follows: First, one chooses the strand of DNA that has the $5'$-$3'$ direction pointing from the beginning of the rod to the end of the rod. We call this the "defining strand." Second, one obtains the director basis. For this purpose, one defines a vector $\mathbf{r}_{C1'-C1'}(s)$ starting at the C1$'$ atom of the purine (R), i.e., A or G, and ending on the C1$'$ atom of the pyrimidine (Y), i.e., C or T

$$\mathbf{r}_{C1'-C1'}(s) = \mathbf{r}_{C1',R}(s) - \mathbf{r}_{C1',Y}(s). \tag{2.9}$$

$\hat{\mathbf{d}}_1(s)$ points in the direction of $\mathbf{r}_{C1'-C1'}(s)$,

$$\hat{\mathbf{d}}_1(s) = \frac{\mathbf{r}_{C1'-C1'}(s)}{|\mathbf{r}_{C1'-C1'}(s)|}. \tag{2.10}$$

$\hat{\mathbf{d}}_3(s)$ is defined normal to the plane formed by $\mathbf{r}_{C1'-C1'}(s)$ and the vector starting at the pyrimidine's C1$'$ atom and ending in the pyrimidine's C6 atom

$$\mathbf{r}_{C1'-C6'}(s) = \mathbf{r}_{C1',Y}(s) - \mathbf{r}_{C6,Y}(s), \tag{2.11}$$

i.e.,

$$\hat{\mathbf{d}}_3(s) = \pm\frac{\mathbf{r}_{C1'-C6'}(s) \times \mathbf{r}_{C1'-C1'}(s)}{|\mathbf{r}_{C1'-C6'}(s) \times \mathbf{r}_{C1'-C1'}(s)|}. \tag{2.12}$$

The sign of $\hat{\mathbf{d}}_3(s)$ is chosen such that it points along the $5'$-$3'$ direction of the defining strand, introduced above. Since the local director basis is an orthogonal right-handed set, $\hat{\mathbf{d}}_2(s)$ is defined as

$$\hat{\mathbf{d}}_2(s) = \pm\, \hat{\mathbf{d}}_3(s) \times \hat{\mathbf{d}}_1(s), \tag{2.13}$$

with the sign equal to that in (2.12). Finally, the origin of the local coordinate frame of reference is located at

$$\mathbf{r}(s) = \frac{1}{2}\, \left(\mathbf{r}_{C1',R}(s) + \mathbf{r}_{C1',Y}(s)\right) + \left(\mathbf{r}_{C1'-C6'}(s) \cdot \hat{\mathbf{d}}_2(s)\right)\, \hat{\mathbf{d}}_2(s). \tag{2.14}$$

The local coordinate systems $\hat{\mathbf{d}}_i(0)$ and $\hat{\mathbf{d}}_i(1)$, $i = 1, 2, 3$, and their respective origins $\mathbf{r}(0)$ and $\mathbf{r}(1)$ are obtained following the convention outlined above, with all vector coordinates given in the LCS.

Application of forces and torques in the tbp. The forces and torques due to the DNA resisting being forced into a loop are determined by the elastic rod model of DNA, as further detailed below, and are applied to selected atoms of each tbp, namely, the nonhydrogen atoms belonging to the bases (Figure 2.2(c)). The sugar and phosphate atoms are not subject to the forces. The force applied to each atom is the sum of the contribution from the total force $\mathbf{N}$ and the total torque $\mathbf{M}$ acting on the respective tbp. $\mathbf{N}$ and $\mathbf{M}$ are defined through their RCS counterparts $\mathbf{N}'$ and $\mathbf{M}'$ in (2.26) and (2.27), respectively. The stress $\mathbf{N}$ due to the DNA loop is applied to the center of mass of the group of participating atoms

$$\mathbf{n_i} = \frac{m_i}{\sum_i m_i}\, \mathbf{N}, \tag{2.15}$$

where $\mathbf{n_i}$ is the contribution of the force $\mathbf{N}$ to atom i, m_i is the mass of the atom, and i runs over all the atoms to which the force is applied (Figure 2.2(c)). The total torque $\mathbf{M}$ due to the DNA loop contributes forces ζ_i to each atom. These forces are defined through

$$\zeta_i = \frac{\mathbf{M} \times \rho_i}{\sum_i |\rho_{\perp i}|^2} \tag{2.16}$$

with $\sum_i \zeta_i = 0$, where $\rho_i = \mathbf{r}_i - \frac{1}{N_{tbp}} \sum_i \mathbf{r}_i$ is the distance from the atom to the geometrical center, $\mathbf{r}_i$ is the atomic position, N_{tbp} the number of atoms subject to the force, i.e., the nonhydrogen atoms in the bases, and $|\rho_{\perp i}| = \frac{|\mathbf{M} \times \rho_i|}{|\mathbf{M}|}$ is the perpendicular distance from atom i to the axis along which the torque is applied (Figure 2.2(d),(e)). Equations (2.15) and (2.16) determine the total force

$$\mathbf{F}_{ext} \;=\; \mathbf{n_i} + \zeta_i \tag{2.17}$$

applied to each atom using (2.2) at every time step of the simulation. The procedure outlined needs to be applied to each tbp separately.

2.2. Elastic rod model of DNA. In this section we describe the application of elasticity theory to DNA loops, along with two numerical methods used to obtain the structure of the protein-bound DNA loop.

Before we start, it is convenient to introduce the RCS, since all quantities in this section are given in this frame of reference. The RCS is determined by the boundary conditions of the loop, which are the key determinants of the loop geometry. These boundary conditions are given by the tbp's, as defined in section 2.1.2. According to Figure 2.1, the elastic rod theory description refers to the RCS frame of reference, which is different from the LCS system used in the MD simulation. We reiterate that the vectors in the RCS appear as primed. In the RCS system (cf. Figure 2.1), the boundary condition at the beginning of the loop is given by a local frame of reference with origin at

$$\mathbf{r}'(0) = 0 \tag{2.18}$$

and coordinate frame

$$\left[\hat{\mathbf{d}}'_j(0)\right]_k = \delta_{jk}, \tag{2.19}$$

where δ_{jk} is the Kronecker delta. One can readily see from Figure 2.1 that the boundary condition at the loop end then has the origin

$$\mathbf{r}'(1) = \frac{1}{l}\ (\mathbf{r}(1) - \mathbf{r}(0))\,, \tag{2.20}$$

and coordinate frame

$$\hat{\mathbf{d}}'_j(1) = \mathcal{O}^{-1}\hat{\mathbf{d}}_j(1), \tag{2.21}$$

where $\mathcal{O}^{-1}$ is the operator needed to transform the director basis $\{d'_1(0), d'_2(0), d'_3(0)\}$ given by (2.19) to $\{d_1(0), d_2(0), d_3(0)\}$ defined through (2.5)–(2.14); i.e., $\mathcal{O}$ is given by the orthogonal matrix

$$\mathcal{O} = \left(\hat{\mathbf{d}}_1(0), \hat{\mathbf{d}}_2(0), \hat{\mathbf{d}}_3(0)\right) \tag{2.22}$$

using an obvious notation. l denotes the length of the DNA loop. Since $\mathcal{O}$ furnishes an orthogonal transformation, it holds that $\mathcal{O}^{-1} = \mathcal{O}^{\mathbf{T}}$. The factor l^{-1} is introduced to normalize the coordinates in order to simplify the mathematical description, as explained below.

2.2.1. Formulation of the elastic rod model of DNA. The application of elasticity theory to DNA is referred to as the elastic rod model of DNA. The model is based on Kirchhoff's theory of elasticity [36, 39, 17], which represents DNA as an elastic rod described through its centerline $\mathbf{r}'(s) = (x'(s), y'(s), z'(s))$. $\mathbf{r}'(s)$ is a three-dimensional curve parametrized by its arclength s, and its cross section, described by a local coordinate system (director set) $\{\hat{\mathbf{d}}'_1(s), \hat{\mathbf{d}}'_2(s), \hat{\mathbf{d}}'_3(s)\}$. The cross sections are stacked along the centerline, with their vectors $\hat{\mathbf{d}}'_1$ and $\hat{\mathbf{d}}'_2$ lying in the plane of the cross section, and the vector $\hat{\mathbf{d}}'_3 = \hat{\mathbf{d}}'_1 \times \hat{\mathbf{d}}'_2$ normal to that plane, i.e., tangential to $\mathbf{r}'(s)$. The geometric features of the elastic rod model are presented in Figure 2.3. The director set $\{\hat{\mathbf{d}}'_1(s), \hat{\mathbf{d}}'_2(s), \hat{\mathbf{d}}'_3(s)\}$ uniquely defines the orientation of the cross section at each point s along the centerline. In the case of DNA, the centerline of the rod follows the axis of the DNA helix, and the cross section is defined following [48] at each Watson–Crick base pair (cf. Figure 2.3(b)) as described in Figure 2.2.

The equations of Kirchhoff's theory [36] were modified to account for the specific physical properties of DNA: (i) intrinsic twist to mimic DNA helicity, (ii) electrostatic charge of the phosphate groups, (iii) bending anisotropy toward the backbone and grooves, (iv) deformability, and (v) sequence-dependent bend and twist.

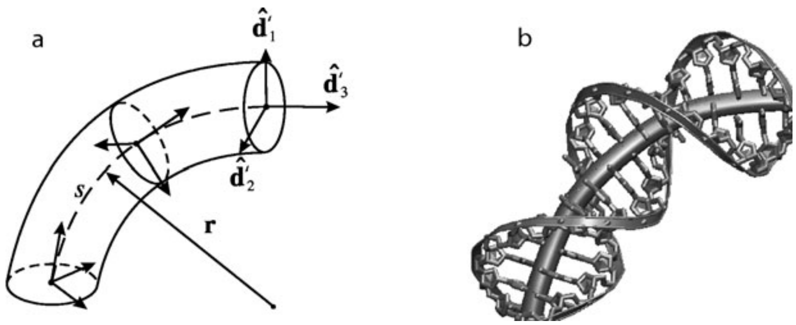

FIG. 2.3. *Elastic rod model of DNA.* (a) *Parametrization of the elastic rod, using the centerline position vector* $\mathbf{r}'(s)$ *and the director basis defined by the unit vectors* $\hat{\mathbf{d}}'_1(s,)$, $\hat{\mathbf{d}}'_2(s)$, *and* $\hat{\mathbf{d}}'_3(s)$, *where* $\hat{\mathbf{d}}'_3 = \hat{\mathbf{d}}'_1 \times \hat{\mathbf{d}}'_2$. (b) *The elastic rod (center tube) fitted to an all-atom structure of DNA.*

We consider the DNA to be inextensible and unshearable, imposing

$$\dot{\mathbf{r}'} = \hat{\mathbf{d}}'_3. \tag{2.23}$$

In this section the "dotted" derivatives, e.g., in $\dot{f}(s)$, are taken with respect to the arclength s, not time. The angular velocity of the local coordinate frame can be written as

$$\dot{\hat{\mathbf{d}}}'_i \; = \; \mathbf{k}' \times \hat{\mathbf{d}}'_i, \tag{2.24}$$

where $\mathbf{k}' = \{K_1, K_2, \Omega\}$ is the vector of strains and has as components the curvatures $K_1(s)$ and $K_2(s)$ and the local twist $\Omega(s)$ of the rod around its axis $\hat{\mathbf{d}}'_3$.[1]

For a proper description of DNA, the model needs to account for the intrinsic twist and curvature of DNA. The shape of ideal, straight DNA is helical with 10.4 bp per helix turn, corresponding to a pitch of 36 Å [14]. Comparing this value to the commonly accepted value $l_p = 500$ Å [26, 61] for the persistence length of DNA shows that DNA is tightly twisted. DNA is also known to be intrinsically curved [19] and twisted depending on its sequence [67, 18, 49]. These effects can be accounted for by including in the description the intrinsic curvatures $\kappa^{\circ}_{1,2}(s)$ and twist $\omega^{\circ}(s)$. The geometry of the rod can then be defined using deviations of curvature and twist from the intrinsic values

$$\kappa_{1,2}(s) = K_{1,2}(s) - \kappa^{\circ}_{1,2}(s), \qquad \omega(s) = \Omega(s) - \omega^{\circ}(s). \tag{2.25}$$

When the geometry of the rod departs from the intrinsic form, elastic forces $\mathbf{N}'(s)$ and torques $\mathbf{M}'(s)$ develop inside the rod:

$$\mathbf{N}'(s) = \sum_{i=1}^{3} N'_i \, \hat{\mathbf{d}}'_i, \tag{2.26}$$

$$\mathbf{M}'(s) = \sum_{i=1}^{3} M'_i \, \hat{\mathbf{d}}'_i, \tag{2.27}$$

where N'_1 and N'_2 are the shear forces, M'_1 and M'_2 the bending moments along the principal axes, N'_3 the force of tension (if $N'_3 > 0$) or of compression (if $N'_3 < 0$) at

[1] The components of $\mathbf{k}'$ exist in the RCS, yet for clarity will not appear primed.

the cross section, and M'_3 the twisting moment. At equilibrium, the elastic forces $\mathbf{N}'$ and torques $\mathbf{M}'$ balance the external forces $\mathbf{f}'$ and torques $\mathbf{g}'$ at every point s of the centerline, obeying

$$\dot{\mathbf{N}}' + \dot{\mathbf{f}}' = 0, \tag{2.28}$$

$$\dot{\mathbf{M}}' + \dot{\mathbf{g}}' + \dot{\mathbf{r}}' \times \mathbf{N}' = 0. \tag{2.29}$$

We adopt the widely used Bernoulli–Euler approximation [36, 39], which assumes linear dependence of the classic elastic torque on the curvature κ_1, κ_2 and twist ω:

$$\mathbf{M}'(s) = A_1\kappa_1\hat{\mathbf{d}}'_1 + A_2\kappa_2\hat{\mathbf{d}}'_2 + C\omega\hat{\mathbf{d}}'_3, \tag{2.30}$$

where A_1 and A_2 are the bending rigidities along the directions of the groove and backbone, respectively, and C is the twisting rigidity. The components of the vectors $\hat{\mathbf{d}}'_i(s)$ can be expressed through four Euler parameters or quaternions [39], $q'_i(s)$, in the elastic rod calculation frame of reference:

$$\hat{\mathbf{d}}'_1 = \ \{{q'_1}^2 - {q'_2}^2 - {q'_3}^2 + {q'_4}^2\,,\ 2(q'_1q'_2 + q'_3q'_4)\,,\ 2(q'_1q'_3 - q'_2q'_4)\}, \tag{2.31}$$

$$\hat{\mathbf{d}}'_2 = \ \{2(q'_1q'_2 - q'_3q'_4)\,,\ -{q'_1}^2 + {q'_2}^2 - {q'_3}^2 + {q'_4}^2\,,\ 2(q'_2q'_3 + q'_1q'_4)\}, \tag{2.32}$$

$$\hat{\mathbf{d}}'_3 = \ \{2(q'_1q'_3 + q'_2q'_4)\,,\ 2(q'_2q'_3 - q'_1q'_4)\,,\ -{q'_1}^2 - {q'_2}^2 + {q'_3}^2 + {q'_4}^2\}, \tag{2.33}$$

and are subject to the constraint

$${q'_1}^2 + {q'_2}^2 + {q'_3}^2 + {q'_4}^2 = 1. \tag{2.34}$$

Use of the Euler parameters avoids the polar singularities that arise when three Euler angles are used instead and, therefore, these parameters are preferred here.

Equations (2.23) and (2.26)–(2.34) form the basis of the Kirchhoff theory of elastic rods. We simplify the equations by making all variables dimensionless:

$$\bar{s} = s/l,\ \ \bar{x}' = x'/l,\ \ \bar{y}' = y'/l,\ \ \bar{z}' = z'/l, \tag{2.35}$$

$$\bar{K}_{1,2} = lK_{1,2},\ \ \bar{\Omega} = l\Omega, \tag{2.36}$$

$$\alpha = A_1/C_\circ,\ \ \beta = A_2/C_\circ,\ \ \gamma = C/C_\circ, \tag{2.37}$$

$$\bar{N}'_i = N'_i l^2/C_\circ,\ \ \bar{M}'_i = M'_i l/C_\circ. \tag{2.38}$$

Here l is the length of the rod and $C_\circ = 3 \cdot 10^{-19}$ erg·cm is the average intrinsic twist. We can now combine (2.23) and (2.28)–(2.34) into a system of nonlinear differential equations of 13th order. Here and below, the bars over the variables defined in (2.35)–(2.38) are dropped for simplicity:

$$\begin{aligned}((\alpha\ddot{\kappa}_1)) = {} & ((2\beta\dot{\kappa}_2\Omega)) - ((\gamma\dot{K}_2\omega)) - \beta\kappa_2\dot{\Omega} + \alpha\kappa_1\Omega^2 - \gamma K_1\omega\Omega + K_1N'_3 \\ & + \Omega\dot{g}'_2 - \dot{f}'_2 - \ddot{g}'_1,\end{aligned} \tag{2.39}$$

$$\begin{aligned}((\beta\ddot{\kappa}_2)) = {} & -((2\alpha\dot{\kappa}_1\Omega)) + ((\gamma\dot{K}_1\omega)) + \alpha\kappa_1\dot{\Omega} + \beta\kappa_2\Omega^2 - \gamma K_2\omega\Omega + K_2N'_3 \\ & - \Omega\dot{g}'_1 + \dot{f}'_1 - \ddot{g}'_2,\end{aligned} \tag{2.40}$$

$$((\gamma\dot{\omega})) = \alpha\kappa_1K_2 - \beta K_1\kappa_2 - \dot{g}'_3, \tag{2.41}$$

$$\dot{N}'_3 = -((\alpha\dot{\kappa}_1))K_1 - ((\beta\dot{\kappa}_2))K_2 - ((\gamma\dot{\omega}))\Omega - g'_1K_1 - g'_2K_2 - g'_3\Omega - \dot{f}'_3, \tag{2.42}$$

$$\dot{q}'_1 = \frac{1}{2}(K_1q'_4 - K_2q'_3 + \Omega q'_2), \tag{2.43}$$

$$\dot{q_2'} = \frac{1}{2}(K_1 q_3' + K_2 q_4' - \Omega q_1'), \tag{2.44}$$

$$\dot{q_3'} = \frac{1}{2}(-K_1 q_2' + K_2 q_1' + \Omega q_4'), \tag{2.45}$$

$$\dot{q_4'} = \frac{1}{2}(-K_1 q_1' - K_2 q_2' - \Omega q_3'), \tag{2.46}$$

$$\dot{x} = 2(q_1' q_3' + q_2' q_4'), \tag{2.47}$$

$$\dot{y} = 2(q_2' q_3' - q_1' q_4'), \tag{2.48}$$

$$\dot{z} = -{q_1'}^2 - {q_2'}^2 + {q_3'}^2 + {q_4'}^2. \tag{2.49}$$

Above we used the notation $\dot{((f\,g))} = \dot{f}g + f\dot{g}$ and its generalization. The solutions to this system correspond to the equilibrium conformations of the elastic rod approximating a DNA loop. The 13 functions that constitute a solution to the system (2.39)–(2.49), $\mathbf{r}'(s)$, $q_{1-4}'(s)$, $\kappa_{1,2}$, $\dot{\kappa}_{1,2}$, ω, and N_3', describe the geometry of the elastic rod and the distribution of the stress and torques along the rod. The elastic rod theory neglects the actual dynamics of loop formation. It states directly the shape of the loop after adjustment to the imposed boundary conditions (2.18)–(2.21), i.e., after equilibration.

For DNA, the external forces $\mathbf{f}'$ and torques $\mathbf{g}'$ introduced in (2.28) and (2.29), respectively, originate mainly from electrostatic interactions. The inclusion of electrostatics in the elastic rod model of DNA is considered in Appendix A. If electrostatic effects are not included in the DNA rod model, the external forces $\mathbf{f}'$ and torques $\mathbf{g}'$ are neglected in (2.39)–(2.42).

The forces $\mathbf{N}'$ and torques $\mathbf{M}'$ at the boundaries $s = 0, 1$ can be obtained from this solution and are communicated to the MD program. The shear forces N_1', N_2' can be obtained by combining (2.23)–(2.26) and (2.30) with (2.29); taking $\dot{\mathbf{g}'} = 0$ in (2.29), one obtains

$$N_1' = -\beta\dot{\kappa}_2 + (1-\alpha)\kappa_1\omega - \alpha\kappa_1, \tag{2.50}$$

$$N_2' = \alpha\dot{\kappa}_1 + (1-\beta)\kappa_2\omega - \beta\kappa_2. \tag{2.51}$$

The force of tension or compression N_3' results directly from the solution of (2.39)–(2.49), and the components of the torques M_i', $i = 1, 2, 3$, are directly obtainable from $\mathbf{k}'(s)$ by virtue of (2.30).

Equations (2.39)–(2.49) can be solved for various given boundary conditions, as explained below. In fact, for proteins that force DNA into loops, (2.39)–(2.49) lead to a boundary value problem, where the equilibrium geometry of the loop is obtained for fixed ends of the DNA loop $\mathbf{r}'(0)$ and $\mathbf{r}'(1)$, and fixed orientation of the cross section at these ends, given by $q_i'(0)$ and $q_i'(1)$, $i = 1, 2, 3, 4$, or equivalently, $d_i(0)$ and $d_i(1)$, $i = 1, 2, 3$. The quantities $\mathbf{r}'(0), \mathbf{r}'(1)$ and $q_i'(0), q_i'(1)$ are known from the all-atom structure of the DNA bound to the protein (see section 2.1.2).

The differential equations (2.39)–(2.49) for specific boundary conditions might yield multiple solutions. It is then necessary to consider the solution that minimizes the elastic energy, since such a solution would be predominantly represented in a thermodynamic ensemble. The elastic energy of each solution is computed in proportion to the square of the geometric deviation from the reference configuration, in accordance with the Bernoulli–Euler approximation (2.30)

$$U = \frac{1}{2}\int_0^l (A_1\kappa_1^2 + A_2\kappa_2^2 + C\omega^2)ds. \tag{2.52}$$

2.2.2. Numerical solution to the elastic rod problem. In order to solve the system of ordinary differential equations (2.39)–(2.49) of the rod problem, we use the boundary value problem solver COLNEW [6] which employs a damped quasi-Newton method to construct the solution to the problem as a set of collocating splines.

The solver needs to be provided with a guess for an approximate solution, and a parameter or set of parameters is gradually changed in an iterative manner in order to reach the solution for the system. There are two methods for iteratively finding the solution for the rod structure depending on the guessed solution provided to the solver. The two methods are presented in the following sections "Initiation" and "Continuation."

We recall at this point that the boundary conditions of the elastic rod problem are stated in the RCS systems through (2.18)–(2.21).

Initiation. When no information on the structure of the loop is available, the initial guess consists of a known exact solution $\mathbf{r}'(s)$ to the system (2.39)–(2.49) characterized through parameters, e.g., boundary conditions, different from the desired ones. The desired solution is obtained in several rounds of computing in which the deviant parameters are adjusted one by one to the correct value. Typically, the initial guess deviates from the desired solution in terms of boundary position $\mathbf{r}'(1)$, orientation $\mathbf{d'_i}(1)$, as well as A_2/A_1, the ratio of the bending rigidities (cf. (2.30)). The computations can be made in the following order:

1. *Translation.* A solution with a deviant position $\mathbf{r}'(1)$ (defined in (2.20)) is assumed for the initiation and gradually changed until the position corresponding to the correct boundary condition is reached.

2. *Rotation.* The initial solution usually also assumes an incorrect orientation of the tbp at $s = 1$, specified by $\left\{\hat{\mathbf{d}}'_1(1), \hat{\mathbf{d}}'_2(1), \hat{\mathbf{d}}'_3(1)\right\}$. This local frame is gradually rotated so that eventually it coincides with the frame imposed by the boundary conditions of the problem. This step is achieved by simultaneously turning the normal $\hat{\mathbf{d}}'_3$ to coincide with the normal of the boundary condition defining base pair and rotation about this normal in order to align the cross sections $(\hat{\mathbf{d}}'_1, \hat{\mathbf{d}}'_2)$.

3. *Rotation about* $\hat{\mathbf{d}}'_3$. The elastic rod model presented here does not permit one to define the linking number of the loop. Therefore, integral turns about the normal $\hat{\mathbf{d}}'_3$ yield the same boundary conditions. Rotations by 2π are performed in order to explore other possible solutions.

4. *Anisotropic flexibility.* In many polymer physics and DNA biology applications the bending rigidities, A_1 and A_2 in (2.30) and (2.52), are assumed to be isotropic, e.g., $A_1 = A_2 = k_BTl_p$, where k_B is Boltzmann's constant, T is the absolute temperature, and $l_p = 500\,\text{Å}$ is the persistence length of the rod. Similarly, the twisting rigidity C is defined in terms of a twisting persistence length $C = k_BTl_{twist}$ using $l_{twist} = 750\,\text{Å}$ [26, 61]. The bending persistence length l_p used above is defined only for isotropically bendable rods. However, anisotropic bending rigidities must be accounted for in order to yield a correct model of DNA. This is evident from the DNA structure: bending toward the grooves should require less energy than bending toward the backbone. Therefore, for the present study we assume anisotropic bending rigidities characterized through $\mu = A_2/A_1 = 4$ [9]; this value reproduces well the DNA persistence length in Monte Carlo simulations [50]. The relaxed structure is described by the intrinsic components $\kappa^{\circ}_{1,2} = 0$ and $\omega^{\circ} = 34.6^{\circ}$ per base pair. Sequence-dependent effects on curvature or twist can be accounted for but are not included in the application presented here. Solutions for anisotropic bending rigidities are usually obtained starting from a solution for isotropic rigidities (i.e., A_2/A_1) and

gradually adjusting rigidities until the desired $\mu = A_2/A_1$ ratio is reached.

In a certain sense the initiation construction of the solution to (2.39)–(2.49) reflects a relaxation process of a DNA loop from an initial form to the desired final form. However, no timescale is linked to the process and, in fact, it is assumed that the desired shape of the DNA loop adjusts itself instantaneously once the boundary conditions and A_2/A_1 ratio have been stated. As noted earlier the description adopted also neglects fluctuations (entropy effect) around the equilibrium solution.

Continuation. When a good guess for the structure is available, the rod calculation may be performed in a single cycle of iteration. For the case of the multiscale method, as explained below, this situation applies. We denote the centerline of the loop by $\mathbf{r}'(s,k)$, where k counts the progress of the simulation as specified below. The solution $\mathbf{r}'(s,k)$ at step k is generally a good initial guess for the solution at the next step $k+\delta k$, since the changes in the boundary conditions, i.e., relative positions $\mathbf{r}'(1)$ and orientation $\mathbf{d}'_\mathbf{i}(1)$ given by (2.18)–(2.21), are expected to be small during single steps if chosen appropriately small. The single iteration cycle performs steps 1 and 2 of "Initiation" (see above) simultaneously. Step 3 of "Initiation" is not performed since it is desirable to keep the topology, i.e., linking number, of the loop unchanged. Furthermore, the loop determined at step k already accounts for the correct ratio A_2/A_1 of bending rigidities; i.e., step 4 of "Initiation" is not necessary. The solution of the rod calculation using the continuation method at step $k+\delta k$ then starts from $\mathbf{r}'(s,k)$ and is obtained by providing the solver with the new target boundary condition $\mathbf{r}'(1,k+\delta k)$, $\hat{\mathbf{d}}'_i(1,k+\delta k)$, obtained from the latest structure of the protein in the MD simulation, and an initial guess for the solution of the system of equations given by the elastic rod solution at step k, $\mathbf{r}'(s,k)$, that corresponds to boundary conditions $\mathbf{r}'(1,k)$ and $\hat{\mathbf{d}}'_i(1,k)$. $\mathbf{r}'(1)$ and $\hat{\mathbf{d}}'_i(1)$ are then simultaneously changed iteratively until the desired boundary conditions $\mathbf{r}'(1,k+\delta k)$, $\hat{\mathbf{d}}'_i(1,k+\delta k)$ are met, yielding a new structure of the loop with geometry described by $\mathbf{r}(s,k+\delta k)$.

The continuation construction reflects the mechanical relaxation of the DNA loop during a certain period of protein dynamics, chosen below (cf. (2.55)) as 10 ps. The multiscale method assumes that this relaxation adapting to the new boundary conditions occurs instantaneously. The neglect of the actual dynamics of DNA and solvent leads to great savings in computing effort—much more than due to mere reduction in atom count (in the presented example of the *lac* repressor-DNA complex, from 700,000 to 230,000), as the underlying complex motion involving, in particular, water molecules and ions is not explicitly described. The simplification comes at the cost of losing the timescales for the loop dynamics stemming from effects of inertia, friction, and other factors [7].

2.2.3. Forces and torques. Once a shape of the loop $\mathbf{r}'(s)$ has been established, one can determine the forces $\mathbf{N}'(0), \mathbf{N}'(1)$ and torques $\mathbf{M}'(0), \mathbf{M}'(1)$ at the loop termini in the RCS frames, according to (2.26)–(2.27), (2.30), and (2.50)–(2.51). In the LCS system the forces are

$$\mathbf{N}(s) = \mathcal{O}^{-1}\,\mathbf{N}'(s), \tag{2.53}$$

and the torques are

$$\mathbf{M}(s) = \mathcal{O}^{-1}\,\mathbf{M}'(s), \tag{2.54}$$

with $\mathcal{O}$ given by (2.22). These forces are needed for $s=0$ and $s=1$ as input for the MD simulation. We recall the property $\mathcal{O}^{-1} = \mathcal{O}^{\mathbf{T}}$.

2.2.4. Equivalent all-atom models of the DNA loop. An important advantage of the elastic rod model of DNA is that it has the capacity to recover atomistic detail from the solutions of (2.39)–(2.49), i.e., from the computed equilibrium shapes of the rod for given boundary conditions. In fact, one can construct a full-atom structure of the entire protein-DNA complex under study. One such example is presented in Figure 1.1 for the case of the application presented here, the *lac* repressor-DNA complex. The obtained all-atom structures can then be employed for MD simulations of the whole complex. The steps for creating a full-atom structure of a DNA loop given a loop geometry are outlined in Appendix B.

2.3. Linking molecular dynamics and elastic rod model. The methods presented here serve to study proteins that bind to DNA, inducing it to loop or coil. The proteins, and the DNA segments in direct contact with the protein, are described by MD, as introduced in section 2.1. The DNA loops formed between the protein-bound DNA segments are described by the elastic rod model, as introduced in section 2.2. Here we present the link between these two descriptions.

Figure 2.4 illustrates the system on which the two computations are done in intertwined steps. The box shown in Figure 2.4 holds the protein, DNA segments bound to the protein, water, and ions; in the case of the *lac* repressor-DNA complex the box includes altogether 230,000 atoms. The molecules in the box are treated by means of atomic-level MD simulations. From the box emanates the DNA loop that is treated via elastic rod theory. The coupling of the two descriptions is now specified.

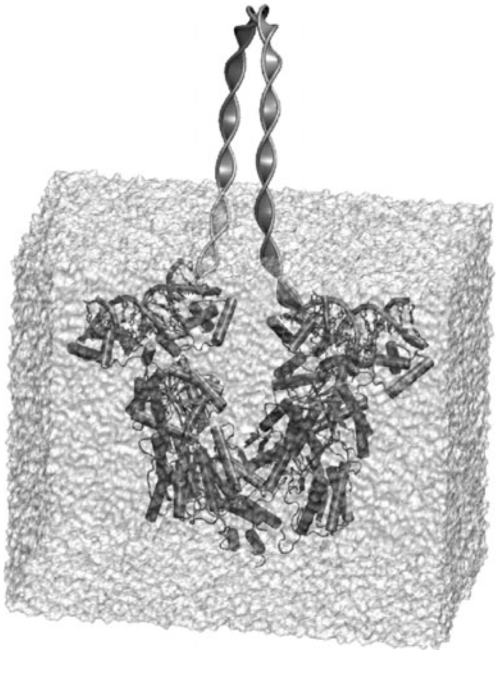

FIG. 2.4. *Multiscale simulation of the* lac *repressor-DNA complex. The box represents the system studied by MD, including the modeled full-atom structure of the* lac *repressor bound to two* 19 *bp DNA segments, placed in a water box with sodium and chloride ions, for a total of* 230,000 *atoms. The loop arising from the box represents the DNA loop modeled by means of elasticity theory. The multiscale method consists in coupling these two methods, as explained in section* 2.3.

The underlying computational strategy of the multiscale method is the following: The MD program calls the elastic rod calculation every 10,000 integration steps:

$$10{,}000\ \delta t = \delta k. \tag{2.55}$$

For $\delta t = 1\,\mathrm{fs}$ this implies $\delta k = 10\,\mathrm{ps}$. The boundary conditions needed for the elastic rod calculation, i.e., position $\mathbf{r}'(1)$ and orientation $\mathbf{d}'_{\mathbf{i}}(1)$, defined through (2.20) and (2.21), respectively, are provided by the MD simulation of the protein-DNA complex, as specified by (2.5)–(2.14).

Once the rod calculation has been provided with the boundary conditions, it needs also to be provided with a guess for the initial solution. For the calculation at step k, the immediate previous solution $\mathbf{r}(s, k - \delta k)$ of (2.39)–(2.49) is taken as the initial guess for the rod calculation, and the continuation method is used to obtain a new solution for the geometry of the rod $\mathbf{r}(s, k)$. The forces $\mathbf{N}$ and torques $\mathbf{M}$ arising for the loop $\mathbf{r}(s, k)$ in the RCS given by (2.26) and (2.27) are then returned to the MD simulation by converting them into the LCS by means of (2.53) and (2.54) and then using (2.15)–(2.17) to apply the corresponding forces to the relevant atoms of each tbp for $10{,}000\,\delta t$, after which a new continuation rod calculation must be made, initiating the next cycle of the intertwined elastic rod and MD simulations. We recall that the adaptation of the DNA loop to the protein dynamics within 10 ps intervals is assumed to be instantaneous.

3. The *lac* repressor protein and its induced DNA loop.

We applied the multiscale method outlined in section 2 to study the *lac* repressor-DNA complex. The *lac* repressor is a tetrameric protein, with each monomer consisting of 360 amino acids. It is formed as a “dimer of dimers,” with each dimer consisting of a core and a DNA binding head group (Figure 1.2). The dimers are associated by a four-helix bundle at the bottom, adopting a “V-shape” conformation. Crystallographic as well as NMR structures have been determined for this protein and were recently reviewed in [11]. None of the available structures describe the loop that the protein induces in the DNA, due to the difficulty of crystallizing proteins with full DNA loops. The *lac* repressor can form loops of 76 bp and 384 bp. In this study we focus on the shorter loop (Figure 1.1). Since the length of the loop is only half the $\sim$147 bp corresponding to the persistence length of DNA [26, 61], enthalpic effects should dominate over entropic effects and, thus, we may neglect the latter, describing the loop by means of the theory of elasticity presented in section 2.2.

3.1. Equilibrating the *lac* repressor protein without the DNA loop.

In order to perform the MD simulation, a full-atom structure of the *lac* repressor-DNA complex was needed. The available crystal and NMR structures of the *lac* repressor describe only parts of the protein in atomistic detail. For example, the structure by Lewis et al. [34] contains the full tetramer but no coordinates for the amino acid side chains. We constructed an all-atom structure of the *lac* repressor-DNA complex employing relevant entries from the Protein Data Bank (PDB) [15]. We used the 1LBI structure [34] as a scaffold, to which we aligned two copies of the 1EFA dimer structure [12], patching the head groups using the 1CJG head group structure [60] and taking the symmetric DNA segments from the 1LBG head group-DNA structure [34]. A complete description of our construction can be found in [10].[2] The obtained model consists of a full-atom structure of the tetramer bound to symmetric DNA operators, as shown in Figure 1.2. The new structure does not contain the DNA loop.

[2] The structure will be made available in a forthcoming publication [64].

The modeled protein structure did not contain buried water molecules that occur *in vivo*. The program DOWSER [28] was used to place a total of 387 water molecules inside the protein and in external crevices of the protein. The program VMD [31] was subsequently used to place the protein model in a box of TIP3 water molecules, with selected water molecules replaced by sodium and chloride ions corresponding to a total ion concentration of 100 mM. The ions were initially distributed according to the electrostatic map obtained with the Poisson–Boltzmann solver DelPhi [29]. The resulting system of 230,000 atoms was minimized for 4000 conjugate gradient steps, then equilibrated using the NAMD2 MD program [32] with the CHARMM22 force fields for the energy parameters [37] for 1.8 ns (nanoseconds) with a 1 fs time step. The simulation proceeded in the so-called NPT ensemble, i.e., with particle number N, pressure P, and temperature T being held fixed. The temperature was fixed at 298.15K and the pressure at 1 atmosphere (NPT) using the Langevin piston method [22] with a damping coefficient of $5\,\mathrm{ps}^{-1}$ and a piston period of 100 fs. The Particle Mesh Ewald (PME) method was used for computing electrostatic forces without cut-off [20]. The grid spacing was kept below 1 Å, and a fourth order spline was used for the interpolation, with the long-range part of the electrostatics being evaluated every fourth step. The van der Waals interactions were cut off at 12 Å via a switching function starting at 10 Å. Full periodic boundary conditions were imposed.

In a first phase of equilibration, the backbone atoms of the protein and DNA were harmonically constrained, except for those in the newly built protein regions connecting the available structures. The constraints were gradually released to allow for the amino acids in the newly built parts to avoid nonphysical configurations. During the equilibration the protein backbone showed an average root mean squared deviation (RMSD) of 1.3 Å with respect to the originally built structure, which is small compared to the results from typical MD simulations starting from crystal structures, and implies that the protein is very stable in the predicted structure and that the equilibration achieved here provides a good starting point for further simulation.

3.2. The DNA loop induced by the *lac* repressor. The elastic rod model was used to build structures of the missing loop that connects the DNA segments that are bound to the *lac* repressor. The modeled and equilibrated *lac* repressor structure, described in section 3.1, contains two protein-bound DNA segments of 19 bp each. As explained above, the tbp's (Figure 1.2) are taken as the ends of the elastic rod, providing the coordinates $\mathbf{r}(0), \mathbf{r}(1)$ and orientation $\hat{\mathbf{d}}_i(0), \hat{\mathbf{d}}_i(1)$, $i = 1, 2, 3$, of the boundaries used in the elastic rod calculation.

In the case of the 76 bp loop, the *lac* repressor binds to two segments of DNA denoted operators O1 and O3. The sequence of the operators has pseudopalindromic symmetry (two-fold symmetry broken by the insertion of a central G-C base pair). This permits orientations of the DNA in each of the head groups in either the 5′-3′ or 3′-5′ direction. Figure 3.2 shows the four possible arrangements of tbp's that result for the possible orientations; these arrangements define the boundary conditions for the rod calculation, and the different topologies of the loop, as suggested in [24]. We use the notation for the orientation of the DNA in the head groups introduced in [62], where I denotes the 5′-3′ direction pointing toward the protein (inside), and O the direction away from the other head group (outside). This yields four possible combinations: II, IO, OI, OO, where the first letter denotes the orientation of the operator O3 and the second the orientation of the operator O1. In the present study, the II and OO topologies are equivalent since the model doesn't presently account for sequence-dependent curvatures and twist of the DNA loop. The loop structures were obtained for all of these sets of boundary conditions, as illustrated in Figure 3.2. In the following section, we consider only the case IO. The other cases will be briefly discussed below.

3.2.1. Adding the DNA loop to the *lac* repressor with IO topology. Given the structure of the equilibrated protein, but not of the DNA loop, one can build a structure of the loop using the initiation method described in section 2.2.2 and illustrated in Figure 3.1.

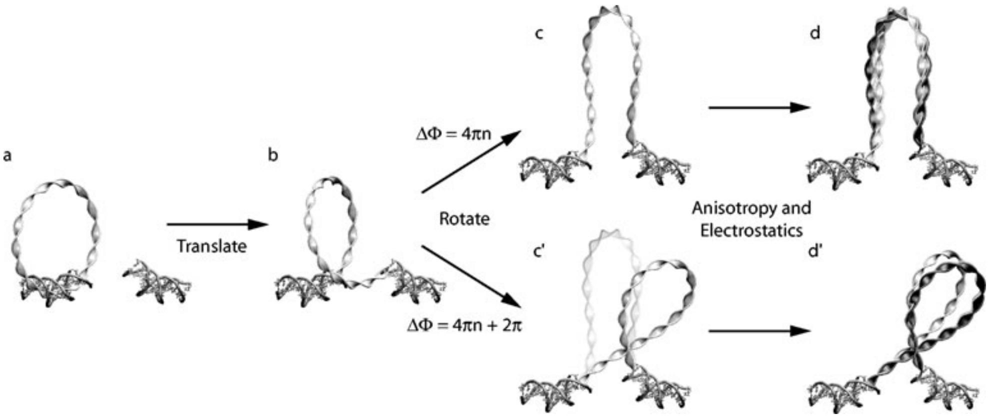

FIG. 3.1. *Initiation method for the elastic rod structure of the DNA loop formed by the* lac *repressor protein.* (a) *Initial solution: a closed circular loop.* (b) *Solution after translation of the end of the rod to the correct coordinates.* (c,c′) *The solutions after the second iteration cycle, rotation of the $s = 1$ end; solutions are shown for the underwound "U" loop* (c) *and overwound "O" loop* (c′). (d) *Solutions after the third and fourth iteration cycles that include the effect of anisotropy and electrostatics. Previous solutions are shown in light gray. The protein-bound DNA segments from the* lac *repressor crystal structure are shown for reference only; they played no other role during the iteration cycles than that of providing the boundary conditions.*

The starting point is a circular closed elastic loop ($\mathbf{r}'(0) = \mathbf{r}'(1)$) shown in Figure 3.1(a), with intrinsic curvature $\kappa^\circ_{1,2} = 0$, constant intrinsic twist $\omega^\circ = 34.6^\circ$ per base pair, constant elastic moduli $A_1 = A_2 = \frac{1}{2}C$, and zero electrostatic charge. At $s = 0$, the loop has the trivial correct position $\mathbf{r}'(0) = 0$ and orientation $\hat{\mathbf{d}}'_i(0)$ given by (2.19) (Figure 3.1(a)). In the first step, the value of $\mathbf{r}'(1) = 0$ is changed so that the end of the rod moves to the terminal base pair in the other head group, corresponding to a translation by 51.3 Å (Figure 3.1(b)). Subsequently, the local frame at $s = 1$ is rotated to meet the orientation of the terminal base pair $\hat{\mathbf{d}}_i(1)$ (Figure 3.1(c)). Rotations by 2π about $\hat{\mathbf{d}}_3(1)$ were performed in order to explore other possible solutions. A first 2π rotation yielded a new solution, represented in Figure 3.1(c′). This loop is overwound (O) with respect to the intrinsic twist by an average of 1.32° per base pair. The former solution is underwound (U) by −1.24° per base pair. A further rotation by 2π returns to the first solution. This can be explained through the occurrence of a self-crossing of the loop during this second rotation; topologically, a rotation by 4π increases the linking number by 1, while a self-crossing reduces it by 1, returning the loop to the original solution. Although nonphysical, self-crossings can occur in the model since volume exclusion is not accounted for.

In the next step, the values of the bending moduli were changed from $A_1 = A_2 = \frac{1}{2}C$ (with $C = 3 \cdot 10^{-19}\,\text{erg·cm}$) to the values $A_1 = \frac{4}{15}C = 0.8 \cdot 10^{-19}\,\text{erg·cm}$ and $A_2 = \frac{16}{15}C = 3.2 \cdot 10^{-19}\,\text{erg·cm}$ [50].

The resulting O loop involves an elastic energy of $23.5k_B$T, while the U loop has a higher elastic energy of $30.43k_B$T (cf. the experimental value of $20k_B$T [30]; we employed (2.52) in the calculation).[3] The elastic stress of the U loop results in forces of 7.6 pN (picoNewtons) pushing the ends of the DNA (and thus the head groups of the protein) away from each other. The O loop induces forces of 7.16 pN pointing toward each other, i.e., in the opposite direction of the forces of the U loop, bringing the head groups together. In evaluating these forces we employed (2.50)–(2.51), and the component N_3 is directly obtained from the solution of (2.39)–(2.49).

The electrostatic effects on the structure of the DNA loop induced by the *lac* repressor-DNA complex are discussed in Appendix A.

3.2.2. Alternative configurations of the DNA loop. We determined also solutions to the DNA loop for other topologies that the *lac* repressor may induce, i.e., II, IO, OI, OO. The resulting structures are shown in Figure 3.2. The values for the energies, excess twist, and resulting forces of these structures are given in Table 3.1.

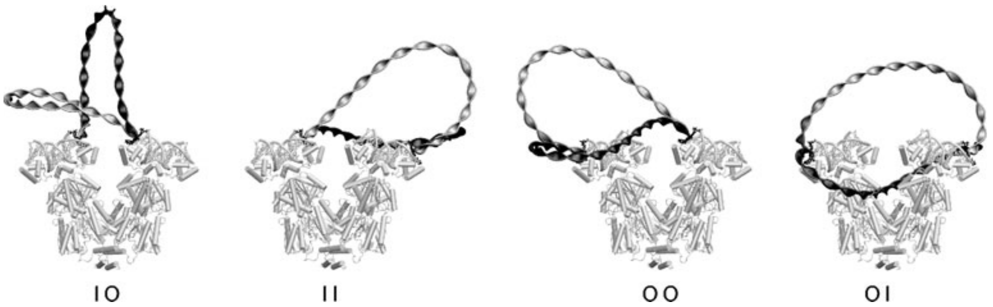

FIG. 3.2. *Other topologies of the DNA loop formed by the* lac *repressor based on orientation of the operators in the head groups. I denotes the* 5′-3′ *direction pointing toward the protein (inside), and O denotes the* 5′-3′ *direction pointing away from the other head group (outside). Notation is adopted from* [62]*. The* II *and* OO *topologies are equivalent in our model.*

TABLE 3.1
Energies and end forces of the loops induced by the lac *repressor for the different possible topologies of the loop. I denotes the* 5′-3′ *direction pointing toward the protein (inside), and O away from the other head group (outside). Notation is adopted from* [62]. ω *denotes the twist, as defined in* (2.25)*,* $U_{elastic}$ *is the elastic energy defined in* (2.52)*,* U_Q *is the electrostatic energy in* (A.5)*,* U_{total} *is the sum of both contributions to the energy as included in* (A.4)*, and* $N(s)$ *is the magnitude of the forces at the end points of the DNA loop* ($s = 0, 1$) *arising from the elastic rod calculation (cf.* (2.28)*). For each topology, two solutions, labeled* a *and* b*, are characterized.*

Topology		ω (deg/bp)	$U_{elastic}$ (k_BT)	U_Q (k_BT)	U_{total} (k_BT)	$N(s=0)$ (pN)	$N(s=1)$ (pN)
IO	a	−1.24	23.52	0.05	23.57	7.71	7.73
IO	b	1.32	30.62	0.42	31.04	7.64	6.72
II	a	−1.25	23.32	0.07	23.39	6.47	6.54
II	b	1.15	19.29	0.05	19.34	5.58	5.6
OO	a	−1.00	19.42	0.03	19.45	5.03	5.05
OO	b	1.01	22.74	0.03	22.78	6.87	6.86
OI	a	−0.23	41.62	0.16	41.79	7.88	8.22
OI	b	−0.04	27.03	0.04	27.07	6.57	6.62

[3] The difference between energy values in [9] and those reported here is due to the MD equilibration of the *lac* repressor structure performed here.

The orientation of binding of the *lac* repressor to the operators has been assumed by many authors to be of the IO kind; see, e.g., [34, 21]. Friedman, Fischmann, and Steitz introduced the "wrapping away" loop [24], which corresponds to II and OO loops. At physiological salt concentrations, these loops have energies comparable to that of the IO loop and to the experimental value, and thus should also be counted among possible *in vivo* configurations of the loop. The IO, II, and OO boundary conditions all show an underwound (U) and an overwound (O) solution (Figure 3.2). The underwound solution yields lower energy for the II and IO loops, whereas the overwound solution yields lower energy for the OO loops. The OI boundary condition yields two underwound loops. One of these loops involves overlap with the *lac* repressor protein and, thus, should be discarded. It is possible that this configuration of the loop would wrap around the protein if the looping DNA segment was longer, as in the "wrapping towards" model introduced by Friedman, Fischmann, and Steitz [24].

3.3. Turning on the loop-protein interaction: The multiscale application. The multiscale simulation for the *lac* repressor-DNA complex was performed using the IO loop with the U topology. Here, we present the results for 1 ns of simulation, which demonstrate the feasibility of the multiscale method for studying protein-DNA complexes.

During the simulation time, the overall structure of the protein remained stable. Here and henceforth, all RMSD values are reported for the backbone atoms with respect to the equilibrated structure. The protein showed an average RMSD of 1.7 Å, a small value for MD simulations of proteins. We analyzed different regions of the protein for structural distortion due to the application of the forces. The core of the protein keeps its structural features and remains in the V-like configuration, with an average RMSD of 1.3 Å. The overall structure of the head groups is also preserved, with an average RMSD of 1.4 Å. The protein-bound DNA maintains its structure as well as its contacts with the protein. The stability of the protein during the simulation shows that the multiscale simulation is capable of describing equilibrium behavior of the protein and suggests that the observed crystal and NMR structures of the *lac* repressor (1LBI [34], 1EFA [12], 1CJG [60], and 1LBG [34]) are relevant, despite the missing DNA loops.

The difference between the RMSD of the overall structure and the individual components of the protein suggests relative motions between its parts (cf. Figure 3.3). The flexible linker regions that connect the core of the protein to the head groups have been proposed to permit high mobility of the latter [24, 34]. This was confirmed,

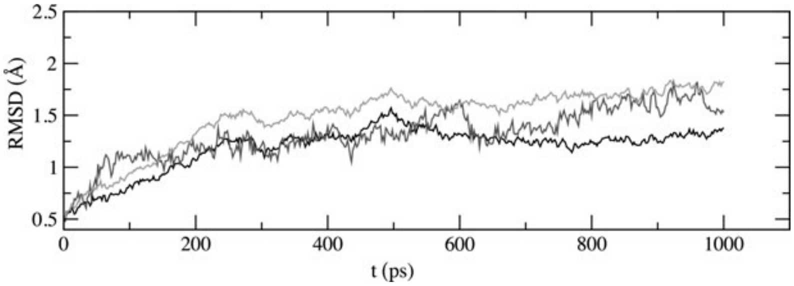

FIG. 3.3. *RMSD of the* lac *repressor structure during the multiscale simulation. Black represents the core of the protein, dark gray one of the head groups, and light gray the whole protein. RMSD is taken with respect to the equilibrated structure.*

as relative motions were found to occur between the head groups and the core. The change in the structure of the protein observed in the simulation suggests that, although the protein is very stable, there is a tendency for the head groups to move in the direction of the forces. Possibly, the *lac* repressor-DNA complex reaches its equilibrium configuration on a longer timescale than covered in the present simulation. A longer multiscale simulation of the complex will be presented in a forthcoming publication [64].

4. Conclusion and outlook. A general multiscale methodology for studying biopolymers in their environment is unlikely to materialize. For the broad range of biomaterials, as well as the wide length and timescales covered by cell processes, each class of problems will require a specific physical model with a specific computational framework [27]. A multiscale methodology for simulating protein-DNA complexes that include looped or coiled DNA is presented here. An example of a cellular mechanism that involves DNA loop formation is gene control. Interestingly, it has been proposed that the mechanics of the processes involved in gene control are shared by all proteins acting on DNA [45]. Therefore, the methodology presented here permits access to wider studies of protein-DNA interaction.

The elastic rod model used for the multiscale methodology provides a universal description of DNA properties and interactions [10]. In our present treatment, electrostatic interactions of DNA in the loop were disregarded. However, these properties may be important for the structure of the DNA loop and should be carefully considered on a case-by-case basis in future treatment. The multiscale model can be further extended, to account for the dynamics of the DNA loops, through implementing a Brownian dynamics model of DNA, as discussed in [7].

Another important aspect of the multiscale methodology is that it has the capacity to recover atomistic detail from the elastic rod calculations; i.e., one can obtain a full-atom structure of the entire protein-DNA complex, as mentioned in section 2.2.4 and outlined in Appendix B. The obtained structures can be employed in future MD simulations. Moreover, protein-DNA aggregates consisting of more than two macromolecules can be modeled. For example, in the case of the *lac* repressor-DNA complex, a second protein, CAP, is known to bind to the *lac* repressor-DNA complex. An MD simulation of the ternary *lac* repressor-CAP-DNA complex can be performed starting from a description furnished by the elastic rod model as described in [10]. Alternatively, one could perform a multiscale simulation of the ternary complex by combining an all-atom simulation of the *lac* repressor-DNA complex and one of the CAP-DNA complex, connected by the elastic rod model description of the rest of the loop. This can be extended to large protein-DNA aggregates where multiple proteins act at far locations on the same DNA loop, as is the case in gene repression, gene transcription, DNA replication, and DNA packing, processes in which the size of the system and long timescales do not permit full-atom simulations.

The suggested multiscale method was successfully applied to a landmark system in modern biology, the *lac* repressor-DNA complex. The simulation on the *lac* repressor-DNA complex suggests mobile head groups and a stable protein core. On the timescale of the simulation, a significant change in overall structure was not observed. A longer multiscale simulation of the *lac* repressor-DNA complex can identify the key mechanical degrees of freedom of the protein. The structure of the *lac* repressor suggests potential candidates: the bottom helix bundle is thought to act as a hinge, since it is attached to the core parts of the protein by flexible linkers. Experimental data suggests that the protein can undergo a change from the V-shape to an open conformation [57], not observed so far in the simulation. The DNA binding head groups are also connected to the core of the protein by flexible linker regions. Predicted flexibility of these groups in response to an elastic strain of DNA has been confirmed by the simulation. To what extent these play a role in repression and why the molecule is designed in this way are some of the issues that will be addressed in a forthcoming publication [64].

Appendix A. Electrostatics of the DNA elastic rod model. In this appendix we consider electrostatic contributions in the framework of the DNA theory of elasticity. We first introduce the implementation of electrostatics in the theory and subsequently discuss the electrostatic effects for the *lac* repressor-DNA complex in particular.

A.1. Changes to the equations of elasticity. Electrostatic interactions cause external forces $\mathbf{f}'$ and torques $\mathbf{g}'$ to arise in DNA (see (2.28) and (2.29)). These forces and torques originate from self-repulsion of the rod, since DNA holds a negative charge in every phosphate of its backbone, and from electrostatic effects arising from the interaction with other charged bodies (e.g., proteins). The electrostatic forces are described by

$$\dot{\mathbf{f}'}_Q(s) = \sigma(s)\mathbf{E}'(\mathbf{r}'(s)), \tag{A.1}$$

where $\sigma(s)$ is the electrostatic charge density of the rod and $\mathbf{E}'$ is the electrostatic field computed using the Debye screening formula

$$\begin{aligned}\mathbf{E}'(\mathbf{r}'(s)) = \frac{1}{4\pi\epsilon\epsilon_\circ}\Bigg(&\sum_i q_i \nabla \frac{\exp(-|\mathbf{r}'(s)-\mathbf{R}'_i|/\lambda)}{|\mathbf{r}'(s)-\mathbf{R}'_i|}\\ &+ 2e\chi {\sum_j}' \nabla \frac{\exp(-|\mathbf{r}'(s)-\mathbf{r}'(s_j)|/\lambda)}{|\mathbf{r}'(s)-\mathbf{r}'(s_j)|}\Bigg).\end{aligned} \tag{A.2}$$

Here the first term describes the interaction of the rod with other charges present in the system, and the second term originates from the self-interaction of the rod, the term $s = s_j$ excluded from the sum; $\lambda = 3\text{Å}/\sqrt{c_s}$ is the Debye screening length in an aqueous solution of monovalent electrolytes of molar concentration c_s at 25°C [40], $\epsilon_\circ$ is the vacuum dielectric constant, $\epsilon = 80$ the dielectric permittivity of water, $\mathbf{R}'_i$ the location of an external charge q_i, and $2e\chi$ represents the electrostatic charge of each DNA phosphate located at $\mathbf{r}'(s_j)$, with susceptibility $\chi = 0.25$, a value observed for a broad range of salt concentrations [40].

In our calculations we used the electrostatic charge density

$$\sigma(s) = \frac{8}{3} Q_{DNA} \sin^4\left(\pi s N_{DNA}\right), \tag{A.3}$$

where N_{DNA} is the number of bp and Q_{DNA} is the total charge of the DNA loop corrected for the counterion condensation, yielding $Q_{DNA} = 2e\chi N_{DNA}$.

When the electrostatic force term $\mathbf{f}'_Q$ in (A.1) is included in (2.28), the solution describing the rod geometry needs to minimize the new energy functional

$$U = U_{elastic} + \left(U_Q - U_{Q(relaxed)}\right), \tag{A.4}$$

where $U_{elastic}$ is the elastic energy in (2.52), U_Q is the electrostatic energy of the loop, and $U_{Q(relaxed)}$ the electrostatic energy of the relaxed form of the loop, i.e., a "ground state" energy of the DNA segment. The electrostatic energy is computed as

$$U_Q = \frac{1}{4\pi\epsilon\epsilon_\circ} \int_0^1 \sigma(s) \left(\sum_i q_i \frac{\exp(-|\mathbf{r}'(s) - \mathbf{R}'_i|/\lambda)}{|\mathbf{r}'(s) - \mathbf{R}'_i|} \right.$$
$$\left. + 2e\chi {\sum_j}' \frac{\exp(-|\mathbf{r}'(s) - \mathbf{r}'(s_j)|/\lambda)}{|\mathbf{r}'(s) - \mathbf{r}'(s_j)|} \right) ds. \tag{A.5}$$

The computation of electrostatic forces can be done for the interaction of the loop with itself as well as with the DNA phosphates in the all-atom structure. This results in the system of equations becoming integrodifferential [10, 66], necessitating the use of a computationally more expensive algorithm in which each step of the iteration cycle becomes its own iterative subcycle. It is desirable to exclude this calculation in order to significantly save computer time. The choice of inclusion of electrostatic effects must be made on a case-by-case basis. In general, for ionic concentrations in the range of physiological conditions (50–150 mM NaCl), the solutions are practically indistinguishable from those obtained without electrostatics. The choice of whether or not to include electrostatic contributions depends mainly on the structure of the loop; e.g., in the case of a near self-crossing, the resulting electrostatic self-repulsion contribution actually dominates over bending and twisting energies. A detailed discussion of the effect of electrostatics is found in [9].

A.2. Changes to the numerical algorithm. The electrostatic interactions are introduced in a separate iteration cycle for the solutions of the loop geometry. The deviant parameter in this case is the "electrostatic weight" w_E, which defines the strength of the electric field through $\mathbf{E}_i = w_E\mathbf{E}$, where $\mathbf{E}$ is the desired electric field, and w_E grows linearly from 0 to 1. Each step of this iteration cycle becomes its own iterative subcycle. The electric field $\mathbf{E}_i$ is computed at the beginning of the subcycle and the equations are solved obtaining the external force $\mathbf{f}'$ from this value of the constant electric field. The field is then recomputed for the new geometry of the loop. Then the cycle starts again with the equations being solved for this new field, until convergence of the rod to a permanent geometry (and, consequently, of the field to a permanent value) is realized. The weight w_E is kept constant throughout a subcycle.

A.3. Electrostatic effects on the loop induced by the *lac* repressor. The physiological ionic concentration of 100 mM NaCl was assumed for a test of electrostatic effects. The external charges in (A.2) were taken from the coordinates of the phosphates of the protein-bound DNA segments in the all-atom structure.

IO topology. For the U loop, the computed energy becomes $23.52k_B$T and the new forces are asymmetric and have values of 7.71 and 7.73 pN (Figure 3.1(d)). For the O loop, the total energy is $30.62k_B$T and forces of 7.64 and 6.72 pN arise. Note that the U loop presents changes of 0.2% in energy and an average 0.8% change on the forces. For the O loop, the energy and forces change by 1.4% and 6.0%, respectively. This difference can be explained by the structures of the loops. The U loop has a planar structure and, thus, it is far enough from "itself" that the electrostatic force of self-repulsion is screened by the ions. In the case of the O loop, the form of the loop (Figure 3.1(c′),(d′)) shows that the DNA is approaching contact and, thus, self-repulsion is expected. Nevertheless, for both contributions the change is small. For higher values of salt concentrations, the effect of electrostatics becomes even smaller. We performed a detailed study of the effect of electrostatics in the loops formed by the *lac* repressor [9]. From the results we conclude that for the implementation of the multiscale method to the *lac* repressor, the electrostatic contribution to the energy may be neglected, significantly speeding the rod computations.

Other topologies of the loop. The results in Table 3.1 show that the electrostatic contribution to the total energy at 100 mM NaCl is very small, resulting in very small changes in the structure of the loop and forces exerted on the protein. Thus, electrostatic effects may be neglected for the case of the *lac* repressor-DNA complex, regardless of the choice of boundary conditions.

Appendix B. Building full-atom structures of DNA loops. An important advantage of the elastic rod model described above is that an idealized crystallographic structure can be assigned to any calculated DNA loop geometry (cf. Figure 2.3(b)). Combining this with the full-atom model of the protein, one can build full-atom models of the protein-DNA complex and potentially employ these models in full-atom simulations. The algorithm for recovering full-atom detail from the elastic rod model is the following: (i) Build ideal bp with Quanta [54] and obtain the local frame of reference $\mathbf{d}_{i,bp}$ as explained in section 2.1.2; (ii) place one such ideal base pair at each cross section s_j along the centerline of the rod solution, according to the desired sequence, centering it at $\mathbf{r}(s_j)$ and aligning it to the local frame of reference $\mathbf{d}_i(s_j)$; (iii) build phosphodiester bonds between bp (this can be efficiently done using a molecule structure builder package, e.g., the `psfgen` plugin of VMD[4]); (iv) minimize the built ideal DNA structure to avoid bad contacts and optimize topologies, using the Auto-IMD [25] feature of NAMD with the CHARMM22 force field [37].

REFERENCES

[1] A. Aksimentiev and K. Schulten, *Extending molecular modeling methodology to study insertion of membrane nanopores*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 4337–4338.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *The Cell*, 4th ed., Garland Science, New York, London, 2002.

[3] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, New York, 1987.

[4] E. Arnold and M. G. Rossmann, *Analysis of the structure of a common cold virus, human rhinovirus-*14*, refined at a resolution of* 3.0 *Å*, J. Mol. Biol., 211 (1990), pp. 763–801.

[5] G. Ayton, S. Bardenhagen, P. McMurtry, D. Sulsky, and G. Voth, *Interfacing continuum and molecular dynamics: An application to lipid bilayers*, J. Chem. Phys., 114 (2001), pp. 6913–6924.

[6] G. Bader and U. Ascher, *A new basis implementation for a mixed order boundary value ODE solver*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 483–500.

[7] A. Balaeff, C. R. Koudella, L. Mahadevan, and K. Schulten, *Modeling DNA loops using continuum and statistical mechanics*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 1355–1371.

[8] A. Balaeff, L. Mahadevan, and K. Schulten, *Elastic rod model of a DNA loop in the* lac *operon*, Phys. Rev. Lett., 83 (1999), pp. 4900–4903.

[9] A. Balaeff, L. Mahadevan, and K. Schulten, *Modeling DNA loops using the theory of elasticity*, Phys. Rev. E (3), submitted.

[10] A. Balaeff, L. Mahadevan, and K. Schulten, *Structural basis for cooperative DNA binding by CAP and* Lac *repressor*, Structure, 12 (2004), pp. 123–132.

[11] C. Bell and M. Lewis, *Crystallographic analysis of Lac repressor bound to natural operator O*1, J. Mol. Biol., 312 (2001), pp. 921–926.

[12] C. E. Bell and M. Lewis, *A closer view of the conformation of the Lac repressor bound to operator*, Nature Struct. Mol. Biol., 7 (2000), pp. 209–214.

[13] H. C. Berg, *The rotary motor of bacterial flagella*, Ann. Rev. Biochem., 72 (2003), pp. 19–54.

[14] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, 5th ed., W. H. Freeman, New York, 2002.

[4] URL: http://www.ks.uiuc.edu/Research/namd/ug/node18.html

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The Protein Data Bank*, Nucl. Acids Res., 28 (2000), pp. 235–242.

[16] P. D. Boyer, *The ATP synthase—a splendid molecular machine*, Ann. Rev. Biochem., 66 (1997), pp. 717–749.

[17] B. D. Coleman, E. H. Dill, M. Lembo, Z. Lu, and I. Tobias, *On the dynamics of rods in the theory of Kirchhoff and Clebsch*, Arch. Rational Mech. Anal., 121 (1993), pp. 339–359.

[18] B. D. Coleman, W. W. Olson, and D. Swigon, *Theory of sequence-dependent DNA elasticity*, J. Chem. Phys., 118 (2003), pp. 7127–7140.

[19] D. M. Crothers, T. E. Haran, and J. G. Nadeau, *Intrinsically bent DNA*, J. Biol. Chem., 265 (1990), pp. 7093–7096.

[20] T. Darden, D. York, and L. Pedersen, *Particle mesh Ewald. An* $N{\cdot}\log(N)$ *method for Ewald sums in large systems*, J. Chem. Phys., 98 (1993), pp. 10089–10092.

[21] L. M. Edelman, R. Cheong, and J. D. Kahn, *Fluorescence resonance energy transfer over* ≈130 *basepairs in hyperstable lac repressor-DNA loops*, Biophys. J., 84 (2003), pp. 1131–1145.

[22] S. E. Feller, Y. H. Zhang, R. W. Pastor, and B. R. Brooks, *Constant pressure molecular dynamics simulation—the Langevin piston method*, J. Chem. Phys., 103 (1995), pp. 4613–4621.

[23] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed., Academic Press, New York, 2001.

[24] A. M. Friedman, T. O. Fischmann, and T. A. Steitz, *Crystal structure of* lac *repressor core tetramer amd its implications for DNA looping*, Science, 268 (1995), pp. 1721–1727.

[25] P. Grayson, E. Tajkhorshid, and K. Schulten, *Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics*, Biophys. J., 85 (2003), pp. 36–48.

[26] P. J. Hagerman, *Flexibility of DNA*, Ann. Rev. Biophys. Biophys. Chem., 17 (1988), pp. 265–286.

[27] S. C. Harrison, *Whither structural biology?*, Nature Struct. Mol. Biol., 11 (2004), pp. 12–15.

[28] J. Hermans, G. Mann, L. Wang, and L. Zhang, *Simulation studies of protein-ligand interactions*, in Computational Molecular Dynamics: Challenges, Methods, Ideas, P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and R. D. Skeel, eds., Lect. Notes Comput. Sci. Eng. 4, Springer-Verlag, Berlin, 1999, pp. 129–148.

[29] B. Honig and A. Nicholls, *Classical electrostatics in biology and chemistry*, Science, 268 (1995), pp. 1144–1149.

[30] W. Hsieh, P. A. Whitson, K. S. Mathews, and R. D. Wells, *Influence of sequence and distance between two operators on interaction with the* lac *repressor*, J. Biol. Chem., 262 (1987), pp. 14583–14591.

[31] W. Humphrey, A. Dalke, and K. Schulten, *VMD—Visual Molecular Dynamics*, J. Mol. Graphics, 14 (1996), pp. 33–38.

[32] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten, *NAMD*2: *Greater scalability for parallel molecular dynamics*, J. Comput. Phys., 151 (1999), pp. 283–312.

[33] A. R. Leach, *Molecular Modelling: Principles and Applications*, Addison Wesley, Essex, UK, 1997.

[34] M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu, *Crystal structure of the lactose operon repressor and its complexes with DNA and inducer*, Science, 271 (1996), pp. 1247–1254.

[35] C. Lopez, P. Moore, J. Shelley, M. Shelley, and M. Klein, *Computer simulation studies of biomembranes using a coarse grain model*, Comput. Phys. Comm., 147 (2002), pp. 1–6.

[36] A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, Dover, New York, 1927.

[37] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, I. W. E. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, *All-hydrogen empirical potential for molecular modeling and dynamics studies of proteins using the CHARMM*22 *force field*, J. Phys. Chem. B, 102 (1998), pp. 3586–3616.

[38] B. A. Maguire and R. A. Zimmerman, *The ribosome in focus*, Cell, 104 (2001), pp. 813–816.

[39] L. Mahadevan and J. B. Keller, *The shape of a Möbius band*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 440 (1993), pp. 149–162.

[40] G. S. Manning, *The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides*, Quart. Rev. Biophys., 2 (1978), pp. 179–246.

[41] J. F. Marko, *DNA under high tension: Overstretching, undertwisting, and relaxation dynamics*, Phys. Rev. E (3), 57 (1998), pp. 2134–2149.

[42] K. S. Matthews, *DNA looping*, Microbiological Reviews, 56 (1992), pp. 123–136.

[43] K. S. Matthews, *The whole lactose repressor*, Science, 271 (1996), pp. 1245–1246.

[44] J. H. Miller and W. S. Reznikoff, eds., *The Operon*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1980.

[45] B. Müller-Hill, *The* lac *Operon*, Walter de Gruyter, New York, 1996.

[46] S. Oehler, E. R. Eismann, H. Krämer, and B. Müller-Hill, *The three operators of the* lac *operon cooperate in repression*, EMBO J., 9 (1990), pp. 973–979.

[47] W. K. Olson, *Simulating DNA at low resolution*, Curr. Op. Struct. Biol., 6 (1996), pp. 242–256.

[48] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.-S. Tung, E. Westhof, C. Wolberger, and H. M. Berman, *A standard reference frame for the description of nucleic acid base-pair geometry*, J. Mol. Biol., 313 (2001), pp. 229–237.

[49] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, *DNA sequence-dependent deformability deduced from protein-DNA crystal complexes*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 11163–11168.

[50] W. K. Olson, N. L. Marky, R. L. Jernigan, and V. B. Zhurkin, *Influence of fluctuations on DNA curvature. A comparison of flexible and static wedge models of intrinsically bent DNA*, J. Mol. Biol., 232 (1993), pp. 530–554.

[51] W. K. Olson and V. B. Zhurkin, *Modeling DNA deformations*, Curr. Op. Struct. Biol., 10 (2000), pp. 286–297.

[52] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. van Oudenaarden, *Multistability in the lactose utilization network of* Escherichia Coli, Nature, 427 (2004), pp. 737–740.

[53] R. Phillips, M. Dittrich, and K. Schulten, *Quasicontinuum representations of atomic-scale mechanics: From proteins to dislocations*, Ann. Rev. Mater. Res., 32 (2002), pp. 219–233.

[54] Polygen Corporation, *Quanta*, Waltham, MA, 1988.

[55] M. Ptashne, *A Genetic Switch*, 2nd ed., Cell Press & Blackwell Scientific Publications, Cambridge, MA, 1992.

[56] T. J. Richmond and C. A. Davey, *The structure of DNA in the nucleosome core*, Nature, 423 (2003), pp. 145–150.

[57] G. Ruben and T. B. Roos, *Conformation of* lac *repressor tetramer in solution, bound and unbound to operator DNA*, Mic. Res. Tech., 36 (1997), pp. 400–416.

[58] T. Schlick, *Modeling superhelical DNA: Recent analytical and dynamic approaches*, Curr. Op. Struct. Biol., 5 (1995), pp. 245–262.

[59] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, Springer-Verlag, New York, 2002.

[60] C. A. E. M. Spronk, A. M. J. J. Bovin, P. K. Radha, G. Melacini, R. Boelens, and R. Kaptein, *The solution structure of Lac repressor headpiece* 62 *complexed to a symmetrical* lac *operator*, Structure Fold. Des., 7 (1999), pp. 1483–1492.

[61] T. R. Strick, J.-F. Allemand, D. Bensimon, A. Bensimon, and V. Croquette, *The elasticity of a single supercoiled DNA molecule*, Science, 271 (1996), pp. 1835–1837.

[62] D. Swigon, B. D. Coleman, and W. K. Olson, *Modeling the lac Repressor-Operator Assembly:* I. *The Influence of DNA Looping on lac Repressor Conformation*, in preparation.

[63] E. Tajkhorshid, A. Aksimentiev, I. Balabin, M. Gao, B. Isralewitz, J. C. Phillips, F. Zhu, and K. Schulten, *Large scale simulations of protein mechanics and function*, in Protein Simulations, Advances in Protein Chemistry 66, V. Daggett, ed., Academic Press, New York, 2003, pp. 195–247.

[64] E. Villa, A. Balaeff, and K. Schulten, Lac *Repressor-DNA Loop Dynamics*, in preparation.

[65] A. V. Vologodskii and N. R. Cozzarelli, *Conformational and thermodynamic properties of supercoiled DNA*, Ann. Rev. Biophys. Biomol. Struct., 23 (1994), pp. 609–643.

[66] T. P. Westcott, I. Tobias, and W. K. Olson, *Modeling self-contact forces in the elastic theory of DNA supercoiling*, J. Chem. Phys., 107 (1997), pp. 3967–3980.

[67] Y. Yang, T. P. Westcott, S. C. Pedersen, I. Tobias, and W. K. Olson, *Effects of localized bending on DNA supercoiling*, Trends Biochem. Sci., 20 (1995), pp. 313–319.