MULTISCALE MODELING OF NUCLEIC ACIDS IN CELLULAR ENVIRONMENTS

By

Asli Yildirim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry—Doctor of Philosophy

2017

ABSTRACT

MULTISCALE MODELING OF NUCLEIC ACIDS IN CELLULAR ENVIRONMENTS

By

Asli Yildirim

Biological cells are highly crowded due to the presence of various macromolecules. Here, the macromolecular crowding effects on DNA structure were investigated via computer simulations. Molecular dynamics simulations of DNA with crowder proteins showed that B-form of DNA is stabilized resulting from non-specific interactions of crowder proteins with DNA sugar-phosphate backbone, while the reduced dielectric response of cellular environments was found to favor A-like conformations as shown by implicit solvent simulations of DNA in reduced dielectric environments. Overall, the results obtained here suggest that different aspects of cellular crowding have opposite impacts on DNA structure.

As the largest molecule in the cell, genomic DNA also occupies a large fraction of the bacterial cell and causes crowding. An experimentally-driven multiscale modeling protocol was developed to study the three-dimensional structure of bacterial chromosomes and their effects on protein diffusion. Using this protocol, three dimensional structures of *Caulobacter crescentus* chromosome at base-pair resolution were generated. Investigation of these models provided insights into chromosome structural variability and organization. Coarse-grained Brownian dynamics simulations of these chromosome models in the presence of proteins, on the other hand, suggest that protein diffusion becomes slower and anomalous when around DNA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

This dissertation aims to comprehend the structure of DNA in cellular environments using computational modeling techniques. The first chapter provides a brief introduction to the DNA structure and macromolecular crowding observed in cellular environments, and explains the modeling approaches that are used in this dissertation. The second and third chapters present the results obtained from molecular dynamics simulations of DNA in environments with reduced dielectric response and protein crowding, respectively. The fourth chapter describes the development of a multiscale modeling algorithm to generate three-dimensional structures of bacterial nucleoids at base-pair resolution. The fifth chapter discusses the effect of nucleoid crowding on the diffusive properties of proteins observed in Brownian dynamics simulations of coarse-grained model systems containing bacterial nucleoid and proteins. Finally, the sixth chapter summarizes the findings and the future directions of the work.

1.1 Structure of DNA

DNA (deoxyribonucleic acid) is a complex biomolecule found in living cells and it carries the biological instructions for life. DNA was first discovered and isolated from human white blood cells by Friedrich Miescher in 1869 [1]. This discovery led many scientists to investigate molecular structure of DNA and the first clues were found by Phoebus Levene in 1919 [2]. He discovered that nucleotides are the building blocks of DNA structure and the major components of these nucleotides are sugar, nitrogenous bases (adenine (A), guanine (G), cytosine (C), thymine (T)) and phosphate (Figure 1.1). Later, Erwin Chargaff expanded Levene's work and proposed two rules about DNA structure; relative A, G, C, T content differs from one species to another, and the total amount of A is equal to the total amount of T as well as the total amount of G is equal to the total

amount of C [3]. These rules provided an additional hint on the facts that DNA is the genetic material rather than proteins and the bases in its structure are paired.

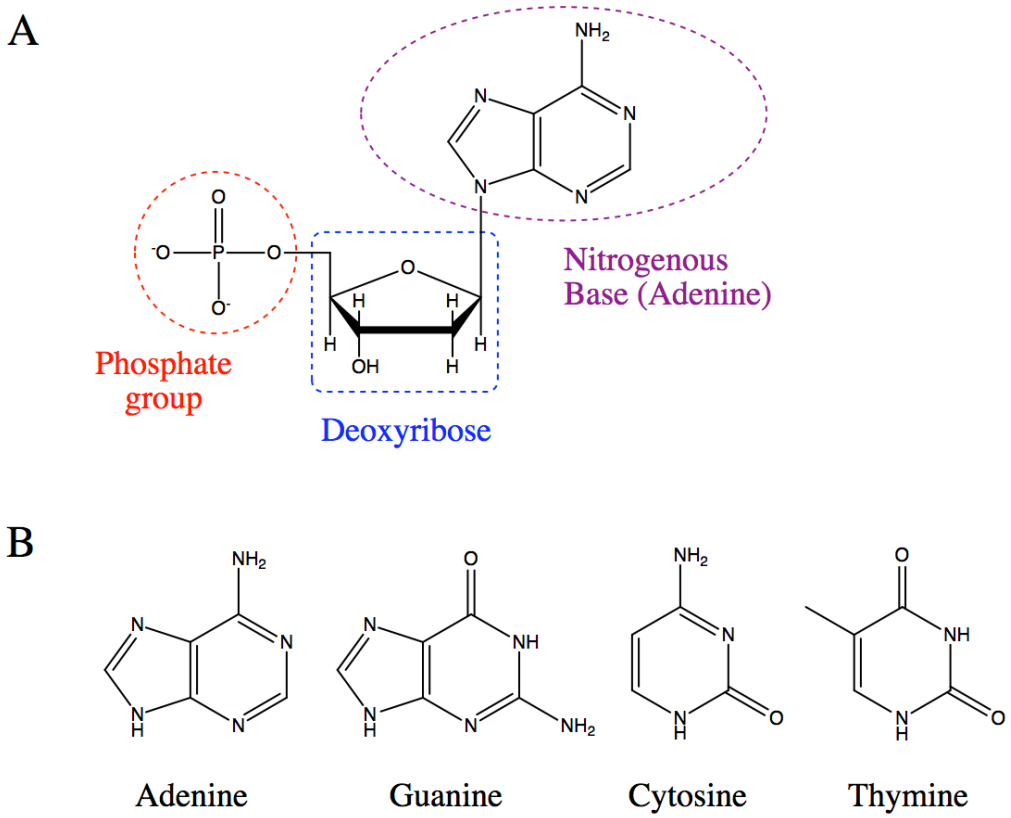

Figure 1.1 Building blocks of DNA structure. (A) Structure of a nucleotide. (B) Nitrogenous bases found in DNA.

In the meantime, Maurice Wilkins and Rosalind Franklin were also investigating the DNA structure by using X-ray crystallography. In 1952, Rosalind Franklin had taken the X-ray diffraction image of DNA which provided crucial information on helical-nature of DNA structure [4]. Combining all the obtained knowledge of DNA structure, Francis Crick and James Watson derived the three-dimensional (3D) double-helical DNA model in 1953 [5]. They proposed that DNA is a right-handed double-stranded helix with the two strands coiled around the helical axis (Figure 1.2*A*). Each strand is composed of nucleotides linked by a phosphodiester bond between

the 3′ carbon atom of the sugar of one nucleotide and the 5′ carbon atom of the next nucleotide's sugar, forming the sugar-phosphate backbone of DNA. The two strands with sequences running in opposite directions, are connected by hydrogen bonds between the purine and pyrimidine bases of nucleotides, constructing the DNA base-pairs, A-T and G-C. They also suggested that the helical diameter of DNA is 20 Å, the distance between the two-consecutive base-pairs in the helical axis direction is 3.4 Å and the rotation angle between them is 36°.

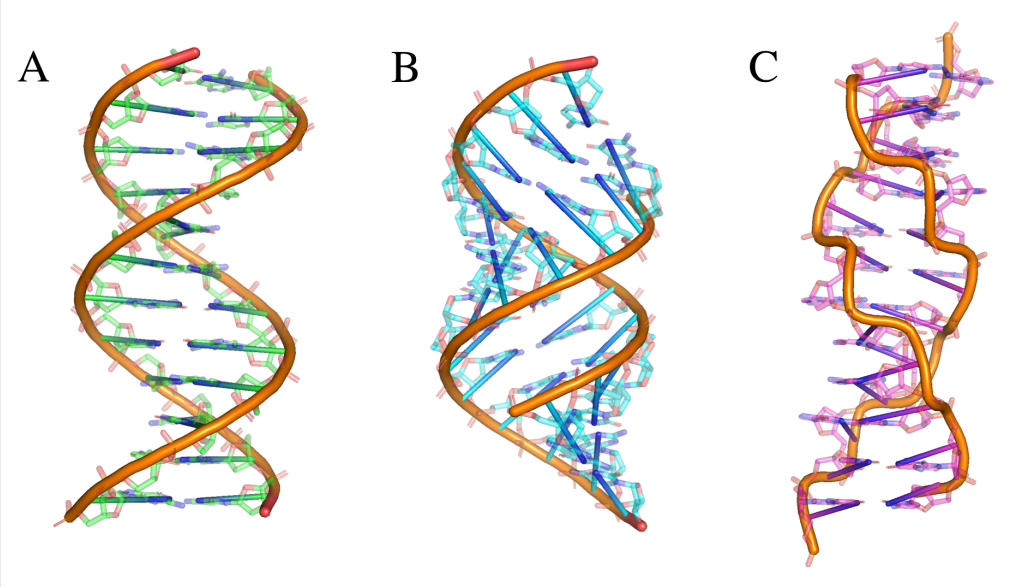

Figure 1.2 Double helical DNA conformations. (A) B-DNA. (B) A-DNA. (C) Z-DNA.

The described DNA model by Watson and Crick is called "B-DNA" (Figure 1.2*A*) and accepted as the most common DNA conformation found in nature. In addition to B-DNA, there are also two other significant conformations of DNA of which existences are experimentally proven: A-DNA and Z-DNA (Figure 1.2*B* and 1.2*C*). When B-DNA is dehydrated, it gets into A-form [6]. A-DNA is much shorter than B-DNA with ~2.6 Å distance between two consecutive base-pairs, also wider as its diameter is suggested to be around 23 Å. This form is also observed

in environments with high alcohol content due to the lower water activity of the environment [7, 8] and in solutions with salt which helps bridging negatively charged phosphate atoms in the DNA backbone [9], therefore forming a more compact structure. Z-DNA, on the other hand, is much different from A- and B- forms since it has a left-handed helical structure. It is also relatively narrower and longer with a smaller helical diameter and a larger distance between the adjacent base pairs. Z-form is mainly found in DNA with alternating purine and pyrimidine sequences and also some proteins were found to bind to Z-DNA [10, 11] specifically.

Similar to proteins, DNA is also assumed to have a primary, secondary and tertiary structures. The DNA sequence is the primary structure, and the 3D DNA conformations described above are the common secondary structures of DNA. Genomic DNA can be billions of base-pairs in length for eukaryotes and it has to undergo conformational compaction to fit into the microscopic cell nucleus. The compact conformation that DNA folds into, therefore, is considered as its tertiary structure. Briefly, genomic DNA in eukaryotes first wraps around eight histone cores and this DNA-histone complex is called nucleosome [12]. Multiple nucleosomes then package together forming the chromatin fibers [12]. The chromatin fibers also undergo further compaction with coiling and looping, finally forming chromosomes [12]. This multilevel compaction allows meters of genomic DNA to fit inside the cell nucleus that has an average diameter of 6 microns.

On the other hand, genomic DNA in prokaryotes is a circular DNA and relatively smaller in size. However, it still needs compaction to fit inside the cell. In contrast to the eukaryotic genome, prokaryotic genome does not have a specific compartment separated from the rest of the cytoplasm with a nuclear membrane. It instead resides inside the cytoplasm and interacts freely with other biomolecules in the cell. The region of the cytoplasm where the genomic DNA lies together with proteins and RNA is called nucleoid. The DNA inside the nucleoid is lack of histones; however, it

still undergoes compaction via negative supercoiling introduced by topoisomerases [12]. Further looping of the supercoiled DNA also takes place with the help of nucleoid-associated proteins (NAPs). The prokaryotic chromosome structure and organization is discussed in more detail in the next section as it is one of the major subjects of this dissertation.

1.2 Bacterial Chromosome Structure and Organization

The bacterial chromosome contains a closed circular DNA with a length of several million base-pairs which would reach millimeters in length if stretched out [13]. In order to fit inside the cell nucleoid which has less than 1 μm^3 volume, the chromosome undergoes multiple levels of compaction (Figure 1.3). The circular chromosomal DNA compacts itself by generating loops with the help of proteins [13]. Bacterial chromosomes do not have histones to wrap around for compaction as eukaryotic chromosomes have, however, there is a class of DNA-binding proteins in bacteria, called NAPs as discussed in Section 1.1, which play the dominant role in compacting and organizing chromosome structure by introducing sharp bends and kinks [14]. Some of the most abundant NAPs in bacteria are H-NS (histone-like nucleoid-structuring), Fis (factor for integration stimulation), IHF (integration host factor) and Smc (structural maintenance of chromosome) proteins [14]. The other crucial compaction mechanism for bacterial chromosomes is DNA supercoiling. Supercoiling of chromosomal DNA is controlled by DNA gyrase and topoisomerase I enzymes which introduce negative and positive supercoiling, respectively. As a result of these compaction mechanisms, bacterial chromosomes get into a bottlebrush-like structure containing several topological domains with an average size of ~10 kb [15].

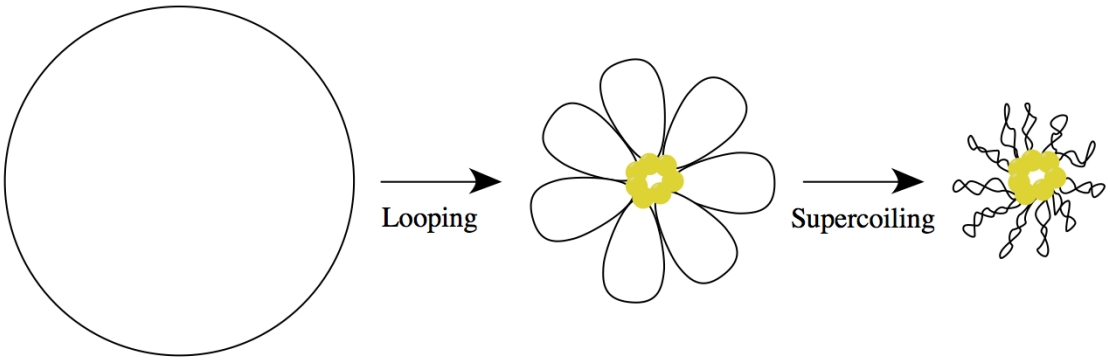

Figure 1.3 Bacterial chromosome compaction.

DNA supercoiling occurs from under- or overwinding of the double-stranded DNA helix. When the DNA helix is underwound, it becomes negatively supercoiled (right-handed) and it is positively supercoiled (left-handed) when overwound. Most of the organisms including bacteria have negatively supercoiled DNA. Besides, there are two forms of supercoiling as illustrated in (Figure 1.4A) which are toroidal and plectonemic supercoiling. Plectonemic supercoiling is more common in nature; it is the dominant supercoiling form observed in bacterial chromosomes as well.

The mathematical description of the degree of supercoiling is found using numerical properties of supercoiled DNA which are linking number (Lk), twist (Tw) and writhe (Wr). Lk gives the number of winding of one DNA strand about the other DNA strand which is always integer and constant for a closed circular DNA. Tw is the number of turns in the DNA helix whereas Wr is the number of times the DNA helix coils around itself and the sum of these two values gives the Lk of DNA (Equation 1.1):

$$Lk = Tw + Wr \tag{1.1}$$

For the relaxed form of DNA, Wr is always zero, therefore Lk is equal to Tw. Lk_o, the linking number for the relaxed B-DNA form is calculated using Equation 1.2:

$$Lk_o = bp/10.5 \quad (1.2)$$

where bp is the number of basepairs in the DNA and 10.5 is the number of basepairs relaxed B-DNA has per turn. ΔLk, the difference between the Lk of a supercoiled DNA and Lk_o (Equation 1.3) gives the change in the total number of turns in the supercoiled DNA. When ΔLk is negative, DNA is negatively supercoiled (underwound).

$$\Delta Lk = Lk - Lk_o \quad (1.3)$$

Finally, the specific linking difference calculated by Equation 1.4, σ, is the measure of the degree of supercoiling. The specific linking difference of chromosomal DNA isolated from a cell is between -0.03 – -0.09 indicating that DNA is negatively supercoiled in cells [16].

$$\sigma = \Delta Lk/Lk_0 \quad (1.4)$$

A typical configuration for a negatively supercoiled plectoneme is shown in Figure 1.4*B*. The average opening angle of a superhelix, α, has been shown to be constant and independent of the specific linking difference of DNA by various electron microscopy and computational studies [17, 18]. The superhelix radius, r, on the other hand, is strongly dependent on σ; it shows a decreasing trend with an increasing |σ| [17, 18]. The pitch of the superhelix is $2\pi p$, and p is defined by Equation 1.5:

$$p = tan\alpha \times r \quad (1.5)$$

Apart from the specific linking difference, α and r have also been suggested to depend on the salt concentration of the environment [19]. Nevertheless, α is mostly around 55° and r varies between 5 – 12 nm at typical cellular conditions with σ range of -0.03 – -0.09 and 0.1 M salt concentration [16-19].

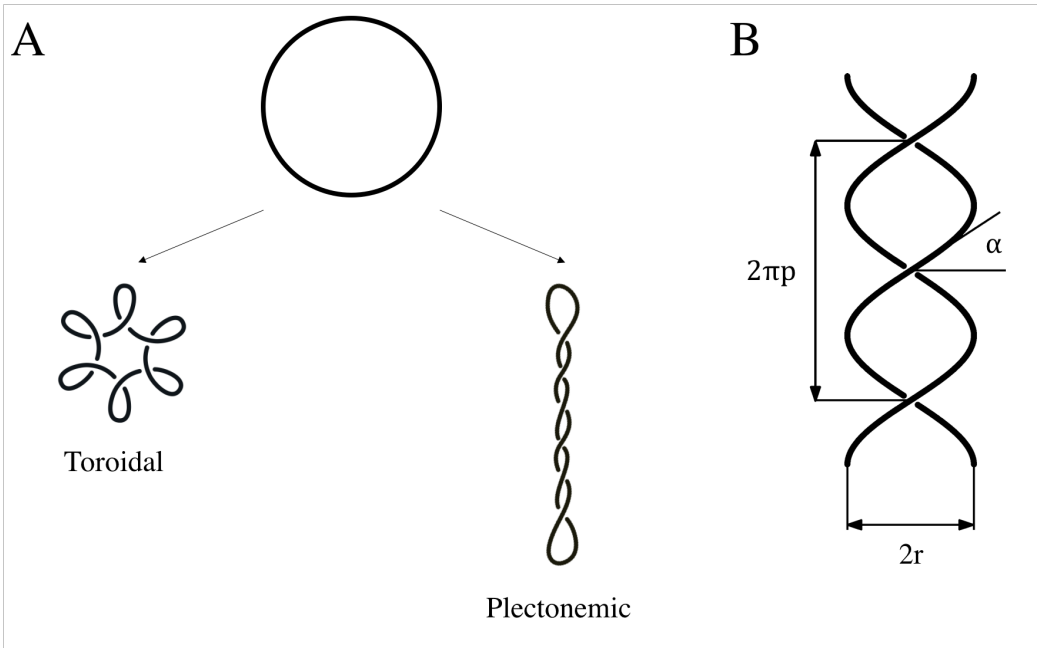

Figure 1.4 Properties of supercoiling. (A) Different types of supercoiling, (B) Plectonemic supercoiling parameters.

In addition to the abovementioned knowledge on bacterial chromosome structure revealed by early studies, the developments in experimental biological techniques have provided new insights into the chromosome structure and organization [13, 20]. A summary of the most common experimental approaches to understand the spatial organization of bacterial chromosomes is illustrated in Figure 1.5. Fluorescence in situ hybridization (FISH) allows visualization of individual genetic loci in fixed cells using fluorescent probes that are able to bind DNA (Figure 1.5A), while the fluorescence repressor-operator system (FROS) technique enables the same visualization in live cells by inserting exogenous sequences into the genome to which fluorescently labeled proteins can bind (Figure 1.5B). Besides these methods allowing the localization of individual loci, genome-wide approaches are also extensively used in investigating chromosome structure and organization. Binding-sites of a DNA-binding protein in a whole genome can be identified by using chromatin immunoprecipitation (ChIP)-based techniques (Figure 1.5C). In ChIP-based experiments, DNA and bound proteins are first crosslinked and then cells are lysed.

Next, the DNA-protein complexes are isolated from the rest of the solution using a purification technique called immunoprecipitation. And lastly, DNA is purified and sequenced. Another genome-wide method, called chromosome conformation capture (3C) [21], allows identifying the frequency of interactions between two different *loci* in a genome. The experiment starts with crosslinking the whole genomic DNA and then cutting it with a restriction enzyme. The resulting DNA products are then ligated. Finally, the ligated fragments are purified and sequenced. The derivative of this experiment (Hi-C) in which high-throughput sequencing is used results in a genome-wide contact matrix which provides insight into the 3D conformation of the chromosome [22].

Using these techniques on *Escherichia coli*, *Caulobacter crescentus*, *Bacillus subtilis,* and *Mycoplasma pneumoniae* revealed that bacterial chromosomes are highly-organized [23-33]. Early FISH studies showed that the *E. coli* chromosome is organized into four macrodomains; Ori, Ter, Left and Right [23] which have also been identified by 3C-based experiments later [30]. FROS studies on *C. crescentus* and *B. subtilis* reported that the origin and terminus of the replication reside at the opposite poles of the *C. crescentus* chromosome [24], and in the *B. subtilis* chromosome, while the origin of replication lies at the cell quarter position, the terminus resides at the mid-cell [25, 26]. Further investigation of the *C. crescentus* chromosome with 3C-based techniques have revealed that two chromosomal arms between the poles intertwined around each other [27] and the chromosome contains 23 different chromosome interacting domains (CIDs) where a *locus* is interacting more frequently with another within the same CID compared to a *locus* from other CIDs [28]. CIDs have also been observed in *B. subtilis* and *M. pneumoniae* chromosomes by Hi-C experiments [29, 31-33]. These CID domains with a size range of 30 – 400 kb are similar to the topologically associating domains (TADs) found in eukaryotic chromosomes.

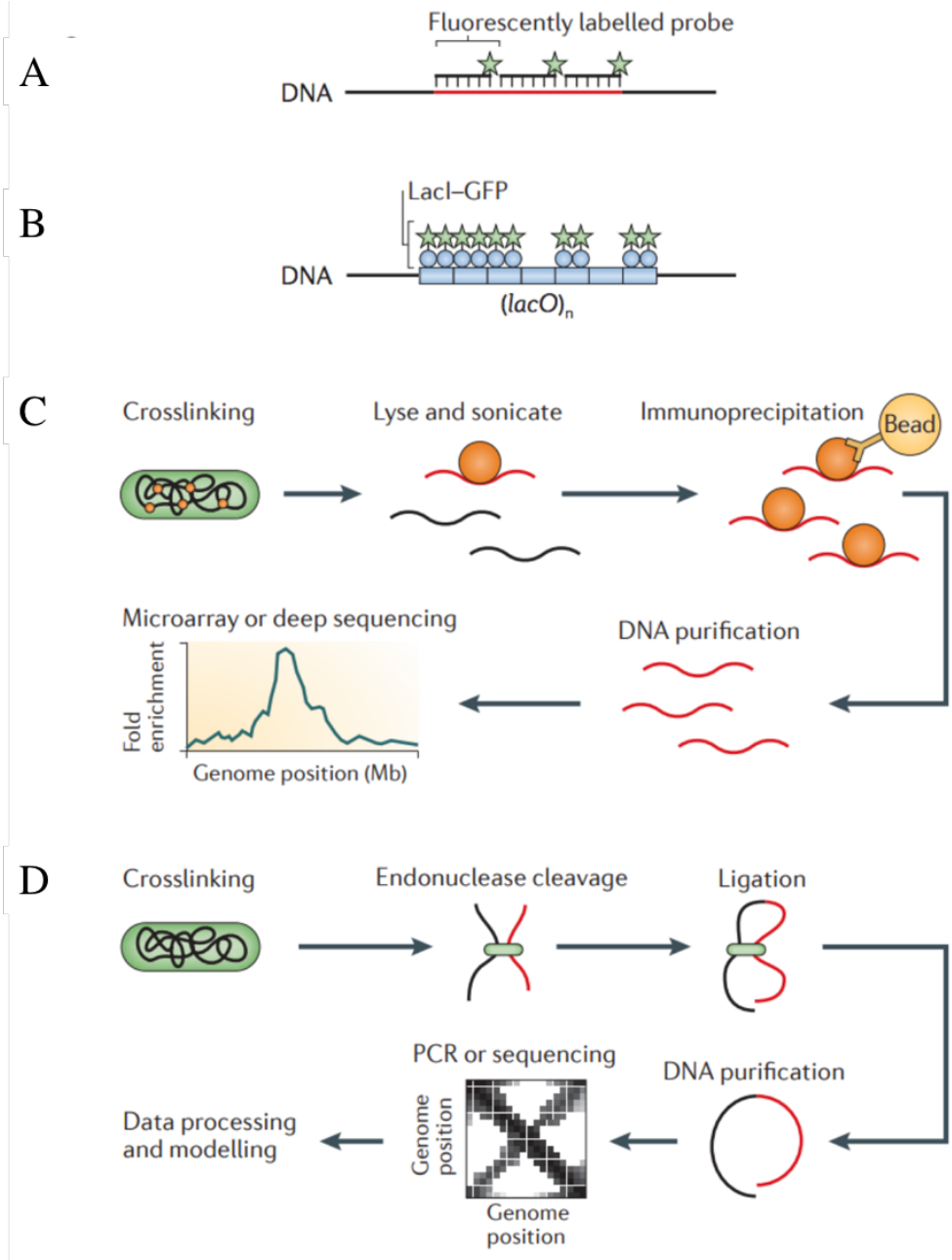

Figure 1.5 Technical advances for studying bacterial chromosomes. (A) Fluorescence in situ hybridization, (B) Fluorescently tagged DNA-binding proteins, including the fluorescence repressor–operator system (FROS), (C) Chromatin immunoprecipitation (ChIP)-based methods, (D) Chromosome conformation capture methods. Adapted by permission from Macmillan Publishers Ltd: [Organization and segregation of bacterial chromosomes] (Nature Reviews Genetics 14, 191–203 (2013)), copyright (2013).

Recent developments in experimental techniques to study chromosome structure and organization have paved the way for 3D modeling efforts for entire chromosomes. Interaction frequencies of two *loci* obtained from 3C-based experiments can be converted to spatial distances based on the assumption that two *loci* that are frequently interacting are spatially close, and when there is no interaction, they lie apart in 3D space. By using the converted distances derived from 3C frequencies, 3D chromosome structures of *C. crescentus* and *M. pneumonia* have been generated at 10-kb resolution [27, 29]. Another modeling strategy where the topological parameters of a circular polymer model consisting of plectonemes were varied in order to find the model that fits the Hi-C data best, provided 3D *C. crescentus* chromosome models at 434 base-pairs resolution [28]. In addition to using 3C contact frequencies, CHIP data on the RNA polymerase binding sites for the *E. coli* genome combined with the genomic DNA topological properties have also been utilized in generating 3D chromosome models of *E. coli* at nucleotide resolution [34]. Overall, the obtained 3D models have enabled better visualization of bacterial chromosomes in 3D space and shed light on the chromosome structure–function relationships.

1.3 Macromolecular Crowding

Macromolecular crowding phenomenon has drawn considerable attention by researchers in recent years since cellular environments are highly crowded and their crowded nature has been found to affect biomolecular structure and dynamics [35]. The cellular cytoplasm contains high concentrations of proteins, nucleic acids and various cosolvents [36]. A typical cell has a macromolecular concentration of 300 – 400 mg/ml which results in a 5 – 40 % fraction of the cell volume occupied by macromolecules [37]. This crowding condition is significantly different from a dilute solution which is an extensively used environment when conducting experiments, hence

the effect of macromolecular crowding on the biological cell functions and processes is mostly ignored *in vitro*.

Studies on macromolecular crowding to date have reported three major effects of cellular crowding: volume exclusion, non-specific interactions with crowders, and reduced dielectric response of the crowded environment. The volume-exclusion effect arises from the less available volume for solutes because of the presence of other crowder macromolecules [38]. The steric repulsion between solutes and other crowders has been shown to entropically favor more compact conformation states due to the confinement [35]. The excluded-volume had been regarded as the most crucial effect of crowding over the years; however, it has been challenged by new crowding studies suggesting that enthalpic contributions resulting from possible intracellular interactions stabilize partially extended conformations in contrast to the excluded-volume effect [39-41]. Therefore, these studies show that nonspecific interactions between solutes and crowders also play a dominant role in determining structure and dynamics of biomolecules. Lastly, since the water is much less available because of its replacement by less polar crowder macromolecules, the reduced dielectric response of the environment is significantly reduced. The dielectric constant of a cellular environment is believed to lie in a range of 20 – 60 which is much lower compared to the bulk water's dielectric constant (ε=80) [42].

The effects of macromolecular crowding specifically on protein structure and dynamics has been extensively studied. Primary studies from Minton *et al*. have shown that the reduction of the accessible volume due to crowders stabilizes the folded native structure of proteins [35, 43]. On the other hand, recent studies have shown that the nonspecific interactions of proteins with other crowder proteins stabilize non-native, partially extended structures [39-41, 44-46]. Several experimental studies have reported the destabilizing effect of the cellular environment on different

proteins with different experimental techniques [39-41]. Pielak *et al*. have studied chymotrypsin inhibitor 2 (Cl2) by using hydrogen-exchange NMR and found that it is destabilized in the presence of lysozyme [40]. Another in-cell NMR study has shown the destabilization of ubiquitin in human cells due to possible interactions with intracellular macromolecules [39]. These experimental findings have also been supported by computational studies; different studies from our group suggest that protein-protein interactions lead to stabilization of unfolded states of proteins [44-46]. Cellular crowding has also been reported to alter the hydration structure around proteins [42]. In addition to the effects of crowding on structural properties of proteins, studies also show the change in diffusive properties of proteins in cellular environments. Diffusion rates of proteins were found to be greatly reduced in the cytoplasmic environment [47-49].

Most of the crowding studies thus far have focused on characterizing the effect of crowding on protein structure and dynamics. The available findings on crowding in the literature are limited and less detailed when it comes to nucleic acids [50, 51]. It has been found that the excluded-volume and different hydration patterns under crowding alter the structures of DNA G-quadruplexes [52-54]. Short DNA duplexes have been found to undergo B- to A- form transition in the presence of alcohols and explicit salts [7, 8, 55-62]. On the other hand, longer DNA duplexes are most stabilized in compact conformations instead of elongated states in the presence of polyethylene glycol (PEG) due to the excluded-volume [63, 64]. In addition to structural observations, there is also some evidence that plasmid DNA has lower diffusion rates due to the crowding [65-68]. However, all in all, the changes in structure and dynamics of DNA have not been investigated as systematically as in protein studies; although DNA duplex structure has been studied in environments with alcohol or additional salt, the effect of explicit cellular crowder macromolecules, such as proteins, on DNA structure has not been explored.

On the other side, in addition to proteins which were generally used as crowding agents in studies due to their high concentration in cells, genomic DNA also occupies a large fraction of the cell volume; hence it is highly probable that DNA does not only experience the effect of crowding in cells, but they also give rise to crowding. Especially in prokaryotes, genomic DNA is not isolated in a nuclear compartment as in eukaryotes, and it freely interacts with the macromolecules in the cytoplasm as discussed in the previous section, suggesting that the nucleoid crowding effect may also play a crucial role in cellular processes.

1.4 Computational Modeling of DNA

Computer simulations play a very significant role in biology today as a complementary tool to experiments. Molecular dynamics (MD) simulations have become one of the most powerful techniques to study complex biomolecular systems since it enables studying the structure and dynamics of biological macromolecules at atomistic detail which many experiments lack [69]. MD simulations provide the time evolution of the positions and velocities of the particles in a biological system by using Newton's law of motion. In MD, first, the force on each particle is calculated from the derivation of the defined potential energy by using Equation 1.6:

$$F_i = -\Delta_i U(\vec{X}) = m_i a_i \tag{1.6}$$

where F_i is the force on particle i, m_i and a_i are the mass and acceleration of particle i, and $U(\vec{X})$ is the potential energy of particle i as a function of its set of coordinates, $\vec{X}$. The integration of Equation 1.6 then provides the new positions and velocities.

The potential energy function used in Equation 1.6 is a set of potentials based on bonded and nonbonded interactions in the system:

$$U(\vec{X}) = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\phi[1 + \cos(n\phi - \delta)] + \sum_{impropers} K_\omega(\omega - \omega_0)^2 + \sum_{electrostatic} \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{vdW} \varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right] \quad (1.7)$$

where, $K_b, K_\theta, K_\phi, K_\omega$ are the force constants of the bonding terms, bonds (b), angles (θ), dihedrals (ϕ) and impropers (ω), and the constants with a subscription *0* denote the respective equilibrium values. Nonbonded terms contain Lennard-Jones and Coulombic terms. The former is for van der Waals interactions with constants ε_{ij}, r_{ij}^{min}, r_{ij} corresponding to the depth of the potential well, distance between particles when the potential is minimum, and distance between particles at that time, and the latter is for electrostatic interactions where q_i and q_j are the charges on particle i and j, r_{ij} is the distance between particles and ε is the dielectric constant of the medium.

The functional form and the set of parameters used in the potential energy is called "force field". For the fully atomistic representations of biomolecules, there are several force fields available in the literature. The most widely used ones are CHARMM [70-72], AMBER [73] and GROMOS [74]. The most accurate force fields for all-atomic DNA models are CHARMM36 [72] and OL15 [75] and parmbsc1 [76] AMBER force fields. MD simulations of all-atom DNA models have been extensively carried out and shed light on conformational preferences of DNA [77, 78], the effect of salt and cosolvents on DNA structure [57, 62], B- ←→ A- form DNA transitions [58-60, 79], hydration and ion density patterns around DNA [80-82]. Reviews on MD studies on DNA structure can be found in the following references [77, 78].

All-atom simulations of biomolecules generally contain explicit all-atom water molecules as well, since biomolecules are mostly solvated with water in nature. The addition of water molecules

to the biological system for solvation increases the number of atoms in the system as well as the non-bonded interactions drastically, therefore requires more computational resources. In order to reduce the computational cost of the explicit solvation, the solvent environment can also be treated implicitly in which a continuous medium represents the solvent instead of explicit water molecules [83]. This way of solvation is less accurate than the explicit solvation; however, the computational cost is much lower. In implicit solvation, the solvation free energy is calculated with the following equation:

$$\Delta G_{solvation} = \Delta G_{electrostatic} + \Delta G_{nonpolar} \tag{1.8}$$

where the solvation energy consists of two parts: electrostatic ($\Delta G_{electrostatic}$) and nonpolar ($\Delta G_{nonpolar}$) contributions. One way of calculating the electrostatic part of the Equation 1.8 is using Generalized Born (GB) theory which is an approximation of Poisson-Boltzmann (PB) method [84]. GB theory calculates the electrostatic component of the solvation free energy with Equation 1.9:

$$\Delta G_{electrostatic} = -k\left(\frac{1}{\varepsilon_{solute}} - \frac{1}{\varepsilon_{solvent}}\right)\sum_{i,j}\frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp\left(-r_{ij}^2/4\alpha_i\alpha_j\right)}} \tag{1.9}$$

where ε_{solute} and $\varepsilon_{solvent}$ are the dielectric constants of the solute and solvent, q is the charge of the atoms, r is the distance between the atoms and α is the Born radii of the corresponding atom. The estimation of Born radii can be done with different approaches which are reviewed in the following references [85, 86]. In this dissertation, the Generalized Born with molecular volume method (GBMV) is used where Born radii are calculated from the integration over molecular volume built from a superposition of atomic functions [87-89].

The second component of Equation 1.8, $\Delta G_{nonpolar}$, which is the contribution to the solvation free energy resulting from nonpolar interactions between solute and solvent as well as the cost of cavity is calculated from the solvent accessible surface area (SASA):

$$\Delta G_{nonpolar} = \sum_i \gamma_i SASA_I \tag{1.10}$$

where γ is the surface tension coefficient.

Classical MD simulations either with explicit or implicit solvation, sometimes is not efficient to sample all conformational states in a rough energy landscape due to the requirement of long simulation times to observe barrier crossings. To overcome this problem, many enhanced sampling methods have been developed including metadynamics, simulated annealing and replica-exchange molecular dynamics (REMD) [90]. REMD, which is also applied within this dissertation, simulates several independent replicas of the system simultaneously at different temperatures and exchanges between these replicas periodically according to the Metropolis criterion [91]:

$$p(X_m \rightarrow X_n) = \begin{cases} 1 & for\ \Delta < 0 \\ \exp(-\Delta) & for\ \Delta \geq 0 \end{cases} \tag{1.11}$$

where p is the probability of swapping between configurations X_m and X_n at temperatures T_m and T_n, and Δ is;

$$\Delta = [\beta_n - \beta_m](E(q_m) - E(q_n)) \tag{1.12}$$

where $\beta = 1/kT_{m(n)}$, $E(q_m)$ and $E(q_n)$ are the corresponding potential energies of configurations X_m and X_n.

The idea behind REMD is that higher temperatures will cross the barriers between local minima easier, therefore span more conformational space and make other local minima that are

not reachable via straight MD available. In this dissertation, the secondary DNA structure under the effect of macromolecular crowding is investigated via REMD simulations of all-atom DNA model in implicit solvent (Chapter 2) and MD simulations in explicit solvent (Chapter 3).

Although fully-atomistic representations of biological systems are the most realistic models to use in MD simulations, all-atom simulations suffer substantially from the computational cost required to reach biologically relevant timescales, especially for the biological systems because they generally contain thousands to millions of atoms. In addition to the enhanced sampling methods, an alternative way to reduce the computational cost of running long MD simulations is to reduce the degrees of freedom in the system by switching to coarse-grained (CG) models where several atoms are represented with one particle. CG models are extensively used in biomolecular studies of proteins, nucleic acids, lipids and carbohydrates [92]. One of the challenges in CG modeling is to develop CG force fields as accurate as the ones used in all-atom simulations. The reported CG force fields in the literature have been developed by extracting CG interactions from all-atom simulations, optimizing parameters in the interaction potentials to reproduce the key experimental data, or combination of these two approaches [92]. These available CG models have a wide variety of simplification levels. For DNA, coarsening levels of models vary from several base-pairs per a single particle [93, 94] to more sophisticated representations with one to three particles per nucleotide [95, 96] as reviewed in the following reference [97]. Representations of CG models with resolutions of one particle per nucleotide and four basepairs per particle for a DNA dodecamer are illustrated in Figure 1.6.

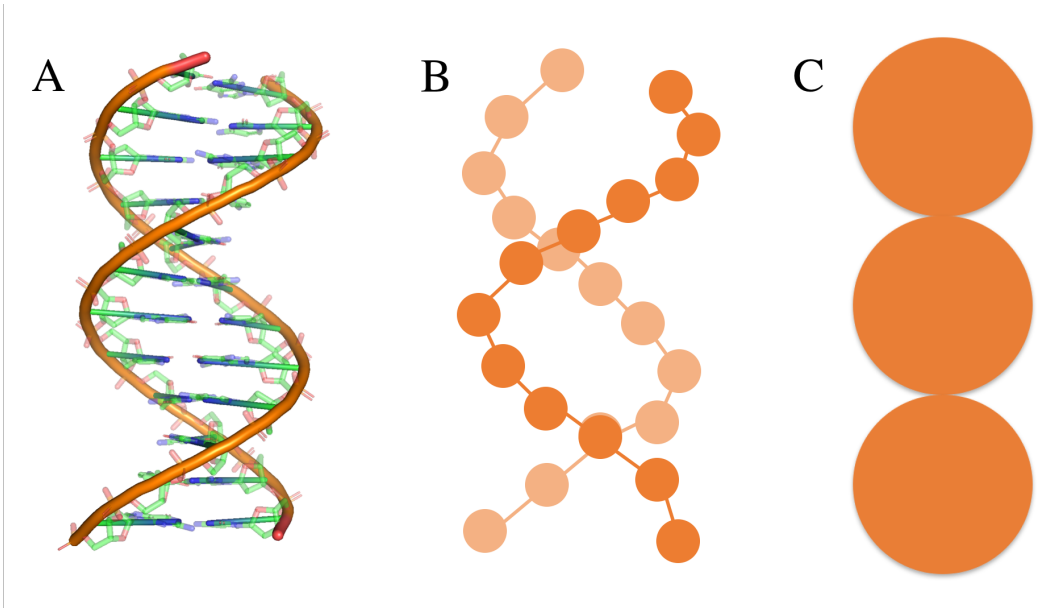

Figure 1.6 DNA models at different resolution. (A) All-atom model. (B) One bead per nucleotide model. (C) One bead per four basepairs model.

The level of detail to include in CG models mainly depends on the phenomenon of interest. Biological processes where chemical interactions or a specific DNA sequence may play a role require finer details in the CG models, whereas for the studies investigating DNA at larger scales such as DNA supercoiling or 3D structure of chromosomes, coarser models could be used. The CG DNA model with a resolution of three particles per nucleotide developed by de Pablo *et al*. [95] has been used to study DNA hybridization, bubble formation and thermodynamics of DNA melting [95, 98, 99]. Simulations of another CG model which uses one particle per nucleotide have provided insight into DNA persistence length and rigidity [96, 97]. On the other hand, worm-like-chain or beads-on-a-string models with particles smaller than the persistence length of DNA, corresponding to several tens of base pairs, have been investigated to understand thermodynamic and kinetic properties of DNA supercoiling [16, 18, 93, 94, 100]. At the extreme end, there are also some CG models available in which hundreds to thousands of basepairs are represented by a single particle to study chromosome structure and organization [28, 101].

In addition to the CG models of biological systems, CG water models have also been developed for solvation [97]. Nevertheless, the implicit treatment of the water environment is more frequently performed in CG simulations in spite of the fact that the application of implicit solvation does not only neglect the interactions between solute and solvent or solvent and solvent, but also excludes the friction and random collisions caused by solvent molecules. This omission can be corrected by addition of two extra terms to the Newton's equation:

$$F_i - \xi v_i + \vec{R}(t) = -\Delta_i U(\vec{X}) - \xi v_i + \vec{R}(t) = m_i a_i \tag{1.13}$$

where ξ is the frictional coefficient, v_i is the velocity of particle i and $\vec{R}(t)$ is a random force vector. The rest of the terms are the same as in Equation 1.6. Equation 1.13 is a stochastic differential equation and called the Langevin equation. The mean of the random force in the equation, $\vec{R}(t)$ is zero and its variance is related to the parameter ξ by fluctuation-dissipation theorem:

$$< \vec{R}(t) >= 0, \qquad < \vec{R}(t)\vec{R}(t') >= 2\gamma k_B T m_i \delta(t - t') \tag{1.14}$$

where γ is the collision frequency and equal to ξ/m_i, k_B is the Boltzmann's constant, T is the temperature, m_i is the mass and δ is the Dirac delta. The collision frequency, γ can be estimated via Stokes' law:

$$\gamma = \frac{\xi}{m_i} = \frac{6\pi\eta\alpha}{m_i} \tag{1.15}$$

where η is the solvent viscosity and α is the radius of the particle i. Considering the magnitude of γ and the integration time step, Langevin motion can be divided into three parts: (1) $\gamma\delta t \leq 1$ where the effect of solvent is minimal, (2) $\gamma\delta t \geq 1$ describes the diffusive limit of the Langevin equation, (3) the intermediate values. The second part describes the Brownian motion which assumes a more

random movement due to the high number of collisions with the solvent. In this case, the momentum relaxation occurs much faster than the inertial relaxation of the particle, and the inertial motion does not affect the motion anymore. Therefore, the right-hand side of the Equation 1.13 can be neglected:

$$F_i + \vec{R}(t) = -\Delta_i U(\vec{X}) + \vec{R}(t) = \xi v_i \tag{1.16}$$

Equation 1.16 is the simplified version of the Langevin equation and the basis of Brownian Dynamics (BD) simulations. In comparison to MD, BD simulations can use larger time steps which allow longer simulations and sample larger conformational space due to the random force which can help crossing barriers.

The most common algorithm to solve Equation 1.16 in order to get new positions of the particles is Ermak & McCammon first-order algorithm [102]:

$$\vec{X}^{n+1} = \vec{X}^n + \frac{\Delta t}{k_B T} D^n F^n + R^n \tag{1.17}$$

where $\vec{X}^n$ and $\vec{X}^{n+1}$ are the initial and new position of the particle, D^n is the translational diffusion coefficient, F^n and R^n are the actual and random force acting on the particle. Later, second- [103] and third-order [104] BD algorithms have also been developed which provide better accuracy.

BD simulations have been extensively applied in biomolecular studies from molecular to cellular levels [105] since large biomolecules in water environment also follow Brownian motion. The CG models for closed circular DNA have also been simulated via BD to investigate DNA supercoiling [16, 18, 93, 94, 100]. In those studies, some modifications were applied in order to also take the torsional motion of DNA into account. Similar to the translational motion (Equation 1.16), the torsional motion is described by Equation 1.18:

$$T_i + \vec{Q}(t) = -\Delta_i U(\vec{\phi}) + \vec{Q}(t) = \xi_r \dot{\phi}_i \tag{1.18}$$

where T_i and $\vec{Q}(t)$ are the actual and random torques acting on the particle i, ξ_r is the rotational friction coefficient and ϕ is the torsional angle of the particle i. T_i is obtained from the derivation of the potential energy with respect to ϕ. Therefore, the torsional displacement can be calculated as:

$$\vec{\phi}^{n+1} = \vec{\phi}^{n} + \frac{\Delta t}{k_B T} {D_r}^n T^n + Q^n \tag{1.19}$$

where $\vec{\phi}^n$ and $\vec{\phi}^{n+1}$ are the initial and new torsion angles of the particle, ${D_r}^n$ is the rotational diffusion coefficient, T^n and Q^n are the actual and random torques acting on the particle. The CG model and the additional torsional motion algorithm described here are also used within this dissertation and the model details can be found in Chapter 4.

Lastly, in addition to the simulation techniques described above, Monte Carlo (MC) method is also widely used for biomolecular studies. It samples the conformational space by random moves as opposed to Newton's or Langevin's law of motion in MD/BD simulations. At each step of the algorithm, a new configuration is generated via randomly changing a structural property and its energy is calculated. The new configuration is then accepted or rejected via the following criteria:

$$p(X \to X_n) = \begin{cases} 1 & for\ U(X_n) \le U(X) \\ \exp(-(U(X_n) - U(X))/k_B T) & otherwise \end{cases}$$

where $U(X_n)$ and $U(X)$ are the energies of the new and old configurations and $p(X \to X_n)$ is the probability of accepting the move. After sufficient number of MC steps run, the Boltzmann distribution of the configurations is obtained. Therefore, this method provides a distribution of conformational states rather than a time evolution as in MD. On the other hand, observing larger

number of configurations compared to MD is possible due to the random moves which prevents getting stuck in local minima. MC sampling is also applied within this dissertation to generate 3D chromosome structures and the details of MC moves can be found in Chapter 4.

In addition to the Chapters 2 and 3 where the results of MD simulations of all-atomic DNA models are discussed, the effect of nucleoid as a crowding agent on diffusive properties of proteins acquired from BD simulations of CG model systems of nucleoid and proteins are presented in Chapter 5. On the other hand, in line with the CG BD simulations, a multiscale modeling algorithm is proposed to generate 3D structures of nucleoids at base-pair resolution using MC and Langevin dynamics simulations (Chapter 4).

CHAPTER 2

Conformational Preferences of DNA in Reduced Dielectric Environments

Asli Yildirim, Monika Sharma, Bradley Michael Varner, Liang Fang, Michael Feig

Adapted from

Journal of Physical Chemistry B 2014, *118*, 10874–10881.

2.1 Abstract

The effect of reduced dielectric environments on the conformational sampling of DNA was examined through molecular dynamics simulations. Different dielectric environments were used to model one aspect of cellular environments. Implicit solvent based on the Generalized Born methodology was used to reflect different dielectric environments in the simulations. The simulation results show a tendency of DNA structures to favor noncanonical A-like conformations rather than canonical A- and B-forms as a result of the reduced dielectric environments. The results suggest that the reduced dielectric response in cellular environments may be sufficient to enhance the sampling of A-like DNA structures compared to dilute solvent conditions.

2.2 Introduction

DNA is an essential biomolecule due to its role in numerous biological processes. Its structure, flexible and sensitive to environmental conditions, is one of the most crucial physical determinants of its biological roles. The major conformations of DNA, B-form and A-form, are well-known from in vitro studies; the B-form is the major conformation in solution [5] while the A-form is seen for GC-rich sequences under low-humidity environments and, sometimes, when bound to other biomolecules [4]. The structural preferences and transitions of DNA between these two forms have been studied extensively for many years not only to gain insight into the structure–function relationship of DNA, but also because DNA is a fascinating biophysical model system that exquisitely balances electrostatic and solvent interactions [7-9, 56-62, 79, 106-119]. Experimental studies on DNA structure have shown that B- to A-form transitions occur under conditions of "low water activity", for example, when ethanol is introduced as a cosolvent [7, 8, 106, 110] and also when salt is added to solution which bridges the phosphate groups in the major

groove of A-DNA [9, 112, 116]. These findings have also been reproduced by computational approaches, especially by molecular dynamics (MD) simulation techniques for the exploration of these conformational preferences in atomistic detail [57-62, 79, 107, 108, 115, 117, 118]. These MD studies showed that DNA maintained its B-DNA form in aqueous solvent, while A-DNA was stabilized in solutions containing explicit salt ions or cosolvents such as ethanol as expected from the experiments [57-60, 115, 117, 118]. The effect of salt on DNA structure is easily rationalized by increased electrostatic screening that allows subsequent phosphate groups along the backbone to come closer to each other so that A-DNA can be formed. It is less clear, however, whether cosolvents affect DNA structure through specific molecular interactions or via a more general physical effect such as an altered dielectric response of the environment.

Reduced dielectric environments are also interesting in the context of understanding DNA structure in the cellular context. Cellular environments are highly crowded systems with high concentrations of proteins, nucleic acids, and numerous cosolvents. It has been reported that 20–30% of the cellular volume is occupied by macromolecules [120]. Previous studies have shown that the effect of such crowding on protein structure and dynamics can be significant [35]. Crowded environments alter the balance between enthalpic and entropic contributions to the conformational free energy either directly through macromolecular interactions or indirectly by modifying solvent properties [35, 39-46, 121-125]. Essentially, the effect of crowding is threefold: (1) volume exclusion by surrounding macromolecules resulting in an entropic penalty for assuming more expanded states [121, 124]; (2) reduced dielectric response of the environment due to the displacement of 20–30% of the water with less polar macromolecules [126-131] and due to slowed dynamics of the water itself [42, 132]; (3) specific interactions between different macromolecules. Therefore, studying the effect of reduced dielectric environments on DNA structure addresses one

major consequence of cellular crowding. Such an approach neglects the effect of specific interactions between DNA and protein crowder molecules as well as the volume-exclusion effect, but it allows for more fundamental insight about how DNA structure is altered by different environments.

In this study, we rely on MD simulations using an implicit continuum dielectric model based on the Generalized Born (GB) methodology [87, 89, 133] to directly observe the effect of reduced dielectric environments on the conformational sampling of DNA. The GB methodology approximates the solvation free energies obtained from Poisson(-Boltzmann) theory in a numerically convenient and computationally efficient manner that is suitable for the application in MD simulations. This methodology has been applied previously to study peptides in reduced dielectric environments [45, 134]. We are focusing here on reduced dielectric constants of 20 and 40 that are compared with $\varepsilon = 80$. The reduced values cover the effective dielectric response of ethanol–water mixtures as well as that of cellular environments. As mentioned above, 20–30% of water in the cell is replaced by macromolecules with an internal dielectric constant of 2–20 [126-130] and cosolvents [131]. Furthermore, water itself was found to have a reduced dielectric constant in crowded environments [42, 132]. As a result, the effective total dielectric constant of cellular environments is assumed to be in the range of 20–60 depending on the degree of crowding [45, 134, 135]. We considered only scalar, static dielectric constants here as the dielectric response of aqueous solvent is isotropic and largely independent of frequency for the range of dynamics considered here (ps–μs). Implicit solvent simulations of DNA using GB models with $\varepsilon = 80$ were previously shown to closely approximate the conformational sampling of DNA seen in explicit solvent [136-139]. Therefore, we have confidence in applying the implicit solvent methodology to nucleic acids.

2.3 Methods

Replica exchange molecular dynamics simulations [91] of $d(CGCGAATTCGCG)_2$ and $d(CGCCCGCGGGCG)_2$ dodecamers were performed to enhance the sampling by accelerating barrier crossings at elevated temperatures. The Generalized Born with molecular volume (GBMV) implicit solvation model [87, 89, 133] was used with dielectric constants 20, 40, and 80. The initial $d(CGCGAATTCGCG)_2$ structure was obtained from X-ray analysis (PDB ID code: 1BNA) [140] and the initial $d(CGCCCGCGGGCG)_2$ structure was obtained by mutating the base sequence in the X-ray structure of 1BNA. Structures were minimized and equilibrated before the replica exchange simulations. Replica exchange simulations were performed using eight replicas between temperatures 300 and 400 K. Although the replicas visit higher temperatures than 300 K, only the sampling at 300 K is considered and reported here. Exchange probabilities of the simulations were around 25–30%. Langevin dynamics [141] was performed using a friction coefficient of 50 ps^{-1} to control the temperature of the system. Simulations were carried out for 50 ns for each replica with a total simulation time of 400 ns and replica exchange was attempted every 10 ps. The first 6 ns of each replica was excluded during the analysis of the simulations. Nonbonded interactions were cut off at 18 Å with a switching function becoming effective at 16 Å and a cutoff at 20 Å was used for the nonbonded list.

All simulations were performed using the CHARMM program package (v c37a2) [142] with CHARMM36 force field [143, 144] for nucleic acids. Replica exchange simulations were carried out using the Multiscale Modeling Tools for Structural Biology (MMTSB) [145] in combination with CHARMM. Cluster analysis of the conformations was carried out with the kclust program in MMTSB. Helicoidal and backbone parameters of dodecamers were analyzed by using the 3DNA

program package [146]. VMD [147] and PyMOL [148] were used for the visualization of structures.

2.4 Results

As test systems, we considered the Drew-Dickerson dodecamer d(CGCGAATTCGCG)$_2$ [140] known to be exceptionally stable in B-form and the GC-rich dodecamer d(CGCCCGCGGGCG)$_2$ [149] which tends toward A-form conformations under conditions of high salt, low humidity, or in the presence of cosolvents. Simulations of both systems were started from the standard B-form experimental structure of the Drew-Dickerson dodecamer and we hypothesized that a reduced dielectric response may favor A-like structures akin to the effect of cosolvents shifting the A/B balance toward A. This is indeed what we observe as described in more detail below, although the A-like conformations we observe here appear to be distinct from salt-induced A-DNA. In the following, results are described and discussed.

DNA conformations are distinguished by the orientation of their bases, captured by helical parameters, and by backbone torsional angles. Both were analyzed from our simulations and will be described in the following. The most distinctive helical parameters to characterize A- and B-type conformations are the base inclination relative to the helical axis, the twist between subsequent base pairs, and the x-displacement of base pairs from the helical axis. Secondary parameters of common interest are the relative displacement between subsequent base pairs along the base pair axis (slide) and along the helical axis (rise) as well as the propeller twist between bases forming a base pair. The average values from the simulation are given in Tables 2.1 and 2.2. Distributions are shown as Figures 2.1 and 2.2. The simulation results are compared with averaged helical parameters for ten A-form and ten B-form DNA duplexes and for the X-ray structures of

the exact sequences that were studied here. In all analyses, only the eight inner base pairs were taken into consideration, because of structural distortions at the duplex ends due to occasional base fraying.

Table 2.1 Helicoidal parameters for the GC-rich dodecamer compared to experimental results.

		Canonical[b]		MD Simulations[c]		
	X-ray[d]	A-DNA	B-DNA	ε=20	ε=40	ε=80
Propeller (deg)	–12.49 (*2.0*)	–9.00 (*7.1*)	–11.72 (*5.8*)	2.16 (*15.5*)	–2.85 (*12.0*)	–7.26 (*14.1*)
Slide (Å)	–1.71 (*0.4*)	–1.67 (*0.4*)	0.17 (*0.7*)	–1.48 (*0.8*)	–1.23 (*0.9*)	–0.52 (*1.4*)
Twist (deg)	29.59 (*3.5*)	30.24 (*4.0*)	34.81 (*5.7*)	27.50 (*10.0*)	29.70 (*7.3*)	36.27 (*17.7*)
X-displacement (Å)	–5.01 (*1.1*)	–4.54 (*1.3*)	–0.13 (*1.1*)	–2.68 (*3.6*)	–2.80 (*2.9*)	–1.19 (*2.5*)
Helical rise (Å)	2.66 (*0.6*)	2.77 (*0.5*)	3.25 (*0.2*)	3.42 (*1.0*)	3.36 (*0.8*)	3.36 (*0.8*)
Inclination (deg)	20.71 (*11.4*)	16.00 (*10.7*)	4.01 (*6.7*)	0.45 (*20.7*)	3.73 (*15.1*)	5.43 (*16.4*)

[a] (a) All values are averaged over nonterminal base pairs with standard deviations given in parentheses. (b) Averages over the A-form structures 3V9D, 3QK4, 2B1B, 1ZEX, 1ZEY, 1ZF1,1ZF6, 1ZF8, 1ZF9, 1ZFA and the B-form structures 2M2C, 4AGZ, 4AH0, 4AH1, 3U05, 3U08, 1VTJ, 3U2N, 3OIE, 3BSE. (c) Averages over snapshots at 300 K from replica exchange MD simulations. (d) PDB ID code for the X-ray structure is 399D.

Table 2.2 Helicoidal parameters for the Drew-Dickerson dodecamer compared to experimental results.

		Canonical[b]		MD Simulations[c]		
	X-ray[d]	A-DNA	B-DNA	ε=20	ε=40	ε=80
Propeller (deg)	–13.34 (*5.9*)	–9.00 (*7.1*)	–11.72 (*5.8*)	-6.93 (*12.5*)	–10.67 (*12.7*)	–11.32 (*13.7*)
Slide (Å)	0.07 (*0.5*)	–1.67 (*0.4*)	0.17 (*0.7*)	–0.86 (*0.8*)	–0.50 (*0.7*)	–0.32 (*0.7*)
Twist (deg)	34.22 (*5.7*)	30.24 (*4.0*)	34.81 (*5.7*)	30.62 (*6.4*)	32.36 (*5.9*)	33.12 (*6.7*)
X-displacement (Å)	–0.23 (*0.5*)	–4.54 (*1.3*)	–0.13 (*1.1*)	–2.33 (*2.8*)	–1.65 (*2.3*)	–1.23 (*2.4*)
Helical rise (Å)	3.29 (*0.1*)	2.77 (*0.5*)	3.25 (*0.2*)	3.24 (*0.7*)	3.25 (*0.6*)	3.18 (*0.7*)
Inclination (deg)	4.02 (*7.2*)	16.00 (*10.7*)	4.01 (*6.7*)	7.26 (*19.1*)	8.00 (*16.7*)	6.30 (*19.8*)

[a] (a) All values are averaged over nonterminal base pairs with standard deviations given in parentheses. (b) Averages over the A-form structures 3V9D, 3QK4, 2B1B, 1ZEX, 1ZEY, 1ZF1,1ZF6, 1ZF8, 1ZF9, 1ZFA and the B-form structures 2M2C, 4AGZ, 4AH0, 4AH1, 3U05, 3U08, 1VTJ, 3U2N, 3OIE, 3BSE. (c) Averages over snapshots at 300 K from replica exchange MD simulations. (d) PDB ID code for the X-ray structure is 1BNA.

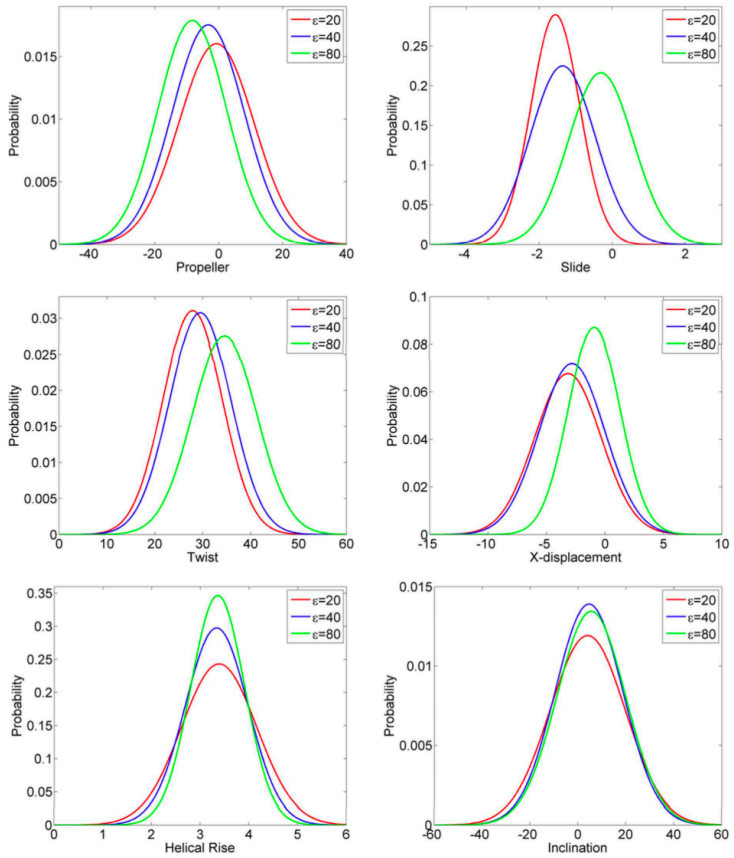

Figure 2.1 Distributions of helicoidal parameters for the GC-rich dodecamer in different dielectric environments with ε=20 (red), 40 (blue) and 80 (green).

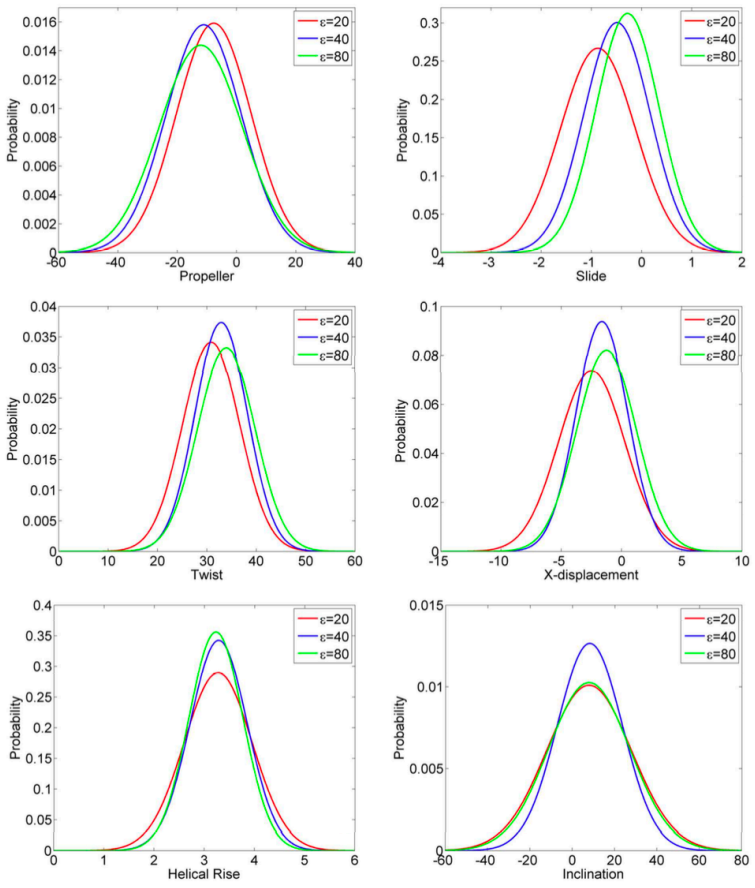

Figure 2.2 Distributions of helicoidal parameters for the Drew-Dickerson dodecamer in different dielectric environments with ε=20 (red), 40 (blue) and 80 (green).

X-displacement is one of the clearest indicators of A- vs BDNA conformations. It is expected to have near zero values for B-DNA and significantly negative values for A-DNA, where bases are displaced from the helical axis. We find that, for both the GC-rich and the Drew-Dickerson dodecamers, reduced dielectric environments resulted in a clear shift toward more negative x-displacement values that are about halfway between canonical B- and A-DNA. The shift was slightly greater for the GC-rich sequence and was equally large at ε = 40 and ε = 20. On the other hand, the Drew-Dickerson dodecamer values changed more gradually as the dielectric constant was reduced. A similar observation was made for the twist angle that changed from values of 33–36°, typical for B-DNA, to 28–31°, typical for A-DNA, as the dielectric constant was changed

from ε = 80 to ε = 20. Again, the change was already complete at ε = 40 for the GC-rich sequence but occurred more gradually for the Drew-Dickerson dodecamer. The inclination of base pairs relative to the helical axis in A-type conformations is typically quite pronounced compared to B-type conformations, which are mostly perpendicular to the helical axis. Interestingly, we find that the average for the GC-rich dodecamer shifted to more B-like values as the dielectric constant was reduced. This appears to be in part due to increased sampling of negative inclination angles at ε = 20 (Figure 2.1). On the other hand, the Drew-Dickerson dodecamer showed a slightly increased inclination angle upon reduction of the dielectric response of the environment. In contrast to what x-displacement and twist values suggest, the changes in base inclination angles are not fully consistent with a classical transition from B-DNA to A-DNA. However, as commented by Dickerson and Ng [150], the inclination angle can be problematic in short helical segments for distinguishing A- from B-form due to the difficulties in separating inclination from local helix bending.

The slide of a base pair along its long axis relative to its neighboring base pair is another measure that can discriminate between A- and B-form DNA structures. While a significant slide is not observed for base pairs in typical B-DNA, A-DNA base pairs have a more pronounced propensity to slide along their long axes. We observed a significant increase in negative slide values for the GC-rich dodecamer at reduced dielectric constants and to a lesser extent for the Drew-Dickerson dodecamers. Another discriminative helical property is helical rise, which is smaller for typical A-DNA, around 2.7–2.8 Å, compared to B-DNA, where it is typically around 3.3–3.4 Å. Interestingly, we found the helical rise to remain unchanged upon reduction of the dielectric constants. Finally, the propeller twist of bases in a base pair with respect to each other is known to be highly sequence dependent with more negative values for A/T base pairs than for C/G

base pairs, but it is also reduced in A-DNA vs B-DNA. As expected, the propeller twist was less pronounced for the GC-rich dodecamers than for the AT-base pair containing Drew-Dickerson dodecamer. There was a trend, however, toward less negative propeller twist values upon a decrease of the dielectric constant. This is again indicative of a transition toward A-like structural features. Overall, the analysis of the helical parameters suggests a tendency toward A-type conformations upon a reduction in the dielectric constant. The tendency toward A-DNA was more pronounced for the GC-rich dodecamer and it manifested itself at higher dielectric constants compared to the Drew-Dickerson dodecamer. The helical parameters in the X-ray structure of the B-form Drew-Dickerson dodecamer agree closely with the simulation results at $\varepsilon = 80$ while the helical parameters of the A-form X-ray structure of the GC-rich dodecamer are reproduced best by the simulation results at $\varepsilon = 20$ and $\varepsilon = 40$ rather than $\varepsilon = 80$.

Among backbone torsion angles, the ribose sugar pseudo-rotation angle is the most significant indicator of A- vs B-DNA conformations. While A-DNA structures commonly have C3′- endo or C2′-exo sugar conformations, sugars in B-DNA structures are more often found in C3′-exo or C2′-endo conformations. The preferred sugar conformations for each base in the GC-rich and the Drew-Dickerson dodecamers in different dielectric environments are depicted in Figure 2.3. As expected from DNA simulations with the CHARMM force field [143], there was extensive sampling of both C2′-endo and C3′- endo conformations at $\varepsilon = 80$ but with an overall preference for C2′-endo conformers. With a decreasing dielectric constant of the environment, sugar conformations in the GC-rich dodecamer were switched largely to C3′-endo or C2′-exo conformations, further indicating a B to A transition for this sequence. In contrast, the ribose sugars in the Drew-Dickerson dodecamer were largely unaffected by a change in the dielectric environment, suggesting that the backbone retained B-like features.

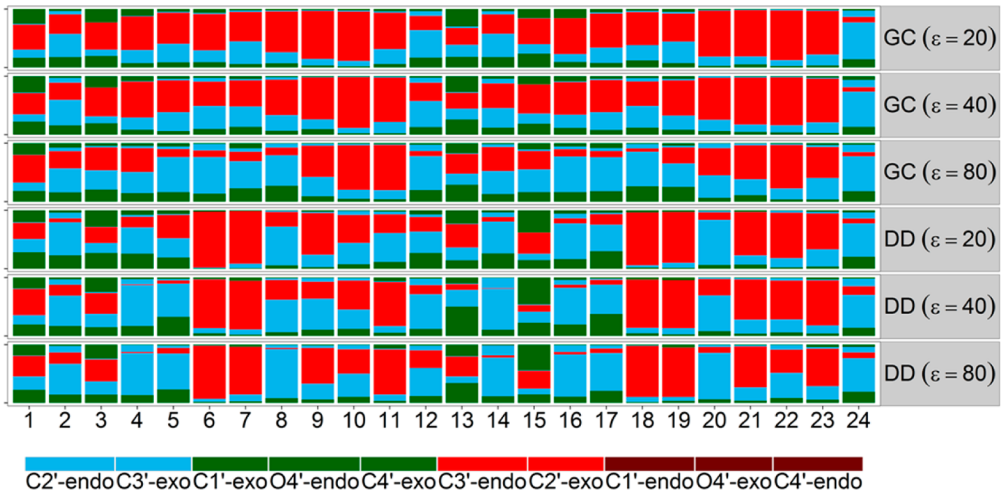

Figure 2.3 Sugar pucker conformations of each base from simulations of the GC-rich dodecamer (GC), the Drew-Dickerson dodecamer (DD) in different dielectric environments with ε = 20, 40, and 80.

We analyzed the ε and ζ backbone torsion angles as an indicator of the relative sampling of BI and BII conformations. A value of ε – ζ near –90° corresponds to the BI conformation whereas +90° characterizes BII conformations. The potential of mean force as a function of ε/ζ is given in Figure 2.4. Frequent BI/BII transitions were observed in both dodecamers for all dielectric constants as expected from the most recent version of the CHARMM nucleic acid force field [143] used in this study. Closer inspection shows that the sampling of BII conformations diminished at ε = 20 and ε = 40 for the GC-rich dodecamer and at ε = 20 for the Drew-Dickerson dodecamer. This suggests that excursions to the less-populated BI conformer may be further suppressed in reduced dielectric environments. We also found two minima at ε = 80 corresponding to noncanonical conformations of the GC-rich dodecamers that appear to be missing at the lower dielectric constants. However, it is likely that this observation is due to limited sampling in our simulations and/or structural distortions due to the implicit solvent model used here (see Figure 2.4).

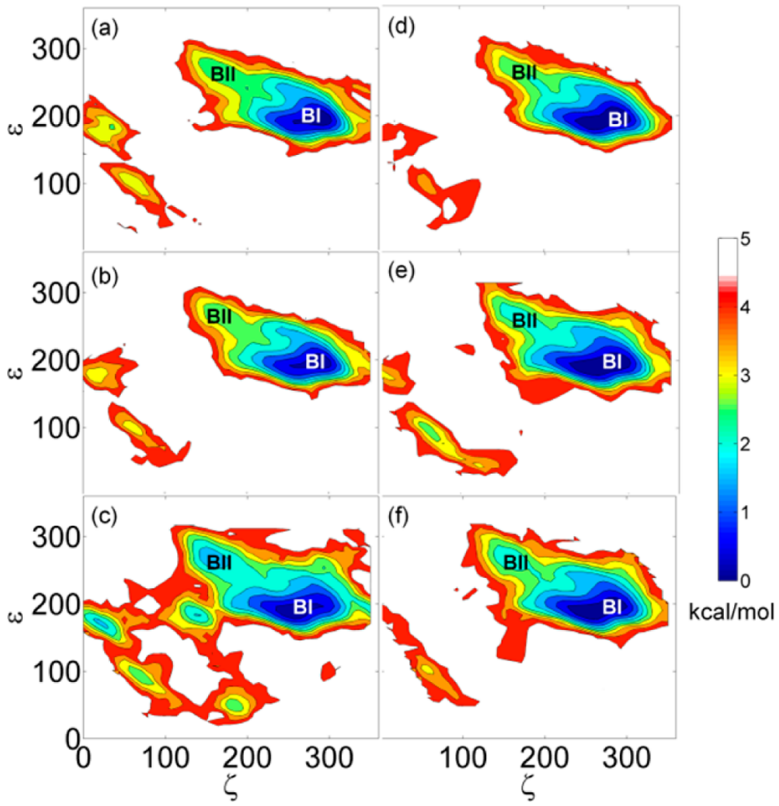

Figure 2.4 Potential of mean force (kcal/mol) from simulations as a function of ε and ζ torsion angles for the GC-rich at ε = 20. (a), ε = 40 (b), and ε = 80 (c) and for the Drew-Dickerson dodecamer at ε = 20 (d), ε = 40 (e), and ε = 80 (f).

Next, we clustered snapshots from each of the simulations to obtain representative structures at different dielectric environments. The resulting structures for both studied dodecamers are shown in Figure 2.5. Using a clustering radius of 3 Å resulted in five major conformations for the GC-rich dodecamer and three major conformations for the Drew-Dickerson dodecamer. The average helical parameters of these clusters were calculated and compared with canonical A-, B-, and C-DNA forms (Tables 2.3 and 2.4). According to this analysis, GC1 and GC2, GC5 and DD3, and DD1 and DD2 have A-like, B-like, and mixed A- and B-type features, respectively, further confirming the prevalence of B-form conformations (GC5 and DD3) at ε = 80 and a shift toward A-like structures at lower dielectric constants. We also observed a minor population of a distorted conformation (GC4) at ε = 80 that may be a simulation artifact. At the lowest dielectric, a new

conformation appears (GC3) that has mixed characteristics of A- and B- type structures. As for the Drew-Dickerson dodecamer, the B-form DD3 conformation persists at ε = 40, but disappears at ε = 20. Instead, A-like structures (DD1 and DD2) appear with a small percentage at ε = 40 while becoming dominant at ε = 20. None of the structures observed here resemble other previously characterized conformations of DNA, such as C-DNA.

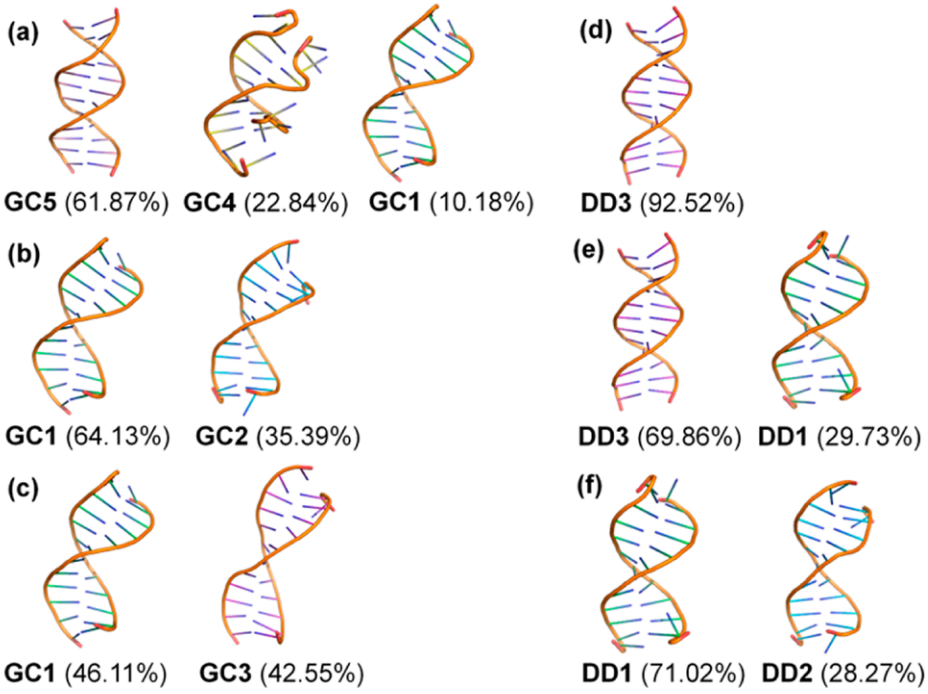

Figure 2.5 Representative conformations from clustering analysis in different dielectric environment for GC-rich dodecamer with ε = 80. (a), 40 (b), and 20 (c) and for the Drew-Dickerson dodecamer with ε = 80 (d), 40 (e), and 20 (f). Cluster populations are given in the parentheses.

Table 2.3 Helicoidal Parameters for GC1, GC2, GC5 clusters compared to canonical A-, B- and C-DNA forms.

	A-DNA	B-DNA	C-DNA	GC1	GC2	GC5
Propeller (deg)	–9.00 (*7.1*)	–11.72 (*5.8*)	-2.01	-2.67 (*11.8*)	–1.78 (*9.7*)	–8.07 (*10.9*)
Slide (Å)	–1.67 (*0.4*)	0.17 (*0.7*)	0.09	–1.23 (*0.5*)	–1.90 (*0.5*)	–0.21 (*0.6*)
Twist (deg)	30.24 (*4.0*)	34.81 (*5.7*)	38.06	30.55 (*5.2*)	27.48 (*5.3*)	34.11 (*5.9*)
X-displacement (Å)	–4.54 (*1.3*)	–0.13 (*1.1*)	0.82	–2.62 (*2.3*)	–3.95 (*2.2*)	–0.88 (*1.7*)
Helical rise (Å)	2.77 (*0.5*)	3.25 (*0.2*)	3.31	3.40 (*0.7*)	3.46 (*0.8*)	3.39 (*0.5*)
Inclination (deg)	16.00 (*10.7*)	4.01 (*6.7*)	-8.07	3.22 (*14.8*)	1.65 (*13.6*)	4.85 (*12.2*)
Minor Groove (Å)	9.92 (*0.7*)	10.77 (*0.9*)	10.6	14.9 (*0.4*)	15.49 (*0.4*)	13.77 (*1.2*)
Major Groove (Å)	7.14 (*2.4*)	17.14 (*0.9*)	16.1	20.6 (*0.9*)	21.70 (*0.9*)	18.04 (*1.3*)

Parameters are calculated for canonical C-DNA form using 3DNA program package [146].

Table 2.4 Helicoidal Parameters for DD1, DD2, DD3 clusters compared to canonical A-, B- and C-DNA forms.

	A-DNA	B-DNA	C-DNA	DD1	DD2	DD3
Propeller (deg)	–9.00 (*7.1*)	–11.72 (*5.8*)	-2.01	-9.65 (*10.9*)	–5.32 (*12.0*)	–12.23 (*12.3*)
Slide (Å)	–1.67 (*0.4*)	0.17 (*0.7*)	0.09	–0.64 (*0.5*)	–1.04 (*0.7*)	–0.41 (*0.5*)
Twist (deg)	30.24 (*4.0*)	34.81 (*5.7*)	38.06	32.47 (*4.8*)	30.41 (*6.2*)	34.12 (*4.9*)
X-displacement (Å)	–4.54 (*1.3*)	–0.13 (*1.1*)	0.82	–1.80 (*2.1*)	–2.55 (*2.8*)	–1.12 (*2.1*)
Helical rise (Å)	2.77 (*0.5*)	3.25 (*0.2*)	3.31	3.23 (*0.6*)	3.22 (*0.7*)	3.21 (*0.5*)
Inclination (deg)	16.00 (*10.7*)	4.01 (*6.7*)	-8.07	7.03 (*17.4*)	5.38 (*20.3*)	4.20 (*18.7*)
Minor Groove (Å)	9.92 (*0.7*)	10.77 (*0.9*)	10.6	14.42 (*0.5*)	14.86 (*0.4*)	13.74 (*0.6*)
Major Groove (Å)	7.14 (*2.4*)	17.14 (*0.9*)	16.1	19.16 (*0.9*)	21.38 (*1.2*)	18.23 (*0.9*)

Parameters are calculated for canonical C-DNA form using 3DNA program package [146].

Finally, we estimated relative conformational free energies using the MMPB/SA approach [118] and compared the effect of reduced dielectric environments to previously described effects of increased salt concentrations on the A- vs B-DNA equilibrium [9, 58, 112, 116, 118]. Calculated free energy differences between A- and B-form structures are given in Tables 2.5 and 2.6 with positive values indicating a preference for the A-form while negative values indicate a preference for the B-form.

The total free energy differences ($\langle E(total) \rangle$) at $\varepsilon = 80$ between simulated B- and A-forms and between simulated B- and canonical A-structures were more positive for the GC-rich dodecamer

(Table 2.5) than for the Drew-Dickerson dodecamer (Table 2.6). This indicates that the A-form is relatively more favorable in aqueous medium for the GC-rich dodecamer as we would expect. Furthermore, at $\varepsilon = 80$, both dodecamers appear to be more favorable in canonical A-DNA than in the A-form structures observed in our simulations. This is also consistent with expectations since past in vitro studies have established the canonical A-DNA structure as the relevant A-form in non-cellular environments instead of the somewhat different A-form reported here. We note that the energetic differences are relatively large (tens of kcal/mol instead of physically probably more realistic values of less than 10 kcal/mol). This is likely a result of the approximate nature of the MMPB/SA methodology and/or limitations in sampling. Overall, we estimate that the errors of the MMPB/SA estimates are on the order of tens of kcal/mol, consistent with previous studies [115, 118, 151]. Nevertheless, the MMPB/SA estimates appear to give at least qualitatively the correct picture.

Addition of salt to the environment is known to favor A- DNA over B-DNA structures [9, 58, 112, 115, 116, 118]. Within the MMPB/SA framework, salt can be considered by solving the Poisson– Boltzmann equation with added salt [118]. While the overall MMPB/SA estimate is prone to uncertainties because, in part, the vacuum force field term is highly sensitive to small conformational changes, and because entropic effects are not fully taken into account, the relative change in the electrostatic solvation term as a result of added salt or a changed dielectric constant can be estimated with much higher accuracy [152]. We find that canonical A-DNA was favored with increasing salt as expected. However, the trend was reversed when the A-like structures from our simulations were considered instead of the canonical forms. Increasing salt concentration appeared to stabilize the B-form over our A-form structures. On the other hand, a reduced dielectric resulted in a stabilization of our A- like structures while canonical A-DNA structures were

destabilized for both dodecamers when compared to ε = 80. This suggests that the effects of increased salt and reduced dielectric environment are different and that, although both changes in the environment appear to lead to A-like structures, the resulting conformational ensembles appear to have distinct features.

Table 2.5 Conformational free energies from MMPB/SA analysis for the GC-rich dodecamer.[a]

	A[b]	B[b]	Canonical A[d]	(B – A)	(B – Can.A)
⟨E(gas)⟩[e]	2181.2 *(4.1)*	2331.1 *(4.1)*	2371.2 *(4.2)*	149.9	-40.1
⟨E(nonpolar)⟩[f]	26.9 *(0.0)*	26.9 *(0.0)*	26.3 *(0.0)*	0.0	0.6
⟨E(PB), ε = 80⟩[g]	-5837.5 *(3.2)*	-5965.5 *(3.5)*	-6069.4 *(3.7)*	-128.0	103.9
⟨E(total), ε = 80⟩	-3629.4 *(4.0)*	-3607.5 *(3.9)*	-3672.0 *(2.3)*	21.9	64.5
⟨ΔE(total), salt: 0.1 M⟩	-40.6 *(0.0)*	-41.1 *(0.0)*	-42.9 *(0.0)*	-0.5	1.8
⟨ΔE(total), salt: 1.0 M⟩	-53.4 *(0.1)*	-54.5 *(0.1)*	-58.5 *(0.1)*	-1.1	4.0
⟨ΔE(total), ε = 20⟩	242.6 *(0.2)*	247.3 *(0.2)*	261.1 *(0.3)*	4.7	-13.8

(a) Standard errors are given in the parentheses. (b) 100 snapshots selected from GC1 and GC2 clusters closest to cluster centers and with minimal distortions at terminal base pairs. (c) 100 snapshots selected from GC5 cluster. (d) Snapshots from 1 ns restrained implicit solvent simulation of canonical A-DNA structure taken at 5 ps intervals starting from 0.5 ns. (e) CHARMM force field energy in vacuum. (f) Solvent-accessible surface area term: γ*SASA+β with γ = 0.00542 kcal/mol/Å2 and β = 0.92 kcal/mol [153]. (g) Solution of Poisson–Boltzmann equation using PBEQ module [142] in CHARMM program with a grid spacing of 0.25 Å.

Table 2.6 Conformational free energies from MMPB/SA analysis for the Drew-Dickerson dodecamer.[a]

	A[b]	B[b]	Canonical A[d]	(B – A)	(B – Can.A)
⟨E(gas)⟩[e]	2742.6 *(5.8)*	2812.8 *(7.3)*	2944.2 *(5.3)*	70.2	-131.4
⟨E(nonpolar)⟩[f]	26.9 *(0.0)*	26.7 *(0.0)*	26.2 *(0.0)*	-0.2	0.5
⟨E(PB), ε = 80⟩[g]	-5818.1 *(3.9)*	-5931.7 *(6.1)*	-6090.8 *(4.6)*	-113.6	159.1
⟨E(total), ε = 80⟩	-3048.6 *(5.5)*	-3092.3 *(3.5)*	-3120.4 *(2.0)*	-43.7	28.1
⟨ΔE(total), salt: 0.1 M⟩	-40.5 *(0.0)*	-41.1 *(0.1)*	-43.1 *(0.0)*	-0.6	2.0
⟨ΔE(total), salt: 1.0 M⟩	-53.1 *(0.1)*	-54.2 *(0.1)*	-59.1 *(0.1)*	-1.1	4.9
⟨ΔE(total), ε = 20⟩	239.2 *(0.2)*	245.0 *(0.3)*	260.4 *(0.3)*	5.8	-15.4

(a) Standard errors are given in the parentheses. (b) 100 snapshots selected from DD1 and DD2 clusters as in Table 2.5. (c) 100 snapshots selected from DD3 cluster as in Table 2.5. (d) Snapshots from 1 ns restrained implicit solvent simulation of canonical A-DNA structure taken at 5 ps intervals starting from 0.5 ns. (e) CHARMM force field energy in vacuum. (f) Solvent-accessible surface area term: γ*SASA+β with γ = 0.00542 kcal/mol/Å2 and β = 0.92 kcal/ mol [153]. (g) Solution of Poisson–Boltzmann equation using PBEQ module [142] in CHARMM program with a grid spacing of 0.25 Å.

2.5 Discussion and Conclusions

The main goal of this study has been to examine the conformational sampling of DNA in reduced dielectric environments to both understand the origin of cosolvent induced A-DNA stabilization as well as explore the possible effects of cellular environments. At ε = 80 we observe

stable sampling of mostly B-DNA structures with the implicit solvent methodology used here in agreement with experimental expectations. This suggests that the implicit solvent methodology can reasonably describe the conformational sampling of DNA in aqueous solutions. Additional simulations were then carried out at lower dielectric constants. This approach emphasizes mean field solvation properties while neglecting specific molecular interactions with the environment. %85 (v/ v) Ethanol/water solutions have dielectric constants of 30 [154]. At the same time, while $\varepsilon = 20$ may be at the low end of what is found in crowded environments, an effective dielectric near $\varepsilon = 40$ is assumed to be a typical value for crowded biological cells [45, 134].

Our results indicate an overall shift toward A-like conformations that was moderate at $\varepsilon = 40$ and became more pronounced at $\varepsilon = 20$ for both studied dodecamers. The tendency to form A-like conformations was greater for the GC-rich dodecamer, but even the classical B-DNA Drew-Dickerson dodecamer exhibited A-like features in low-dielectric environments. The transition to canonical A-form DNA was nearly complete for the GC-rich dodecamer, but the Drew-Dickerson dodecamers retained many B-like features even at $\varepsilon = 20$. However, even for the GC-rich dodecamer at $\varepsilon = 20$, the resulting structures differ somewhat from canonical A-form. MMPB/SA analytical results further show that increased salt appears to destabilize the A-like structures seen in our simulations, while it stabilizes canonical A-DNA. This suggests the important conclusion that salt effects and reduced dielectric environments, for example, by reducing "water activity" through cosolvents or crowding, affect DNA structure in similar but different ways. The results are not fully consistent with fiber diffractions studies where DNA remains in canonical B-DNA form with its two hydration layers [155]. However, the fiber environment is likely not as packed as crowded environments, and hence, the effective dielectric environment may not actually be reduced dramatically. Furthermore, DNA–DNA interactions and the presence of salt in fibers may

further explain differences between the fiber experiments and our findings. Specific interactions of DNA with proteins in crowded cellular environments as well as volume exclusion effects were also not considered here and may significantly modulate the results reported here. Nevertheless, any tendency toward A-DNA features under crowded cellular conditions could have significant energetic consequences for protein–DNA interactions since most DNA bound to proteins is in B-form.

It is clear that further studies are needed to fully understand DNA structure in crowded environments and in the presence of cosolvents. One specific question to be addressed is whether protein–DNA interactions could counteract the apparent tendency toward A-form DNA and modulate these findings. Furthermore, the studies performed here involved short oligomers vs the very long polymeric DNA found in vivo. Further investigations are also required to see whether longer DNA molecules would exhibit the same tendencies. It is our hope that the results reported will also stimulate new experimental studies of DNA structure under crowded conditions none of which are available to our knowledge.

2.6 Acknowledgements

Funding from NSF (MCB 1330560) and NIH (R01 GM092949) is acknowledged.

CHAPTER 3

Protein Interactions Stabilize Canonical DNA Features in Simulations of DNA in Crowded Environments

Asli Yildirim, Nathalie Brenner, Robert Sutherland, Michael Feig

Submitted to

Journal of Physical Chemistry B

3.1 Abstract

The effect of protein crowding on the conformational preferences of DNA is described from fully atomistic molecular dynamics simulations of systems containing a DNA dodecamer surrounded by protein crowders. Analysis of the simulations show that DNA structures prefer to stay in B-like conformations in the presence of the crowders. The preference for B-like conformations results from non-specific interactions of crowder proteins with the DNA sugar-phosphate backbone. The results are complementary to a previous study of DNA in reduced dielectric environments suggesting that the reduced dielectric response of cellular environments and non-specific interactions with protein crowders have opposite impacts on DNA structure under in vivo conditions.

3.2 Introduction

Biological cells are highly crowded environments due to the presence of various macromolecules. The macromolecular crowding in cells plays a crucial role in biological processes as it may alter the structure and dynamics of biomolecules [35]. A typical biological cell has a concentration of biomolecules in the range of 300 – 400 mg/ml [156], corresponding to a macromolecular volume fraction of 20 – 30 % [120]. Such an environment is substantially different from dilute solutions, the frequently considered environment in most biological experiments. Recent studies have begun to consider the effects of cellular crowding and have shed light on its effects on the structure and function of biomolecules [43, 122, 157-160]. Three essential crowding effects have been reported from experiments[161] and simulations [162]: (1) the volume exclusion effect has been suggested to favor more compact conformations based on entropic arguments, thereby generally stabilizing

more compact states [163, 164]; (2) non-specific interactions between biomolecules and surrounding protein crowders have led to the destabilization of native states [44, 46, 165] as well as reduced diffusion [166]; and (3) altered solvation properties including reduced dynamic and dielectric properties [42] have implied a reduced hydrophobic effect [134, 167].

While much attention so far has been on proteins, nucleic acids are also affected by macromolecular crowding [50, 51]. G-quadruplex DNA structure assumes a parallel-G quadruplex form under crowded environments due to the excluded volume effect as well as alterations in the hydration of DNA [52-54]. Long DNA duplexes undergo a collapsing transition in the presence of polyethylene glycol (PEG) in solution, which can also be explained by the volume exclusion effect favoring states that are more compact [63, 64]. The negatively charged protein bovine serum albumin (BSA) similarly causes a compaction of large DNA molecules due to the volume exclusion effect and repulsive electrostatic interactions [168]. Short DNA duplexes, on the other hand, have been extensively investigated by both experimental techniques and computer simulations in terms of co-solvent and salt effects [7, 8, 55-62]. The DNA duplex is well-known to be most stable in the B-form [169] in aqueous solution and in A-form in environments with depleted water and for certain sequences [4]. High concentrations of salt can also induce the B- to A- form transition by bringing the negatively charged phosphate groups of DNA closer [112, 116, 170-172] while the addition of ethanol favors the A-form due to reduced electrostatics [8, 57, 59, 110, 118, 173, 174]. More recently, the effect of reduced dielectric environments on DNA as one aspect of cellular crowding was investigated and has also been shown to favor non-canonical A-form structures in implicit solvent simulations [167]. However, the effect of explicit protein crowder molecules on DNA duplex structures is still not well understood.

Here, we describe fully atomistic molecular dynamics (MD) simulations of DNA dodecamers in the presence of explicit protein crowders in order to investigate how DNA structure and stability may be affected under such conditions. We find a general tendency of the DNA to favor the B-form in crowded environments, which is in contrast to the shift towards A-form DNA observed in the simpler reduced dielectric environments [167]. The stabilization of B-DNA appears to be due to non-specific protein-DNA interactions. We also observe, some alterations in the hydration structure and ion distributions around DNA under crowded conditions. The results are described in detail and discussed in the following after outlining the computational methods used in this study.

3.3 Methods

MD simulations of Drew-Dickerson ($(CGCGAATTCGCG)_2$) and GC-rich ($(CGCCCCGCGGGCG)_2$) dodecamers in crowded protein environments were carried out using the CHARMM program package (v41a1) [175] with the CHARMM36 force field [71, 72]. The initial Drew-Dickerson dodecamer structure was obtained from the X-ray structure (PDB: 1BNA)[140], and the initial GC-rich dodecamer structure was obtained by mutating the base sequence in the X-ray structure of the Drew-Dickerson dodecamer using the MMTSB Tool Set [145]. The Drew-Dickerson dodecamer is very stable in B-form[140], while the GC-rich dodecamer prefers to stay in A-form in environments with co-solvents or high salt concentrations [149]. For each dodecamer, a dilute system without crowders (0% crowder fraction) and three systems with different protein crowder volume fractions (20%, 30%, 40%) were prepared. Protein G (PDB: 1PGB) [176] was selected to be used as a crowder protein due to its small size and stability in computer simulations. In our simulations, we used neutral protein G molecules, which were obtained by the protonation

of D36, D40, E19 and E42. Protein G is not known to specifically interact with DNA and the neutral form minimizes electrostatic interactions with the highly charged DNA to focus on the more general crowding effects. The crowded systems (20%, 30%, 40%) consisted of one dodecamer and 8 protein G molecules, whereas the dilute systems only contained one dodecamer. Simulation box sizes were varied between 53.2 – 61.3 Å to obtain the abovementioned crowder volume fractions. Simulation conditions of the systems are given in Table 3.1. The initial crowded systems were set up by randomly placing the DNA dodecamer and the crowder proteins in the simulation box. All systems were solvated with explicit TIP3P [177] water molecules. To neutralize the DNA dodecamer, 22 sodium ions were added to the systems. In order to keep the ion molality of all systems the same, 6 and 12 additional pairs of sodium and chloride ions were added to 30 % and 0/20 % systems, respectively. Therefore, all systems had 0.45 mol/kg ion molality.

The initial systems were minimized for 1000 steps using the adopted basis Newton Raphson (ABNR) algorithm and were subsequently heated by running simulations without using any restraints at 50K, 100K, 150K, 200K, 250K for 4 ps and at 298K for 10 ps. Productions runs were carried out at 298 K in the NVT ensemble for 1 μs with a 2 fs time step. The SHAKE algorithm [178] was used to constrain bond lengths involving hydrogen atoms. Temperature control was obtained by a Langevin thermostat with a 0.01 ps^{-1} friction coefficient. Lennard-Jones and direct electrostatic interactions were cut off at 12 Å with a switching function becoming effective at 10 Å. Electrostatic interactions were calculated from particle-mesh Ewald[179] summation using 1 Å grid spacing. All simulations were performed using periodic boundary conditions. For the crowded systems, five independent simulations were carried out starting from different initial orientations.

For the dilute systems, simulations were replicated three times starting from different initial velocities for the atoms. The total length of all of the simulations reported here is 36 μs.

The analysis of the helicoidal and backbone parameters of the dodecamers were performed by using the 3DNA program package [146]. Radial distribution functions and 3D volume densities were analyzed by using in-house scripts. All the other analysis was carried out using the Multiscale Modeling Tool for Structural Biology Tool Set (MMTSB) [145] in combination with CHARMM [175]. Only the last 700 ns of the simulations were analyzed because of larger variations in the helicoidal parameters during the first 300 ns. Only the inner eight base-pairs were taken into consideration to ignore structural distortions due to base fraying. VMD [180] and PyMOL [181] were used for visualization.

Table 3.1 Simulation conditions.

DNA Dodecamer	Box Size (Å)	Protein Vol (%)	Protein Conc. (g/L)	Ion Molarity (M)	Ion Molality (mol/kg)	Simulation Length (μs)
Drew-Dickerson	54.62	0	0.00	0.45	0.45	1
Drew-Dickerson	61.02	20	362.49	0.32	0.45	1
Drew-Dickerson	57.90	30	424.31	0.29	0.45	1
Drew-Dickerson	53.21	40	546.68	0.24	0.45	1
GC-rich	56.26	0	0.00	0.43	0.45	1
GC-rich	61.31	20	357.37	0.32	0.45	1
GC-rich	57.74	30	427.84	0.29	0.45	1
GC-rich	53.18	40	547.61	0.24	0.45	1

3.4 Results

Microsecond-scale molecular dynamics simulations of DNA dodecamers with and without protein crowders were carried to study the effect of crowding on DNA structure. We focused our analysis on helical properties including base geometries, groove widths and DNA bending, backbone torsions, interactions with crowder proteins, correlations between protein contacts and helical properties, and water and ion distributions around DNA.

3.4.1 Helical properties

Snapshots from the simulations were clustered to identify major conformations. Representative structures for each of the major clusters are depicted in Fig. 3.1. Generally, the helices stayed intact with only minor base fraying at the termini, which is common in simulations of short DNA fragments. The structures generally resemble B-DNA structures for both sequences with only minor changes in the presence of the crowders.

Helicoidal parameters for both, Drew-Dickerson and GC-rich dodecamers were averaged from the simulations. They are summarized in Tables 3.2 and 3.3, respectively. Helicoidal parameters for crystal structures of the respective dodecamers as well as canonical A- and B-forms of DNA, averaged over ten A- and B-form crystal structures each, are provided for comparison. Average properties for each of the clusters shown in Fig. 3.1 are given in Tables 3.4 and 3.5 (*Supplementary Information*). The more detailed analysis of the base geometries also indicates that both dodecamers remained close to B-DNA. The Drew-Dickerson dodecamer also remained reasonably close to the respective crystal structure (1BNA) but there are larger deviations between the simulation results and the crystal structure of the GC-rich dodecamer. The crystal structure for the GC-rich dodecamer sequence used here was reported to overall have A-form but show some B-form features for terminal base-pairs [149]. The reason we see deviations from the crystal structure

could be that we started the simulations from canonical B-form and since the GC-rich dodecamer may be stable in this form to some degree, we did not observe the full transition to A-form in the simulation length we used here. In the presence of the protein crowders, the helical parameters generally did not change much. Statistically significant changes are increased X-displacement and base inclination for both dodecamers and a slight increase in the helical twist angle and a slight reduction in slide of base-pairs along the base-pair axis for the GC-rich dodecamer. The increased x-displacement and base inclination point towards A-DNA but the values upon crowding still remained much closer to canonical B-DNA than A-DNA. The increased twist angles, on the other hand are consistent with a shift towards more canonical B-DNA structures.

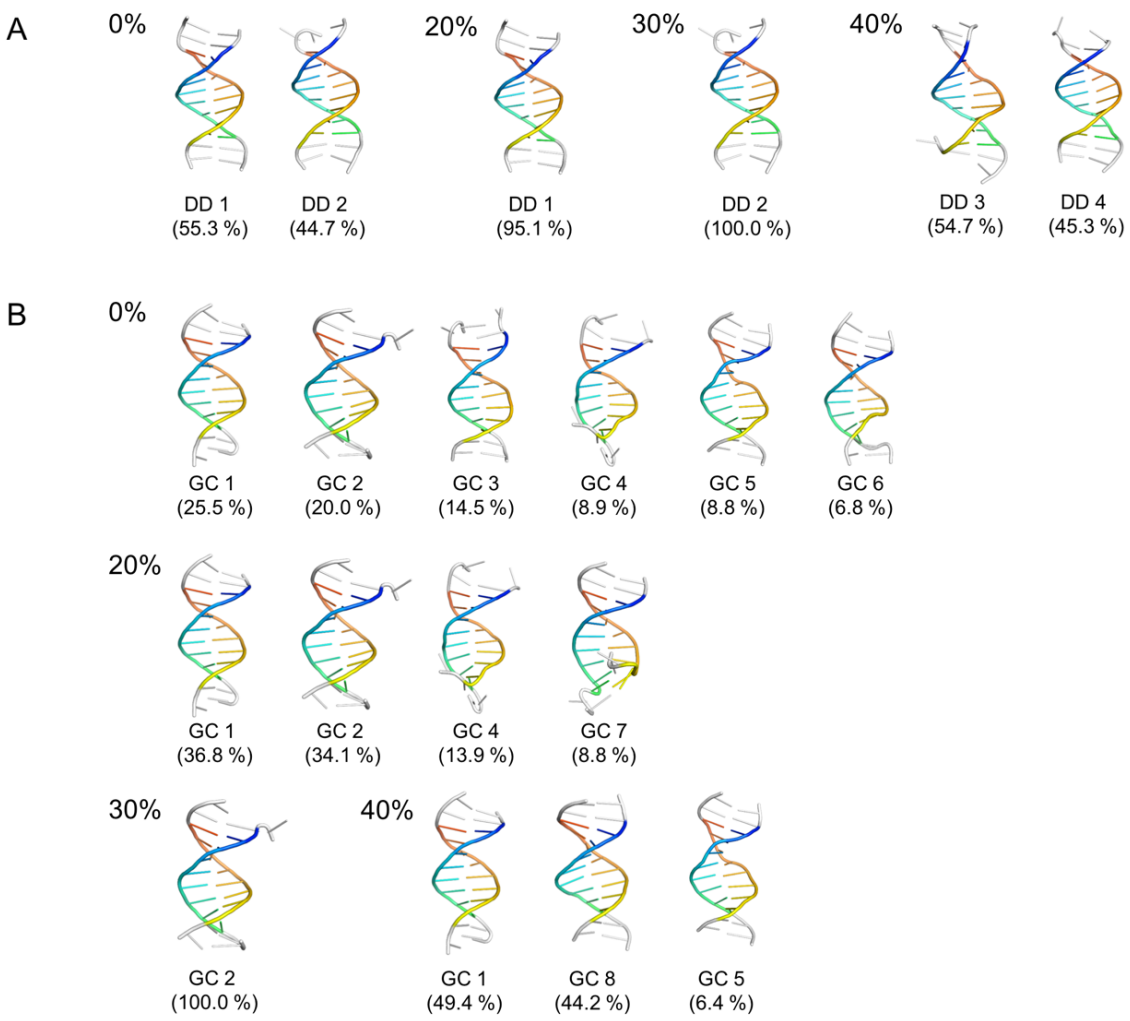

Figure 3.1 Representative conformations from clustering simulation snapshots for the Drew-Dickerson (A) and GC-rich (B) dodecamers with and without crowders. Cluster populations are given in parentheses.

Table 3.2 Average Helical Parameters for the Drew-Dickerson Dodecamer.

		Canonical		Simulations in crowded environment			
	X-ray (1BNA)	A-DNA	B-DNA	0 %	20 %	30 %	40 %
Slide (Å)	0.07 (*0.20*)	-1.62 (*0.06*)	0.16 (*0.08*)	0.26 (*0.03*)	0.25 (*0.03*)	0.21 (*0.02*)	0.25 (*0.02*)
Twist (deg)	34.22 (*2.13*)	30.34 (*0.58*)	34.70 (*0.70*)	33.42 (*0.13*)	34.06 (*0.14*)	33.96 (*0.16*)	33.43 (*0.14*)
X-displacement (Å)	-0.23 (*0.20*)	-4.50 (*0.18*)	-0.20 (*0.13*)	-0.59 (*0.06*)	-0.60 (*0.06*)	-0.72 (*0.06*)	-0.79 (*0.06*)
Helical rise (Å)	3.29 (*0.05*)	2.68 (*0.08*)	3.25 (*0.02*)	3.24 (*0.00*)	3.27 (*0.01*)	3.25 (*0.01*)	3.27 (*0.01*)
Inclination (deg)	4.02 (*2.73*)	17.78 (*1.56*)	4.34 (*0.77*)	10.87 (*0.13*)	10.62 (*0.19*)	11.15 (*0.34*)	12.59 (*0.35*)
z_p(Å)	-0.23 (*0.07*)	2.06 (*0.07*)	-0.33 (*0.04*)	-0.07 (*0.08*)	0.12 (*0.04*)	0.14 (*0.03*)	0.22 (*0.04*)
Minor groove (Å)	10.32 (*0.46*)	15.72 (*0.12*)	10.77 (*0.12*)	13.50 (*0.03*)	13.42 (*0.08*)	13.08 (*0.07*)	13.73 (*0.08*)
Major groove (Å)	17.34 (*0.33*)	12.94 (*0.39*)	17.14 (*0.12*)	16.47 (*0.07*)	16.54 (*0.06*)	16.39 (*0.08*)	16.31 (*0.06*)

Averages over all base-pairs excluding the first and last two terminal base-pairs with errors given in parentheses from block averaging. Canonical values were averaged over the A-form structures 3V9D, 3QK4, 2B1B, 1ZEX, 1ZEY, 1ZF1, 1ZF8, 1ZF9, 1ZFA and the B-form structures 2M2C, 4AGZ, 4H0, 4AH1, 3U05, 3U08, 1VTJ,3U2N, 3OIE, 3BSE.

Table 3.3 Average Helical Parameters for the GC-rich Dodecamer.

		Canonical		Simulations in crowded environment			
	X-ray (399D)	A-DNA	B-DNA	0 %	20 %	30 %	40 %
Slide (Å)	-1.71 (*0.16*)	-1.62 (*0.06*)	0.16 (*0.08*)	0.32 (*0.10*)	0.02 (*0.05*)	0.02 (*0.04*)	0.00 (*0.04*)
Twist (deg)	29.59 (*1.34*)	30.34 (*0.58*)	34.70 (*0.70*)	32.47 (*0.45*)	32.93 (*0.23*)	33.34 (*0.16*)	33.11 (*0.20*)
X-displacement (Å)	-5.01 (*0.41*)	-4.50 (*0.18*)	-0.20 (*0.13*)	-0.46 (*0.15*)	-1.03 (*0.10*)	-1.09 (*0.09*)	-1.10 (*0.10*)
Helical rise (Å)	2.66 (*0.22*)	2.68 (*0.08*)	3.25 (*0.02*)	3.26 (0.*03*)	3.23 (*0.02*)	3.26 (*0.01*)	3.27 (*0.02*)
Inclination (deg)	20.71 (*4.33*)	17.78 (*1.56*)	4.34 (*0.77*)	10.01 (*0.37*)	10.70 (*0.34*)	11.34 (*0.28*)	11.34 (*0.50*)
z_p(Å)	1.56 (*0.35*)	2.06 (*0.07*)	-0.33 (*0.04*)	-0.21 (*0.02*)	-0.20 (*0.02*)	-0.23 (*0.02*)	-0.18 (*0.02*)
Minor groove (Å)	16.22 (*0.47*)	15.72 (*0.12*)	10.77 (*0.12*)	14.91 (*0.13*)	14.54 (*0.06*)	14.52 (*0.06*)	14.76 (*0.10*)
Major groove (Å)	13.14 (*2.63*)	12.94 (*0.39*)	17.14 (*0.12*)	16.26 (*0.14*)	16.53 (*0.09*)	16.22 (*0.06*)	16.27 (*0.09*)

see Table 2.

We further analyzed the displacement of phosphorus atoms relative to the horizontal plane passing between base-pairs in a base-pair step (z_p) and major/minor grooves (Tables 3.2 & 3.3). The z_p parameter is very different between the two forms of DNA. While B-DNA has values around -0.3 Å, the parameter is mostly larger than 2.0 Å for A-DNA. This parameter does not show a trend upon crowding for the GC-rich dodecamer, while the Drew-Dickerson dodecamer had slightly larger values in crowded environments, again indicating a slight tendency towards A-DNA geometries but still remained much closer to canonical B-DNA values. Minor and major groove widths also did not change significantly upon crowding but we note that minor groove widths were generally overestimated compared to canonical B-DNA values. This is a general feature of the CHARMM force field that was used here [182]. Finally, we analyzed the helical bending angles (see Table 3.6, *SI*) which also did not show a significant change upon crowding.

3.4.2 Sugar conformations and backbone torsions

A key feature of nucleic acid backbone is the ribose pucker conformation. A-form DNA is known to prefer C3'-endo and C2'-exo conformations whereas B-form DNA is characterized by C3'-exo and C2'-endo conformations. The simulation results shown in Fig. 3.2 show that the sugars of both dodecamers generally remain in C3'-exo and C2'-endo conformations. As expected, C3'-endo and C2'-exo sugar conformations are more prominent for the GC-rich dodecamer (see Fig. 3.2B). Again, there is no major change upon crowding, but in the GC-rich dodecamer, sugars slightly shift to C3'-exo and C2'-endo sugar conformations up to 30% crowding, but then revert back to more A-form conformations at 40 % crowder concentrations.

We further analyzed torsion angles along the phosphate backbone. χ and δ angles are the most distinctive backbone angles to distinguish between A- and B- form DNA. We constructed potentials of mean force (PMF) as a function of δ/χ from the simulations (Fig.

3.3). The separation between A- and B-DNA torsion angles is readily apparent. Consistent with the ribose puckers and helical geometries, there is more sampling of B-DNA torsion angles for both dodecamers. While there is little change in the sampling of the major A- and B-form, the presence of crowders appears to affect the sampling of minor conformations with A-like δ values around 80 degrees and B-like χ values around -100 degrees. Sampling in this region is significantly reduced in both dodecamers upon crowding (see Fig. 3.3). This region corresponds to a conformation where bases stay in the same orientation relative to the sugar as in B-form but they are slightly more exposed to the environment, and apparently, this conformation is largely prevented by crowder proteins. The sampling of ε and ζ torsion angles distinguishes between BI/BII forms. A similar trend is observed where crowding reduces the sampling of minor states outside the major BI/BII basins (see Fig. 3.12, *SI*). Based on this analysis, it appears that one effect of protein crowders may be to focus the sampling of DNA conformations on the major conformations.

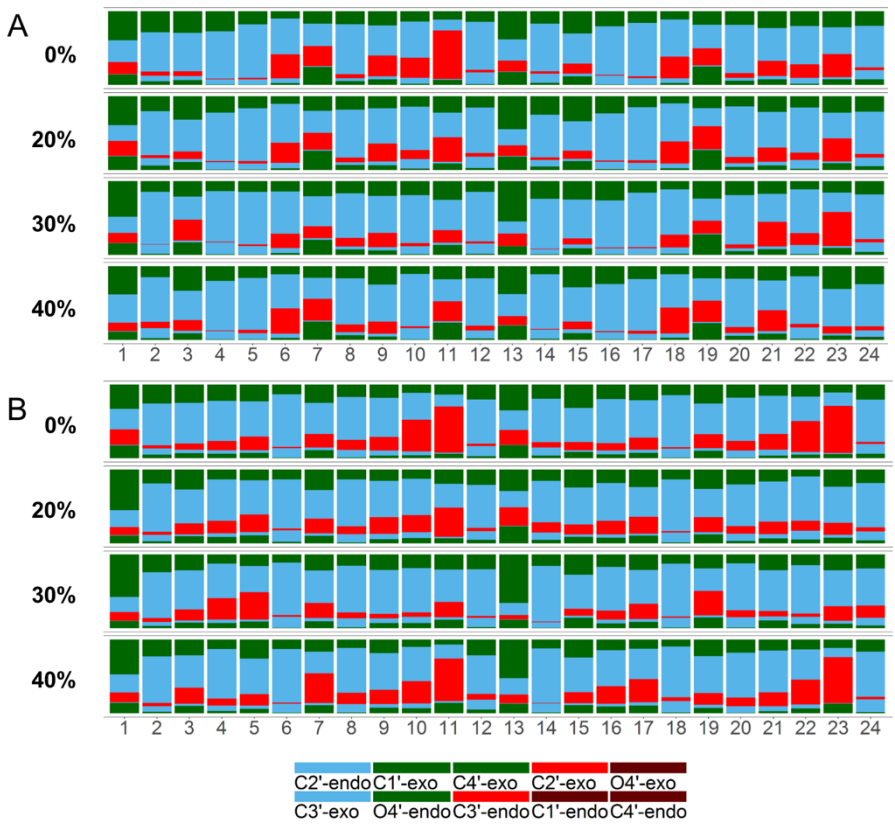

Figure 3.2 Sugar pucker conformations for each base of the Drew-Dickerson dodecamer (A) and the GC-rich dodecamer (B) from simulations at different protein concentrations.

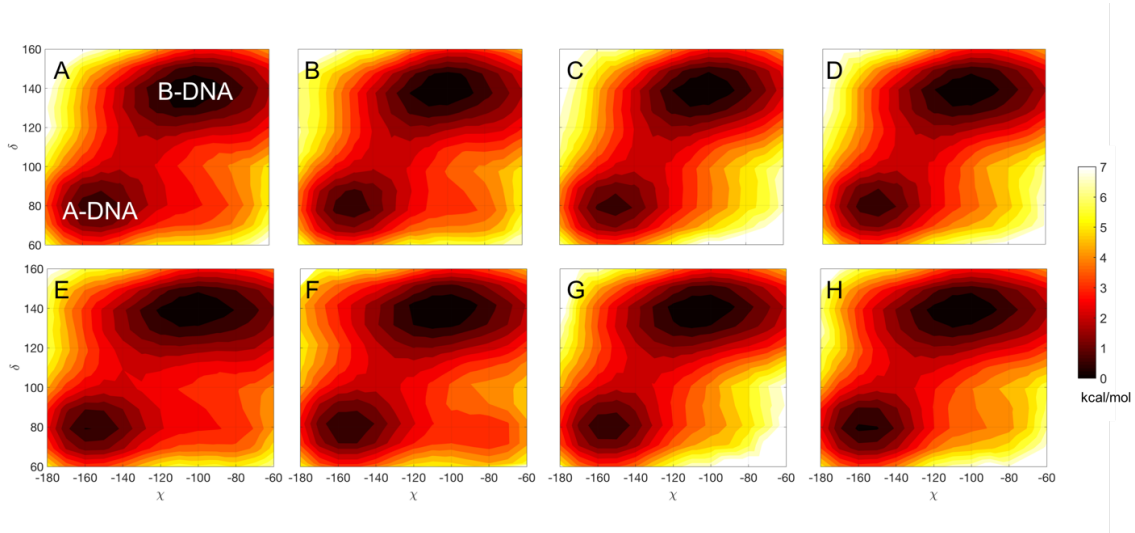

Figure 3.3 PMF (kcal/mol) as a function of δ and χ backbone angles for the Drew-Dickerson dodecamer at 0 % (A), 20 % (B), 30 % (C) and 40 % (D) protein concentrations, and for the GC-rich dodecamer at 0 % (E), 20 % (F), 30 % (G) and 40 % (H) protein concentrations.

3.4.3 DNA-protein interactions

Protein G is not known to interact specifically with DNA but under highly crowded conditions, interactions are unavoidable. Fig. 3.4 shows where contacts between protein G and DNA occur based on minimum distances between the major/minor grooves and sugar/phosphate groups of the DNA with different residues of protein G. More detailed contact analysis between individual base-pairs and protein G residues is shown in Figures 3.13 and 3.14 (*SI*) for the Drew-Dickerson and GC-rich dodecamers, respectively. Most of the contacts are between the DNA sugar-phosphate backbone and protein residues 15-30, mostly in the α-helix of protein G, as well as residues at the N-terminus and near the C-terminus. Contacts involving the DNA grooves, a typical mode of interaction for DNA-binding proteins were not common with protein G. The interactions partially involve electrostatic attraction between the DNA phosphate and certain lysine residues (K4, K28, K31, and K50), but sugar oxygens O3' and O4' as well as phosphate oxygens also form hydrogen bonds with other polar protein residues. Representative snapshots of protein G-DNA interactions are shown in Fig. 3.5. As would be expected, the contacts between the proteins and DNA increase with crowder concentration and crowding seems to increase sugar-phosphate-protein contacts more for the Drew-Dickerson dodecamer than for the GC-rich dodecamer.

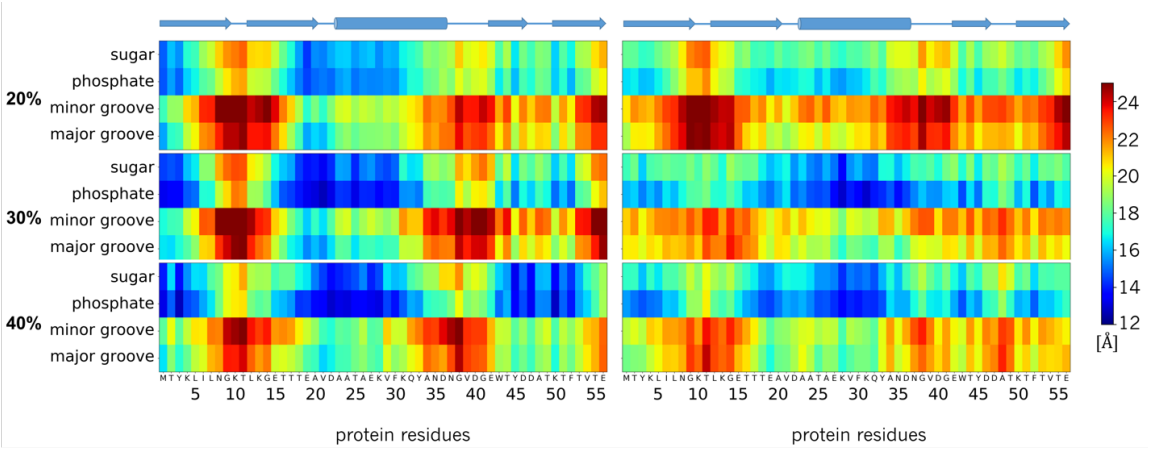

Figure 3.4 Average minimum heavy atom distances between crowder protein residues and DNA major groove, minor groove, sugar and phosphate backbone for the Drew-Dickerson dodecamer (left) and the GC-rich dodecamer (right) at different protein concentrations. The secondary structure of protein G is indicated on top for reference.

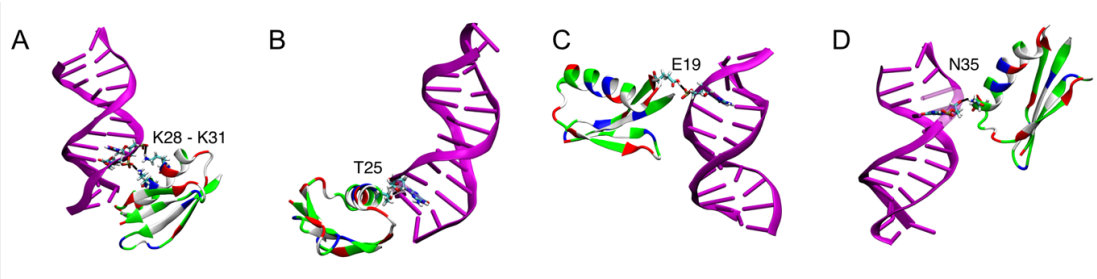

Figure 3.5 Representative structures showing interactions of crowder proteins with the Drew-Dickerson phosphate group (A) and sugar (B), and the GC-rich phosphate group (C) and sugar (D). Interacting residues are shown in licorice representation.

3.4.4 Correlations between DNA-protein contacts and DNA helix properties

To investigate in more detail whether the close contacts of the crowder proteins with the DNA have the potential to perturb DNA structure, we analyzed correlations between DNA – protein contacts and helicoidal properties of DNA as well as backbone torsion and pseudorotation phase angles. First, we examined the effect of close contacts on the helicoidal parameters listed in Tables 3.2 and 3.3. We found that a higher number of close protein contacts corresponded to a narrower range of sampled values for all of the helicoidal parameters (Figures 3.6 – 3.9, Figures 3.15 – 3.18, *SI*). Among these parameters, slide (Fig. 3.6), x-displacement (Fig. 3.7), helical rise (Fig. 3.8) and z_p (Fig. 3.9) values showed a clear shift towards B-form values with increasing number of contacts. These parameters focus on the displacement of bases along the

x- (x-displacement) and y- (slide) axes and of phosphates along the base-pair axis (z_p). All of the values approach zero with crowding. This suggests that DNA bases and phosphates undergo less displacement as a result of crowding. On the other hand, rotations of base-pairs about helical (twist) or base-pair axes (inclination) do not show a distinct shift towards any canonical values (Fig. 3.15, 3.16, *SI*). Major and minor groove widths do not seem to be affected by contacts except for the GC-rich dodecamer, where there appears to be a clear tendency towards larger minor groove values, i.e. values more similar to A-DNA (Fig. 3.17, 3.18, *SI*).

Similar to the helicoidal parameters, backbone torsion angles also fluctuate in a narrower range upon crowding (Fig. 3.18 – 3.25, *SI*). This suggests that protein – DNA interactions limit the conformational fluctuations of DNA backbone. Particularly, δ and χ angles shift towards B-form values upon higher number of protein contacts, explaining a decrease in the sampling of non-canonical conformations shown in Fig. 3.3. Finally, pseudorotation angles move to B-form values with protein contacts which lead to C3′-exo and C2′-endo sugar pucker conformations (Fig. 3.26, *SI*).

The results discussed here are most pronounced for the Drew-Dickerson dodecamer. In the GC-rich dodecamer, the fluctuations of helicoidal parameters and backbone angles are reduced less and the tendency to sample A-form values further complicates the picture. Overall, our results suggest that the interactions of protein crowders with DNA sugar/phosphate backbone shown in the previous section result in a stiffer DNA backbone, which also prevents larger base/base-pair displacements, and therefore restricts the conformational space of DNA. Although it appears that there is not a specific tendency towards one of the major forms of DNA upon crowding, there is a distinct effect of protein crowders on DNA structure by narrowing the conformational sampling to canonical structures.

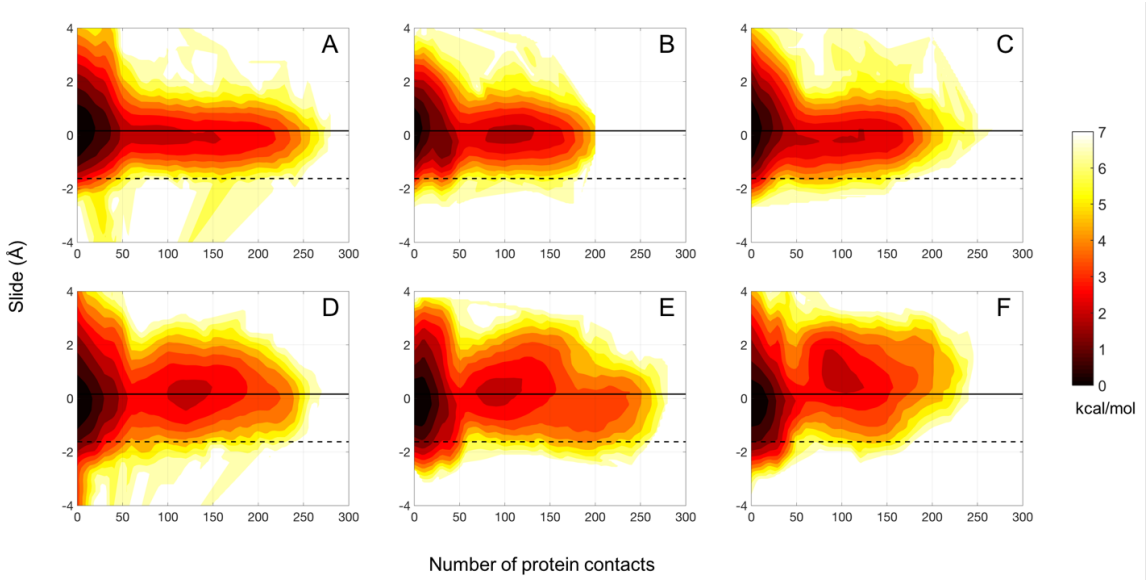

Figure 3.6 PMF (kcal/mol) as a function of slide and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the slide and x-displacement values for canonical B- and A-forms, respectively.

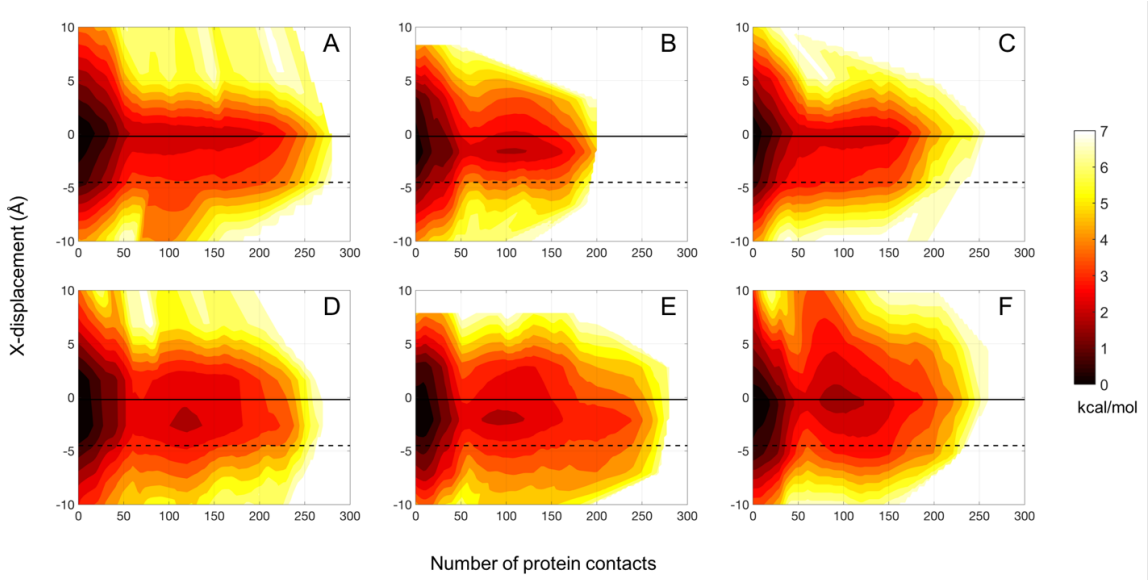

Figure 3.7 PMF (kcal/mol) as a function of x-displacement and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations.

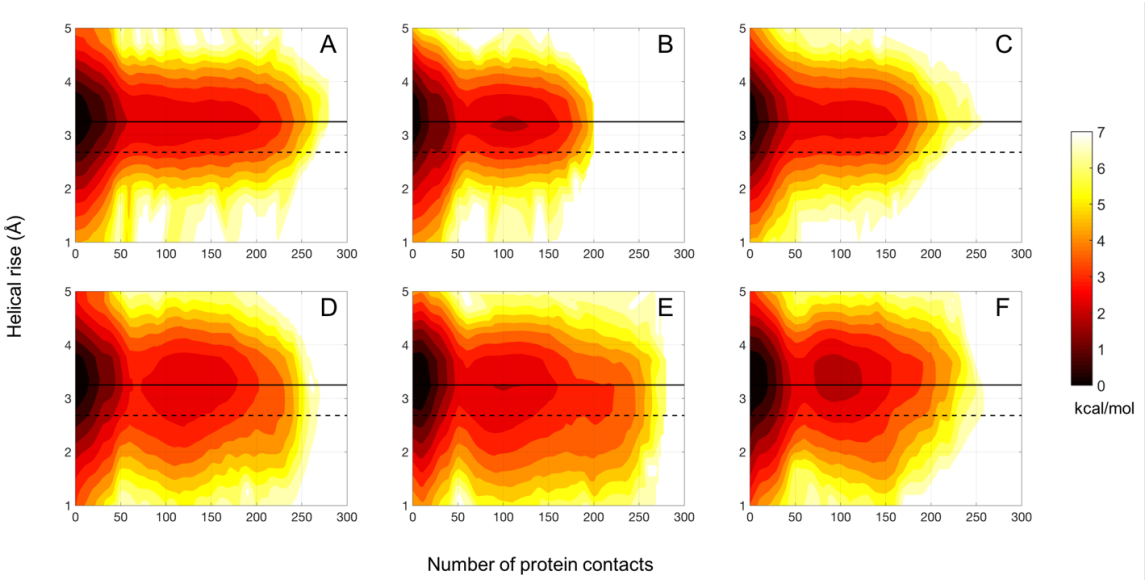

Figure 3.8 PMF (kcal/mol) as a function of helical rise and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations.

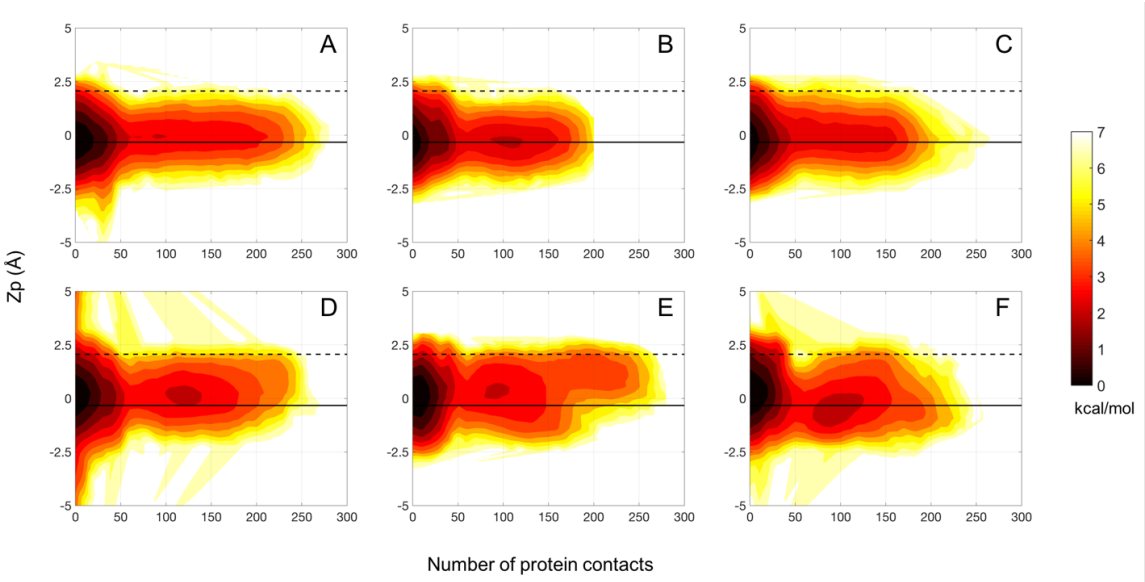

Figure 3.9 PMF (kcal/mol) as a function of Zp and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations.

3.4.5 Hydration and ion distributions around DNA

Water and ions are integral parts of DNA structures. We analyzed hydration patterns and sodium ion distributions around DNA as a function of crowding. Conditional water radial distribution functions (RDF) were obtained for water oxygen distances to the closest heavy atoms in DNA, normalized by the corresponding accessible volume at each distance and the bulk water density (0.034 $Å^{-3}$) (see Figure 3.10A). The analysis shows that the first hydration shell is almost unaffected by the level of crowding, but the RDF decreases beyond the hydration shell significantly as a function of crowding. This observation is similar to what has been reported previously for the hydration around proteins under crowded environments[42].

Sodium RDFs were calculated in the same way as the water RDFs but normalized by the ion density of the system (0.002 $Å^{-3}$). There are two peaks in the sodium RDFs corresponding to ions in direct contact with the DNA (around 2.5 Å and largely in the minor groove) and ions interacting with the DNA through water (around 4.5 Å)[80, 183, 184]. While the direct contact peak is not affected significantly by crowding, the second peak shows a greater dependence on crowding. At the highest crowder fractions the second peak is significantly reduced in both dodecamers (see Fig. 3.10B) and the ion density is reduced further at larger distances similar to the reduction in hydration upon crowding. The effect of crowding on the ion distributions also impacts the DNA neutralization as a function of distance (Figure 3.10C). 76 % of the DNA phosphate groups are neutralized as suggested by counterion condensation theory at around 9 Å for the dilute system, however, it takes up to 11 – 12 Å to reach 76 % DNA neutralization under crowding conditions. It is interesting, that despite the impact of crowding on the second peak of the ion distribution, the counterion condensation is affected less for distances less than 6 Å. As this may seem counterintuitive, the reader is reminded that the RDF is normalized by the available volume and the overall ion density, at constant ion molality, whereas Fig. 3.10C simply describes the net neutralization of the DNA by the ions. The extended distance to reach

76% charge neutralization upon crowding may seem to challenge counterion condensation theory. However, the protein crowders, despite being net neutral, can provide additional charge neutralization by orienting basic lysines near the DNA surface as described above to compensate for the reduced neutralization by the sodium ions.

Finally, the 3D distributions of sodium ions around the Drew-Dickerson and GC-rich dodecamers are compared in Fig. 3.11. The sodium ion networks in the major and minor grooves of DNA are largely preserved for both dodecamers with little changes upon crowding. However, additional densities become apparent further away from the DNA at different locations upon crowding. This additional ordering is a result of crowder proteins interacting with the DNA and coordinating ions near the DNA. A snapshot showing a crowder protein interacting with the DNA and orienting a sodium ion at the same time is shown in Figure 3.27 (*SI)*.

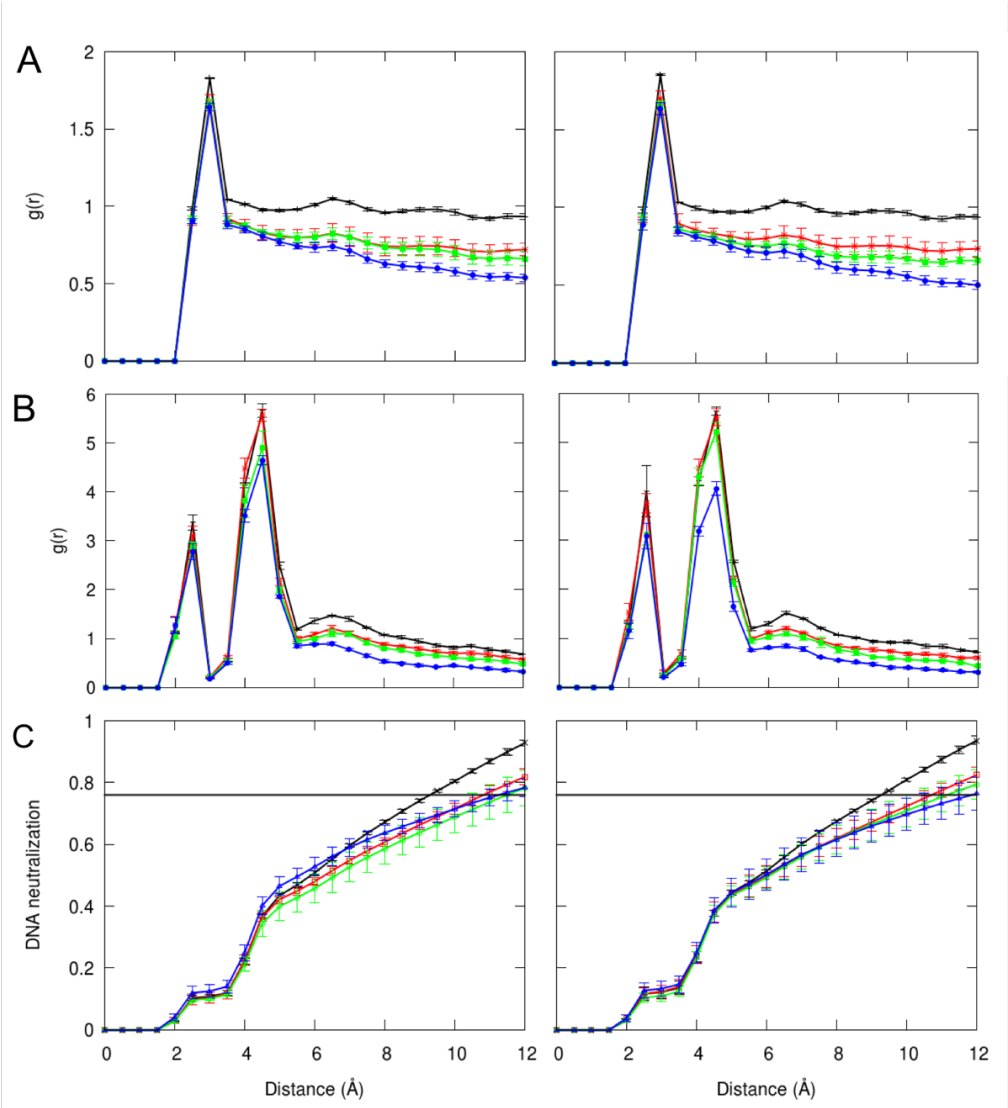

Figure 3.10 Radial distribution functions for water (A), sodium ions (B) and DNA neutralization fractions (C) as a function of distance from the closest heavy atoms of the Drew-Dickerson dodecamer (left) and the GC-rich dodecamer (right). Line colors indicate different protein concentrations (black: 0 %, red: 20 %, green: 30 %, blue: 40 %). The horizontal black line in C indicates the counterion condensation value of 76% of the ions to be condensed on the surface of the DNA. Error bars indicate the calculated standard errors from five independent replicate simulations.

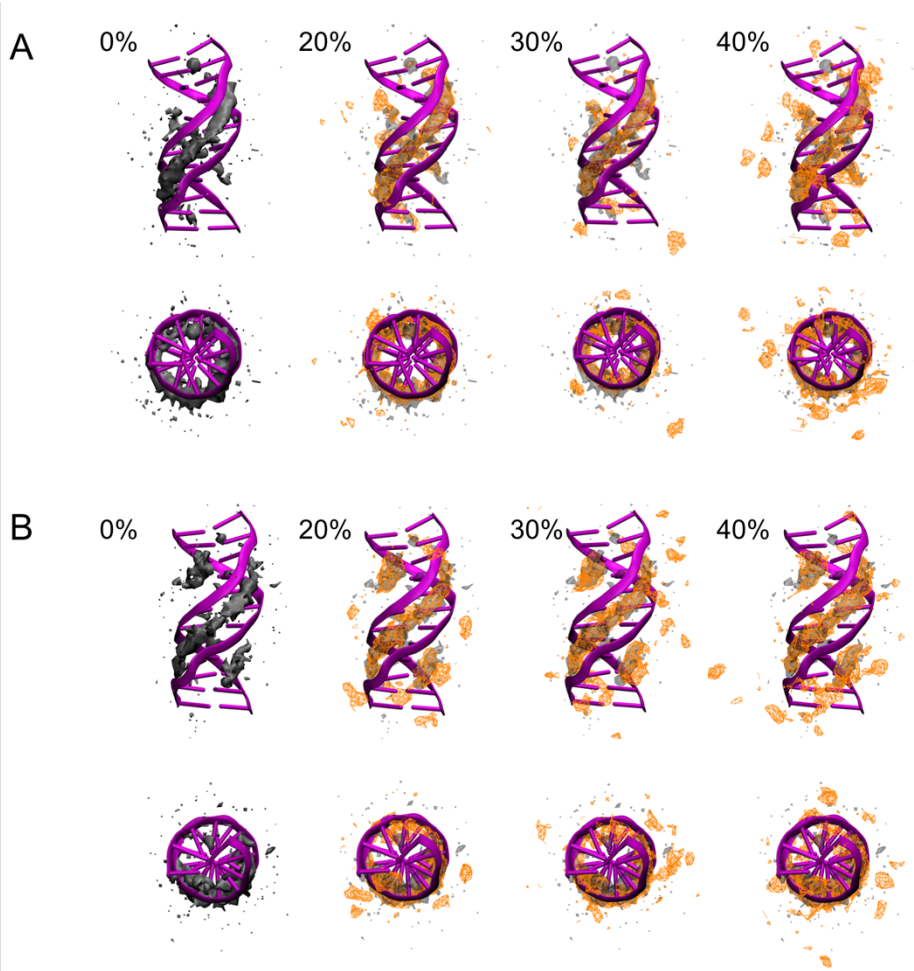

Figure 3.11 3D sodium ion densities around the Drew-Dickerson dodecamer (A) and the GC-rich dodecamer (B) at different protein concentrations. The ion density observed in 0 % crowding is shown with a transparent representation for comparison with ion densities (orange) in crowded solutions. Density contours are shown at a level of 0.002 $Å^{-3}$. The top and bottom figures represent front and top views of DNA.

3.5 Discussion and Conclusions

In this study, we investigated the effect of protein crowding on the conformational preferences of DNA duplexes. In a previous study, we examined one aspect of cellular crowding, namely a reduced dielectric response of the environment due to the less available water and its slowed dynamics. We found an overall shift towards A-like conformations for DNA under as a result of a reduced dielectric response of its environment [167]. Here, we included protein crowders and solvent explicitly to test whether the same conclusions would be found. Although some of base parameters moved slightly towards A-like values upon crowding, B-DNA was largely

maintained in contrast to our previous findings. This suggests that a reduced dielectric response of crowded environments and interactions with crowder proteins have different effects on DNA conformational preferences with a net effect of not altering canonical B-DNA structures much. We found that the crowder proteins mostly interact with DNA via its phosphate-sugar backbone as previously observed in non-specific binding of proteins to DNA [185]. These interactions arise from the electrostatic interactions between negatively charged phosphate oxygens and positively charged amino acid residues as well as the polar interactions between phosphate and/or sugar oxygens and side chains of polar amino acid residues. Previous studies have shown that DNA can undergo structural deformations from its B-form towards A-type helix as a result of forming complexes with specific DNA binding proteins [186-190], but we did not see such an effect here. It does appear, however, that for the system studied here, the presence of the protein crowders limits the conformational space of DNA to more canonical structures, mostly in B-form, both for the backbone torsions and the helical parameters. However, the narrowed conformational sampling appears to have little effect on the overall structural averages. Such a crowding effect on DNA structure may be understood in similar ways as protein native state stabilization due to the volume exclusion effect [35, 164, 191, 192] where the reduced space due to crowders limits the ability to widely sample conformational space. This would mean that protein crowding *in vivo* helps stabilize the biologically most relevant forms of DNA.

We also studied hydration patterns and ion densities around DNA in protein crowding. The first hydration shell around DNA is largely unaffected by crowding, while the water densities beyond the first solvation shell significantly reduced compared to the bulk water density under crowding effect. This result is very similar to the hydration shell around proteins upon crowding [42]. This further confirms that, protein crowding in cells generally does not alter the first hydration shell around biomolecules. However, sodium densities around DNA are affected

already when interacting with DNA through water. Only the direct-contact first peak in the sodium-DNA RDF appears to be unaffected by crowding. Moreover, the charge neutralization by ions is altered upon crowding with the classical counter-ion condensation threshold reached at larger distances from the DNA than under dilute conditions. This suggests that proteins have to play an increasing role in neutralizing DNA under highly crowded conditions.

In conclusion, the results obtained here shed light on the effect of protein crowding on DNA structure. We found that the crowder proteins mostly assist DNA to stay in canonical B-like conformations, limiting excursions to non-canonical conformations rather than a clear shift in the overall, average structure as suggested by a simple dielectric model of cellular environments. We hope that this hypothesis will motivate new experimental efforts to characterize DNA structure under crowded conditions. We expect that reduced conformational dynamics upon crowding may be observable via NMR spectroscopy. Another testable hypothesis is the altered ion distribution predicted by our simulations, which could be amenable to the ion-counting experiments recently carried out by the Herschlag group [193-196].

3.6. Acknowledgements

Funding from NSF (MCB 1330560) and NIH (R01 GM092949) is acknowledged.

3.7 Supplementary Information

Table 3.4 Parameters and RMSD values from the canonical B-form structure for the individual clusters of the Drew-Dickerson dodecamer.

	DD1	DD2	DD3	DD4
Slide (Å)	0.07 (*0.00*)	0.34 (*0.00*)	0.19 (*0.00*)	0.38 (*0.00*)
Twist (deg)	33.67 (*0.01*)	34.63 (*0.01*)	33.12 (*0.01*)	32.73 (*0.02*)
X-displacement (deg)	-1.09 (*0.00*)	-0.30 (*0.01*)	-0.94 (*0.01*)	-0.42 (*0.01*)
Helical rise (Å)	3.21 (*0.00*)	3.30 (*0.00*)	3.24 (*0.00*)	3.29 (*0.00*)
Inclination (deg)	12.89 (*0.03*)	8.95 (*0.03*)	13.14 (*0.04*)	11.07 (*0.05*)
z_p(Å)	-0.06 (*0.00*)	-0.32 (*0.00*)	-0.14 (*0.00*)	-0.29 (*0.00*)
Minor groove (Å)	13.63 (*0.01*)	12.77 (*0.01*)	13.86 (*0.01*)	13.69 (*0.01*)
Major groove (Å)	16.10 (*0.01*)	16.52 (*0.01*)	16.27 (*0.01*)	16.97 (*0.01*)
RMSD (Å)	1.82 (*0.00*)	1.37 (*0.00*)	2.00 (*0.00*)	1.70 (*0.00*)

All values are averaged over all base-pairs excluding the first and last two terminal base-pairs with standard errors given in the parentheses.

Table 3.5 Parameters and RMSD values from the canonical B-form structure for the individual clusters of the GC-rich dodecamer.

	GC1	GC2	GC3	GC4
Slide (Å)	0.12 (*0.00*)	-0.30 (*0.00*)	0.69 (*0.01*)	0.44 (*0.01*)
Twist (deg)	34.36 (*0.01*)	33.13 (*0.01*)	34.63 (*0.03*)	33.02 (*0.03*)
X-displacement (deg)	-0.67 (*0.01*)	-1.69 (*0.01*)	0.53 (*0.01*)	-0.20 (*0.01*)
Helical rise (Å)	3.30 (*0.00*)	3.20 (*0.00*)	3.37 (*0.00*)	3.33 (*0.00*)
Inclination (deg)	8.80 (*0.03*)	12.12 (*0.03*)	5.90 (*0.08*)	9.42 (*0.07*)
z_p(Å)	0.05 (*0.00*)	0.34 (*0.00*)	-0.21 (*0.01*)	-0.08 (*0.01*)
Minor groove (Å)	14.05 (*0.01*)	14.55 (*0.01*)	14.00 (*0.02*)	14.78 (*0.01*)
Major groove (Å)	16.34 (*0.01*)	16.13 (*0.01*)	16.47 (*0.01*)	16.82 (*0.02*)
RMSD (Å)	1.62 (*0.00*)	2.28 (*0.00*)	1.68 (*0.01*)	2.05 (*0.01*)
	GC5	GC6	GC7	GC8
Slide (Å)	0.44 (*0.01*)	-0.06 (*0.01*)	-0.11 (*0.02*)	0.04 (*0.00*)
Twist (deg)	30.98 (*0.06*)	31.97 (*0.03*)	30.90 (*0.41*)	32.72 (*0.02*)
X-displacement (deg)	-0.55 (*0.02*)	-1.41 (*0.01*)	-0.94 (*0.03*)	-1.19 (*0.01*)
Helical rise (Å)	3.36 (*0.01*)	3.24 (*0.00*)	3.07 (*0.02*)	3.25 (*0.00*)
Inclination (deg)	12.55 (*0.14*)	13.05 (*0.06*)	12.00 (*0.30*)	12.99 (*0.04*)
z_p(Å)	0.03 (*0.01*)	0.19 (*0.01*)	-0.20 (*0.03*)	0.17 (*0.01*)
Minor groove (Å)	15.68 (*0.02*)	15.11 (*0.01*)	14.94 (*0.03*)	15.09 (*0.01*)
Major groove (Å)	16.82 (*0.02*)	16.69 (*0.02*)	15.47 (*0.05*)	16.25 (*0.01*)
RMSD (Å)	2.58 (*0.01*)	2.47 (*0.01*)	3.70 (*0.02*)	2.25 (*0.00*)

All values are averaged over all base-pairs excluding the first and last two terminal base-pairs with standard errors given in the parentheses.

Table 3.6 Bending angles for both DNA dodecamers (degrees).

		Canonical		Simulations in crowded environment			
	X-ray	A-DNA	B-DNA	0%	20%	30%	40%
Drew-Dickerson	170.11	122.72 (*3.57*)	163.05 (*5.94*)	155.24 (*0.47*)	155.55 (*0.65*)	155.50 (*0.76*)	152.80 (*0.87*)
GC-rich	100.77	122.72 (*3.57*)	163.05 (*5.94*)	156.85 (*1.09*)	153.00 (*1.12*)	153.03 (*0.91*)	152.18 (*1.04*)

Bending angle is defined as the angle between the center of masses of three sections of basepairs: Section 1: 3 – 5, Section 2: 6 – 7, and Section 3: 8 – 10. Standard errors given in the parentheses. Statistical errors of the averages over the simulations are estimated from block averaging by comparing results for 100 ns segments from the simulations. Canonical values are averaged over the A-form structures 3V9D, 3QK4, 2B1B, 1ZEX, 1ZEY, 1ZF1, 1ZF8, 1ZF9, 1ZFA and the B-form structures 2M2C, 4AGZ, 4H0, 4AH1, 3U05, 3U08, 1VTJ,3U2N, 3OIE, 3BSE. For X-ray structure values; 1BNA and 399D are used for Drew-Dickerson and GC-rich dodecamers, respectively.

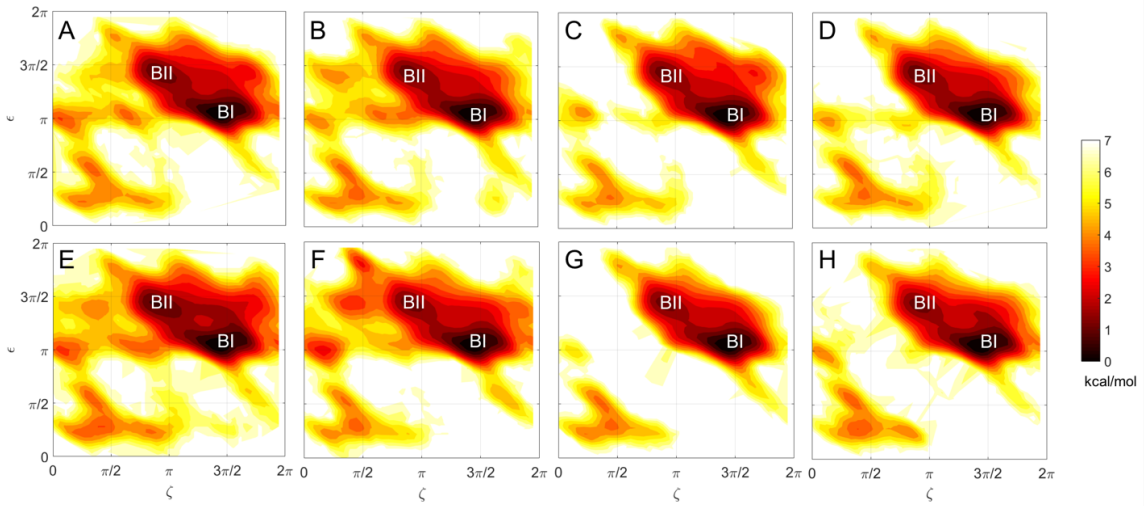

Figure 3.12 Potential of mean force (kcal/mol) as a function of ε and ξ backbone angles for the Drew-Dickerson dodecamer at 0 % (A), 20 % (B), 30 % (C) and 40 % (D) protein concentrations, and for the GC-rich dodecamer at 0 % (E), 20 % (F), 30 % (G) and 40 % (H) protein concentrations.

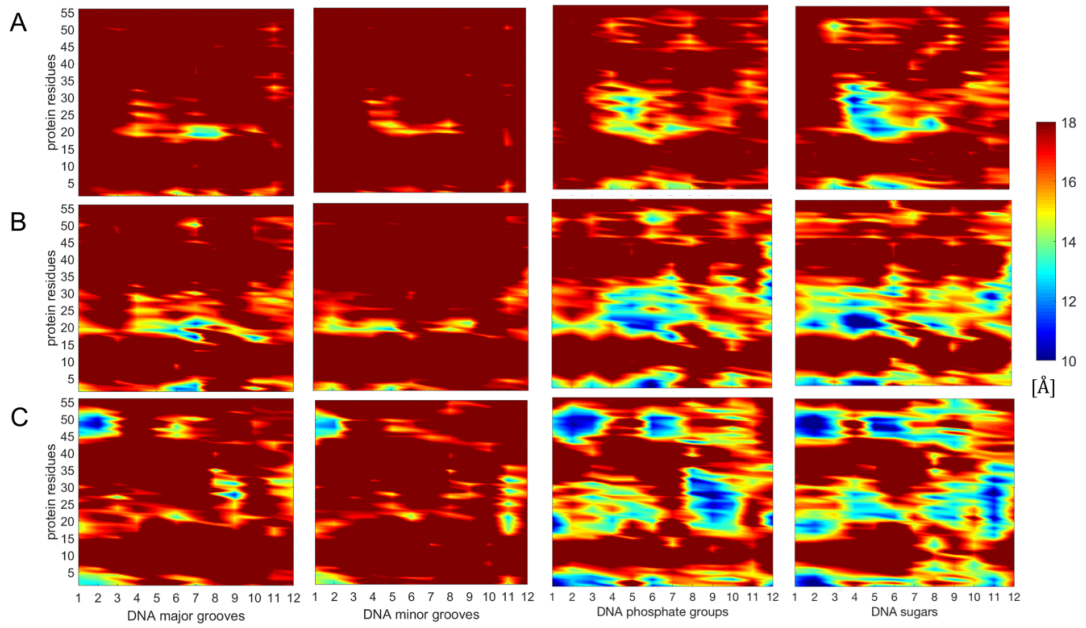

Figure 3.13 Average minimum distances between the crowder protein residues and the major groove, minor groove, sugar and phosphate backbone for the individual base-pairs of Drew-Dickerson dodecamer at 20% (A), 30% (B) and 40% (C) protein concentrations. Distances were calculated from the heavy atoms only.

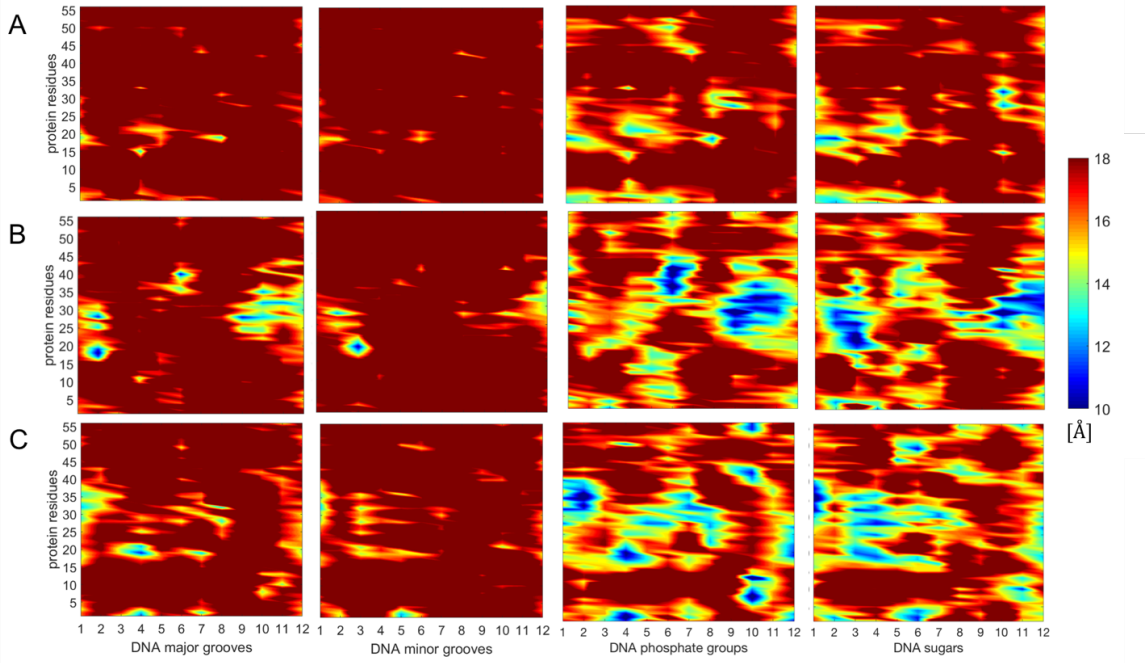

Figure 3.14 Average minimum distances between the crowder protein residues and the major grooves, minor grooves, sugar and phosphate backbone for the individual base-pairs of GC-rich dodecamer at 20% (A), 30% (B) and 40% (C) protein concentrations. Distances were calculated from the heavy atoms only.

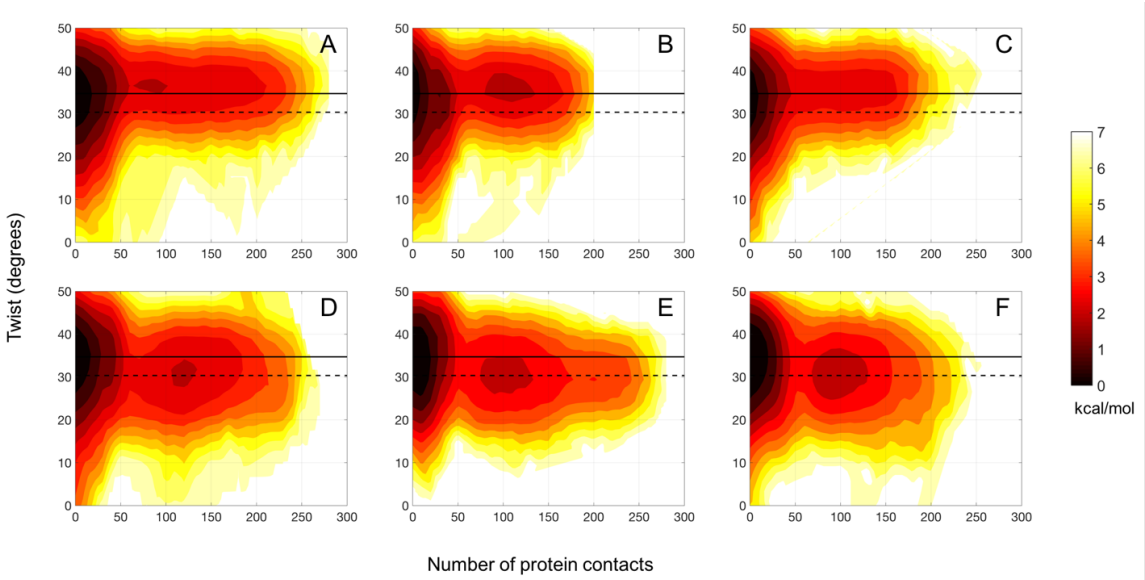

Figure 3.15 Potential of mean force (kcal/mol) as a function of twist angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

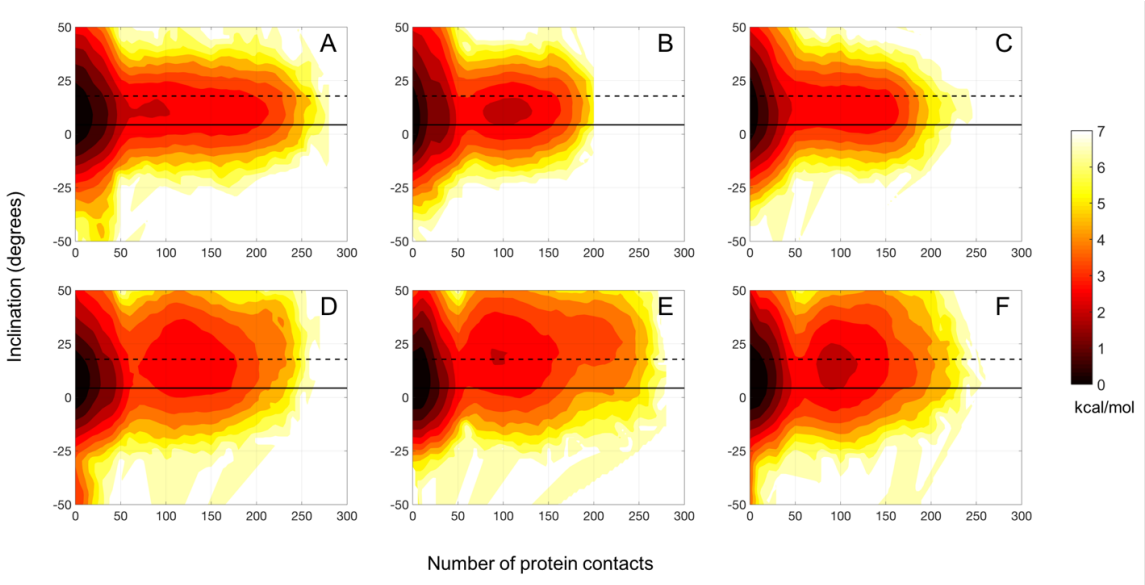

Figure 3.16 Potential of mean force (kcal/mol) as a function of inclination angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

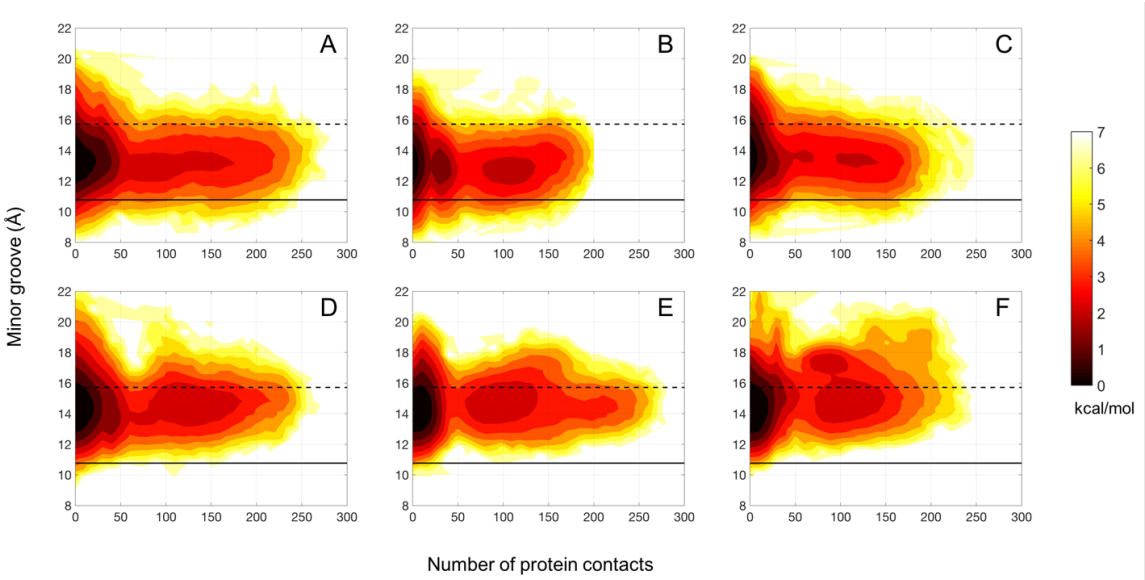

Figure 3.17 Potential of mean force (kcal/mol) as a function of minor groove and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

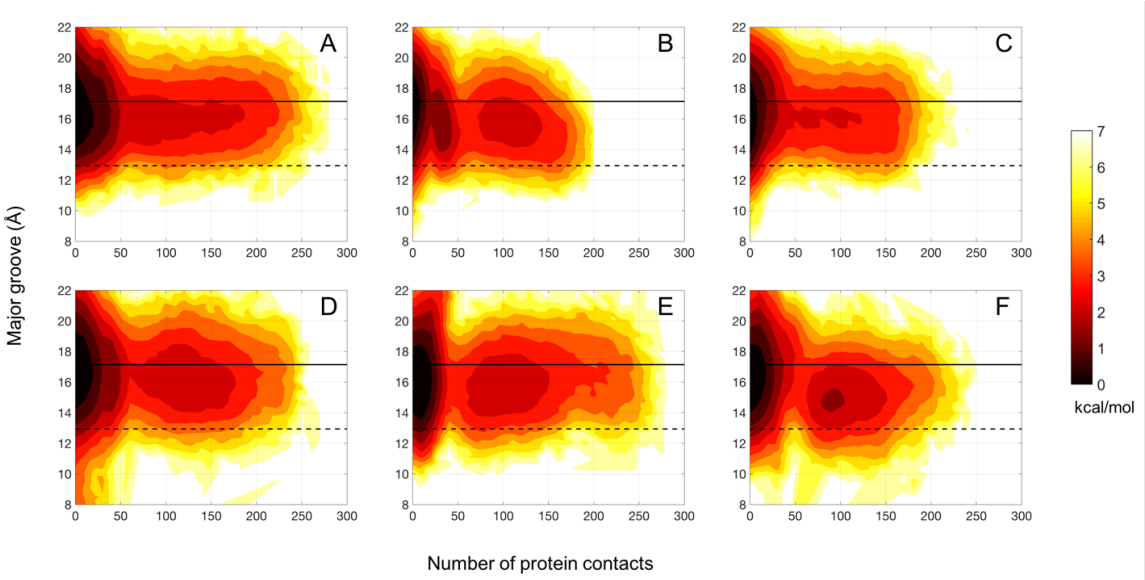

Figure 3.18 Potential of mean force (kcal/mol) as a function of major groove and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

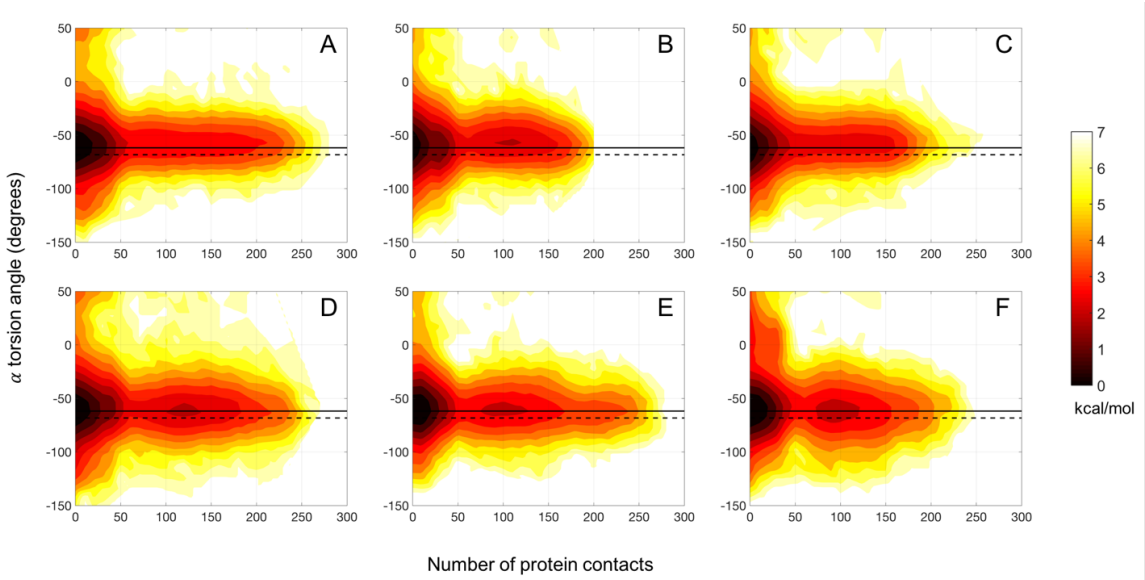

Figure 3.19 Potential of mean force (kcal/mol) as a function of α backbone torsion angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

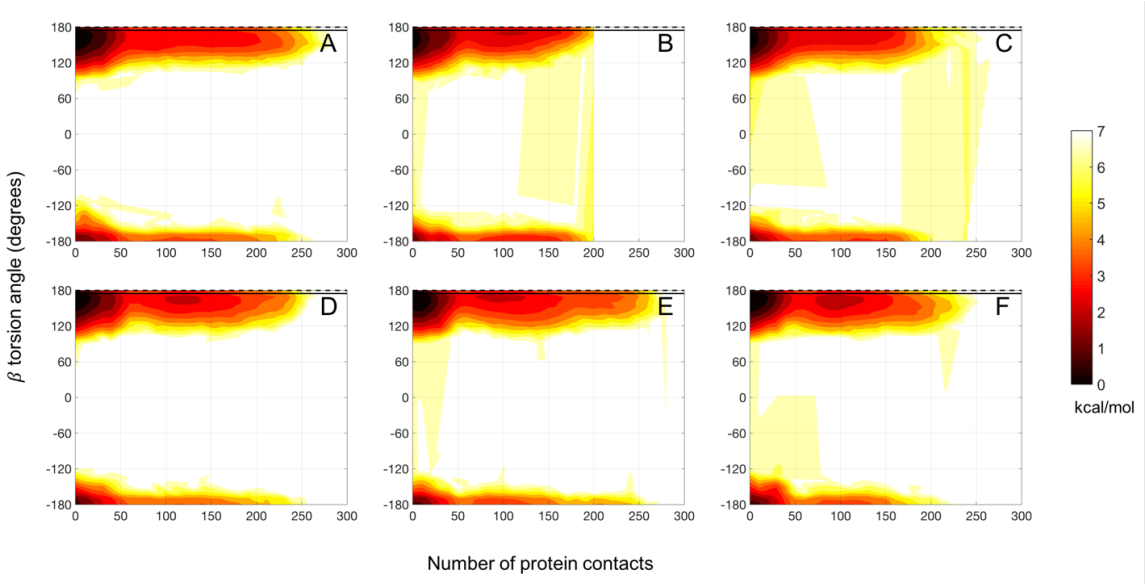

Figure 3.20 Potential of mean force (kcal/mol) as a function of β backbone torsion angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

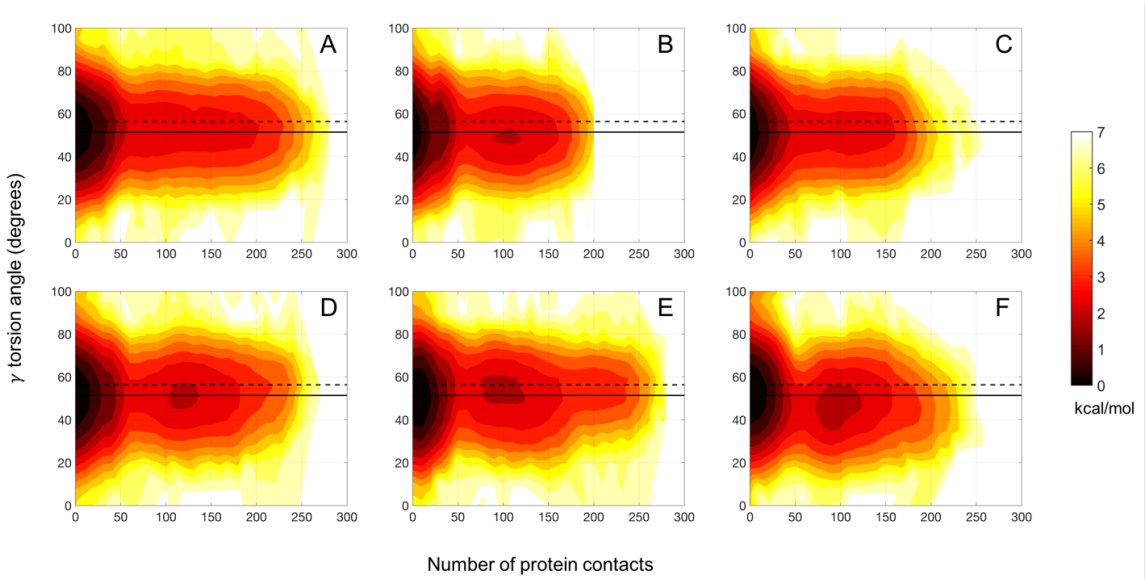

Figure 3.21 Potential of mean force (kcal/mol) as a function of γ backbone torsion angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

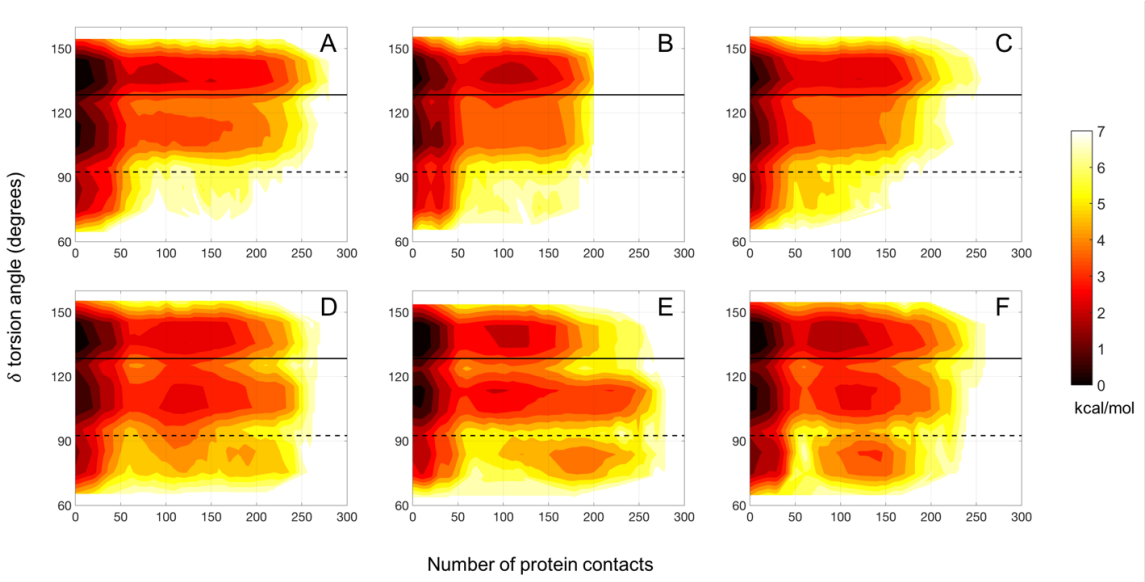

Figure 3.22 Potential of mean force (kcal/mol) as a function of δ backbone torsion angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

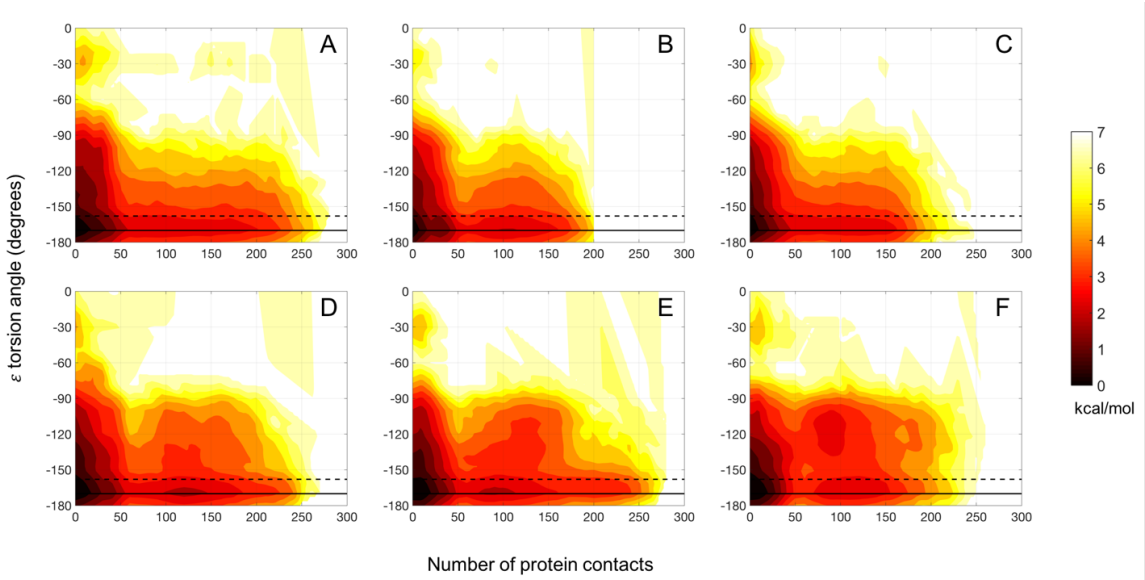

Figure 3.23 Potential of mean force (kcal/mol) as a function of ε backbone torsion angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

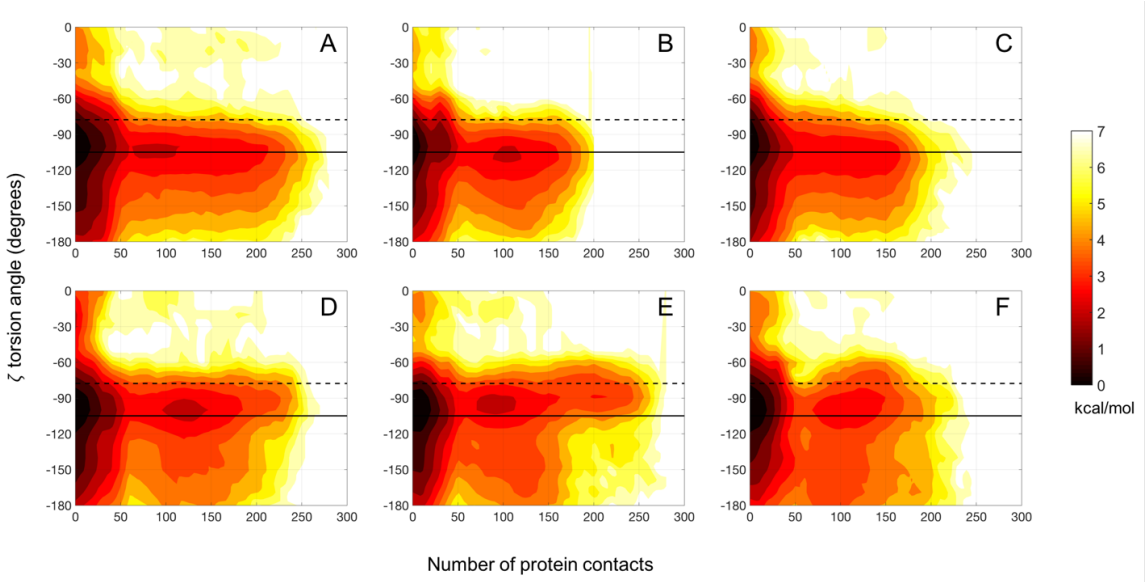

Figure 3.24 Potential of mean force (kcal/mol) as a function of ζ backbone torsion angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

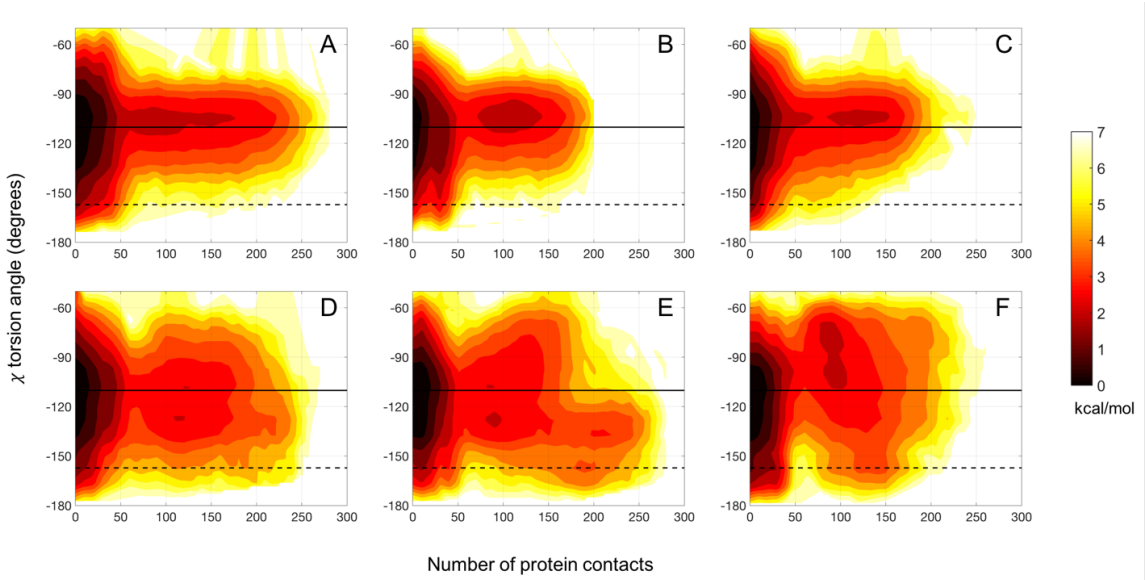

Figure 3.25 Potential of mean force (kcal/mol) as a function of χ backbone torsion angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

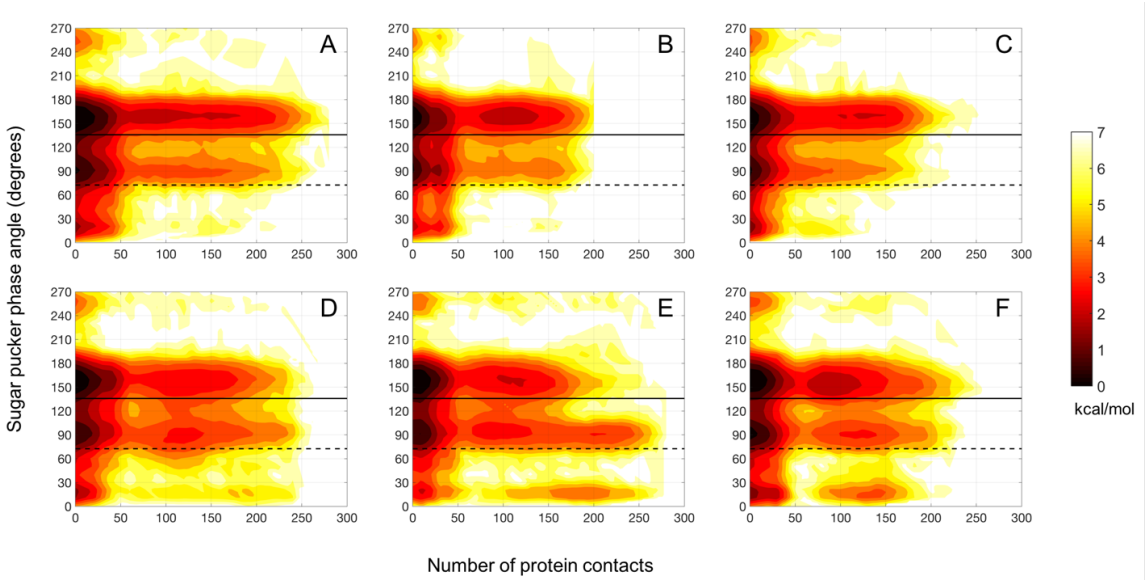

Figure 3.26 Potential of mean force (kcal/mol) as a function of sugar pucker phase angle and number of protein contacts for the Drew-Dickerson dodecamer at 20 % (A), 30 % (B), 40 % (C) protein concentrations, and for the GC-rich dodecamer at 20 % (D), 30 % (E), 40 % (F) protein concentrations. A contact is defined when the minimum distance between the heavy atoms of crowder proteins and DNA phosphate groups is less than 5 Å. Solid and dashed lines indicate the canonical B- and A-form values, respectively.

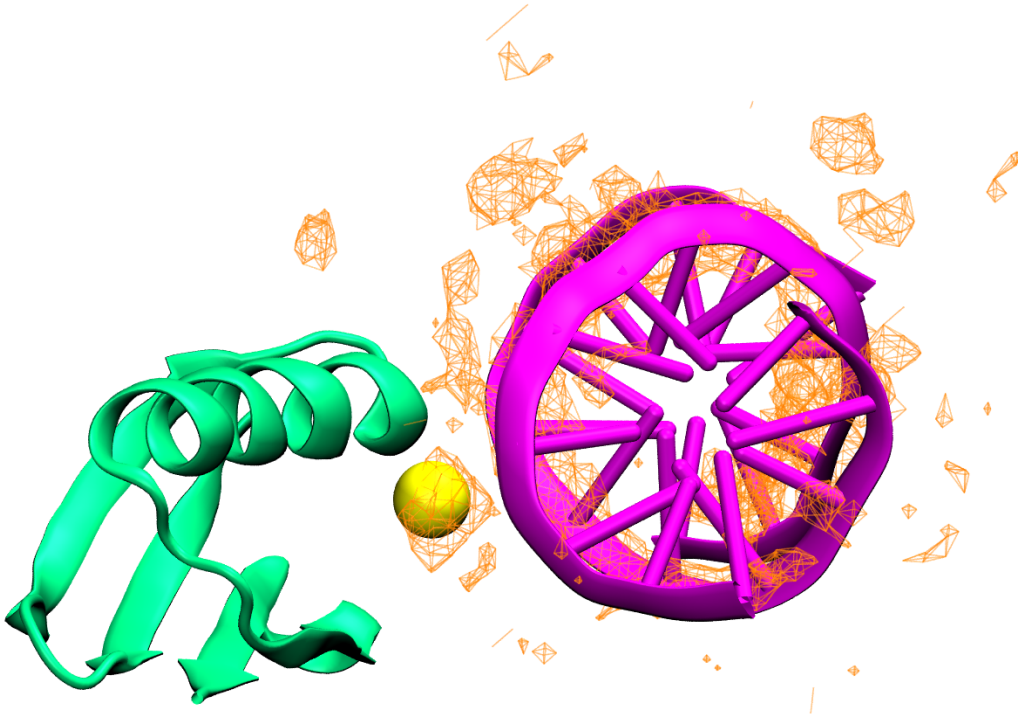

Figure 3.27 A snapshot showing a crowder protein interacting with the Drew-Dickerson DNA and orienting a sodium ion at the same time.

CHAPTER 4

High-Resolution 3D Models of *Caulobacter crescentus* Chromosome Reveal Genome Structural Variability and Organization

Asli Yildirim, Michael Feig

Submitted to

Nucleic Acids Research

4.1 Abstract

High-resolution three-dimensional models of Caulobacter crescentus nucleoid structures were generated via a multi-scale modeling protocol. Models were built as a plectonemically supercoiled circular DNA and by incorporating chromosome conformation capture based data to generate an ensemble of base pair resolution models consistent with the experimental data. Significant structural variability was found with different degrees of bending and twisting but with overall similar topologies and shapes that are consistent with Caulobacter crescentus cell dimensions. The models allowed a direct mapping of the genomic sequence onto the three-dimensional nucleoid structures. Distinct spatial distributions were found for several genomic elements such as AT-rich sequence elements where nucleoid associated proteins (NAPs) are likely to bind, promoter sites, and some genes with common cellular functions. These findings shed light on the correlation between the spatial organization of the genome and biological functions.

4.2 Introduction

Much is known about the gene organization within genomes, but the detailed three-dimensional (3D) structure of chromosomes has so far remained elusive. Gene-structure-function relationships at the DNA level are poorly understood as is the role of chromosomal structure in many cellular processes such as DNA transcription, replication and segregation. Two experimental approaches, fluorescence in situ hybridization (FISH) [197] and the recently introduced chromosome conformation capture (3C) techniques [21], especially the whole-genome sequencing variants (Hi-C) [22], have opened up new possibilities for understanding how chromosomes are folded inside the cell. FISH measures the spatial distance between two DNA segments in single cells providing a direct visualization of the relative positioning of different *loci* in a given chromosome. Hi-C methods generate genome-wide contact

probabilities between *loci* based on cross-linking from population of many cells providing information about spatial proximities between genomic elements that reflects 3D structure ensembles.

High-resolution structural insight of bacterial chromosomes has been derived from FISH [197-209] and 3C-based techniques [32, 210-215]. Bacterial chromosomes form highly compact structures induced by DNA supercoiling and further stabilized by the binding of nucleoid associated proteins (NAPs) [216, 217]. Further investigation of the global organization of bacterial chromosomes have revealed certain features of highly-organized chromosomal structure [200, 202, 218]. In *Caulobacter crescentus*, the origin and terminus of replication are located at opposite poles of a longitudinally organized chromosome, referred to as an *ori-ter* configuration [200]. This configuration has also been confirmed from 3C experiments on the *C. crescentus* genome [211, 214]. Chromosome Conformation Capture Carbon Copy (5C) study reported an ellipsoidal chromosome with periodically arranged arms confirming longitudinal organization [214]. Furthermore, a Hi-C study on the same bacterium further revealed that *C. crescentus* chromosome consists of multiple chromosome interacting domains (CIDs) with highly expressed genes found at domain boundaries [211]. On the other hand, the chromosome of *Escherichia coli* contains four macrodomains (Ori, Ter, Left and Right) with their localization dependent on different stages in the cell cycle [197]. Ori and Ter domains were also identified by 3C-based studies[210]. In slow-growing *E. coli* chromosome, the origin and terminus of replication are located near mid-cell and right and left chromosomal arms reside in separate cell halves [199, 202], so-called left-*ori*-right pattern, whereas the chromosome adopts an *ori-ter* configuration in fast-growing *E. coli* where these macrodomains localize at opposite poles of the cell [203]. The *Bacillus subtilis* chromosome is organized into a similar configuration alternating between *ori-ter* and left-*ori*-right patterns depending on the

cell cycle phase as in *E. coli* [201, 219]. CIDs have also been observed in the *B. subtilis* chromosome [212, 215].

NAPs play a crucial role in the observed organization of bacterial chromosomes [217, 220]. One of the most abundant NAPs, H-NS that is largely found in *E. coli*, has been reported to bridge different segments of the genome [221, 222] and control supercoiling of the DNA [223]. This suggests an active role in nucleoid organization. HU, another ubiquitous NAP in bacteria, has been found to influence chromosome compaction by wrapping the DNA around itself analogous to histones in eukaryotes thereby forming short-range interactions [211, 220]. Integration host factor (IHF) and factor for inversion stimulation (Fis) proteins help chromosome compaction by introducing DNA bending [217, 220]. Additionally, SMC (structural maintenance of chromosomes) proteins, which are believed to contribute to chromosome compaction [220], have recently been shown to affect collinearity of chromosomal arms in *C. crescentus* rather than regulating overall compaction [211].

Since bacterial genomes are not segregated into a special compartment as in eukaryotic cells, they occupy a large portion of the cell and interact extensively with the intracellular environment. Therefore, the spatial organization of a bacterial genomic DNA can play a major role in the regulation of biological functions in the cell. This idea is affirmed by a recent study where the statistical analysis of genome conformation capture data for *E. coli* suggested that operons from the same regulons and genes in the same biological pathway tend to be close to each other in 3D space, thereby maximizing compactness of the genome [224]. Furthermore, genes that are spatially close in 3D genomic structure tend to be co-expressed and their protein products are prone to form more protein-protein interactions [224]. A previous computational study, where the chromosome is modeled as a worm-like polymer chain with interacting sites corresponding to genes that are regulated by same transcription factors, showed that chromosomes form different topological orderings that increase local concentration of

interaction sites, therefore co-localize the co-regulated genes in 3D space, suggesting a correlation between gene co-regulation and 3D chromosome organization [225]. A very recent Hi-C study on the *Mycoplasma pneumoniae* chromosome has provided the first evidence of the correlation between 3D chromosome organization and transcriptional regulation [213]. They found that genes that are in the same CID have higher co-expression levels than genes between different domains [213]. In addition, independent of CIDs, they also reported high co-expression levels for spatially close genes [213]. Despite the recent progress, a more detailed understanding of the structural organization of bacterial chromosomes is still lacking.

The data generated with 3C-based methods lends itself to experimentally-driven modeling of 3D chromosome structures where restraints defined from 3C studies are used. Initial modeling approaches have used the generally accepted assumption that 3C-based contact probabilities are inversely related to the average distances between *loci* pairs [226]. Further calibration is possible by comparing 3C-based contact probabilities with average distances obtained via FISH. Such a calibration curve was obtained for *C. crescentus* where contact probabilities based on Chromosome Conformation Capture Carbon Copy (5C) for 112 different *loci* from the flagellated pole of swarmer cells were compared with FISH data [200]. This allowed for a direct conversion of the 5C contacts to distances, which could then be used as constraints during modeling. Initially, such an approach based on 5C data was used to generate models for the C. crescentus chromosome at 13-kilobase (kb) resolution which provided first insights into its spatial organization in 3D space [214]. Restraint-based modeling based on contact probabilities between loci pairs has also been applied for generating 3D structures of eukaryotic chromosomes or their subsections [227-232].

The generation of chromosome models simply based on satisfying restraints from 3C contact frequencies that are converted to distances via calibration against FISH data is seemingly straightforward, but it has been pointed out that this approach is problematic [226,

233-237]. 3C-based contacts stem from cross-links that can only form if two *loci* come within a certain contact threshold. This means that a given contact frequency for a certain pair of *loci* only reflects in what fraction of cells those *loci* come closer than the contact threshold, instead of directly reporting on the average distance between two *loci* as FISH does when averaged over many cells. In other words, 3C-based studies only give information about the low-end of the distance distribution while FISH considers the mean of the entire distribution. This has prompted efforts to employ population-based modeling techniques to generate ensembles that reflect cell-to-cell variability and explicitly consider the short-range sensitivity of 3C-based methods when matching the cumulative contact map from the experimental contacts [238-241].

Another issue is that the experimental data even with Hi-C is too sparse to fully determine 3D structures beyond kb resolution. This leaves an important role to computational methods to compensate for a lack of resolution in the experimental data by including general topological features and packing constraints as part of the model building protocol. The rationale for this strategy is similar to the well-established protocols for the determination of macromolecular structures based on restraints from nuclear magnetic resonance (NMR) data. In the case of NMR, such assumptions e.g. about peptides being polymers with a backbone and side chains with certain molecular bonding geometries are key to obtaining atomic resolution structures from data that is otherwise effectively at much lower resolutions. In the case of bacterial chromosomes, one can apply knowledge about a plectonemic structure made up of supercoiled segments with branching points and long persistence lengths to generate high-resolution models when combined with the experimental data. Le et al. generated 3D structures for the C. crescentus genome at 434 base pair (bp) resolution by modeling the chromosome as a circular polymer consisting of plectonemes and structures that best fit the Hi-C data were selected after varying model parameters [211]. These models further extended the knowledge on 3D

organization of *C. crescentus* chromosome gained previously and shed light on the CIDs and their organization.

Using a different strategy, Hacker *et al.* very recently described a model of the *E. coli* chromosome at nucleotide resolution [242]. Starting with a multi-scale polymer model that captures the plectonemic topology and physical properties of double-stranded and supercoiled DNA, spatially resolved models were generated primarily based on RNA polymerase (RNAP) binding data from CHIP-chip experiments [243]. RNAP binding sites obtained from the ChIP-chip data were used to identify highly-transcribed regions which were then modeled as plectoneme-free, i.e. not supercoiled, regions in the chromosomal structure. The modeling was then further guided to match the distributions of the beads that mapped to RNAP binding sites to the projected 2D distribution of RNAP as well as the distribution of the rest of the beads to the 2D distribution of HU proteins obtained from single-molecule fluorescence experiments for E. coli [244]. These models allowed, for the first time, the ground-breaking investigation of physical properties a bacterial chromosome at the nucleotide level and a direct mapping of genome sequence to structure.

In this study, we employed a similar multi-scale modeling protocol to encode the plectonemic and supercoiled topology of bacterial DNA via coarse-grained (CG) models at different resolutions up to the bp level but using 3C-based contact frequency matrices to guide the modeling. This protocol was used to generate structural ensembles for the *C. crescentus* chromosome where extensive Hi-C data [211] is available. The Hi-C data provides direct information about relative spatial distances between *loci* under the consideration of dynamics and population-based variations. Similar to the models generated by Hacker et al. for E. coli, the models presented here for C. crescentus are thus believed to provide an accurate picture of where gene loci are located in individual chromosome structures and how such distributions vary between cells and as a result of chromosome dynamics. As the models from Hacker et al.,

the models generated here also allowed a direct mapping of the genomic sequence onto the generated 3D structures and a detailed analysis of how the mapping of genomic sequences onto the bacterial chromosome structures may be related to biological function encoded by the corresponding genes.

4.3 Materials and Methods

Experimentally driven high-resolution models of the *C. crescentus* genome were generated as a hyper-branched polymer of supercoiled DNA segments forming plectonemic rosettes. The multi-scale modeling protocol is illustrated in Figure 4.1 and described in full detail in the *Supplementary Information (SI)*. Briefly, initial Monte Carlo (MC) sampling of a segment-based plectonemic model was guided by distance restraints based on Hi-C interaction frequencies [211] between specific pairs of *loci*. The distance restraints were derived by using a calibration curve obtained by Umbarger et al. [214] which is a polynomial function to map between interaction frequencies from 3C-based data and expected distances. Simply, such mapping was possible by comparing the available average spatial distances of loci pairs measured by fluorescence microscopy data for the C. crescentus chromosome and the corresponding interaction scores for pairs with similar genomic site-separations.

In the initial sampling round, the general topology of branches extending from a central ring was fixed, but branches were allowed to reconnect and move. Next, 15-bp CG models were constructed by wrapping higher resolution beads around the segments in the initial plectonemic models. The 15-bp CG models were then further refined via molecular dynamics (MD) simulations using a CG interaction potential that accounts for the elastic properties of DNA. An example of a C. crescentus model at 15-bp resolution is shown in Figure 4.2. In total, 1,050 models were generated covering a range of branch segment lengths and number of supercoiled loops (microdomains). The resulting models were subsequently reweighted to generate an ensemble of structures where not just the average distance between two loci but

also the distribution of contacts in the models is maximally consistent with the Hi-C scores. The reweighted ensemble was then used for all further analysis.

The final models were also used to reconstruct models at base-pair resolution by taking advantage of the long persistence length of double-stranded DNA (see Supplementary Data). The base-pair resolution model for the structure shown in Figure 4.2 is given in Figure 4.13. Projections of the beads in the bp-resolution models and 15-bp resolution models did not show any difference (Supplementary Figure 4.13B vs Figure 4.5A, Pearson's coefficient: 1.00, Slope: 1.00, Intercept: 0.00). Therefore, 15-bp resolution models were used for further analysis throughout this study.

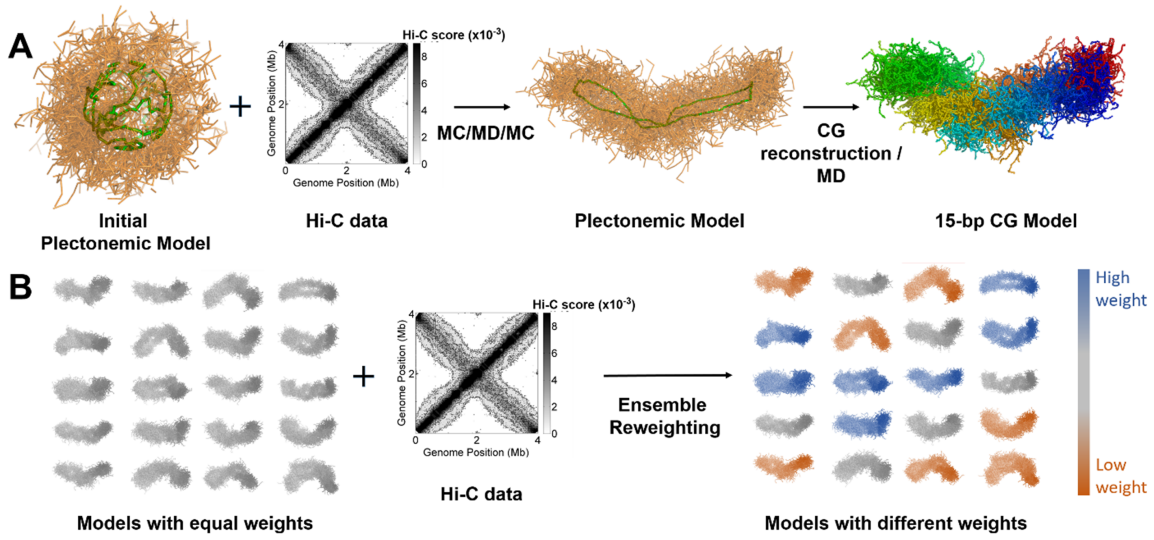

Figure 4. 1 Multi-scale modeling procedure during model generation based on Hi-C data. (A) Plectonemic and CG model generation. (B) Model reweighting.

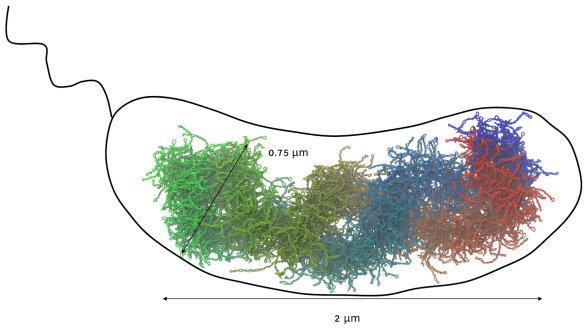

Figure 4. 2 3D structure of C. crescentus chromosome at 15-bp resolution projected onto a C. crescentus cell with typical dimensions.

4.4 Results

4.4.1 Structural Characterization of *C. crescentus* Chromosome Models

An ensemble of chromosome models for C. crescentus were generated as described above and in more detail in the Supplementary Data. To characterize the ensemble, we initially compare contact maps with the experimental data. The contact map based on the number of contacts within a distance threshold (see SI) is shown in Fig. 4.3. This map compares to the Hi-C scores and is in good agreement with the Hi-C contact map with a Pearson's correlation coefficient of 0.88 (Slope: 1.62, Intercept: 0.04) (Figure 4.3). Both contact maps exhibit the same characteristic two diagonals that reflect two chromosomal arms in an ellipsoidal shape interacting with each other [211, 214]. Previously, the inspection of the Hi-C interaction map revealed that C. crescentus chromosome is organized into 23 chromosome interacting domains (CIDs) which appear as triangles along the main diagonal [214]. The boundaries of the CIDs were identified by comparing the interaction preferences of loci from its left- and right-hand side; because when a locus is at the border of a CID, it strongly interacts either with its left- or right-hand side, whereas it interacts with both sides when it is in the middle of CID [211]. The triangles corresponding to CIDs are less apparent in our contact map, but the analysis of the interaction preferences of loci (SI) was carried out in the same manner as was done previously

[211]. The results are also shown in Figure 4.3. We found 21 CIDs, 18 of those match the CIDs identified previously from experiment. These results further confirm that our models are consistent with the information gained from the Hi-C experiment.

The average distance contact map at 10-kb resolution from the models is also shown in Supplementary Figure 4.15B and compared with the distance contact map converted from Hi-C interaction frequencies [211] by using the calibration curve derived by Umbarger et al. [214] (Figure 4.15A). The average distance map from the models shows an excellent agreement with the Hi-C derived map with a Pearson's correlation coefficient of 0.98 [211,214]. One significant difference between the two distance maps are apparent contacts between the origin (at 0 Mb) and the middle of the genome (at 2 Mb) in the experimental map. In the 3D organization of the chromosome, the middle of the genome (2 Mb) resides at the opposite pole from the origin, therefore these interactions are unlikely. The Hi-C scores for those regions were below the cutoff for contacts to be considered significant and likely result from replication and segregation of the origin to the opposite pole in some cells during the experiment [214]. On the other hand, if these contacts were included in the modeling protocol, a large fraction of highly bent structures would be necessary to satisfy average distances of ~650 nm between the opposite poles of the genome structure. Such bent structures are not consistent with the typical cell dimensions of C. crescentus (see below).

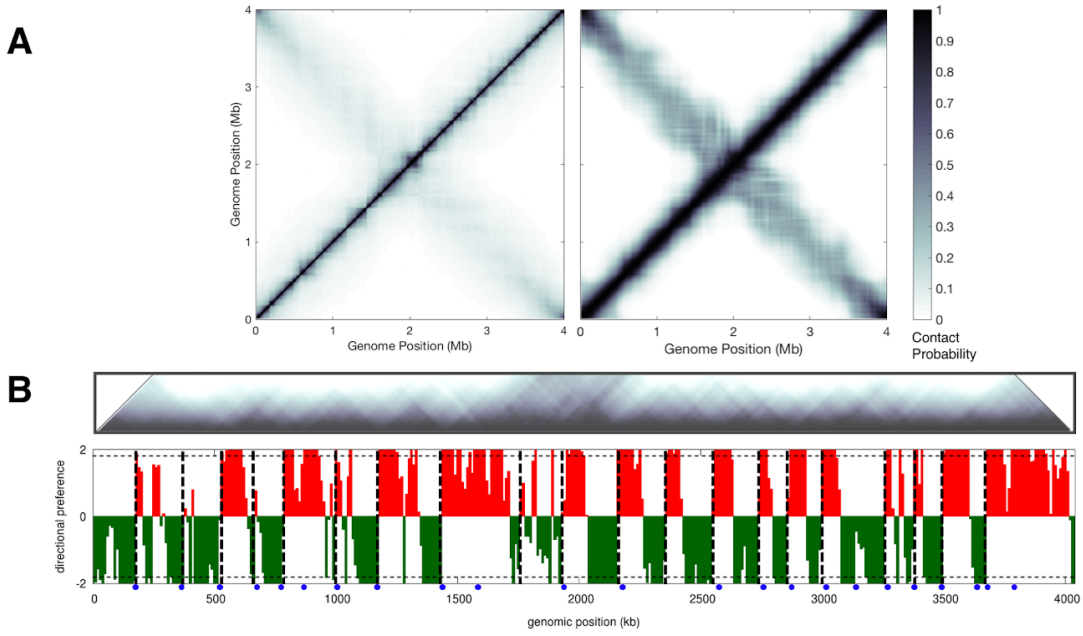

Figure 4. 3 Contact maps. (A) Hi-C score map (left panel) from Le et al. [211], average contact map of models by using cross-linking probability function used in the reweighting procedure (right panel). Hi-C scores were normalized to 1 for comparison. (B) The main diagonal of the average contact from models rotated 45° clockwise. The bottom panel shows the directional preferences of different loci (green: left-, red: right-hand side preference). Dashed lines show the boundaries of CIDs found from our models, blue dots indicate CID boundaries found in the study by Le et al [211].

The intramolecular distances can be further compared with recently reported distance measurements between 51 fluorescent-labeled DNA segments in the *C. crescentus* chromosome by Hong *et al*. [245]. We note that these distances were not used in the modeling or for generating the calibration curve between Hi-C interactions scores and spatial distances. Therefore, a comparison of the models with this data provides important independent validation [226]. The comparison of the projected average distances from the models with the experimental distance measurements is shown in Figure 4.4A. It can be seen that the models generally reproduce the experimental distances. For longer distances, the models have somewhat larger values than the experimental data suggesting an overall slightly too expanded shape. When distances of DNA segments only from the origin of replication are compared, the models produce similar distances up to 500 nm but become more extended for larger distances

compare to the available FISH data [200, 245]. This may be a consequence of modeling the DNA based on the Hi-C experiments that become insensitive to contacts beyond ~485 nm (see *SI*). This threshold stems from the fact that 3C-based experiments result in a flat distributions of interaction frequencies with long tails that correspond to the fragments that are in contact very frequently or infrequently. However, intermediate interaction frequencies are relatively noisy. Umbarger et al. have used the ~ 485 nm threshold to distinguish frequent contact frequencies from less reliable intermediate contacts but differences in the phases of the cells used in the FISH and Hi-C experiments can also explain different dimensions of the chromosome structures.

Another explanation for the different dimensions of the chromosome structures could be the differences in the phases of the cells used in the FISH and Hi-C experiments. FISH experiments were based on swarmer cells that were not allowed to grow FISH experiments were based on swarmer cells that were not allowed to grow [200, 245], whereas the Hi-C data were collected from cells allowed to grow for 0, 10, 30, 60 or 75 minutes [211] which would be expected to result in larger cell sizes [246] and, hence, allow for more extended chromosome structures.

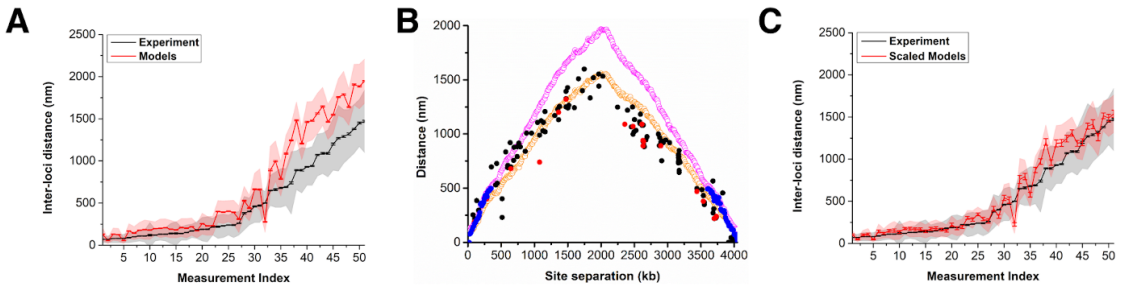

Figure 4. 4 Compatibility of the models with experimentally measured distances from FISH in the study by Hong et al. [245] (A) Distances for loci pairs from Hong et al. and from the models (B) Distance between the origin of replication and different DNA segments with different genomic distances from Viollier et al. [200] (black), Hong et al. [245] (red), models (pink), and scaled models (orange). The distances obtained from the calibration curve and used as restraints during the modeling are shown in blue. (C) Distances for loci pairs from Hong et al. and from models scaled by 0.8.

In order to test if our models would still be compatible with the Hi-C data if they had overall smaller dimensions consistent with the FISH data, we scaled the representative models of clusters obtained from the clustering analysis (see Figure 4.7) by 0.8 followed by 50 ns MD refinement to relax the models. This generated models that reproduce the correlation observed between distance and site-separation of *loci* pairs from experiment (orange curve in Figure 4. 4B) and, as expected, the FISH data also matched better (Pearson's correlation coefficient: 0.98, slope: 1.09, root-mean-squared-error: 15.73) (Figure 4.4C). The resulting average distance map after scaling is still in excellent agreement with the experimental data, but the Pearson's correlation coefficient of 0.96 with respect to the Hi-C data is slightly worse than for the unscaled models (Figure 4.14C).

Next, we analyzed the spatial dimensions of our models. Projections of the beads in the models onto their principal axes are shown in Figure 4.5A. The average length of the core structure is around 2 μm, with the edges extending to around 2.5 μm. This is slightly less than the experimentally observed *C. crescentus* cell lengths of ~2.5 – 3 μm [246]. The projection onto the short axes reflects the curved shape of most of our models. Therefore, we also determined the width perpendicular to the local axis of a curved line fit to the nucleoid models

(see *SI*). The resulting widths shown in Figure 4.4C are around 0.7 μm with a maximum extent to about 0.8 μm. For comparison, experimental cell widths are around ~0.8 μm [246]. Our models fit just inside the known dimensions of *C. crescentus* cells without imposing such a requirement during the modeling protocol. If the models are scaled to better fit the FISH data (see above), the dimensions of the chromosome become somewhat smaller (Figure 4.15, *SI*). In either case, our results imply that the genomic DNA fills out the majority of the cellular volume with the outermost parts of the DNA able to come close to the cellular membrane. These results are consistent with previous work showing the level of nucleoid compaction in different bacteria where nucleoids have been imaged by electron microscopy [247, 248]. Although the morphology of the nucleoid depends upon the organism as well as environmental conditions and growth rate [248, 249], nucleoids are generally found to occupy a large fraction of the cell cytoplasm [247, 248, 250]. In particular, the distribution of the nucleoid-associated HU protein in *C. crescentus* suggests that the nucleoid exhibits a diffuse morphology that extends to most of the cell volume [250]. Interestingly, the distribution of HU proteins in *E. coli* obtained from single-molecule fluorescence data [244] as well as the distribution of DNA in 3D models of *E. coli* nucleoid [242] are quite similar to the DNA distribution on the shortest axis of the *C. crescentus* nucleoid models (Figure 4.5A, green curve). This indicates that the level of nucleoid compaction for *E. coli* and *C. crescentus* are similar. We also compared the nucleotide densities from our models with the 3D models for the *E. coli* nucleoid [242] (Figure 4.5B). We compared absolute nucleotide densities rather than nucleotide counts to better compare with E. coli chromosome models since the number of nucleotides in C. crescentus and E. coli chromosomes is different and the distribution of nucleotides along the axis could be differentOur models show slightly less DNA densities which is consistent with *C. crescentus* having a smaller genome by about 600 kb compared to *E. coli* while the overall cell dimensions

are very similar. However, the overall distribution patterns of the DNA density along the short axis from both our and *E. coli* models was found to be very compatible.

Figure 4.5D shows the bending angle distribution observed for our models. The peak is between 150 and 160 degrees, which is comparable to the experimental bending angle of 162.9o with a standard deviation of 8.46o observed for the overall C. crescentus shape [246]. Our models show a slightly wider distribution with smaller bending angles, however we note that the average experimental bending angle we use for comparison is for swarmer cells [246]; therefore, the difference could again be because of the cells from different cell stages used in the Hi-C experiment.

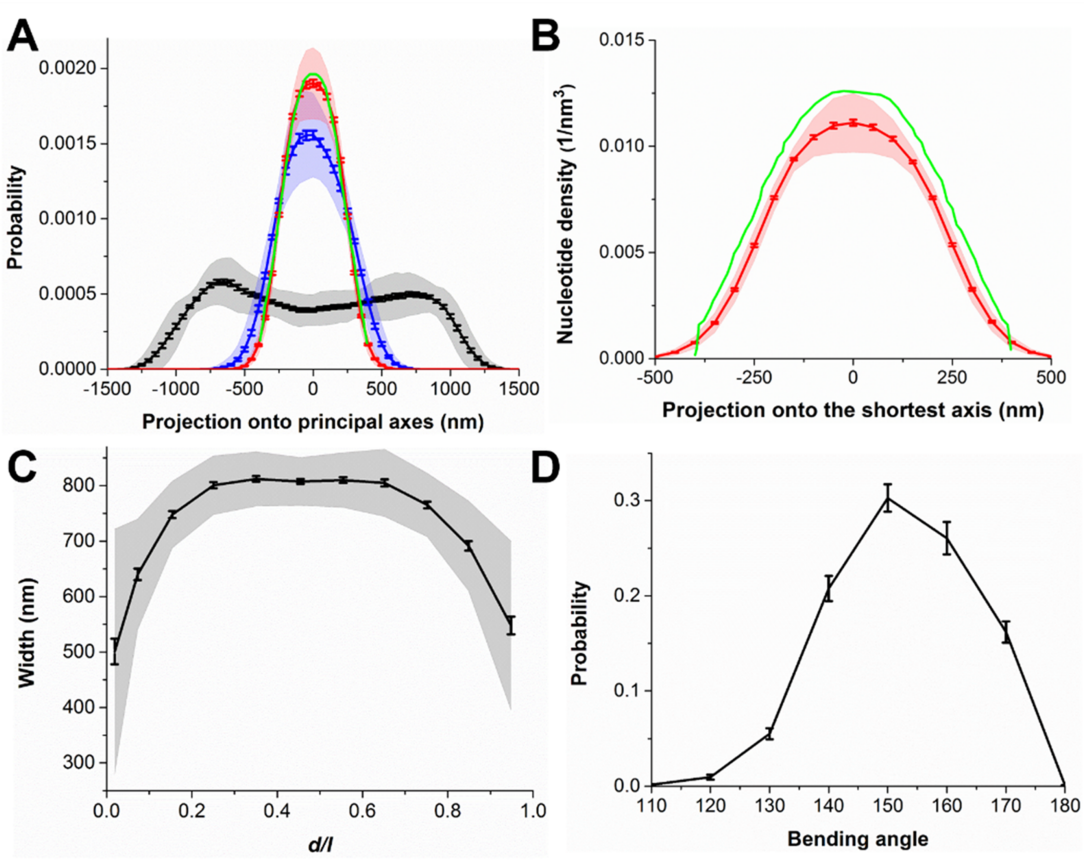

Figure 4. 5 Dimensions of the models. (A) Projections of beads in the models onto their longest principal axis (black), the second principal axis (blue), the shortest principal axis (red) and the DNA distribution based on HU proteins in E. coli nucleoid short axis reproduced by digitizing the figure 4A-ii in the article from Stracy et al. [244] (green). In order to compare the E. coli data with our projections, the relative x-axis values given in the experimental results were multiplied by half the C. crescentus width (400 nm) to obtain comparable distributions, and the curve was normalized again to keep the area under the curve 1. (B) Nucleotide density projection onto the shortest axis from our models (red) and 3D models of E. coli chromosome [242] (green) reproduced by digitizing the figure 5A in the article from Hacker et al. [242]. The probability densities of DNA in E. coli models in the reference were converted to nucleotide densities by multiplying the probability densities by the total number of nucleotides in E. coli and divided by the width and length of their models. (C) Width of the models vs. the positions along the longest principal axis, p, normalized to the length of the models along the x-axis, l. Here, 1 represents the pole that contains the replication of origin while 0 represents the opposite pole. (D) Distribution of bending angles for the models. The shaded areas indicate the standard deviations and the error bars indicate the standard errors based on the ensemble averages.

Since the bacterial nucleoids occupy a large portion of the cell, extensive interactions with the intracellular environment are unavoidable and proteins as large as RNAPs and ribosome subunits have been found to penetrate into nucleoids to initiate co-transcriptional translation

[244, 249, 251]. On the other hand, complete ribosomes and polysomes are strongly segregated from the nucleoid [252, 253] presumably excluded from the nucleoid based on their size. We analyzed our models to determine whether they are compatible with the experiments, i.e. whether the structures are porous enough for RNAP and ribosome subunits to enter while excluding assembled ribosomes. Figure 4.6A shows the accessible cavity volume as a function of macromolecular radius. Essentially molecules with radii up to 10 nm are able to penetrate the chromosome structures at significant fractions whereas molecules larger than 15 nm are almost fully excluded. Table 4.1 compares accessible cavity sizes for a typical size protein, RNAP, the 30S and 50S ribosomal subunits, and entire ribosomes based on their hydrodynamic radii [254]. Despite its size, RNAP can access half of the volume available to small proteins and the ribosomal subunits can still access about a quarter of the volume suggesting that both RNAP and ribosomal subunits are able to penetrate and interact extensively with the DNA. However, assembled ribosomes can only access 10% of the space accessible to small proteins consistent with the experiments that find ribosomes to be largely excluded from the nucleoid [252, 253]. If the scaled models that better fit the FISH data are used, the overall size of the cavities is reduced but the general conclusions remain the same (Table 4.1, Figure 4.16, *SI*). Figure 4.6B further contrasts the much-reduced volume accessible to ribosomes in our models compared to the volume that can be occupied by RNAP. Overall, although the results here depend on the parameters used in the analysis, they suggest that larger molecules have more difficulties accessing the nucleoid interior and seem to agree with the finding that the porous nucleoid structure would allow mobile RNAP to diffuse relatively easily within the nucleoid in contrast to ribosomes [244]. Our results can also be compared with the void distribution in the recent models for the *E. coli* chromosome [242]. The *E. coli* models appear to be less compact since voids with radii up to 40 nm were identified in that study and entire ribosomes with radii of about 12 nm appear to fit comfortably within most of the structure according to

the distribution of void sizes [242] despite the higher DNA density compared to our models (Figure 4.4B). The void size distribution in our models appears to be more consistent with the experimental data than in the *E. coli* models, but a direct comparison between our analysis and the work by Hacker *et al.* is complicated by likely differences in how exactly the voids are calculated as well as slightly different DNA topological parameters such as higher superhelical densities in the *E. coli*.

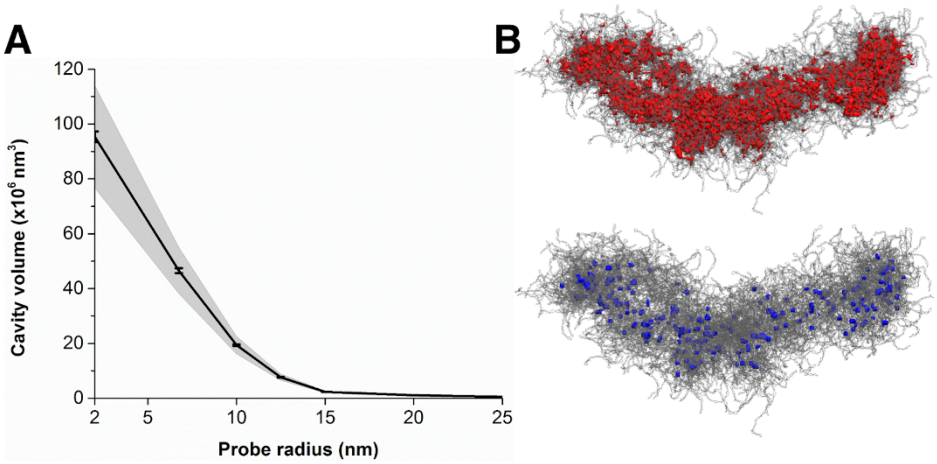

Figure 4. 6 Cavities in the models. (A) Distributions of the cavities for different probe radii in the models. The shaded area indicates the standard deviations and the error bars indicate the standard errors based on the ensemble averages. (B) Cavities that could be occupied by RNAP (red) and ribosomes (blue).

Table 4.1 Nucleoid cavity volumes accessible to proteins of different sizes. Ensemble-averaged volumes with standard errors given in the parentheses. Hydrodynamic radii of the molecules were calculated by HYDRPRO [254] based on PDB structures for RNA polymerase (4KMU), the 30S subunit (5NO3), the 50S subunit (5ADY), and the ribosome (4V4Q).

	Hydrodynamic radius (nm)	Cavity volume (x10^6 nm^3)	
		Original models	Scaled models
Average size protein	2.0	95.4 (*1.5*)	86.0 (*1.8*)
RNA polymerase	6.9	45.5 (*0.7*)	25.5 (*0.4*)
Ribosomal subunit 30S	8.7	28.6 (*0.4*)	11.1 (*0.2*)
Ribosomal subunit 50S	10.0	19.3 (*0.3*)	5.7 (*0.1*)
Ribosome	11.7	10.6 (*0.1*)	2.1 (*0.04*)

4.4.2 Structural variability in the ensemble

All of the nucleoid models generated by our protocol have overall similar shapes. The origin of replication is at one end of the structure and the prominent feature are two arms that are wound around each other in most structures following a sinusoidal pattern with 1.5 – 2 period repeats. The overall shape is similar to lower-resolution models reported previously by Umbarger *et al*. [214]. Many structures exhibit bending that generally follows the curved rod shape of *C. crescentus* cells [246] as already discussed above. Beyond these overall features, there are significant structural variation at the more detailed level. Clustering of the ensemble based on pairwise mutual similarity resulted in 27 different major groups. Representative structures for each cluster are shown in Figure 4.7A, population percentages are given in Table 4.2, and individual contact maps for each cluster that show deviations and many different patterns are shown in Figure 4.17 (*SI)*. The weights of different clusters were optimized in the final step of the ensemble generation to match the resulting contact distributions to experimental Hi-C contact frequencies as described in the *SI*. The weights of the clusters show that some clusters with lower population percentages have higher weights. This could be because in our modeling protocol the initial sampling may be biased and not fully reflect the correct relative populations.

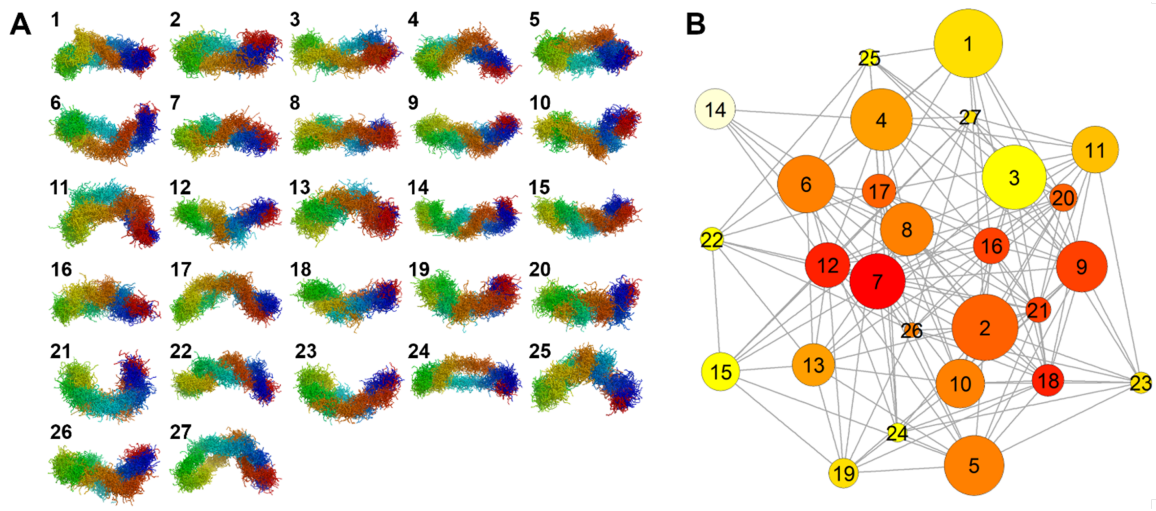

Figure 4. 7 Clustering of our models and possible inter-conversions between clusters. (A) Representative structures for ensemble clusters. (B) Inter-conversion between clusters based on targeted molecular dynamics. Colors indicate the number of connections to other clusters (from red to white, descending order) and the size of the circles corresponds to the weights of the clusters (0.005 – 0.113). The indices correspond to the structures shown in (A)

Table 4.2 Features of major clusters of nucleoid structures. Structural characteristics are averaged over all members in each cluster. The population percentages refer to the models in the unweighted ensemble. Standard errors are given in parentheses.

Cluster	Population %	Weight	Medial axis length (nm)	Width (nm)	Bending angle (degrees)	Bending direction (degrees)	Arm-twisting (degrees)	Arm-crossings
1	2.10	0.113	2684.7 (*52.1*)	762.7 (*12.6*)	151.5 (*2.1*)	-12.9 (*9.1*)	1.27 (*0.61*)	0.07 (*0.23*)
2	2.48	0.103	2549.9 (*33.3*)	753.8 (*9.6*)	148.3 (*2.3*)	6.9 (*17.4*)	-0.26 (*0.53*)	-0.49 (*0.25*)
3	2.00	0.095	2666.8 (*30.1*)	723.8 (*10.2*)	163.2 (*1.6*)	-110.0 (*21.3*)	3.39 (*0.92*)	-0.40 (*0.26*)
4	3.05	0.071	2648.1 (*34.3*)	745.6 (*8.9*)	153.9 (*1.9*)	-42.8 (*10.6*)	-0.92 (*0.48*)	0.26 (*0.17*)
5	1.71	0.068	2441.7 (*33.0*)	738.9 (*10.3*)	159.8 (*2.7*)	-76.1 (*14.5*)	-1.40 (*0.95*)	0.11 (*0.34*)
6	2.38	0.056	2649.9 (*37.5*)	752.8 (*10.1*)	150.8 (*1.8*)	21.5 (*19.4*)	-1.62 (*0.72*)	0.88 (*0.25*)
7	2.29	0.045	2761.9 (*30.8*)	747.5 (*9.5*)	160.0 (*1.6*)	64.8 (*21.5*)	0.49 (*0.91*)	-0.26 (*0.30*)
8	0.57	0.044	2839.1 (*28.0*)	746.7 (*21.7*)	160.5 (*3.3*)	10.8 (*14.3*)	2.55 (*0.58*)	-0.79 (*0.16*)
9	2.29	0.042	2870.2 (*31.2*)	746.7 (*9.1*)	145.5 (*1.6*)	90.9 (*8.7*)	4.38 (*0.78*)	-1.18 (*0.22*)
10	8.86	0.033	2762.6 (*14.1*)	725.8 (*5.2*)	151.6 (*0.9*)	124.2 (*4.1*)	-1.17 (*0.39*)	0.12 (*0.16*)
11	2.67	0.031	2774.0 (*36.1*)	823.6 (*14.0*)	130.8 (*2.1*)	-61.9 (*20.1*)	0.96 (*0.61*)	0.01 (*0.16*)
12	3.52	0.029	2755.8 (*28.4*)	725.4 (*9.4*)	143.0 (*1.5*)	124.8 (*8.5*)	1.52 (*0.57*)	-0.11 (*0.20*)
13	4.10	0.028	2694.0 (*27.0*)	723.7 (*7.9*)	159.2 (*1.6*)	-179.7 (*10.6*)	-0.62 (*0.54*)	0.38 (*0.17*)
14	3.33	0.028	2888.4 (*26.3*)	747.4 (*7.7*)	149.9 (*1.5*)	-65.9 (*11.3*)	-3.17 (*0.63*)	0.87 (*0.19*)
15	4.38	0.027	2791.3 (*24.3*)	721.7 (*6.9*)	144.1 (*1.5*)	-24.1 (*8.2*)	4.05 (*0.57*)	-1.28 (*0.19*)
16	8.57	0.026	2745.4 (*15.2*)	718.4 (*5.8*)	151.9 (*1.0*)	-47.3 (*6.9*)	-0.68 (*0.37*)	-0.16 (*0.14*)
17	1.05	0.025	2964.0 (*51.3*)	760.0 (*13.5*)	147.6 (*3.7*)	-100.3 (*30.9*)	2.92 (*1.07*)	-0.37 (*0.32*)
18	5.43	0.024	2729.7 (*18.3*)	734.0 (*6.5*)	147.5 (*1.4*)	-125.2 (*9.7*)	-1.31 (*0.38*)	0.64 (*0.17*)
19	4.19	0.023	2707.7 (*21.7*)	745.0 (*7.2*)	142.5 (*1.4*)	14.5 (*7.8*)	0.24 (*0.51*)	-0.53 (*0.17*)
20	2.00	0.019	2682.5 (*44.3*)	733.3 (*11.2*)	149.6 (*2.5*)	53.1 (*17.0*)	-3.29 (*0.52*)	1.00 (*0.20*)
21	4.48	0.018	2698.8 (*33.4*)	782.1 (*8.5*)	135.9 (*1.2*)	-97.8 (*13.1*)	0.56 (*0.56*)	0.14 (*0.17*)
22	8.00	0.018	2815.1 (*17.5*)	754.0 (*5.8*)	135.2 (*1.1*)	-108.1 (*9.2*)	3.10 (*0.46*)	-0.56 (*0.12*)
23	4.57	0.011	2795.9 (*23.6*)	774.2 (*7.1*)	133.8 (*1.3*)	-9.9 (*6.6*)	-3.39 (*0.50*)	0.48 (*0.15*)
24	2.57	0.007	2888.3 (*31.9*)	744.4 (*7.6*)	143.8 (*1.7*)	-133.5 (*16.6*)	-2.61 (*0.58*)	0.77 (*0.18*)
25	2.67	0.007	2749.5 (*29.5*)	731.4 (*9.8*)	145.7 (*1.5*)	35.7 (*5.9*)	-0.70 (*0.68*)	0.26 (*0.23*)
26	3.71	0.006	2763.0 (*24.9*)	787.2 (*8.2*)	132.5 (*1.8*)	15.2 (*7.8*)	-0.50 (*0.54*)	-0.36 (*0.15*)
27	4.76	0.005	2851.8 (*24.5*)	734.0 (*6.4*)	116.8 (*2.5*)	2.9 (*9.4*)	0.38 (*0.58*)	-0.22 (*0.17*)

Table 4.2 summarizes the clusters along with their optimized weights and selected structural properties based on cluster averages. The different clusters are primarily distinguished by different degrees of twisting of the arms and bending. Bending angles range from 116.8° (for the most bent cluster 27) to 163.2° (for the least bent cluster 3). The most strongly bent structures tend to have the lowest weights and as they do not seem to fit well into reported bending angles of 162.9̊±08.5 for the *C. crescentus* cell shape [246], they are either rare outliers or modeling artifacts. The bending directions with respect to the two arms also vary in different clusters. 12 clusters show one direction, while others are bent in the opposite direction. Clusters with negative bending direction angles tend to bend towards the nucleoid arm which holds high-index *loci*, in contrast, the positive bending direction results in clusters with bending towards the low-index *loci* arm. In terms of the twisting patterns of the arms, all clusters seem to have partial twisting of arms around each other as observed previously [214].

We further quantified the degree of twisting by calculating the average twisting of one arm around the other and the number of crossings of the arms (Table 4.2) to understand differences between clusters. Positive twist values correspond to right-handed twisting and negative values indicate left-handed twisting. Both directions are present in the models. Although the overall twisting angles sum up to values near 0°, the twist angles fluctuate between -20° – 20° along the nucleoid axis with different patterns in different clusters (Figure 4.18, SI). Right-handed twisting is more dominant at the poles for all clusters, while there is more variation in the twisting pattern near the centers of the nucleoid. The number of arm-crossings also differs among the clusters. Some clusters (1 and 11) show, on average, less arm-crossing values indicating that the arms prefer to remain on different sides of the nucleotide without crossing in these structures. On the other hand, clusters 9 and 15 have larger twisting and crossing values, therefore these clusters tend to

have their arms more extensively intertwined. Although the clusters differ in terms of bending and arm-twisting patterns, they all have similar medial axis lengths (see *SI* for the medial axis definition) and widths, consistent with the reported cell size dimensions of *C. crescentus* [246] as discussed in the previous section (Table 4.2).

As described in the *SI*, we generated models with different branch lengths and number of microdomains. Interestingly, each cluster contains a mixture of different internal topologies in terms of the branch length and the number of microdomains (Table 4.3, *SI*). This suggests that the overall structure in our models is not sensitive to the detailed topology of the DNA. An exponential distribution of microdomain sizes in E. coli chromosome was reported by Postow et al. [15]. A comparison of that data with the microdomain size distributions in our models is shown in Figure S19. Generally, the distributions are similar and the exponential distribution is reproduced in our models. However, our models have overall smaller microdomains and lack the very large microdomain sizes (50 – 60 kb) seen in E. coli. Additionally, the total lengths of the branch segments in a microdomain were also reported for plasmids at ~ 7 kb size by Boles et al. [17]. The total length of the branch segments in the microdomains with similar sizes from our models are in good agreement with the reported branch lengths except for the smallest number of microdomains (Figure S19).

An ensemble of nucleoid models with different structures as reported here would be expected based on cell-to-cell variations. However, nucleoid structures are known to be highly dynamic. For example, studies on *E. coli* nucleoid dynamics in living cells have shown the possibility of global and local nucleoid dynamics during a cell-cycle [198, 255]. Therefore, it is an interesting question that to what extent the different conformations for the *C. crescentus* nucleoid are interconvertible without encountering significant topological barriers that could not be overcome

without unwinding and expanding the nucleoid structures. We carried out targeted MD simulations between all pairs of representative cluster structures to test which interconversions are likely feasible based on simple energetic criteria (see *SI*). We find that many structures appear to be in fact interconvertible, at least at the overall topological level (see Figure 4.6B). There are enough pair-wise connections for all of the structures to be connected either directly or indirectly via intermediates. Some structures stand out as central hub structures with a high number of connections. It appears, however, that the number of connections is not strongly correlated with the cluster weights which may suggest that many of the hub structures are less stable intermediates.

While the simulations do not provide meaningful insight into relative energetics or kinetics due to the biased and coarse-grained nature of the targeted MD simulations, the general conclusion is that in addition to cell-to-cell variability, the nucleoid structures could be dynamically sampling a wide variety of structures within a single cell. A better understanding of nucleoid dynamics is clearly an area that would benefit from further studies, both experimentally and via simulations.

Finally, we analyzed our models with respect to macrodomain formation. We previously showed that the investigation of the average contact map revealed 21 CIDs (Fig. 4.3) which is in agreement with the number of CIDs identified by Hi-C experiment [211]. To test if the same conclusion can be achieved when hierarchical clustering of the CG beads is performed based on pairwise distances. Similarly, the results show a varying number of macrodomains in the range of 16 to 34 depending on the cluster (Table 4.3, SI) and macrodomains consisting typically of 100-350 kb. As an example, the macrodomains of cluster 1 are shown in Figure 4.8.

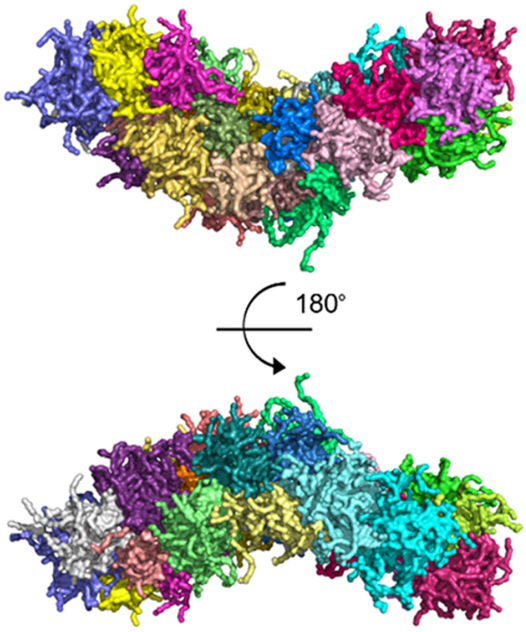

Figure 4. 8 Macrodomains based on hierarchical clustering for cluster 18.

4.4.3 Genome-structure mappings

Our models of the *C. crescentus* nucleoid are of sufficiently high-resolution to directly map the genetic sequence onto the 3D structure. This allows an investigation into possible correlations between the DNA structure and the genome sequence that it represents. In order to facilitate the analysis of projections onto variable 3D structures, we will primarily discuss projections of sequence features onto the long axis of the nucleoid models. All of the results were further averaged over the entire ensemble of models taking into account the weights of the clusters that different models belong to.

We began by analyzing the spatial locations of basic sequence features. The distribution of AT- and GC-rich sections is shown in Figure 4.9 in comparison with the positions of all base pairs in the model. The distributions of AT-rich sites are significantly different from the distribution of all base pairs with enrichment in the central region towards the origin of replication (*Cori*) and

depletion opposite *Cori* but no preference was found for GC-rich sites. Most NAPs are known to bind to AT-rich sites[217], therefore an enrichment of AT-rich sites from the center to *Cori* may imply enhanced binding of NAPs in that region. Integration host factor (IHF), another NAP, is also reported to assist in maintaining a compact genome by introducing U-turns to DNA [217]. We find a statistically significant preference for binding near the center of the nucleoid (Figure 4.9B). Many of our models are bent and such bending is primarily facilitated by kinking in the central region. Therefore, it could be that IHF is involved in stabilizing a bent nucleoid structure to better fit into the curved bacterial envelope of *C. crescentus*. If this hypothesis is correct, we would expect straightening of the DNA when the IHF gene is disabled.

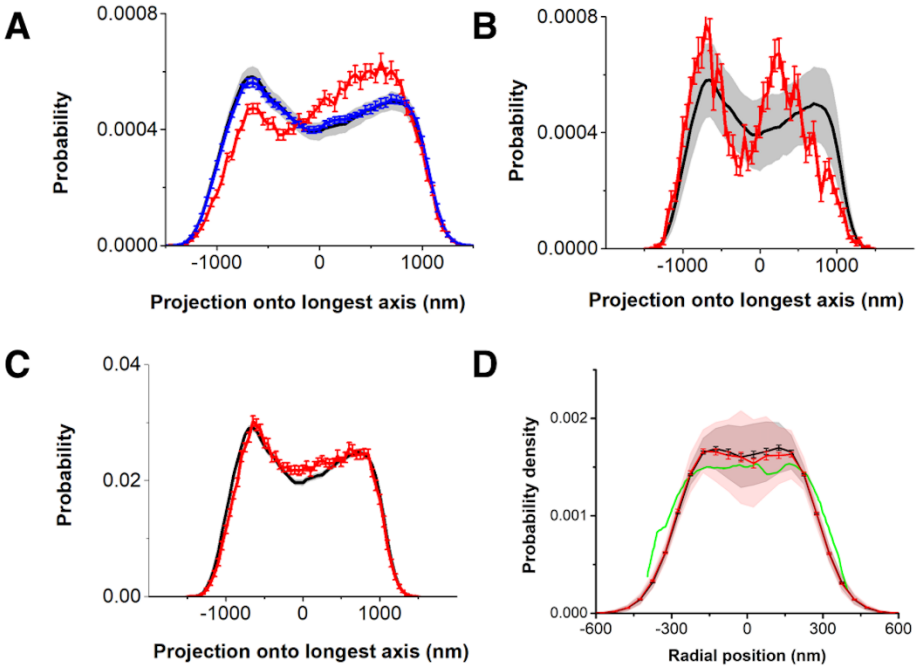

Figure 4. 9 Projections of genomic sequence features onto the 3D nucleoid structures. (**A**) Sites with least 70% is AT-sites (red) or GC-sites (blue) within a 20-bp window compared with the positions of all base pairs (black), (**B**) IHF-binding sites (red), (**C**) Promoter sites (red). (**D**) Radial positions of the promoter (red) and all (black) sites in the models from the nucleoid center (black) and the distribution of bound-RNAP in E. coli nucleoid short axis reproduced by digitizing the figure 4A-ii in the article from Stracy et al. [244] (blue). In order to compare the E. coli data with our projections, the relative x-axis values given in the experimental results were multiplied by half the C. crescentus width (400 nm) to obtain comparable distributions, and the curve was normalized again to keep the area under the curve 1. Gray shaded areas in A, B, and C indicate standard deviations obtained from distributions for the same number of loci that were randomly selected 200 times. The shaded area in D indicates the standard deviations and all the error bars indicate the standard errors based on the ensemble averages.

Promoter sites show only a slight preference for the middle part of the nucleoid along the longest axis (Figure 4.9C). Recently, the distribution of RNAP in *E. coli* nucleoid showed that RNAPs that are specifically bound to DNA tend to stay closer to the edge of the nucleoid [244]. The positions of the bound RNAPs on the nucleoid were later used in the *E. coli* nucleoid modeling and the generated 3D models reproduced the experimental distribution [242]. Here, we also

analyzed the radial distributions of the promoters along the medial axis of the models (see *SI*) in order to compare with the bound RNAP distribution found for *E. coli* [244] (Figure 4.9D). We note that in our modeling protocol we did not impose the RNAP distribution as a constraint. The promoter distribution is quite similar and a slight depression in the bound-RNAP distribution in the center relative to the distribution of nucleotides (see Figure 4.5A) was also observed for the promoter distribution of our models. These results suggest that promoters might be pre-arranged to reside on the outside of the nucleoid to facilitate transcription. However, a similar distribution was found for the radial distribution of all beads which suggests that there is not a clear special pre-arrangement of promoters that would favor positions on the 'outside' of the chromosome structure.

A study by Fang *et al*. analyzed groups of genes in *C. crescentus* that are co-expressed [256]. 76 different gene modules were clustered according to their expression profiles [256]. In order to test whether co-expression correlates with spatial co-localization, we mapped the beginning of the operon for the corresponding genes in each module onto our models and calculated spatial proximity from the average pair-wise distances of all genes in a given module. We note that we only included the genes with unique operons for each module in the analysis. The same analysis was also performed considering genomic distances rather than spatial distances. The resulting z-scores for the modules for which genomic separations of operons are not different than the genomic separations of randomly selected ones ($-1<z\text{-score}<1$) are shown in Figure 4.10A. Out of total 43 modules, z-scores of 3 modules are skewed towards negative values indicating that a subset (about 7%) of the modules of co-expressed genes is also co-localized. Additionally, in order to avoid any linear sequence effects, we also analyzed only the gene-pairs that are at least 500 kb apart in genomic sequence and we found that the percentage of the modules that have negative z-scores is

increased to ~20 %. However, we also observed that a similar fraction of co-expressed gene modules resulted in positive z-scores. Overall, the results show that although the genes that are co-expressed seem to have some non-random distributions, a strong tendency for co-expressed gene pairs to be spatially closer was not found. Z-scores of the models are listed in Supplementary Table 4.4. While co-expression is expected to involve the regulatory elements at the beginning of operons, we also examined whether the observed co-localization remains valid if we analyze the positions of the end of each gene in a given module. Figure 4.10B shows that there is no significant difference to the analysis of the beginning of the operon for each gene.

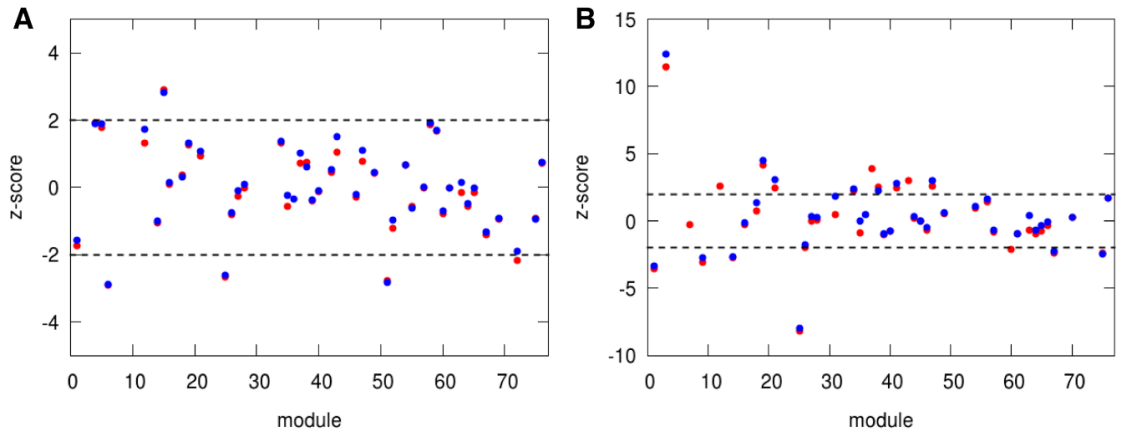

Figure 4. 10 Correlation between co-localized and co-expressed genes. (A) Z-scores for each module of intra-gene distances for co-expressed genes vs. random genes for the end of genes (red) or the beginning of the corresponding operons (blue). (B) Same plot as in (A) but only for gene or operon pairs that are separated at least by 500 kb in the genomic sequence.

To further examine a possible relation between gene co-localization and protein product co-localization, we compared with another experimental study on *C. crescentus* where the localization of ~300 proteins was analyzed [257]. We mapped the corresponding genes according to their protein product localization onto our models. We found weak evidence that genes whose products are near the poles may be located closer to *Cori*, on one end of the structure (Figure 4.11), but we could not correlate proteins localized in the center of the cell with an enhanced localization of their

genes in the center of the nucleoid (Figure 4.11). These results are again consistent with the findings from *E. coli* chromosome models where no localization have been found for the genes of which protein products are co-localized [242]. While this does not contradict the idea that proteins are initially synthesized close to where the gene is located, it suggests that the memory of the synthesis site is largely lost due to protein diffusion and/or other cellular transport processes in the experimental study.

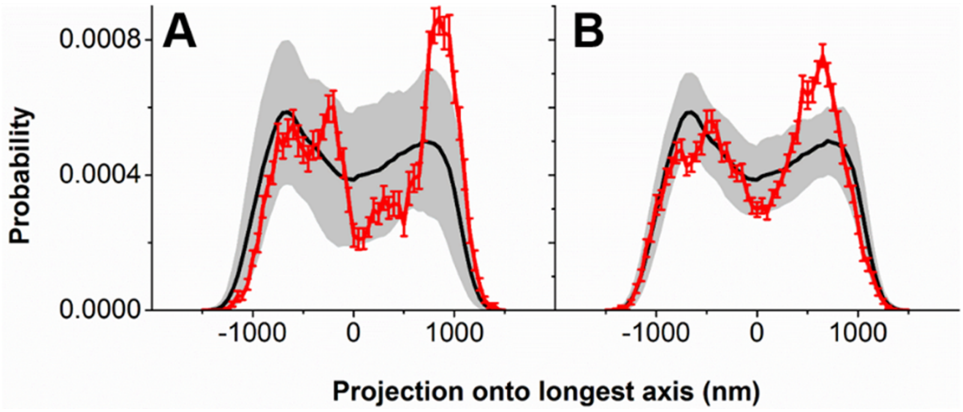

Figure 4. 11 Correlation between gene co-localization and protein product co-localization. Distributions of the genes of the proteins which are experimentally found to be at central **(A)** and polar (B) locations [257]. Central genes include Midband and Central Focus genes whereas polar genes correspond to polar and bipolar genes in the experimental analysis [257]. Gray areas indicate the standard deviations as calculated in Figure 8. The error bars indicate the standard errors based on the ensemble averages.

Finally, we analyzed the spatial organization of the genes as a function of their cellular functions to test the hypothesis that the localization of certain genes on the nucleoid may be correlated to where they are needed within the cell. Groups of functionally related genes that were distributed significantly different from a random distribution are shown in Figure 4.12. Genes for regulatory functions, and proteins with unknown functions ('hypothetical proteins') were found to localize in the middle of the nucleoid. In contrast, translation, metabolism and replication genes appear to have a tendency to localize near the poles of the nucleoid while transcription genes tend

to be clustered around the *Cori* region. Genes involved in cell division are both near the center and near *Cori*. One can rationalize why replication genes would be near *Cori* and why cell division genes are near the center since Fts genes form rings attached to the membrane to constrict the cell during cell division and parA/parB accumulate near the poles before cell division in *C. crescentus* swarmer cells [258-261]. The preferential localization of metabolic genes opposite *Cori* could contribute to enhanced metabolic efficiency by bringing proteins involved in metabolic cascades closer together. Based on our data, we generated a diagram in Figure 4.12B to summarize relative preferences for genes with different functions along the nucleoid. We assume that the resulting protein products would also be enhanced or suppressed accordingly given that transcription and translation is local to a given gene location and that diffusion in crowded environments is relatively slow [262]. In general, our findings based solely on the distribution of genes on the nucleoid structure are consistent with findings for *E. coli* where a special organization of genes with similar biological pathways was found [224] and also with the study by Junier *et al*. that found a correlation between co-localization of genes with similar transcriptional regulations and the 3D structure of the chromosome [225]. However, it is clear that further experimental validation is necessary to fully understand a possible correlation between gene location on the nucleoid and its function. Experimentally, this could be accomplished for example by analyzing phenotypes for bacteria with shuffled gene distributions.

In order to ensure that the localizations of genomic elements observed in our models are really an effect of the 3D organization of the chromosome rather than the relative locations of genomic features in linear sequence, we also checked their linear sequence localization. In general, it is difficult to distinguish specific patterns in the distributions of the gene loci along the linear sequences. (Figures 4.20 – 4.22).

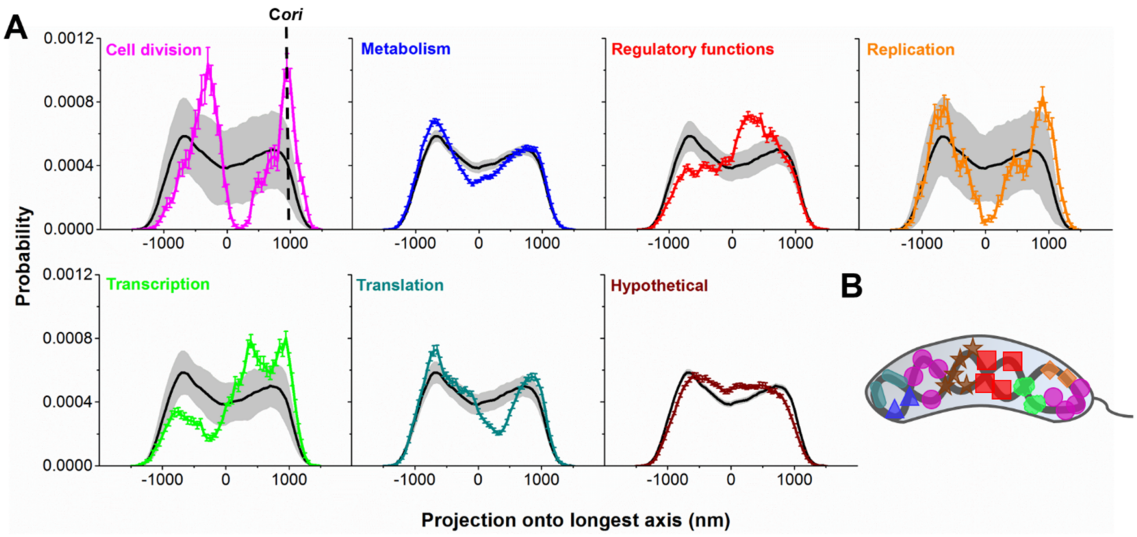

Figure 4. 12 Projections of functionally related genes onto the 3D nucleoid structures. (**A**) Localization of functionally related genes compared to the distribution of all the genes (black) with the standard deviations in gray as calculated in Figure 8. The error bars indicate the standard errors based on the ensemble averages. (**B**) Schematic representation for the proposed functional localization of genes in C. crescentus. Colors are as in (A).

We repeated the mapping analysis for the central hub clusters in Figure 4.7B (clusters 2, 7, 8, 9, 12, 16, 17, 18, and 21) as well as all of the individual clusters to see if there is any difference in the localization of the genomic elements by nucleoid structure and/or connectivity. However, we did not find significantly different results from the results for the overall ensemble (Figure 4.23 – 4.25, *SI*). This suggests that despite differences in the nucleoid structure, the overall genome localization is largely preserved.

4.5 Discussion

In this study, we developed an advanced multi-scale modeling method that is based on data extracted from 3C-based experiments for generating an ensemble of 3D structures of a bacterial chromosome at high-resolution. The key idea is to use information about the plectonemic and

supercoiled organization of bacterial chromosome in combination with the specific experimental restraints that allowed us to generate higher-resolution models compared to models that only use the experimental data. This strategy is similar to the refinement of proteins via NMR where intermediate resolution restraints from the experiment are combined with the knowledge that all proteins are polypeptide chains with certain limited topologies to generate atomistic models. We applied this approach to the *C. crescentus* chromosome where extensive data sets from Hi-C experiments are available [211]. Previously, low-resolution models at 13-kb and 434-bp were constructed for *C. crescentus* chromosome using the 5C [214] and Hi-C [211] experimental data, respectively. The 434-bp models generated by Le et al. [211] have been the starting point for understanding the organization of genomic DNA based on the Hi-C based experimental data. Here, we further extend this work by developing models at higher resolution that includes DNA supercoiling. We are primarily discussing models with 15-bp resolution, slightly more than one helical turn, but because the DNA structure is fairly stiff on such short length scales, we were also able to generate base-pair resolution models with a reasonable degree of accuracy to the extent that the 15-bp resolution model is accurate.

Another important aspect of our modeling involves the application of experimental restraints by considering the population-based nature of 3C-based experiments. In the first step of modeling, we used soft distance restraints to bring a random polymer model to an average possible structure from where it can deviate to other alternative structures. After the initial generation of plectonemic models and the reconstruction of 15 bp CG models, all the simulations were performed without applying any restraints which allowed models to have fluctuations in their local topologies. Further reweighting of the generated models with significant structural variability enabled obtaining an overall ensemble of chromosome structures that possibly reflects the possible cell-to-cell structural

variability observed in 3C-based experiments. Our modeling protocol resulted in 27 different model clusters with different weights. We found that the average distance map of the reweighted ensemble is in very good agreement with the distance map derived from the experiment, on the other hand, individual clusters do not show that much consistency indicating that we were able to generate an ensemble that contains structural variability but also is compatible to the experimental distances in average. In this aspect, our modeling protocol which resulted in structurally variable conformations is different than previous studies on C. crescentus chromosome which focused more on generating a single consensus chromosome structure with small variations [211, 214].

Investigation of the physical properties of the chromosome models showed that our models are compact enough to fit within the cell envelope of *C. crescentus* but at the same time they are porous enough for proteins as large as RNAPs and ribosome subunits to penetrate and diffuse through the nucleoid as previously reported [244, 249, 251]. In addition, we explored if the models are compatible with the experimental data reported for *C. crescentus* that was not used while building the models [245]. The distances calculated from the models showed agreement with the experimental distances for the given DNA segment pairs but with some deviations for distances larger than ~1000 nm. We argued that this deviation might result from the fact that the Hi-C experiment were performed on cells at different cell phases [211] whereas FISH data was collected only from swarmer cells that were stopped from cell growth [200, 245]. However, if we scale our models and reanalyze their compatibility to the FISH distances showed that the scaled models still agree with the experimental data and preserve their global topologies. This suggests that there is significant room for the chromosome to expand and shrink as the cell size varies during the cell cycle.

Our 3D models are similar to previous shape outlines for the *C. crescentus* genome derived based on the 5C data [214] but we also see significant differences in topology. We found that there is different degree of arm-twisting and bending in different clusters. Some of the structures exhibit extreme bending that is not fully compatible with previously reported data for *C. crescentus* [246]. These models may be artifacts of our modeling procedure, however as the modeling suggests relatively low weights, these structures could be rare outliers that are difficult to detect experimentally if they occurred.

We found significant structural variability in our ensemble. This is expected and interpreted generally as cell-to-cell variation of chromosome structure. However, our analysis suggests that different structures may interconvert to a significant degree within the same cell. In fact, it appears that all of the clusters we observe are inter-convertible to each other either directly or through intermediate structures. This is an intriguing finding, although such remodeling of chromosome structures would require cooperation and activity by NAPs which was not considered here. It is clear that there is much room for further studies and the nucleoid structures presented here are an ideal starting point to further examine nucleoid dynamics via simulations.

The high-resolution models enabled mapping the genomic sequence onto the 3D structure. This allowed us to analyze the distribution of generic sequence elements as well as specific genes. We found non-random distributions of AT-rich sequences and other NAP-binding motifs that hint at the role of NAPs in stabilizing the nucleoid structure. We also observed an apparent spatial organization of genes involved in certain functions and we hope that these results will further motivate new research. Experimental tests for a proposed non-random distribution of genes could involve gene or operon randomizations with the hypothesis that randomly distributed genes could affect metabolic efficiency and/or regulatory processes.

High-resolution nucleoid models were also recently generated by Hacker *et al*. for the *E. coli* chromosome [242]. They generated chromosome models at one-nucleotide resolution for the first time. Here we apply similar approach to generate C. crescentus chromosome initially at 15-bp resolution but then extended the resolution to the one-base pair level. One major difference between the modeling protocols by Hacker et al. and us is that our models rely primarily on Hi-C based contacts, which we took into account via an ensemble approach that considers the nature of the measured crosslinking contacts. In the work by Hacker et al., the modeling protocol is guided by experimentally found 2D radial distributions of RNAP to generate models that reproduce these experimental distributions [244]. This approach would be expected to give nucleoid models that are physically realistic, but since RNAP radial distributions are projected onto 2D, the specific positions of the genomic loci in 3D space might not be as well defined as there were no constraints guiding those positions, or their relative distances, in the modeling protocol. On the other hand, E. coli chromosome models include plectoneme-free (relaxed DNA) and plectoneme-abundant (supercoiled DNA) regions based on the ChIP-chip data on RNAP for E. coli [243]. In our C. crescentus chromosome models, we did not specifically identify supercoiled-free regions. However, we assume that some parts may be unwound for active transcription however and is sufficient space for supercoils for unwinding. Because only modeled the segments lying between plectoneme rosettes that form the central ring of the model as single double stranded DNA strand, it is possible that our model is topologically less accurate. However, despite the significant differences in the modeling protocols, it is exciting that the overall structural and mapping analyses showed consistent conclusions in both C. crescentus and E. coli chromosomes.

Our methodology is easily applicable to other bacteria as additional 3C-based data set are becoming available and expand the ability to study the direct connection between the 3D

distribution of genes and their function that should be part of a complete analysis of genotype-phenotype relationships.

4.6 Acknowledgements

The authors would like to thank Dr. Aaron Dinner for sharing experimentally observed bending angles for the *C. crescentus* cell shape and Dr. Aleksei Aksimentiev for sharing potential mean force data for DNA-DNA interactions at different ion concentrations. Authors would also like to thank Dr. Liang Fang for initial contributions in developing the modeling protocol. Computer resources were used at XSEDE facilities [TG-MCB090003] and at Michigan State University's High-Performance Computing Center. This work was supported by the National Institutes of Health [GM092949]; and the National Science Foundation [MCB 1330560].

4.7 Supplementary Information

4.7.1 Hi-C interaction frequencies and conversion to distances

Normalized Hi-C interaction frequency matrices available from Le *et al.* [211] were first converted to z-scores to distinguish frequent interactions from infrequent ones. Contact frequencies that have high z-scores correspond to pairs which interact frequently, while low z-scores indicate infrequent or no interactions. To generate z-scores, the mean and the standard deviation of the frequencies were first calculated row-by-row from the matrix. Z-scores were then obtained for each element in a row by using the corresponding mean and standard deviations of the row in which that element resides. Z-scores were converted to spatial distances by using the calibration curve derived for the *C. crescentus* genome by Umbarger *et al*. [214]. This resulted in a set of distances between *loci* pairs of the genome with an effective resolution of about 10 kb.

4.7.2 Plectonemic Model

Models for the *C. crescentus* chromosomal DNA were based on a hyperbranched polymer chain consisting of ring beads forming a central ring and branch beads extending from the central ring to mimic the rosette model of DNA (Figure 4.26). Each ring bead is connected to two neighbour ring beads and one branch bead, and each branch bead is connected to either one, two or three other branch beads, forming a hyperbranched structure (Figure 4.26). Each ring bead along with all of the attached branch beads represent one microdomain. The number of ring beads (and therefore microdomains) was varied between 200 and 700 as the number of microdomains in a bacterial genome with a 4 Mbp-size is believed to be in that range [263]. The segments between ring beads represent double stranded DNA and connect individual microdomains, whereas the segments between branch beads represent supercoiled DNA (Figure 4.26). NAPs are presumed to bind at the junctions in order to assist microdomain formation and stabilize the plectonemic structure but such proteins were not included in the model. The length of the connectors between individual microdomains was set to vary around ~20 nm, following a previous polymer model of the *C. crescentus* chromosome [211] while providing sufficient space for the binding of NAPs that cover on average about 25-bp (~ 8.5 nm). The length of the segments between branch beads essentially reflects the persistence length of supercoiled DNA and was set to 40 – 100 nm in our models (corresponding to 270 – 670 base pairs per branch segment).

The radii of the ring beads and the cylindrical segments between them were set to 1 nm, corresponding to the radius of double-stranded DNA. The radii of the branch beads, and the segments connecting those, depend on the geometry of supercoiled DNA. The radius of supercoiled DNA depends on a number of factors such as the superhelical density (σ) and salt concentration. The radius of the supercoiled DNA increases with reduced superhelical density

(smaller magnitude) as observed by electron microscopy [264] and found in theoretical calculations [265]. DNA extracted from cells has a superhelical density in a range of -0.03 – -0.09 [266]. However, when considering the effect of NAPs on DNA, σ is believed to be at the lower end in magnitude, around -0.025 [267]. When σ was set to a small value to include the effect of NAPs, our plectonemic models did not fit into the known dimensions of *C. crescentus* cells with the corresponding superhelical radius. Therefore, we assumed an average value of σ =-0.06, corresponding to a 5 nm supercoiled DNA radius at 0.1 M monovalent salt concentration [265]. In our models, we set the radius to a slightly larger value (6.5 nm) to account for proteins that may be bound to the DNA such as H-NS, HU or IHF which have equivalent spherical radii of about between 1.5 – 1.8 nm [268].

Bond and angle restraints: Neighbouring beads were connected via a harmonic bond potential:

$$E_{branch-bond} = k_{branch-bond} \times (l - l_b)^2 \quad (4.1)$$

$$E_{ring-bond} = k_{ring-bond} \times (l - l_r)^2 \quad (4.2)$$

where l is the connecting segment length, l_b is the equilibrium branch segment length set to 40-100 nm (l_r is the ring segment length set to 20 nm) and $k_{branch\text{-}bond}$ is the force constant (and $k_{ring\text{-}bond}$ is the force constant in the ring-bond potential), set to values of $k_{branch\text{-}bond}$= 0.05 kcal/mol/nm^2 and $k_{ring\text{-}bond}$=1.0 kcal/mol/nm^2, respectively. These values kept the deviations of the segment lengths to less than 10%. Deviations of 10% limit the pitch angle to fluctuate between 50° – 70° as reported in previous studies [264, 269]. Furthermore, the limited segment lengths and corresponding pitch angle variations in the supercoiled DNA allow the reconstruction of the higher-resolution DNA representation assuming a fixed number of bases per segment.

The stiffness of the chain was maintained by a one-sided harmonic angle potential that prevents unfavorable sharp angles between segments:

$$E_{angle} = k_{angle} \times (\theta - \theta_0)^2 \qquad if\ \theta < 90^{\circ} \tag{4.3}$$

where $\boldsymbol{\theta}$ is the angle between successive segments between beads, $\boldsymbol{\theta_0}$ is the equilibrium angle set to 90° and k_{angle} is the force constant set to 0.05 kcal/mol/degrees^2 to keep angles between segments larger than 90°.

The chosen bond force constants set the overall energetic scale of the plectonemic model and the following potentials were adjusted to lead to a balanced distribution of potential energies while still maintaining 10% fluctuations in the segment lengths with the full potential.

Segment overlap: To keep supercoiled DNA segments separated, a restraint potential was applied to beads and to the segments connecting two beads. In all-atom molecular dynamics (MD) simulations, the distance between two supercoiled DNA segments was observed to fluctuate between 2.5 – 5 nm in 0.1 M NaCl solution (the same ion concentration as considered here) [270]. Therefore, the closest contact between two segments was penalized with a one-sided harmonic function:

$$E_{overlap} = k_{overlap} \times (d - d_0)^2 \qquad if\ d < d_0 \tag{4.4}$$

where d is the minimum distance between two beads and/or segments with d_0 set to 2 nm and $k_{overlap}$ is the force constant set to 1 kcal/mol/nm^2 to prevent segment-segment distances of less than 2 nm. The minimum distance between segments was calculated as follows:

First two lines which the segments i and j lie on were defined using parametric equations:

$$L_i = P_{i_0} + s\left(P_{i_1} - P_{i_0}\right) = P_{i_0} + s\mathbf{u} \tag{4.5}$$

$$L_j = P_{j_0} + t\left(P_{j_1} - P_{j_0}\right) = P_{j_0} + t\mathbf{v} \tag{4.6}$$

where P_{i0} and P_{j0} are the starting points and P_{i1} and P_{j1} are the end points (the starting points of segment $i+1$ and $j+1$) of segments i and j, and s and t are real numbers. These equations also define

the segments i and j with $0 \leq s \leq 1$ and $0 \leq t \leq 1$. In addition, a vector was defined between two points in L_i and L_j:

$$\mathbf{w}(s,t) = L_i(s) - L_j(t) \tag{4.7}$$

The shortest length of the vector **w** gives the minimum distance between the two lines. The vector **w** is at its minimum length when it is perpendicular to both **u** and **v** vectors. Therefore, s_c and t_c values which give the minimum distance were found by satisfying two equations:

$$\mathbf{u}.\mathbf{w} = 0 \tag{4.8}$$

$$\mathbf{v}.\mathbf{w} = 0 \tag{4.9}$$

If s_c and t_c were between 0 and 1, then the length of **w** also gave the minimum distance between the segments i and j. However, if any or both of the values were outside the range, then additional steps were taken in order to obtain new s_c and t_c values within 0 and 1. If s_c<0 or s_c>1, it was set to 0 or 1, respectively. Then, the following quadratic equation was used in order to find t_c:

$$|\mathbf{w}.\mathbf{w}| = (P_{i_0} - P_{j_0} + s_c\mathbf{u} - t_c\mathbf{v}).(P_{i_0} - P_{j_0} + s_c\mathbf{u} - t_c\mathbf{v}) \tag{4.10}$$

This equation is also minimized when the length of **w** is minimum. Hence, taking the derivative of Eq. (10) with respect to t and setting it to 0 gave the value for t_c since s_c is known. If $0 \leq t_c \leq 1$, these s_c and t_c values resulted in the minimum length of **w**. If not, this time t_c was set to 0 or 1 and s_c was recomputed by taking the derivative of Eq. (10) with respect to s. In case both s_c and t_c values were outside of the range, then minimum distance between the segments were calculated using either starting (s_c=0 or t_c=0) or end points (s_c=1 or t_c=1) of the segments depending on the side of the range s_c and t_c lie on.

Spatial constraints: All models were confined in a sphere with a 4 μm diameter to limit sampling to more compact conformations. A one-sided harmonic potential was used:

$$E_{confinement} = k_{confinement} \times (r - r_0)^2 \qquad if\ r > r_0 \tag{4.11}$$

where r is the distance of the bead from origin, r_0 is the radius of the confined volume set to 2 μm and $k_{confinement}$ is the force constant. $k_{confinement}$ was set to 0.0001 kcal/mol/μm^2 that kept all beads within the confinement sphere. This confinement sphere is significantly larger than the cell dimensions as *C. crescentus* cells have a length of ~2.5 – 3 μm and a width of ~0.8 μm[246]. A spherical confinement was used instead of cylindrical confinement since we did not want to presume the overall shape of the *C. crescentus* chromosome *a priori*.

Polymer branching: The degree of branching was chosen based on an average number of 2.94±1.2 branching points observed for 7 kb plasmids [264]. This value would correspond to ~1700 branching points for the 4 Mb length of the *C. crescentus* genome. The number of total branch segments in our models are around ~10,000, but varying due to different numbers of microdomains and the length of segments between branch beads in different models. 1,700 branch points would correspond to a percentage of about 20% for models with ~10,000 branch beads, therefore, we constrained our models for around 20% of the branch beads to be three-way branching points.

The degree of branching in the model was restrained via a harmonic potential;

$$E_{branching} = k_{branching} \times (brp - brp_0)^2 \qquad (4.12)$$

where *brp* is the branching percentage, brp_0 is the expected branching percentage, set to 20% and $k_{branching}$ is the force constant which was set to 10^6 kcal/mol/brp^2. The large force constant value was necessary to obtain the targeted branching percentage (Figure 4.27). The branching percentage was defined as;

$$brp = \frac{number\ of\ branch\ beads\ with\ 3\ connections}{number\ of\ all\ branch\ beads} \times 100 \qquad (4.13)$$

Restraints based on Hi-C contacts: Hi-C contact frequency z-scores larger than 0.75 were considered significant and pairs with corresponding distances of less than ~485 nm were restrained with a harmonic potential during the initial model generation [214]:

$$H(d) = k_h \times (d - d_0)^2 \quad (4.14)$$

where d is the distance between the pair, d_0 is the distance derived from the Hi-C contact frequencies and k_h is the harmonic force constant set to 0.005 kcal/mol/nm^2 which amounts to a relatively weak guiding force so that the generated models are generally consistent with the Hi-C derived distances but retain significant conformational heterogeneity (Figure 4.28).

Z-scores smaller than -0.85 were also considered significant to indicate a complete lack of interaction [214]. *Loci* pairs with such z-scores were, therefore, restrained to remain apart with a lower-bound potential;

$$L(d) = \begin{cases} 0 & if\ d \geq d_l \\ k_l \times (d - d_l)^2 & else \end{cases} \quad (4.15)$$

where d_l was set to 925 nm, which is the distance beyond which interaction counts are not distinguishable from noise anymore. The lower-bound force constant, k_l, was also set to 0.005 kcal/mol/nm^2. Using z-score thresholds to decide which contact frequencies will be used for restraints resulted in 17,412 harmonic and 15,631 lower-bound restraints.

Model generation: The generation of plectonemic models was carried out by running a custom-written program that sampled plectonemic models using a simulated annealing Monte Carlo (MC) algorithm. Random initial conformations were initially subjected to 2 x 10^6 MC steps with a move set that consisted of either translation of ring and branch beads (attempted 9 out of 10 steps) or the reconnection of branch beads to build one, two, or three connections with other branch beads (attempted 1 out or 10 steps). During the initial MC run, all of the potentials described above except the segment overlap potential were applied so that segment crossing would be possible to

achieve rapid equilibration. The initial temperature was set to 1,500,000 K and gradually decreased to 0 K by reducing the temperature 0.00001 % at each step. This cooling schedule resulted in ~85% initial acceptance rates. The configurations after the initial MC runs established the overall connectivity and were then equilibrated further with MD simulations using the CHARMM program package (c41a2) [142] for 5 ns with a 0.002 ps time step by applying the same potentials used as in the initial MC run (except for the branching potential since the overall connectivity could not change during MD). Further equilibration with MD was carried out to accelerate convergence. The temperature of the system was set to 298 K controlled by a Langevin thermostat with a friction coefficient of 5 ps^{-1}. Lastly, the structures after the MD run were subjected to another round of 2 x 10^6 MC steps, in which the overlap potential was applied together with all the other potentials (again except the branching potential in order to maintain the established overall connectivity) to remove and prevent crossings of the beads and the connecting segments. In the second MC run, the temperature was reduced by 0.0001 % from 150,000 K to 0 K resulting in initial acceptance rates of ~70%. The plectonemic models converged after 1 x 10^6 steps in MC runs, and after 2 ns in the MD step. (Figure 4.29).

The overall MC/MD/MC procedure required 24 hours of runtime to generate one model with 8 cores on recent Intel CPUs and was repeated 25 times for each set of number of microdomains and branch segment lengths with different initial random conformations. Six different numbers of microdomains (200, 300, 400, 500, 600, and 700) and seven different branch segment lengths (40 nm, 50 nm, 60 nm, 70 nm, 80 nm, 90 nm and 100 nm) were explored. Thus 1,050 simulations were carried out in total resulting in 1,050 models.

4.7.3 15-bp coarse-grained model

In this level, *C. crescentus* chromosomal DNA was modeled as a closed worm-like chain of segments with an equilibrium length of 5 nm each representing 15 base pairs and a radius of 1 nm, the width of double-stranded following previous studies by Chirico *et al.* [93], Jian *et al.* [94] and Huang *et al.* [271]. The total energy of the chain consists of the following five terms:

Stretching energy was computed as:

$$E_{stretching} = \sum_{i=1}^{N} \frac{\sigma}{2l_0} (l_i - l_0)^2 \quad (4.16)$$

where N is the number of segments, l_i is the segment length, l_0 is the equilibrium length of the segment set to 5 nm, and σ is the stretching modulus of DNA which was set to 1000 pN [272, 273].

Bending energy was expressed as:

$$E_{bending} = \sum_{i=1}^{N} \frac{g}{2} \theta_i^2 \quad (4.17)$$

where N is the number of segments, θ_i is the angular displacement of segment i relative to segment i+1. g is the bending rigidity constant, set to 9.82 k_BT, corresponding to 20 rigid segments per 100 nm Kuhn length (since our segment length is 5 nm) as estimated by Frank-Kamanetskii *et al.* [274].

Torsional energy was specified as:

$$E_{torsion} = \sum_{i=1}^{N} \frac{C}{2l_0} \left(\Phi_{i,i+1} - \Phi_0\right)^2 \quad (4.18)$$

where N is the number of segments, l_0 is the equilibrium segment length set to 5 nm, $\Phi_{i,i+1}$ is the torsion angle between segments i and i+1, Φ_0 is the equilibrium torsion angle set to 0° [93] and C is the torsional rigidity constant of DNA set to 43.17 kcal*nm/mol [275]. In order to calculate the torsion angle, a body-fixed coordinate system (f_i, v_i, u_i) was defined for each segment and Euler

angles (α, β, γ) were used to describe the torsional angle between segment i and segment i+1: $\Phi_{i,i+1} = \alpha_{i,i+1} + \gamma_{i,i+1}$.

Electrostatic energy was computed according to the Debye-Hückel (DH) potential:

$$E_{electrostatic} = \frac{v^2 {l_0}^2}{Dm^2} \sum_{i,j}^{Nm} \frac{\exp(-\kappa r_{ij})}{r_{ij}} \quad (4.19)$$

where N is the number of segments, l_0 is the equilibrium segment length, v is the linear charge density, D is the dielectric constant of water, κ^{-1} is the Debye length, and r_{ij} is the distance between sections i and j. The linear charge density, v, and the Debye length, κ^{-1}, were set to 6.08 e/nm and 0.961 nm for 0.1 M salt concentration, respectively [276, 277]. Each segment is divided into m sections. The value of m was set to 2 in our simulations since no change in electrostatic interactions was observed in the simulations when m≥2 (Figure 4.30). Electrostatic interactions between DNA segments in our models calculated from the DH potential were also compared with the interaction potential of DNA segments calculated from all-atom simulations [278] and the DH potential was found to give comparable electrostatic interactions between DNA segments (Figure 4.30).

Segment overlap energy was specified using a soft-core potential [279] of a modified Lennard-Jones potential that only has a repulsive term:

$$E_{volume-exclusion} = \lambda\varepsilon \frac{1}{\left[\alpha(1-\lambda)^2 + \left(\frac{r_{ij}}{r_0}\right)^6\right]^2} \quad (4.20)$$

where λ is a scaling factor between 0 and 1 (λ=1 means a standard Lennard-Jones potential), ε is the depth of the potential well, r_{ij} is the minimum distance between segments i and j, r_0 is the distance at which the potential reaches its minimum set to 2 nm, and α is a positive constant which was set to 0.5. We chose λ=0.8, ε=0.5 kcal/mol as sufficient values to avoid passing segments through another during the simulations.

Model generation: Supercoiled CG models were constructed with a custom-written program where the CG beads in the higher resolution model were wrapped around the segments in the plectonemic model. Wrapping of the higher resolution segments was performed by taking the pitch angle of the supercoiled DNA into account, which is around 61.5° at a superhelical density of σ=-0.06 [264, 265]. The exact same sequence mapping in the plectonemic models was also achieved in CG models by modifying pitch angles according to the observed fluctuations in the plectonemic segment lengths. The resulting CG models had 269,529 beads to represent the entire *C. crescentus* genome (for the 4,042,929 base pairs in the NA1000 variant). The average pitch angle from the CG models was found to be 62.4° with a standard deviation of 9.4°.

We carried out Langevin dynamics simulations for our CG models by considering both translational and torsional dynamics at 298 K with a friction coefficient of 10 ps^{-1}. Simulations were performed with the CHARMM program [142] version c41a2 which already has the equations for translational motion, however we had to implement the torsional motion equations by following the derivations in previous work [93]. Simulations were run for 200 ns with a 0.2 ps time step. Electrostatic and soft-core repulsive Lennard-Jones interactions were cut off at 20 nm with a switching function becoming effective at 25 nm and the non-bonded list was cut off at 30 nm. Convergence of the 15-bp CG models was achieved after 150 ns (Figure 4.29). The total time required to run 200 ns MD simulations for one model is ~48 hours with a single core on a modern Intel CPU.

4.7.4 Base-pair resolution models

Double-stranded DNA has a long persistence length of tens of nm. Therefore, it is possible to approximately reconstruct base-pair resolution from coarse-grained models. The C. crescentus nucleoid models are modeled at a resolution of 15 base pairs, which corresponds to a DNA segment

of about 5 nm length. The centripetal Catmull-Rom spline interpolation was applied to generate DNA at base-pair resolution along a smooth curve through the 15-bp coarse-grained beads. The method was tested on the DNA from crystal structures of the nucleosome core particle (see Figure 4.31). It can be seen that the actual base pair locations (green) are closely reproduced by the spline interpolation (red) between the 15-bp beads (orange), although this method cannot capture local distortions of the DNA structure (most pronounced at the upper right corner of the structure shown in Supplementary Figure 4.31). The overall RMSD between the true base pair centers of mass and the interpolated base pairs is 2.6 Å.

4.7.5 Model reweighting

Hi-C experiments provide ensemble-averaged contact frequencies, therefore converting frequencies to spatial distances and using these distances to generate chromosome structure models is not sufficient to generate a structurally variable ensemble of structures, this method rather results in a single consensus model. Recently, population-based modeling efforts have been made to generate an ensemble of structures by applying maximum entropy principle in combination with minimal structural assumptions. Here, we followed an alternative approach which is a combination of restraint and population based modeling methods. In contrast to many restraint-based modeling studies, here we applied relatively weak force for the distance restraints in our modeling rather than rigid forces to overcome the limitation of generating single chromosome models. The assumption we took here is that, chromosome structures in different cells have a similar topology in general terms, but vary significantly in local structural organizations. Our modeling was able to generate an ensemble of structures that are highly variable. To further agree with the experiments, we reweighted our models in order to obtain an ensemble of structures that, on average, matches the Hi-C data but explicitly considers cell-to-cell variability. Since the degree of structural

variability for each contact is not available in Hi-C data, the structural variability was approximated by estimating standard deviations for the experimentally-derived distances. In order to do that, first we estimated a function for how the likelihood of crosslink formation varies by distance (Figure 4.32):

$$Crosslink\ probability = \begin{cases} 1 - \dfrac{d \times (1-c)}{cutoff_1} & if\ d < cutoff_1 \\ c - \dfrac{c \times (d - cutoff_1)}{(cutoff_2 - cutoff_1)} & if\ cutoff_1 < d < cutoff_2 \\ 0 & if\ d > cutoff_2 \end{cases} \tag{4.21}$$

where d is the distance between *loci* pair, $cutoff_1$ and $cutoff_2$ are the distance thresholds and c is a positive constant. Next, we approximated standard deviations for distances by using the function below (Figure 4.32):

$$Standard\ deviation = \begin{cases} \sigma_0 + \dfrac{d \times (\sigma_1 - \sigma_0)}{scutoff_1} & if\ d < scutoff_1 \\ \sigma_1 + \dfrac{(d - scutoff_1) \times (\sigma_2 - \sigma_1)}{(scutoff_2 - scutoff_1)} & if\ scutoff_1 < d < scutoff_2 \\ \sigma_2 + (d - scutoff_2) \times s & if\ d > scutoff_2 \end{cases} \tag{4.22}$$

where d is the distance between *loci* pair, $scutoff_1$ and $scutoff_2$ are the distance thresholds, σ_0, σ_1 and σ_2 are initial standard deviations for each condition in the function and s is the scale factor.

For each experimentally-derived distance, we assumed a distribution of distances observed in the population with an average which is actually the experimentally-derived distance and a standard deviation estimated with Eq. (22). Then, we scaled the Gaussian distribution with the cross-linking probability function in Eq. (21), and from obtained scaled distributions, we

calculated new z-scores for each distance and tried to match the z-score/distance calibration curve used for converting z-scores of the Hi-C scores to distances (see *Conversion of Hi-C interaction frequencies to distance restraints* section) by simultaneously optimizing the variables in those functions. When we set $cutoff_1$, $cutoff_2$, and c to 100 nm, 380 nm and 0.54 in Eq. (21), respectively, and set $scutoff_1$, $scutoff_2$, σ_0, σ_1, σ_2 and s to 100 nm, 300 nm, 10 nm, 30 nm, 42 nm and 0.1 in Eq. (22), respectively, we were able to obtain a z-score/distance curve that is in an excellent agreement with the calibration curve used for converting z-scores of the Hi-C scores to distances (Figure 4.32). We note that, 380 nm threshold for crosslinking seems large to have a formaldehyde crosslinking, however it is still within possible distance range when one considers the large proteins that are interacting with the DNA such as RNAP or SMC complexes, the maximum length of a 10-kb DNA fragment and its dynamic nature. Compared to previous studies which set 1 and 0 for the contact probability below and beyond 200 nm, we here use a relatively more complex function where probability decreases with distance (it decreases to less than ~ 35 % probability beyond 200 nm).

After the assignment of standard deviations for each experimentally-derived average distance, we reweighted our models in order to match the average and the estimated standard deviation for each distance better. In other words, we aimed to obtain an ensemble that agrees with the experimental data better as well as retains the structural variability between models (Figure 4.32). The reweighting was carried out using a MC algorithm in a custom-written program. Initial weights of models were set to 1, and then subjected to 100,000 MC steps. In these MC steps, weight of a randomly selected model was altered by adding a random value between -0.2 – 0.2 and the average and standard deviation of distances were recalculated. If the total deviation of the averages and standard deviations of the distances from the experimentally-derived averages and

initially estimated standard deviations is lower, then new weights of the models were accepted. Ten independent MC runs were carried out. The final weight of each model was then calculated by taking the average of the ten different weights obtained from these ten independent runs. Final weights of all models were normalized so that the weights sum up to 1. All the subsequent structural and mapping analyses were carried out by taking the weights of the models into consideration.

4.7.6 Generation of contact maps

To generate contact maps, the segments in a model were first divided into 10-kb bins. This resulted in 405 bins, each having 666 segments. Next, the distances were calculated between all possible segment pairs in a bin (666 x 666) and the minimum distance was used in the crosslinking probability function described above to calculate the crosslinking probability for the given pair of bins. To generate distance contact maps, for each bin, the center of mass of the 666 segments were calculated. The center of mass points for each bin were used to calculate the distance between each pair of bins, resulting in 164,025 (405x405) distances for each model. The distances between each pair were then averaged over the models and those average distances were used to generate the contact map. Distances between the bins that are above 925 nm were not taken into account during averaging to match how the contact maps were generated based on the experimental data.

4.7.7 Identification of CIDs

The identification of chromosome interacting domains (CIDs) was carried out by following the same approach as in the study by Le et al. [211]. We investigated the crosslinking probabilities of a given bin with bins at its left- and right- hand side up to 100 kb and compared the crosslinking probability distributions of left- and hand-side by calculating t-values which is referred as

"directional preference" in the main text. Negative preference means that a locus is interacting with bins to the left (and positive preference corresponds to right-hand side preference). When there is a sharp change in the preference for a locus (from negative to positive or vice versa), a CID boundary is identified.

4.7.8 Structural Analyses

Structures were clustered with a K-means algorithm using a radius of 30 nm using the Multiscale Modeling Tools for Structural Biology (MMTSB) [145]. Interconvertibility was tested by using the TREK module [280] in CHARMM [142] which constructs smooth minimum energy paths between given start and end points. We used the same CG force field as for the final MD-based refinement step. Paths were generated between all pairs of clusters by using representative structures closest to the respective cluster centers. If a path was found between two states where the energy along the path did not exceed the energies of the initial or final states by more than 10 kcal/mol, those two states were considered interconvertible. Alternative thresholds of 2.5, 5 or 7.5 kcal/mol were also explored as well as shown in Figure 4.33 and no significant change was observed in the interconversion map of the clusters.

Cavity analysis was performed using *Voss Volume Voxelator* (3V) web server (http://3vee.molmovdb.org) [281] using solvent extractor option with a 20 nm outer probe radius. Hydrophobic radii of the macromolecules of interest were used as inner probe radius.

Macrodomain analysis was performed by clustering bp sites at 150-bp resolution for the representative models for each clusters using MMTSB [145] for hierarchical clustering with Xu's optimal clustering criterion [282] while generating sub-clusters only for clusters larger than 420 kb which is the largest reported macrodomain by Le *et al.* [211].

The overall shape of the models was analyzed as follows: First, all models were aligned to their principal axes so that the longest principal axis coincides with the x-axis, second longest principal axis coincides with the y-axis and the shortest principal axis corresponds to the z-axis. Next, the models were divided into 100 nm slices in the longest axis (x-axis) direction as shown in Figure 4.34. The center of masses of the segments were then calculated for each individual slice, and these center of mass points were joined with lines as shown in Figure 4.34. The total length of the lines gave the medial axis length of the models reported in Table 4.2. The widths of the models were estimated by finding the width around the medial line (red arrow in S Figure 4.34B) of each new slice defined by the lines passing through the two-consecutive center of mass points (black dashed lines in Figure 4.34B) where 95% of the segments were found inside. Bending angles of the models were found by calculating the angle between the vectors formed by the initial and midpoint of the medial axis and the midpoint and the end point of the medial axis (Figure 4.34). Bending directions were also calculated from those initial, mid and end points of the medial axis. First, a vector from the midpoint to the end point of the medial axis in the model ($\overrightarrow{BC}$ vector in Figure 4.34D), and another vector from the midpoint to the initial point of the axis were defined ($\overrightarrow{BA}$ vector in Figure 4.34D). Next, all models were aligned so that models' $\overrightarrow{BC}$ vectors as shown in Figure 4.34D fell onto the x-axis. After alignment, $\overrightarrow{BA}$ vector of each model then gave the bending direction in yz plane.

The medial axes constructed for the models were also used in calculating the radial distribution of the promoters. Radial distances were calculated as follows: First the point on the medial axis that is at a shortest distance from the promoter site was found. Then the projections of the promoter site and the point on the axis onto the shortest principal axis of the model were compared in order to figure out the side the promoter site resides relative to the axis. If the position of the promoter

site on the shortest axis is less positive than the point's position, then we assigned a negative sign to the calculated distance.

In order to calculate the degree of twisting of the chromosomal arms in the structures, we first reduced the resolution of the models to 50-kb which resulted in a ring with 81 segments that shows the overall twisting pattern of the chromosome (Figure 4.35). Then each bead in one arm was paired up with a bead in the opposite arm based on their indices (bead 1 – bead 81, bead 2 – bead 80, bead 3 – bead 79 and so on.). This step resulted in 40 bead pairs. The dihedral angles between each consecutive bead pairs gave the arm-twisting angles. The arm-crossing values were calculated from the 50-kb models based on the previously reported method by Klenin *et al.* (Method *2a* in the reference article) [283].

4.7.9 Sequence mapping onto 3D models

The *C. crescentus NA1000* genome sequence was obtained from the National Center Biotechnology Information (NCBI) website (Reference Sequence: NC_011916.1) and mapped onto the structures. AT/GC rich sites were defined as regions in the sequence where at least 70% of the base pairs is AT or GC within a 20-bp window using a custom-written program. Promoter sites in the sequence were obtained from the PromBase database [284]. Operons were found with DOOR[285]. IHF-binding motifs ((A/T)ATCAANNNNTT(A/G)) were obtained from a previous study [217]. Gene annotations and classifications of the genes were obtained from the Microbial Genome Database (MBGD) [286]. Gene annotations for the proteins which are found to be central and polar were obtained from previous work [257].

The statistical significance of projected gene distributions was estimated by comparing the distributions of selected *loci* with the variation in the distribution for the same number of *loci* that were selected randomly multiple times (200 times for projections and 10,000 times for the co-

expression/co-localization analysis). Since all models were reoriented so that the longest principal axis coincides with the x-axis, the positions of the segments of interest on the longest axis of the models were obtained from the x coordinates for the beads contained by the segments.

4.7.10 Graphics and visualization

Graphs were generated using Origin (OriginLab, Northampton, MA) and MATLAB R2016b (The MathWorks, Inc., Natick, MA). Network maps were generated with R (R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/). Models were visualized with PyMOL (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC) and animations were generated using VMD [147].

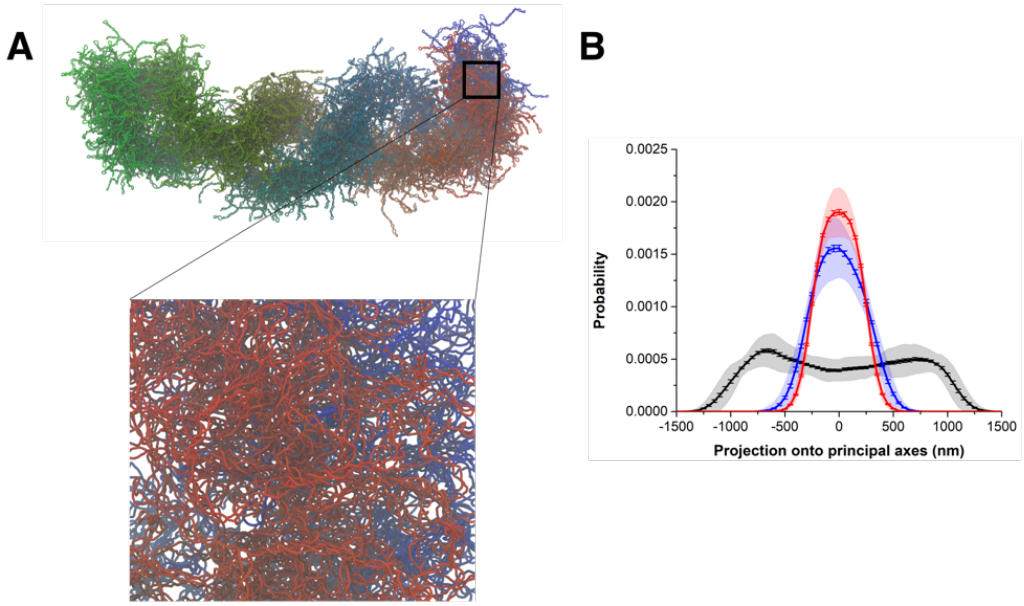

Figure 4. 13 Base-pair resolution models.

(A) 3D structure of *C. crescentus* chromosome at base-pair resolution. (B) Projections of beads in the bp-resolution models onto their longest principal axis (black), the second principal axis (blue), the shortest principal axis (red).

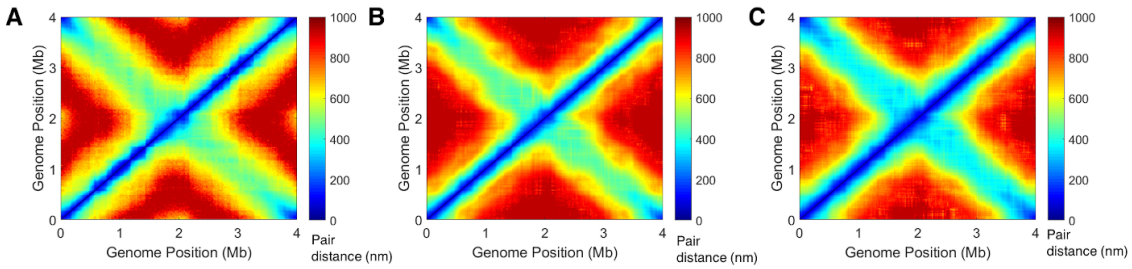

Figure 4. 14 Contact maps between loci for C. crescentus genome.

(**A**) Distance map generated based on spatial distances derived from Hi-C data by Le *et al*. (1) by using the calibration curve derived by Umbarger *et al*. (2) (**B**) Average distance map from the weighted ensemble of models (Pearson's correlation coefficient: 0.98). (**C**) Average distance map for the scaled models (Pearson's correlation coefficient: 0.96).

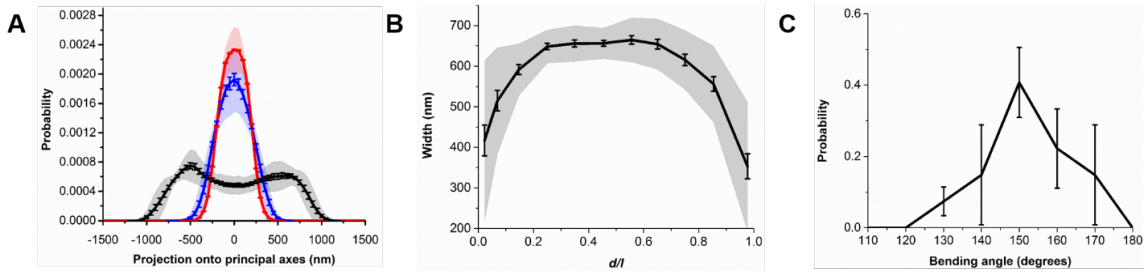

Figure 4. 15 Dimensions of the scaled models.

(**A**) Projections of beads in the models onto their longest principal axis (black), second longest axis (blue), and shortest axis (red). (**B**) Width of the models vs. the positions along the longest principal axis, *d*, normalized to the length of the models along the x-axis, *l*. Here, 1 represents the pole that contains the replication of origin while 0 represents the opposite pole. (**C**) Bending angle distribution of the models. Shaded area indicates standard deviations. Error bars indicate standard errors.

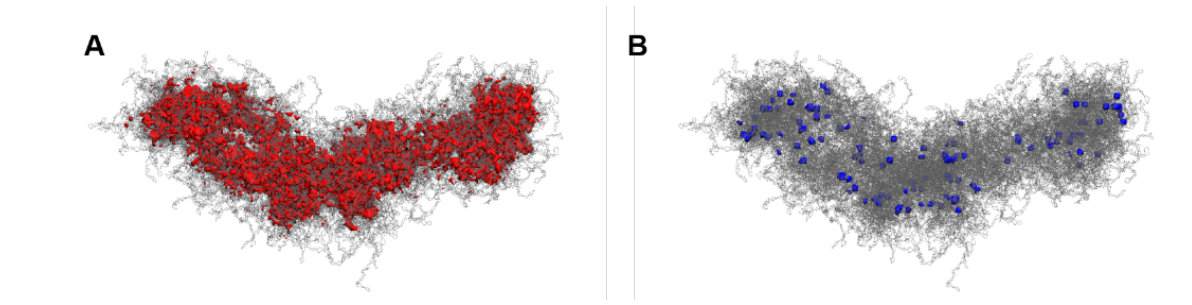

Figure 4. 16 Cavities in the scaled models.

Cavities that could be occupied by (**A**) RNAP and (**B**) ribosomes.

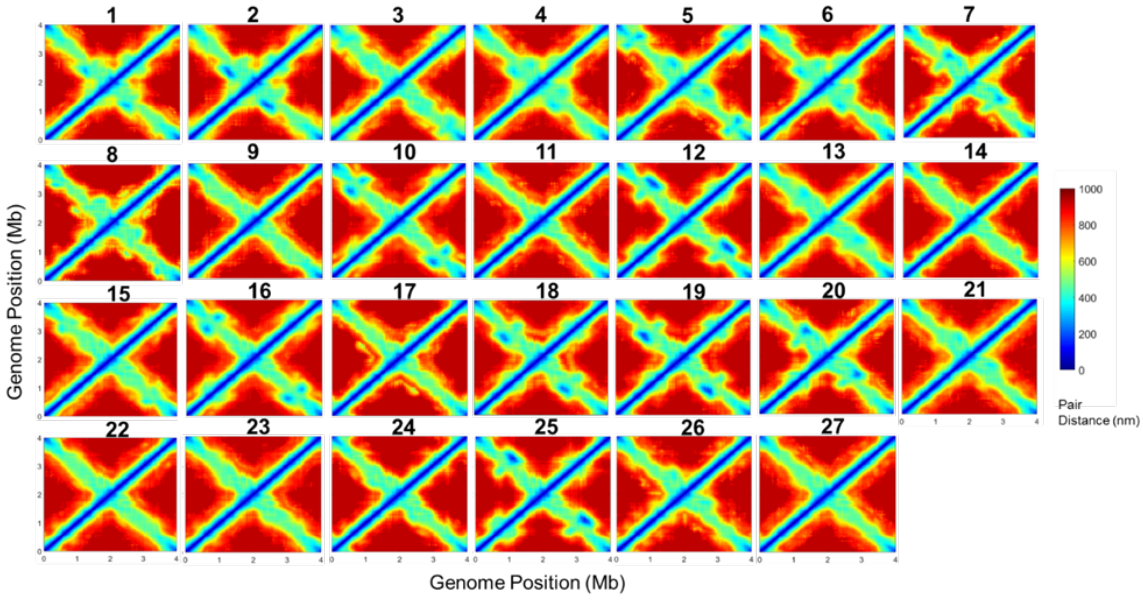

Figure 4. 17 Average distance maps for each 27 individual clusters.

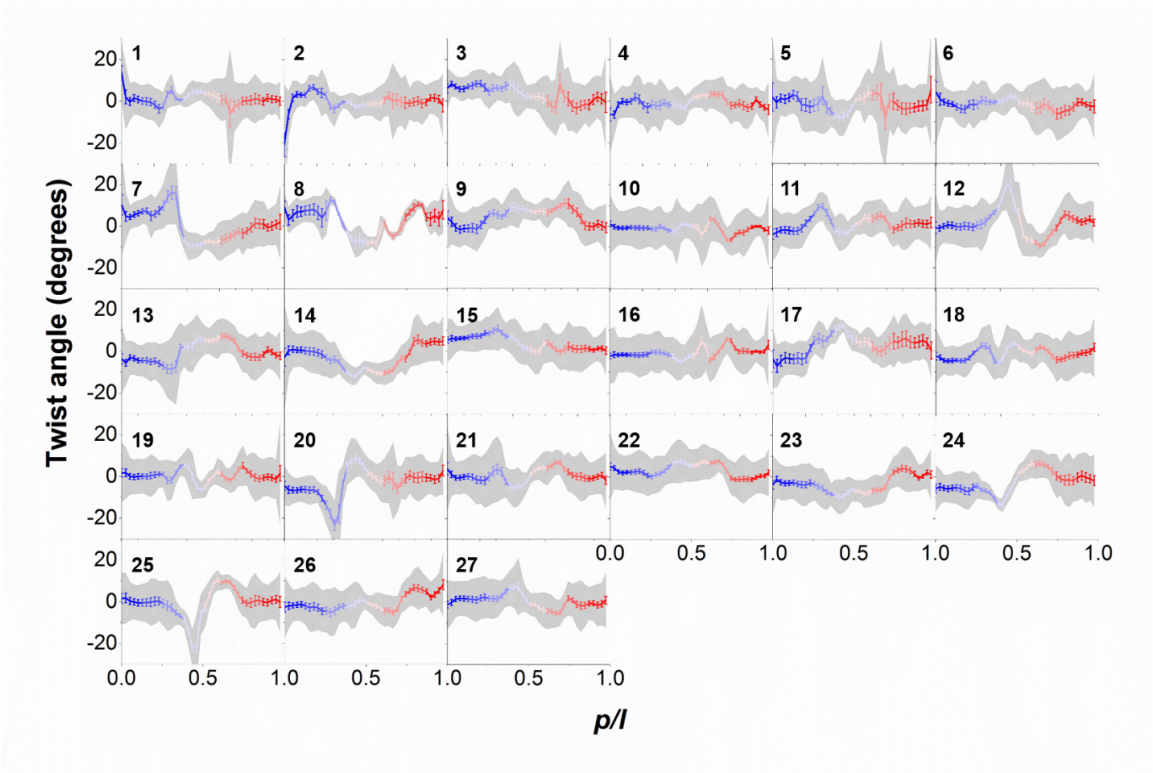

Figure 4. 18 Twist angles of the arms for all 27 clusters along the longest principal axis, p, normalized to the length of the models along the x-axis, l.

Here, 1 represents the pole that contains the replication of origin (red) while 0 represents the opposite pole (blue). Gray shadows and error bars indicate the standard deviations and standard errors, respectively.

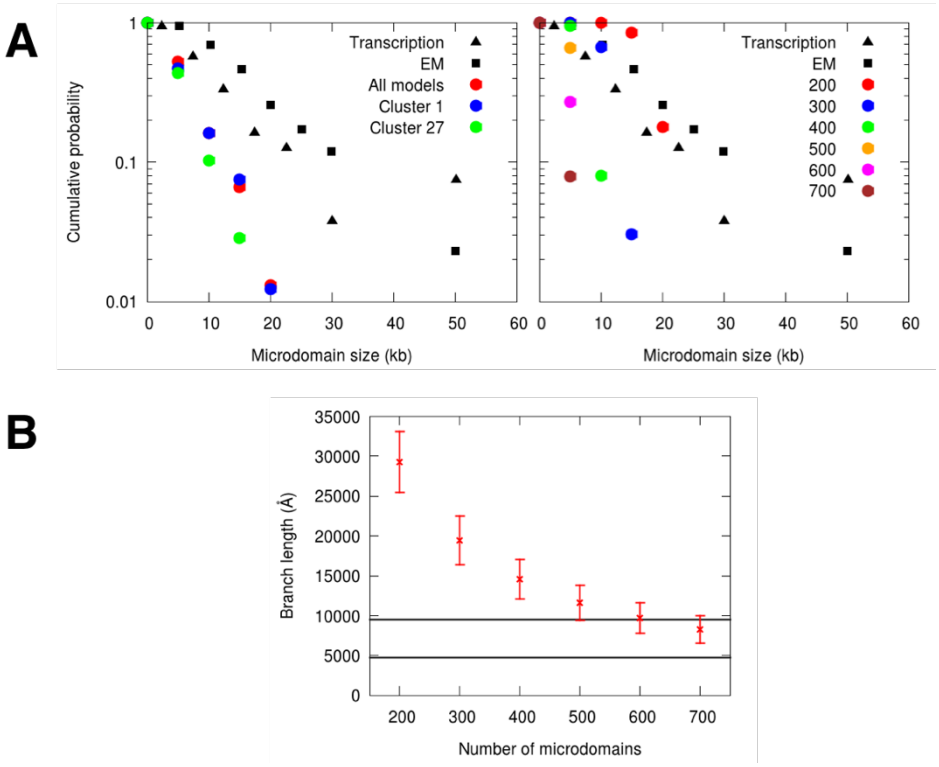

Figure 4. 19 Microdomain and branch length properties of the models.

(A) Microdomain size of the models in Cluster 1 and 27 (left panel), with different numbers of microdomains (right panel) in comparison with the domain sizes in *E. coli* chromosome measured by electron microscopy and transcriptional microarray experiments (39). (B) Average branch lengths of microdomains in the models with different number of microdomains. Upper and lower horizontal lines represent average branch lengths measured in 7-kb 3.5-kb plasmids, respectively (4).

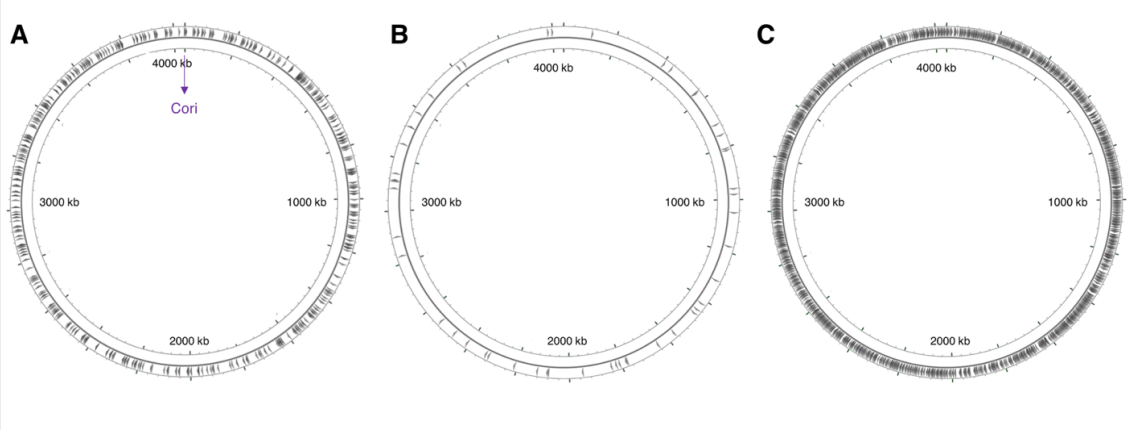

Figure 4. 20 Circular chromosome maps for genomic sequence features.

(A) AT-sites, (B) IHF-binding sites, (C) Promoter sites.

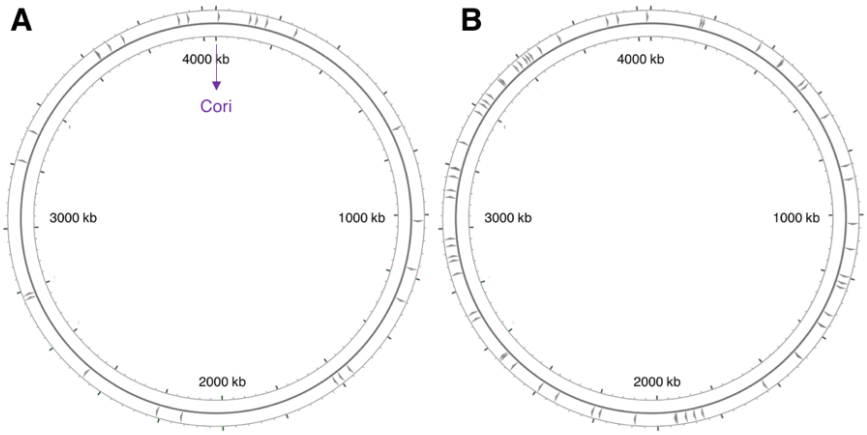

Figure 4. 21 Circular chromosome maps for genes for which protein products were found to be (A) central and (B) polar in previous work [256].

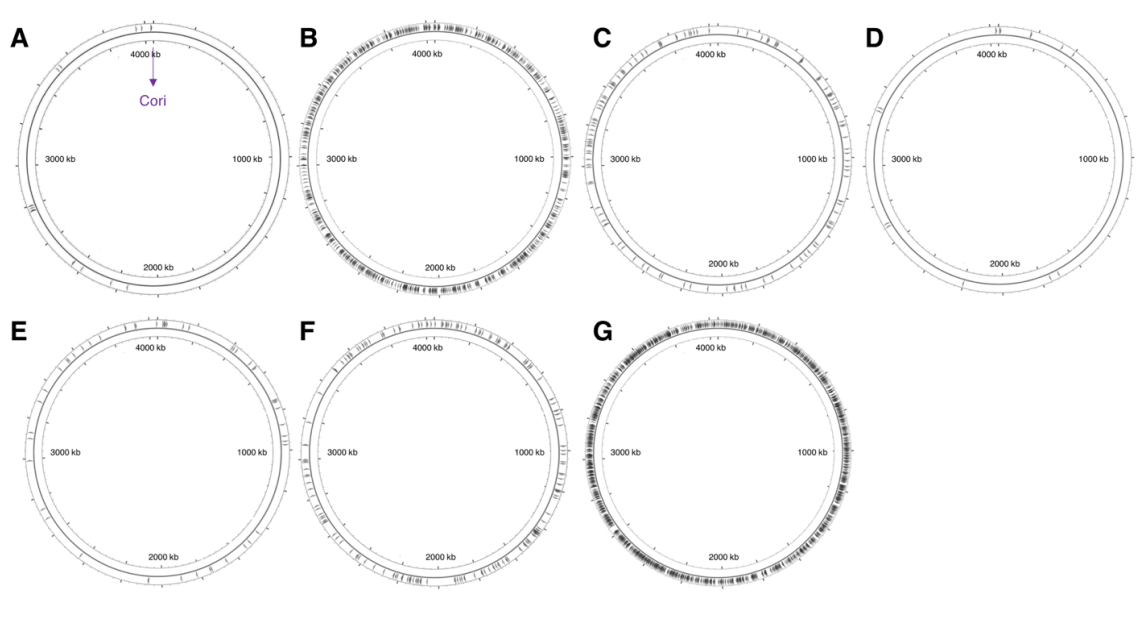

Figure 4. 22 Circular chromosome maps for functionally related genes.

(A) Cell division, (B) metabolism, (C) regulatory functions, (D) DNA replication, (E) transcription, (F) translation, (G) hypothetical genes.

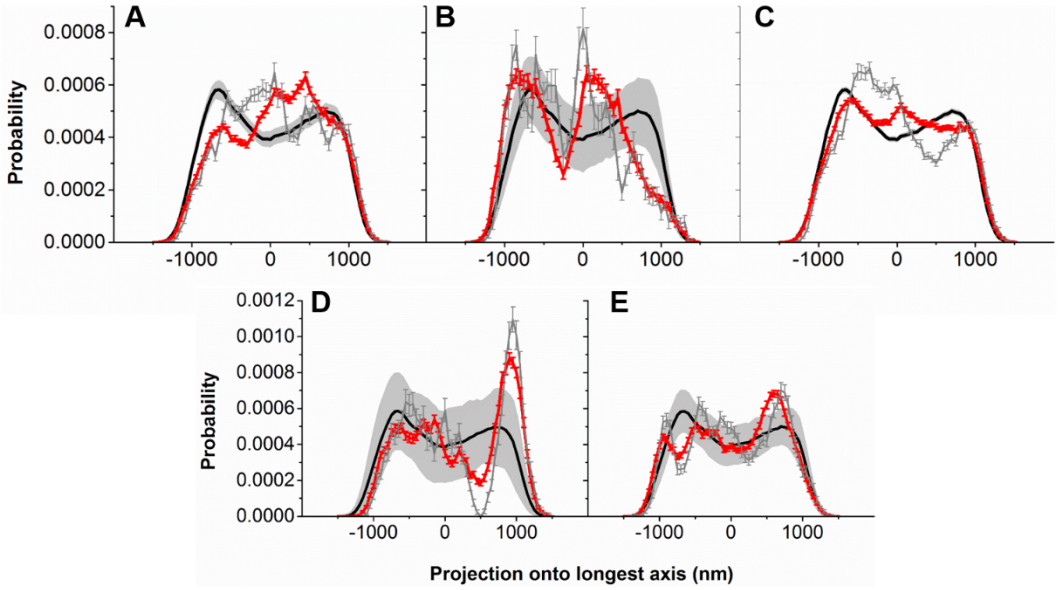

Figure 4. 23 Projections of genomic sequence features onto the 3D nucleoid structures for the average of central hub clusters (solid red), and cluster 18 only (gray).

Positions of all base pairs are shown with solid black line. (**A**) Sites with least 70% in AT-sites within a 20-bp window. (**B**) IHF-binding sites. (**C**) Promoter sites. (**D**) Distributions of the genes of the proteins which are experimentally found to be central and (**E**) polar in previous work (36). Gray shadows indicate standard deviations obtained from distributions for the same number of *loci* that were randomly selected 200 times. Error bars indicate the standard errors of the ensemble averages.

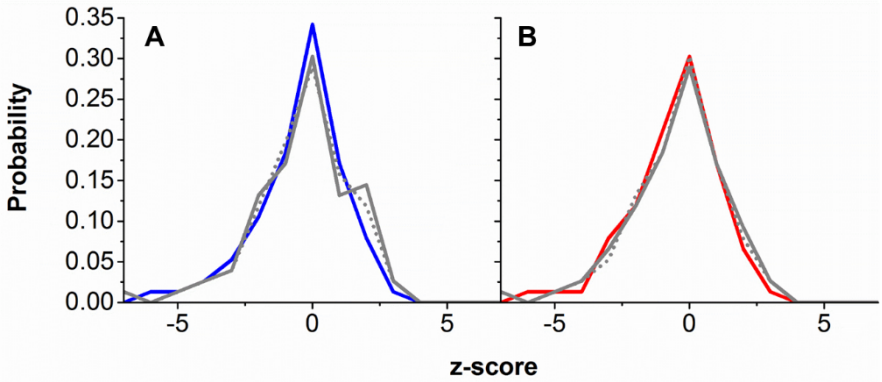

Figure 4. 24 Comparison of the correlation between co-localized and co-expressed genes for the average of all clusters (solid red or blue), average of central clusters (solid gray), cluster 18 only (dashed gray).

(**A**) Z-score distribution of intra-gene distances for co-expressed genes vs. the beginning of the corresponding operons (blue). (**B**) Z-score distribution of intra-gene distances for co-expressed genes vs. random genes for the end of genes (red).

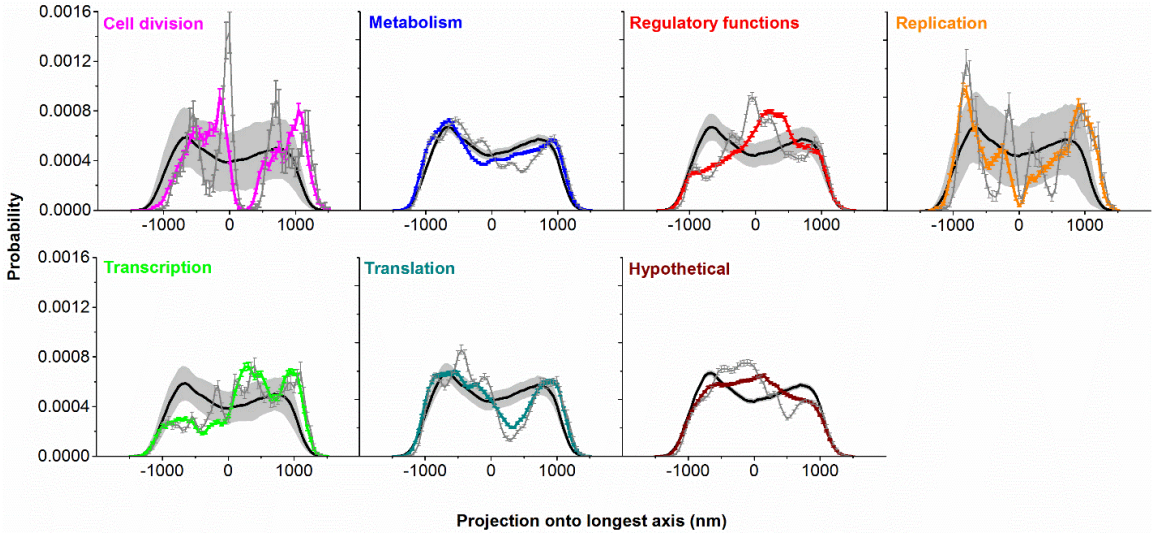

Figure 4. 25 Projections of functionally related genes onto the 3D nucleoid structures for the average of central clusters (solid colors), and cluster 18 only (gray).

Positions of all base pairs are shown with solid black line. Gray shadows indicate standard deviations obtained from distributions for the same number of *loci* that were randomly selected 200 times. Error bars indicate the standard errors of the ensemble averages.

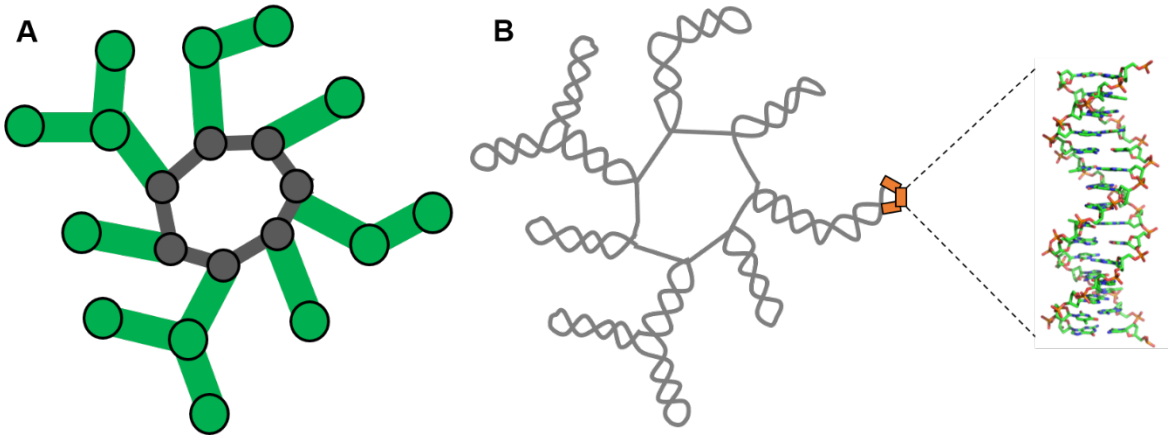

Figure 4. 26 Plectonemic and supercoiled model of bacterial chromosomal DNA.

(**A**) Plectonemic model with ring beads (gray) and branch beads (green). (**B**) Supercoiled model built on the plectonemic model. Each segment in the supercoiled model represents a 15-bp stretch of DNA (orange segments).

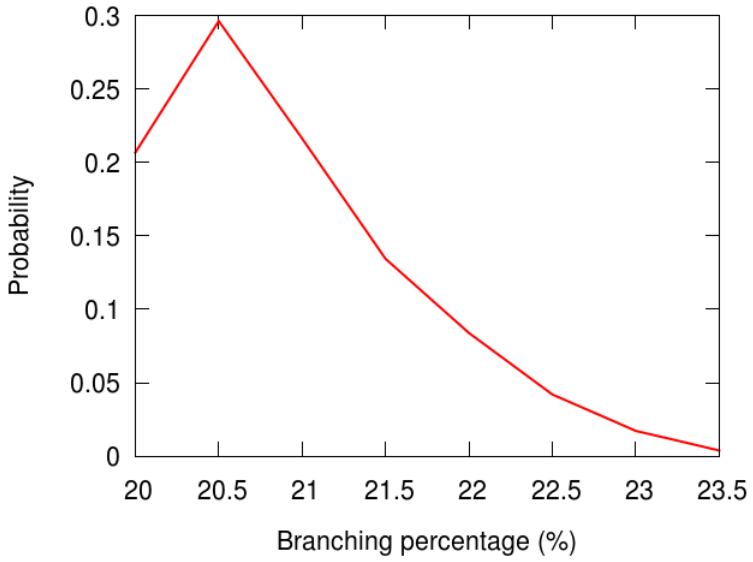

Figure 4. 27 Branching percentage distribution of plectonemic models.

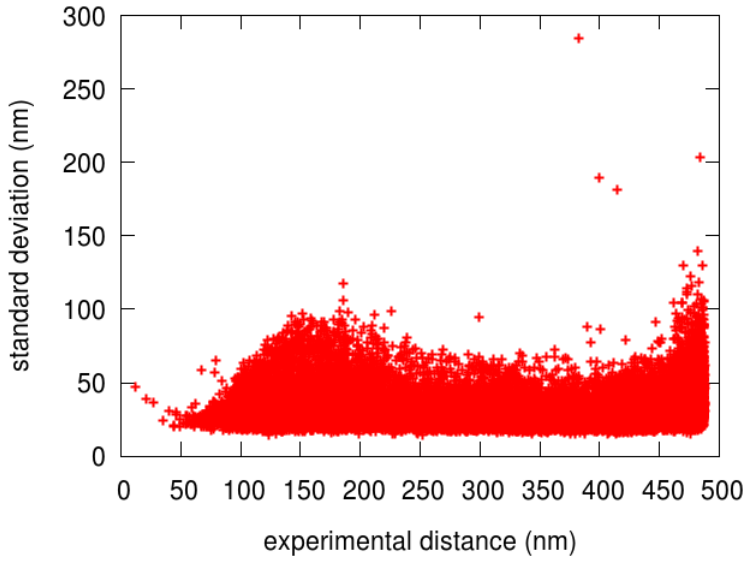

Figure 4. 28 Standard deviations of average distances of plectonemic models for each experimental restraint distances.

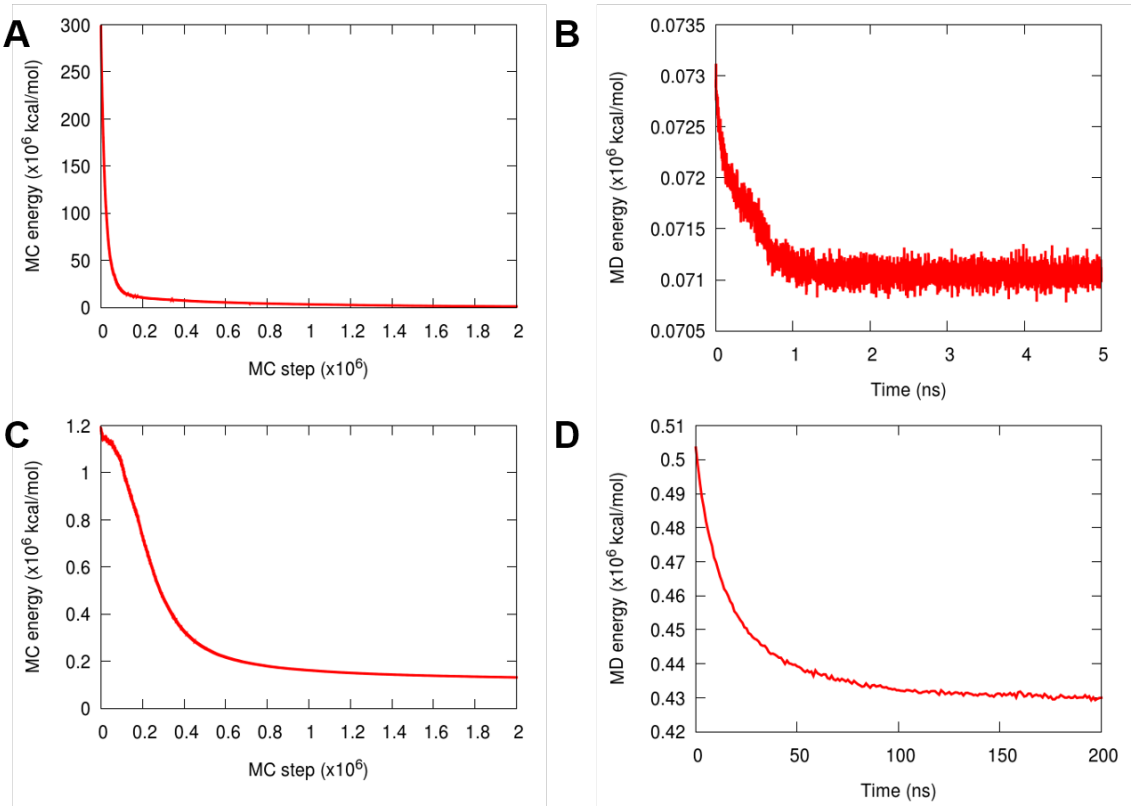

Figure 4. 29 Convergence of energies of the models during MC and MD runs.

Convergence of the models in terms of (**A**) Energy in the first MC run. (**B**) Energy in the first MD run. (**C**) Energy in the second MC run. (**D**) CG force-field energy for double stranded DNA in the last MD refinement run.

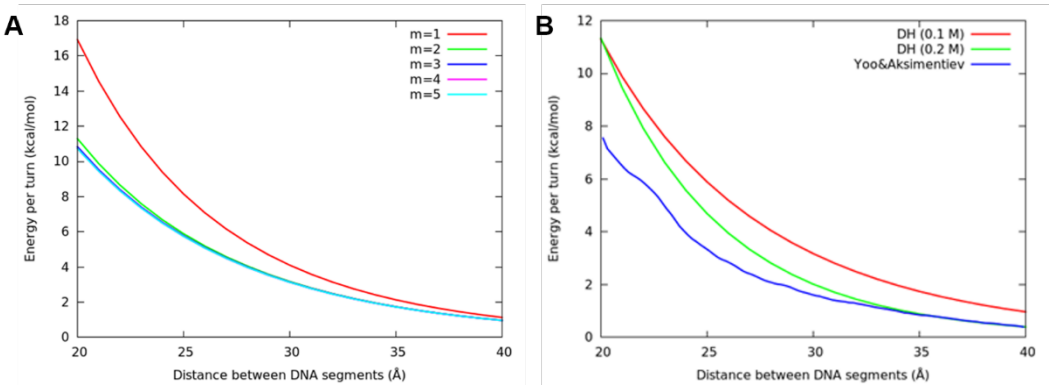

Figure 4. 30 Debye-Hückel potential parameters.

(**A**) Comparison of the electrostatic interaction potentials with different number of point charges per segment (m). Results here are calculated using the Debye-Hückel approximation in 0.1 M salt concentration. (**B**) Comparison of pairwise PMFs from Debye Hückel calculations in 0.1 M and 0.2 M salt concentrations and all-atom calculations in 0.2 M salt concentration by Yoo et *al*. (22). The potential by Yoo et *al.* was normalized to match the 0.2 M DH potential at 40 Å.

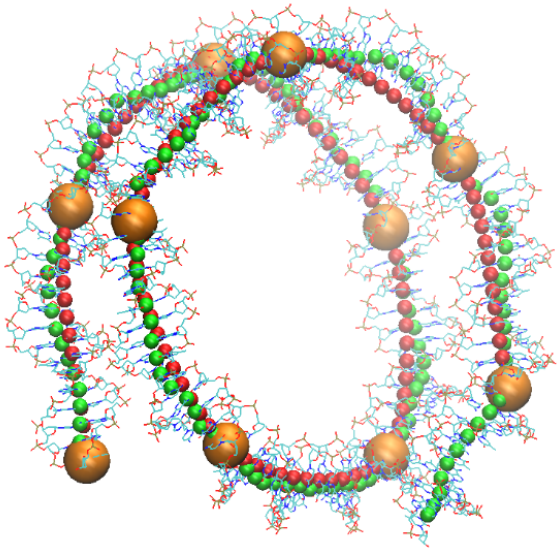

Figure 4. 31 DNA from the nucleosome core particle structure (PDB: 1EQZ).

The atomistic structure is shown in line representation colored by atom types. A 1-bp and 15-bp coarse-grained model based on the centers of mass of base pairs is shown in green and orange, respectively. A reconstructed 1-bp model using the 15-bp CG model as input is shown in red.

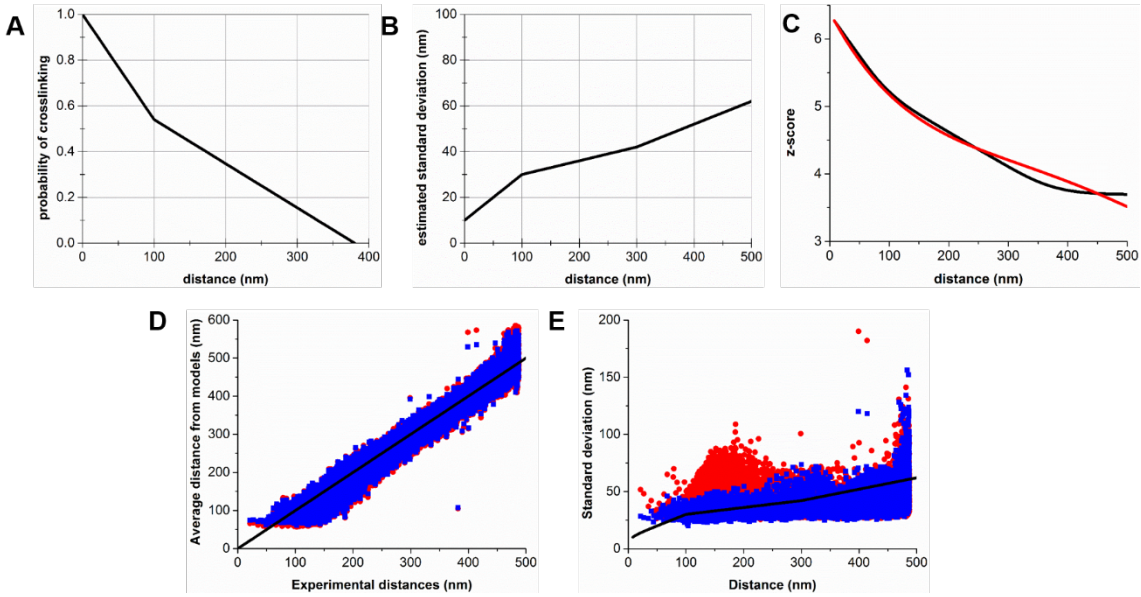

Figure 4. 32 Reweighting parameters and results.

(**A**) Estimated probability of crosslinking as a function of pair distance. (**B**) Estimated standard deviations of distances. (**C**) Calibration curve derived from experimental data (black) and from estimated standard deviations in (B) (red). (**D**) Average distances from models before (red) and after (blue) reweighting. (**E**) Standard deviations from models before (red) and after (blue) reweighting.

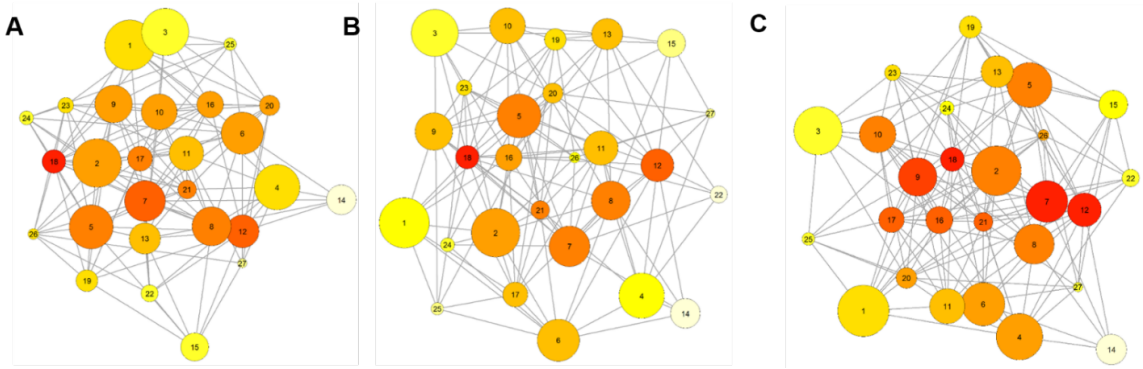

Figure 4. 33 Possible interconversions between clusters based on targeted molecular dynamics with (A) 2.5 kcal/mol, (B) 5 kcal/mol and (C) 7.5 kcal/mol thresholds.

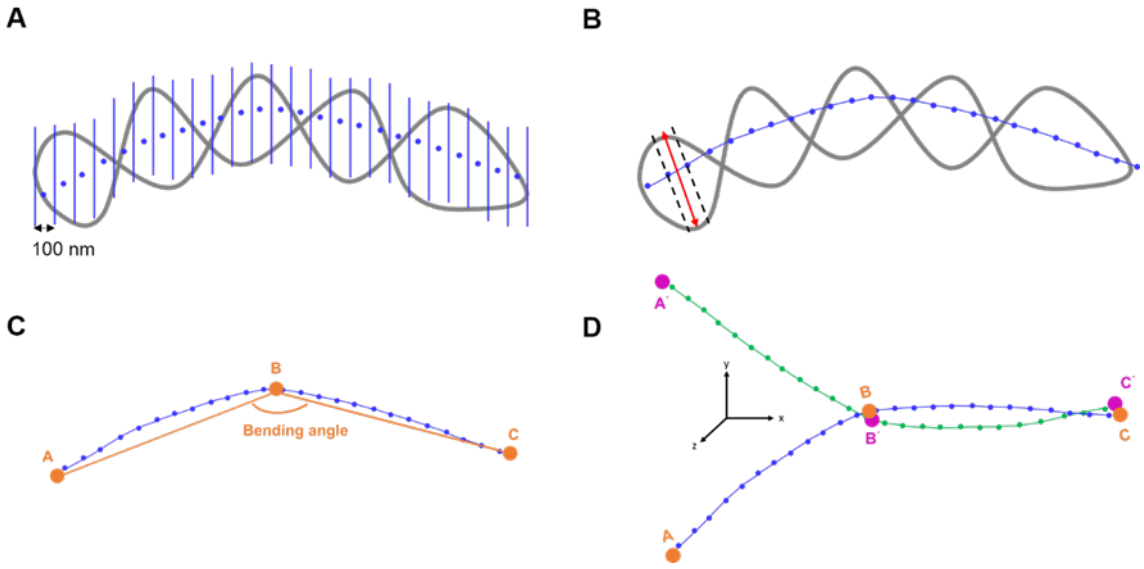

Figure 4. 34 Definition of model shape parameters.

(**A**) The model was divided into slices based on the x coordinates of the beads in the model. Blue points are the center of masses of the segments in each slice. (**B**) Each center of mass point was joined with lines to obtain medial axis of the model. Red dashed arrow shows the width of the slice lying between the second and the third center of masses. (**C**) Bending angle calculation is carried out by finding the initial point of the medial axis (point A), midpoint (point B) and end point (point C) and calculating the angle between those points. (**D**) Bending direction is calculated by aligning models so that their $\overrightarrow{BC}$ vectors coincide with the x axis, and the $\overrightarrow{BA}$ vector of each model gives the bending direction in yz plane.

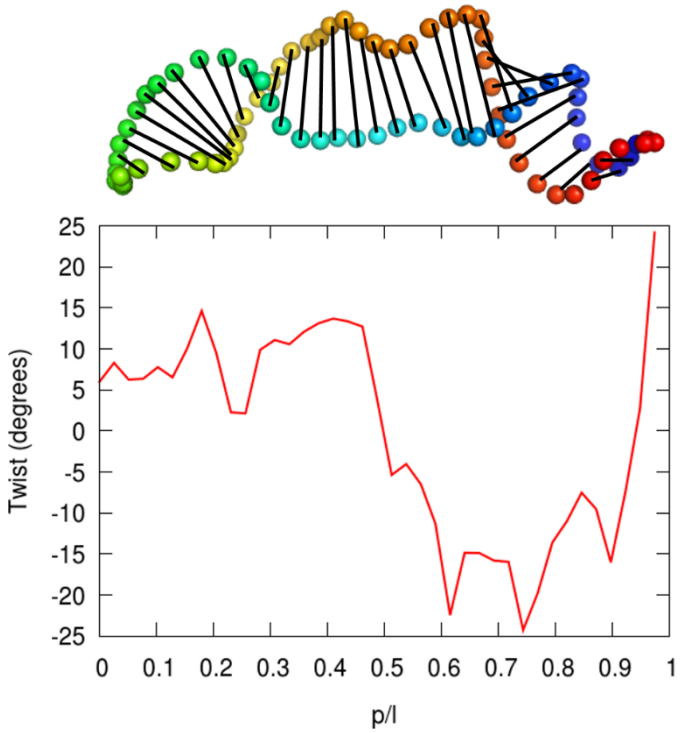

Figure 4. 35 A representative 50-kb model and its arm-twisting along the nucleoid.

Each bead is paired up with the opposite bead in the other arm and the twisting angle is calculated step by step. The bottom graph shows the twisting angle of the arms for the model along the nucleoid.

Table 4.3 Average number of microdomains and branch segment lengths (nm), and the longest principal axis lengths (nm) for each cluster. Standard errors are given as separate columns.

Cluster	Longest principle axis length (nm)	Standard error	Number of microdomains	Standard error	Branch length (nm)	Standard error
1	2200.0	43.0	459.1	40.4	70.9	5.1
2	2065.4	27.6	396.2	37.0	69.6	3.6
3	2288.1	28.0	366.7	32.6	72.4	4.0
4	2165.6	31.8	412.5	29.4	70.9	3.6
5	2083.3	36.2	411.1	47.8	60.6	4.0
6	2142.0	25.1	388.0	30.2	70.8	4.3
7	2329.2	34.0	483.3	31.1	74.6	4.6
8	2366.7	47.7	483.3	74.9	70.0	9.7
9	2345.8	25.9	520.8	31.3	75.0	4.4
10	2266.1	10.4	441.9	17.8	70.6	2.0
11	2117.9	31.7	510.7	30.6	71.1	3.9
12	2204.1	18.0	421.6	25.4	71.4	3.5
13	2245.3	19.7	334.9	23.5	69.1	3.2
14	2370.0	23.8	480.0	26.8	74.9	3.5
15	2247.8	17.4	463.0	23.9	65.4	2.8
16	2251.1	11.9	397.8	19.0	69.3	2.1
17	2450.0	38.1	536.4	38.8	71.8	5.0
18	2220.2	16.4	464.9	21.9	68.2	2.6
19	2184.1	22.0	429.5	25.3	71.6	3.2
20	2202.4	30.6	471.4	40.9	67.1	4.0
21	2090.4	25.9	429.8	27.5	68.3	2.6
22	2195.2	12.3	479.8	17.8	68.2	2.2
23	2172.9	17.7	439.6	22.6	74.8	2.8
24	2324.1	23.0	507.4	23.8	67.4	4.1
25	2242.9	21.8	514.3	34.8	70.0	4.3
26	2126.9	19.9	474.4	24.0	65.1	3.1
27	2230.0	19.8	502.0	23.1	68.2	2.8

Table 4.4 Calculated z-scores for different modules. Module names are adapted from previous study [256].

	z-scores							
	Proteins		Proteins (>500 kb)		Operons		Operons (> 500 kb)	
Module	Genomic	Spatial	Genomic	Spatial	Genomic	Spatial	Genomic	Spatial
yellowgreen	-0.89	-1.74	-0.25	-3.54	-0.85	-1.58	-0.41	-3.36
yellow4	-1.90	-1.84	-1.99	-2.30	-1.83	-1.76	-2.00	-2.15
yellow	-2.34	2.06	0.68	11.43	-2.24	2.21	0.80	12.38
white	0.21	1.91	4.81	8.20	0.21	1.88	4.74	8.16
violet	0.59	1.79	2.37	5.95	0.60	1.89	2.44	6.20
turquoise	-0.06	-2.91	1.64	-14.10	-0.11	-2.88	1.80	-13.75
thistle2	-1.86	-0.92	-0.89	-0.31	-1.79	-0.81	-1.14	-0.20
thistle1	-2.61	0.12	-2.39	1.86	-2.70	0.12	-2.43	1.99
tan	-2.15	-1.51	-0.85	-3.10	-2.01	-1.40	-0.65	-2.75
steelblue	-3.49	-0.13	-2.75	1.71	-3.45	-0.14	-2.74	1.77
skyblue3	-3.38	-0.92	-2.69	-0.36	-3.28	-0.87	-2.14	0.26
skyblue2	0.08	1.31	0.49	2.60	0.55	1.73	1.11	3.17
skyblue1	-1.29	-0.04	-1.90	0.01	-1.24	0.04	-1.91	0.15
skyblue	-0.13	-1.05	-0.38	-2.76	-0.09	-1.00	-0.37	-2.64
sienna3	0.29	2.89	4.50	10.33	0.32	2.82	4.48	10.77
salmon4	0.14	0.09	-0.57	-0.28	0.14	0.15	-0.55	-0.14
salmon	-5.38	-6.32	3.49	-10.48	-5.11	-6.38	3.74	-10.27
saddlebrown	0.55	0.36	0.17	0.74	-0.21	0.31	0.37	1.39
royalblue	-0.16	1.26	0.55	4.13	-0.14	1.32	0.70	4.52
red	-3.04	0.27	-2.19	2.65	-2.96	0.27	-2.11	3.49
purple	0.40	0.94	-0.61	2.47	0.41	1.08	-0.50	3.10
plum2	-3.47	-2.37	-3.01	-3.25	-3.45	-2.31	-3.02	-3.19
plum	-3.90	-2.00	-3.38	-0.46	-3.91	-2.24	-3.34	-0.45
plum1	-1.14	1.29	1.63	5.86	-1.09	1.35	1.63	6.09
pink	-0.29	-2.66	0.73	-8.21	-0.26	-2.60	0.76	-7.97
palevioletred3	0.08	-0.81	-0.14	-1.97	0.07	-0.74	-0.12	-1.81
paleturquoise	0.05	-0.26	0.58	-0.02	0.05	-0.11	0.60	0.35
orangered4	-0.51	-0.03	-0.89	0.04	-0.50	0.09	-0.85	0.25
orangered3	-3.84	-1.37	-3.13	0.68	-3.78	-1.42	-3.19	0.69
orange	-5.73	-3.48	-2.83	-5.32	-5.68	-3.46	-2.73	-4.92

Table 4.4 (*cont'd)*

navajowhite2	-1.55	-0.59	0.59	0.45	-2.86	-0.45	-0.41	1.83
midnightblue	-1.52	-0.39	-1.84	-0.42	-1.48	-0.28	-1.78	-0.08
mediumpurple3	-1.47	-2.40	2.35	-2.34	-1.46	-2.37	2.41	-2.15
mediumpurple2	-0.30	1.33	-0.71	2.25	-0.32	1.37	-0.71	2.36
mediumorchid	-0.57	-0.57	-0.69	-0.92	-0.51	-0.24	-0.08	-0.01
maroon	-3.37	-2.69	-3.12	-3.71	-0.42	-0.34	0.47	0.51
magenta	-0.43	0.73	0.96	3.87	-0.50	1.01	1.23	5.39
lightyellow	-0.64	0.74	-0.57	2.53	-0.67	0.61	-0.66	2.22
lightsteelblue	-0.13	-0.39	-0.58	-1.03	-0.14	-0.37	-0.57	-0.94
lightsteelblue1	0.49	-0.13	0.07	-0.77	0.49	-0.11	0.09	-0.73
lightpink4	-1.14	0.90	-0.79	2.46	-1.14	1.01	-0.77	2.79
lightgreen	0.85	0.45	1.01	1.07	0.87	0.53	1.05	1.40
lightcyan1	-0.30	1.04	0.69	2.98	0.22	1.50	1.28	3.73
lightcyan	-1.30	-0.38	0.13	0.23	-1.26	-0.40	0.13	0.33
lightcoral	-6.58	-4.79	0.00	0.00	-6.64	-4.91	0.00	0.00
lavenderblush3	-0.34	-0.28	-0.63	-0.65	-0.34	-0.20	-0.62	-0.46
ivory	-0.67	0.77	-0.02	2.62	-0.14	1.10	0.27	3.03
indianred4	-1.46	-1.92	-2.13	-2.76	-1.43	-2.07	-2.16	-2.71
honeydew1	0.76	0.42	0.55	0.51	0.76	0.44	0.58	0.62
grey60	-3.59	-0.32	-1.44	2.54	-3.54	-0.13	-1.54	2.99
greenyellow	-0.47	-2.77	4.30	-4.90	-0.46	-2.82	4.30	-4.51
green	0.97	-1.23	1.04	-4.81	0.97	-0.97	1.14	-4.03
floralwhite	-1.40	-1.57	-1.09	-2.67	-1.37	-1.55	-1.03	-2.53
firebrick4	0.46	0.66	0.77	0.94	0.48	0.66	0.78	1.06
darkturquoise	0.16	-0.58	1.25	-0.88	0.20	-0.61	1.27	-0.91
darkslateblue	-1.30	-0.02	0.12	1.46	-1.29	0.08	0.12	1.64
darkseagreen4	0.55	-0.02	-0.35	-0.82	0.56	0.00	-0.07	-0.69
darkred	0.18	1.85	1.10	6.08	0.19	1.92	1.10	6.47
darkorange2	0.62	1.68	2.89	4.97	0.63	1.69	2.86	5.25
darkorange	0.54	-0.79	0.83	-2.10	0.66	-0.70	1.06	-1.66
darkolivegreen4	0.75	-0.02	-0.53	-0.99	0.77	-0.01	-0.53	-0.95
darkolivegreen	-2.29	-3.88	2.54	-6.08	-2.26	-3.92	2.61	-5.93
darkmagenta	0.43	-0.16	0.11	-0.65	0.47	0.14	0.66	0.41
darkgrey	-0.83	-0.56	-0.34	-0.93	-0.83	-0.48	-0.36	-0.71
darkgreen	0.70	-0.17	0.64	-0.77	0.70	-0.02	0.65	-0.36
cyan	-1.06	-0.38	-0.43	-0.32	-1.00	-0.36	-0.40	-0.08
coral2	0.51	-1.40	0.78	-2.38	0.51	-1.34	0.77	-2.23

Table 4.4 (*cont'd)*

coral1	-1.71	-0.03	-1.92	0.41	-1.67	-0.05	-1.91	0.41
brown4	0.68	-0.95	1.41	-1.71	0.67	-0.92	1.44	-1.60
brown2	1.03	0.82	0.19	0.24	1.02	0.82	0.21	0.26
brown	-5.22	-1.13	-3.20	-1.16	-5.08	-1.01	-3.21	-0.31
blue	0.78	-2.17	2.17	-10.24	0.69	-1.91	2.03	-8.42
blue2	-1.23	-0.75	-2.00	-1.07	-1.26	-0.73	-1.98	-1.01
black	-4.22	-3.01	-3.35	-7.52	-4.04	-2.91	-3.27	-7.00
bisque4	0.51	-0.91	-0.44	-2.37	0.49	-0.94	-0.65	-2.48
antiquewhite4	0.10	0.71	0.67	1.72	0.13	0.75	0.67	1.74

CHAPTER 5

Protein Diffusion Around Bacterial Nucleoid

Asli Yildirim, Emma Ford, Tadashi Ando, Yuji Sugita, Michael Feig

5.1 Introduction

Macromolecular crowding in cellular environments plays a crucial role for the structure and dynamics of biomolecules and, accordingly, their biological functions [35, 38, 160, 287]. Therefore, a full comprehension of cellular structure and function requires a systematic understanding of the macromolecular crowding effect. Macromolecular crowding by proteins has been extensively studied to understand the impact of their high concentration in the cell on structure and dynamics of other macromolecules. Studies have shown that protein diffusion which is one of the vital biological processes that governs many cellular functions, is substantially different in crowded environments than *in vitro* [47, 49, 160, 288-290]. Proteins under the effect of macromolecular crowding have been found to move much more slowly and undergo anomalous diffusion where the relationship between mean-square displacement and time is not linear anymore [47, 49].

On the other hand, although genomic DNA also occupies a significant fraction of the cell, little is known about its role as a crowding agent. Notably, in bacteria, nucleoids do not reside in a particular compartment in contrast to the nucleus of a eukaryotic cell and, therefore, the presence of the nucleoid may affect all of the components inside the cell. Compared to proteins, genomic DNA is one very long molecule that can act as a labyrinth which may fasten the diffusion of protein by limiting the possible directions that the protein can go, therefore indirectly guiding its path during diffusion. Very recently, a study has shown that the target search of the *lac* repressor is accelerated in the packed nucleoid environment [291]. On the other hand, the nucleoid may only overcrowd the environment and cause anomalous diffusion as observed as an effect of protein crowding; the available studies are not sufficient to rule out this possible effect yet.

Herein, we use a recently built high-resolution DNA model at a size of 10-kb, an average size of topological microdomains observed in bacterial nucleoids, to study the nucleoid crowding effect on diffusion of proteins with different sizes based on a coarse-grained (CG) formalism in combination with Brownian dynamics (BD) simulations. We use a closed circular DNA model at 15 basepairs (bp) resolution for the nucleoid microdomain and hard-sphere models with various sizes to represent proteins in our systems. We compare the effect of the DNA structure on diffusion with the effects of protein crowding.

5.2 Methodology

BD simulations with hydrodynamics interactions (HI) of CG systems containing a closed circular supercoiled DNA, and neutral proteins were carried out using the GENESIS program package [292]. Proteins were modeled as hard-spheres, and the DNA model was constructed as a worm-like chain of 5-nm long segments with a 1-nm radius each representing 15-bp of double-stranded DNA. The details of the DNA model and the energy potentials for bonded and non-bonded interactions used here can be found in the 15-bp CG model section in Chapter 4. Here, the stretching and bending potentials were used for the DNA model and instead of the torsional potential, an additional dihedral angle potential was applied to avoid DNA unwinding:

$$E_{dihedral} = \sum k_{\phi}\left[1 + \cos\left(\phi - \phi_0\right)\right] \tag{5.1}$$

where ϕ is the dihedral angle formed by four sequential segments, ϕ_0 is the equilibrium dihedral angle, and k_{ϕ} is the force constant. ϕ_0 is set to 209° which is the average dihedral angle calculated from the initial structures, and k_{ϕ} is set to 6.476 kcal/mol which roughly matches the torsion potential used in the previous procedure. Non-bonded interactions between DNA-DNA, DNA-

protein, and protein-protein particles were calculated using Debye-Hückel potential for electrostatic interactions and the segment-overlap potential described in Chapter 4 to avoid passing one particle through another.

The DNA chain consists of 604 segments corresponding to a total length of 9-kb which lies in the microdomain size range observed in bacterial chromosomes [293]. The number of proteins in the system varies depending on the protein size and the crowding concentration. Nine different systems with 20 %, 30 %, and 40 % protein volume fractions and 2.5, 5 and 10 nm protein radii were constructed. The volume fractions used here is within the range of the volume fraction of proteins observed in cells which is believed to be in the range of 5 – 40 % [37] and the proteins with 2.5, 5 and 10 nm radii represent an average size protein observed in bacterial cells, a larger size protein such as RNA polymerase and very big proteins such as ribosome. The simulation conditions for the systems can be found in Table 5.1. All systems with different protein sizes and crowding concentrations were also repeated without DNA for comparison (Table 5.2). The simulations were carried out for 1 μs at 298 K with a 2 ps timestep. Nonbonded interactions were cut off at 30 nm with a switching function becoming effective at 35 nm, and the non-bonded list was cut off at 40 nm.

Table 5.1 Simulation conditions for the systems containing proteins and DNA.

System	Protein Radius (nm)	Number of Proteins	Protein Vol (%)	Number of DNA particles	DNA Vol (%)	Box Size (nm)	Length (μs)
PD1	2.5	1769	20	604	1.6	83.34	1
PD2	2.5	2654	30	604	1.6	83.34	1
PD3	2.5	3538	40	604	1.6	83.34	1
PD4	5.0	221	20	604	1.6	83.34	1
PD5	5.0	332	30	604	1.6	83.34	1
PD6	5.0	442	40	604	1.6	83.34	1
PD7	10.0	28	20	604	1.6	83.34	1
PD8	10.0	42	30	604	1.6	83.34	1
PD9	10.0	55	40	604	1.6	83.34	1

Table 5.2 Simulation conditions for the systems containing only proteins.

System	Protein Radius (nm)	Number of Proteins	Protein Vol (%)	Box Size (nm)	Length (μs)
P1	2.5	1914	21.6	83.34	1
P2	2.5	2799	31.6	83.34	1
P3	2.5	3683	41.6	83.34	1
P4	5.0	239	21.6	83.34	1
P5	5.0	350	31.6	83.34	1
P6	5.0	460	41.6	83.34	1
P7	10.0	30	21.6	83.34	1
P8	10.0	40	31.6	83.34	1
P9	10.0	57	41.6	83.34	1

5.3 Results and Discussion

The effect of the presence of DNA in the environment on protein diffusion was investigated by BD simulations of protein systems with DNA at different crowding concentrations. Initial configurations of PD4 – PD6 systems are shown in Figure 5.1.

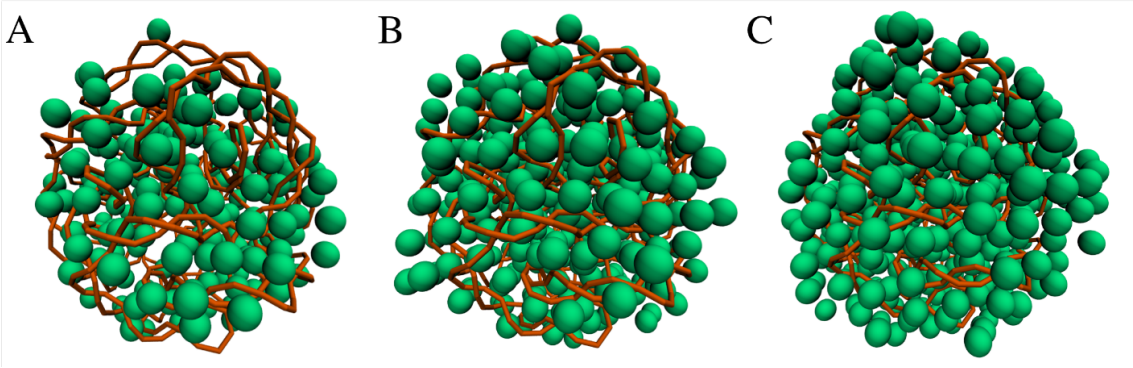

Figure 5.1 Initial configurations of the simulated systems containing DNA (orange) and proteins with 5-nm radius (green). (A) 20 % crowding (PD4), (B) 30 % crowding (PD5), (C) 40 % crowding (PD6).

Mean square displacements (MSD) of proteins with various sizes and different crowding conditions were plotted as a function of time and they are depicted in Figure 5.2. Protein diffusion becomes slower for larger proteins as well as higher crowding ratios as expected. The MSD of small proteins with 2.5 nm radius shows a slight decrease over time in the presence of DNA in all systems with different crowding conditions (Figure 5.2*A*). This decrease is also observed in larger proteins; however, it is more pronounced for the proteins with 10-nm radius (Figure 5.2*B* and *C*). Besides, it has been observed that the MSD of small proteins shows a less linear relationship with time compared to the MSDs of bigger proteins. This can be seen more clearly from the logarithmic scale MSD vs. time plots (Figure 5, right panels). Overall, the more linearity in the MSD vs. time relationship and the larger decrease in the MSD when DNA is in the system observed for larger proteins may suggest that the effect of nucleoid crowding is effective in both long and short time regimes for larger proteins; however, smaller size proteins are only affected in longer-time lag.

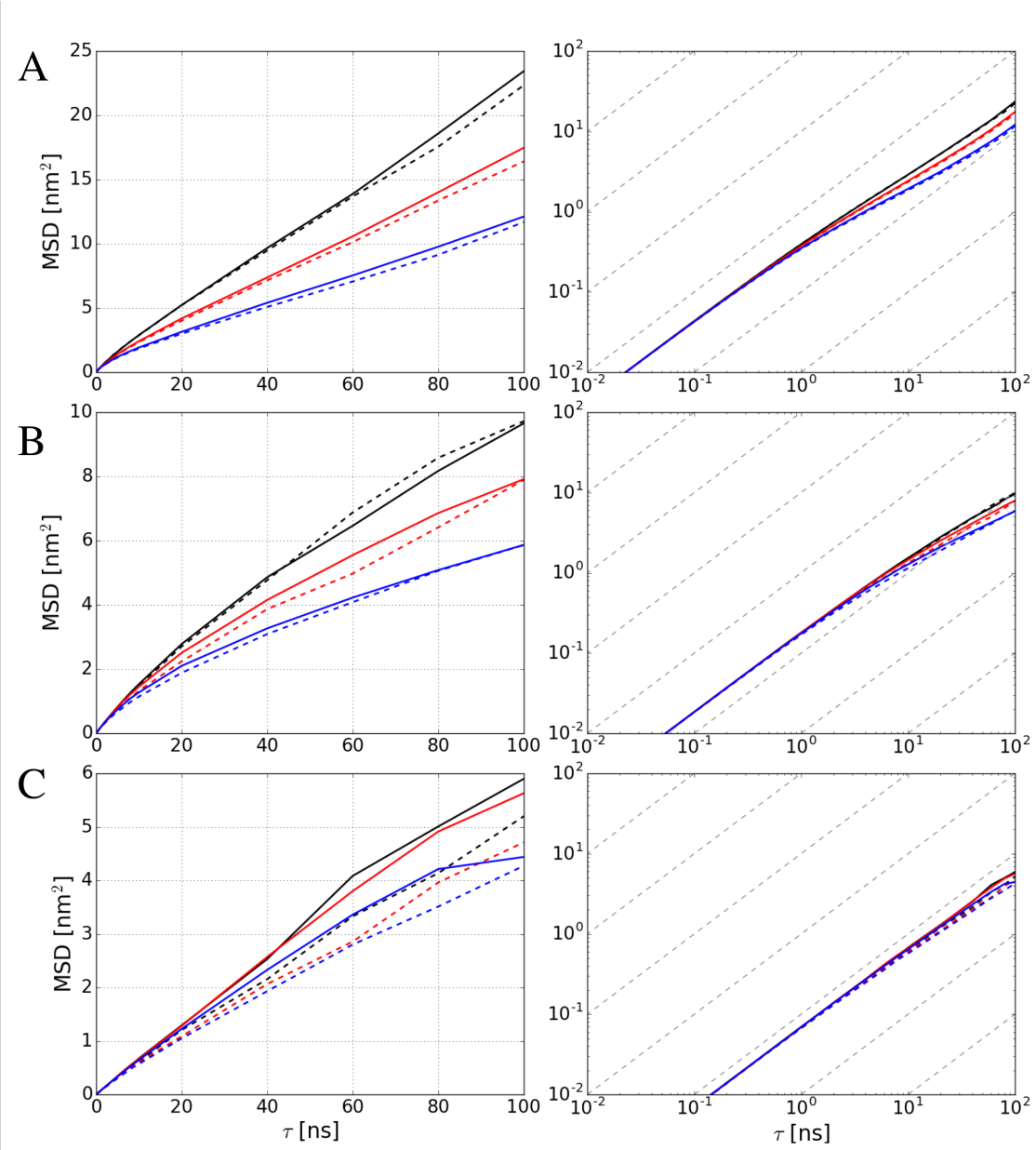

Figure 5.2 Linear (left panel) and logarithmic scale (right panel) plots for MSDs of proteins vs time in the systems with 20 % (black), 30 % (red) and 40 % (blue) crowding. (A) 2.5-nm protein radius, (B) 5-nm protein radius, (C) 10-nm protein radius. Dashed lines are for the corresponding systems with DNA (PD1 to PD9) and solid lines are for P1 – P9 systems.

Next, the diffusion constants were calculated from the MSDs of proteins using the equation

$$D = \lim_{t\to\infty} \frac{MSD(t)}{6t} + \frac{k_B T \xi}{6\pi\eta L} \tag{5.2}$$

where t is time, k_B is the Boltzmann constant, T is the temperature, L is the box length, η is the shear viscosity of water and ξ is the correction factor which is equal to 2.83729 [294]. Since the slope of the MSD vs. time line is different in shorter and longer times for all systems as shown in Figure 5.2, diffusion constants were calculated from the slopes of 1 – 10 ns and 10 – 100 ns time regimes separately and the results are given in Table 5.3 for the DNA + protein systems and in Table 5.4 for the protein systems.

Table 5.3 Diffusion constants of the proteins in systems PD1 – PD9 (Å²/ns). Standard errors are given in the parentheses.

	1 – 10 ns			10 – 100 ns		
Radius	**20 % Crowding**	**30 % Crowding**	**40 % Crowding**	**20 % Crowding**	**30 % Crowding**	**40 % Crowding**
2.5 nm	5.42 *(0.09)*	4.47 *(0.07)*	3.60 *(0.05)*	4.33 *(0.08)*	3.44 *(0.05)*	2.56 *(0.03)*
5.0 nm	3.27 *(0.14)*	2.96 *(0.10)*	2.63 *(0.07)*	2.53 *(0.09)*	2.02 *(0.06)*	1.76 *(0.04)*
10.0 nm	1.95 *(0.16)*	1.81 *(0.12)*	1.78 *(0.11)*	1.68 *(0.15)*	1.62 *(0.10)*	1.54 *(0.08)*

Table 5.4 Diffusion constants of the proteins in systems P1 – P9 (Å²/ns). Standard errors are given in the parentheses.

	1 – 10 ns			10 – 100 ns		
Radius	**20 % Crowding**	**30 % Crowding**	**40 % Crowding**	**20 % Crowding**	**30 % Crowding**	**40 % Crowding**
2.5 nm	5.44 *(0.10)*	4.59 *(0.07)*	3.73 *(0.05)*	4.54 *(0.08)*	3.57 *(0.05)*	2.69 *(0.03)*
5.0 nm	3.35 *(0.14)*	3.19 *(0.11)*	2.87 *(0.08)*	2.40 *(0.09)*	2.11 *(0.06)*	1.73 *(0.04)*
10.0 nm	1.95 *(0.17)*	1.96 *(0.15)*	1.90 *(0.12)*	1.90 *(0.16)*	1.85 *(0.14)*	1.69 *(0.09)*

The comparison of diffusion constants from shorter and longer time regimes show that diffusion is much slower in 10 – 100 ns time range which is an indication of different short and long-time diffusion rates with anomalous behavior. Also, both in the shorter and longer time regimes, proteins have slightly lower diffusion rates in the presence of DNA.

Lastly, in addition to the analysis of the diffusional behavior of proteins under crowding, the spatial distribution of proteins around DNA has also been investigated (Figure 5.3). For the proteins with 2.5 and 5 nm radii (Figure 5.3*A* and *B*), proteins move to the locations which are at a larger distance from the DNA center of mass with increasing crowding. This could be because of the larger number of proteins in the system with increasing crowding, which can also be visually confirmed from Figure 5.1, and limited space inside the nucleoid microdomain. One interesting result is, however, there are two different peaks and a depletion in between for the distribution of proteins with 10-nm radii (Figure 5.3*C*). The depletion shown with a green arrow is seen around ~320 nm which corresponds to the distance of the DNA surface from its center of mass. This result shows that big proteins do not stay around DNA surface, they reside either inside or outside the DNA. This could result from the DNA topology which do not allow larger proteins to sit on the surface due to the lacking void volume.

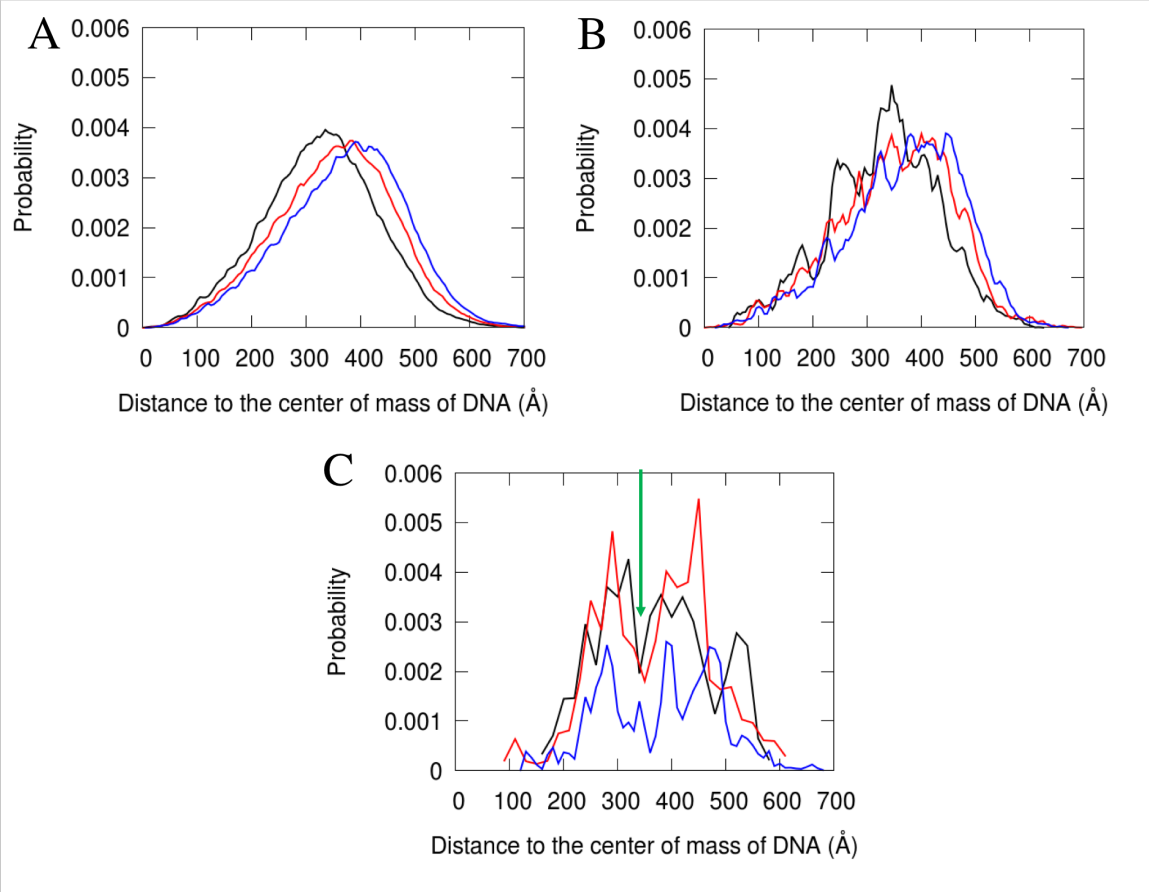

Figure 5.3 Distributions of the distances of the proteins from the center of mass of DNA in the systems with 20 % (black), 30 % (red) and 40 % (blue) crowding. (A) 2.5-nm protein radius, (B) 5-nm protein radius, (C) 10-nm protein radius. The green arrow in C indicates the DNA surface.

5.4 Conclusions and Future Work

In this study, we have performed BD simulations of CG systems containing proteins and DNA at a size of bacterial nucleoid microdomain and have observed slightly slower diffusion for the proteins in the systems where DNA is also present. However, there are some potential concerns:

1) The volume fraction of DNA used here is 1.6 % which is far less than the actual nucleoid volume fraction for bacterial cells which is reported to be around 10 % [295]. This is because of the DNA model used here which is much less compact than the bacterial

nucleoid. The extended model of DNA did not allow using smaller box size that could lead to higher DNA concentration. Therefore, the crowding effect primarily caused by DNA might not have been captured in our simulations.

2) Although the procedure presented in Chapter 4 was followed to generate the DNA model used here, there is one difference from that method: Here, no experimental data was used for the model construction. Use of the recently-built experimentally driven bacterial chromosome models instead of an arbitrary model would result in more accurate DNA models in terms of nucleoid folding and packing since the experimentally-driven models were constructed using the genome-wide *loci* interactions.
3) The diffusion constants calculated here are at least two-fold larger than the protein diffusion constants found from the Stokesian dynamics (SD) simulations of CG systems with similar crowding conditions as well as molecular dynamics (MD) simulations of all-atom systems [48, 49]. In addition to far-field HI, SD also calculates near-field HI as well which is also captured in fully atomistic MD simulations with explicit solvent. Here, we used BD because of its much lower computational cost, however the difference between the diffusion constants calculated here and in the previous studies is probably resulting from neglecting near-field HI which could play a significant role in protein diffusion.

In summary, although the DNA volume fraction considered here is much smaller than the actual volume fraction of nucleoids in bacterial cells, the preliminary results of this study show a slight difference in the rate of protein diffusion in systems with and without DNA, as well as different diffusion rates at shorter and longer time regimes due to the anomalous behavior. Once the abovementioned improvements are made, the models and methodology used here can provide insights into the effects of nucleoid macromolecular crowding on protein diffusion.

CHAPTER 6

Conclusions and Future Outlook

The aim of this dissertation is to contribute to the understanding of the behavior of DNA both as a molecule experiencing macromolecular crowding and as a crowder. In this regard, computer simulations of DNA in crowded environments were performed to investigate the macromolecular crowding effect on DNA structure. On the other hand, the DNA crowding effect on other macromolecules in the cell was also examined by developing a multiscale modeling algorithm to generate three-dimensional (3D) bacterial chromosome models at high-resolution and computer simulations of DNA models constructed using this algorithm together with proteins were carried out to study the nucleoid crowding impact on protein diffusion.

In Chapter 2, one aspect of the macromolecular crowding, the reduced dielectric response of the environment due to the less available water and its slowed dynamics, was investigated to explore its effects on canonical B-form DNA structure. Results show that DNA structure tends to shift towards A-like conformations in reduced dielectric environments as observed in previous studies of DNA in solvents containing cosolvents that have lower dielectric constants than water [57, 59, 117]. Although the reduced dielectric response of the environment due to the crowding favors A-form of DNA, macromolecular crowding has two other significant effects which are the volume-exclusion effect that favors more compact forms of macromolecules due to the less available space in the environment and the non-specific interactions with crowders which challenge the former impact by stabilizing more extended states due to the interactions. In order to study these effects on DNA structure as well, all-atom MD simulations of DNA in the presence of explicit protein crowders were carried out (Chapter 3). Results suggest that the less available space due to crowders, as well as the electrostatic and polar interactions with them, confine and stabilize B-form of DNA which actually modulates the tendency toward A-like conformations due

to the reduced dielectric response of the environment. In conclusion, these studies provided insights into each aspect of macromolecular crowding and its effect on short DNA duplexes.

The DNA models used in the studies mentioned above consist of 12 basepairs (bp), however, DNA are more abundant in a much longer polymeric form *in vivo,* such as plasmids and chromosomes. As a future work, it would be useful to study the effects of cellular crowding on longer DNA molecules. On the other hand, DNA is a highly negatively charged molecule which actively interacts with other charged molecules electrostatically. The protein crowders used in our simulations were neutral, but it would also be interesting to see how negatively or positively charged crowder proteins affect DNA structure. Hopefully, the results described here will also stimulate experimental efforts to characterize DNA structure under crowded conditions using techniques such as Nuclear magnetic resonance (NMR) spectroscopy.

The focus of the following chapters is investigating the macromolecular crowding effect caused by genomic DNA on other macromolecules in cells. Within this context, an experimentally-driven multiscale modeling algorithm to generate realistic bacterial chromosome structures in high-resolution was developed (Chapter 4). Chromosome conformation capture (3C) techniques [21, 22] which are recently developed to study 3D structure of chromosomes also paved the way for computational modeling of chromosome structures. Integration of the available genome-wide contact information of *Caulobacter crescentus* genome [28] with the topological properties of closed circular DNA in the modeling protocol developed here allowed generating *C. crescentus* chromosome structures at 15-bp resolution which are compatible with the available experimental data for this chromosome. High-resolution chromosome structures enabled investigating structural properties of a bacterial chromosome in base-pair detail as well as genome structure and function relationship. The methodology developed here is the pioneer approach to generate chromosome

structures at basepair resolution by using 3C-based data and readily applicable to other bacteria as additional 3C-based data sets are becoming available. Currently, data sets for the chromosomes of *C. crescentus*, *Escherichia coli*, *Bacillus Subtilis* and *Mycoplasma Pneumoniae* bacteria are available [27-33]. Obtaining high-resolution chromosome structures for different bacteria hopefully will help us to gain a global understanding of bacterial chromosomes and their relation to the regulation of gene expression, therefore the biological functions of cells, by enabling direct mapping of the genomic sequence onto the 3D bacterial chromosome structures.

In addition to the investigation of the chromosome structure itself, the constructed high-resolution chromosome models could also be used to investigate their effect on other macromolecules in the cell. A high-resolution DNA model generated by the modeling algorithm discussed above was further used to examine its effect on protein diffusion; but the DNA model was constructed without using any experimental data. The preliminary results of Brownian dynamics simulations of the DNA model with proteins indicate slower motion of proteins around genomic DNA (Chapter 5). However, this work still needs improvements such as using DNA models generated by incorporating the available experimental data and also considering hydrodynamic interactions as discussed in more detail in Chapter 5.

In the end, the ultimate goal of these studies is to expand our ability to generate computational biomolecular systems as accurate as possible to study cellular structure and function in detail. Using high-resolution chromosome models not only with proteins but with all macromolecules in the cell would help us to construct complete realistic models of cellular systems. Previous computational models for cellular systems contained proteins, metabolites, ions, and solvents, however, did not include DNA which could be because of the requirement of bigger system size as well as the unavailability of realistic genomic DNA models [47, 48]. I hope the experimentally-

driven high-resolution chromosome modeling methodology described here will contribute to the computational efforts to understand how all the cellular material works together in harmony and ultimately, how cells, the basic biological units of life, function.

REFERENCES

REFERENCES

1. Dahm, R., *Discovering DNA: Friedrich Miescher and the early years of nucleic acid research*. Human Genetics, 2008. **122**(6): p. 565-581.

2. Levene, P.A., *The structure of yeast nucleic acid. IV. Ammonia hydrolysis*. Journal of Biological Chemistry, 1919. **40**(2): p. 415-424.

3. Chargaff, E., *Chemical Specificity of Nucleic Acids and Mechanism of Their Enzymatic Degradation*. Experientia, 1950. **6**(6): p. 201-209.

4. Franklin, R.E. and R.G. Gosling, *Molecular configuration in sodium thymonucleate*. Nature, 1953. **171**(4356): p. 740-1.

5. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*. Nature, 1953. **171**(4356): p. 737-8.

6. Franklin, R.E. and R.G. Gosling, *The Structure of Sodium Thymonucleate Fibres. I. The Influence of Water Content*. Acta Crystallographica, 1953. **6**: p. 673-677.

7. Ivanov, V.I., et al., *The B to A transition of DNA in solution*. J Mol Biol, 1974. **87**(4): p. 817-33.

8. Zimmerman, S.B. and B.H. Pheiffer, *A direct demonstration that the ethanol-induced transition of DNA is between the A and B forms: an X-ray diffraction study*. J Mol Biol, 1979. **135**(4): p. 1023-7.

9. Xu, Q., R.K. Shoemaker, and W.H. Braunlin, *Induction of B-A transitions of deoxyoligonucleotides by multivalent cations in dilute aqueous solution*. Biophys J, 1993. **65**(3): p. 1039-49.

10. Herbert, A. and A. Rich, *The biology of left-handed Z-DNA*. J Biol Chem, 1996. **271**(20): p. 11595-8.

11. Herbert, A. and A. Rich, *Left-handed Z-DNA: structure and function*. Genetica, 1999. **106**(1-2): p. 37-47.

12. Campbell, M.K. and S.O. Farrell, *Biochemistry*. 2006, USA: Thomson Brooks/Cole.

13. Wang, X., P. Montero Llopis, and D.Z. Rudner, *Organization and segregation of bacterial chromosomes*. Nat Rev Genet, 2013. **14**(3): p. 191-203.

14. Dillon, S.C. and C.J. Dorman, *Bacterial nucleoid-associated proteins, nucleoid structure and gene expression*. Nat Rev Microbiol, 2010. **8**(3): p. 185-95.

15. Postow, L., et al., *Topological domain structure of the Escherichia coli chromosome*. Genes Dev, 2004. **18**(14): p. 1766-79.

16. Vologodskii, A.V. and N.R. Cozzarelli, *Conformational and thermodynamic properties of supercoiled DNA*. Annu Rev Biophys Biomol Struct, 1994. **23**: p. 609-43.

17. Boles, T.C., J.H. White, and N.R. Cozzarelli, *Structure of plectonemically supercoiled DNA*. J Mol Biol, 1990. **213**(4): p. 931-51.

18. Vologodskii, A.V., et al., *Conformational and thermodynamic properties of supercoiled DNA*. J Mol Biol, 1992. **227**(4): p. 1224-43.

19. Ubbink, J. and T. Odijk, *Electrostatic-undulatory theory of plectonemically supercoiled DNA*. Biophys J, 1999. **76**(5): p. 2502-19.

20. Le, T.B. and M.T. Laub, *New approaches to understanding the spatial organization of bacterial genomes*. Curr Opin Microbiol, 2014. **22**: p. 15-21.

21. Dekker, J., et al., *Capturing chromosome conformation*. Science, 2002. **295**(5558): p. 1306-11.

22. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome*. Science, 2009. **326**(5950): p. 289-93.

23. Niki, H., Y. Yamaichi, and S. Hiraga, *Dynamic organization of chromosomal DNA in Escherichia coli*. Genes Dev, 2000. **14**(2): p. 212-23.

24. Viollier, P.H., et al., *Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication*. Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9257-62.

25. Teleman, A.A., et al., *Chromosome arrangement within a bacterium*. Curr Biol, 1998. **8**(20): p. 1102-9.

26. Webb, C.D., et al., *Bipolar localization of the replication origin regions of chromosomes in vegetative and sporulating cells of B. subtilis*. Cell, 1997. **88**(5): p. 667-74.

27. Umbarger, M.A., et al., *The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation*. Mol Cell, 2011. **44**(2): p. 252-64.

28. Le, T.B., et al., *High-resolution mapping of the spatial organization of a bacterial chromosome*. Science, 2013. **342**(6159): p. 731-4.

29. Trussart, M., et al., *Defined chromosome structure in the genome-reduced bacterium Mycoplasma pneumoniae*. Nat Commun, 2017. **8**: p. 14665.

30. Cagliero, C., et al., *Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription*. Nucleic Acids Res, 2013. **41**(12): p. 6058-71.

31. Marbouty, M., et al., *Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging*. Mol Cell, 2015. **59**(4): p. 588-602.

32. Marbouty, M., et al., *Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms*. Elife, 2014. **3**: p. e03318.

33. Wang, X., et al., *Condensin promotes the juxtaposition of DNA flanking its loading site in Bacillus subtilis*. Genes Dev, 2015. **29**(15): p. 1661-75.

34. Hacker, W.C., S. Li, and A.H. Elcock, *Features of genomic organization in a nucleotide-resolution molecular model of the Escherichia coli chromosome*. Nucleic Acids Res, 2017. **45**(13): p. 7541-7554.

35. Zhou, H.X., G. Rivas, and A.P. Minton, *Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences*. Annu Rev Biophys, 2008. **37**: p. 375-97.

36. Fulton, A.B., *How crowded is the cytoplasm?* Cell, 1982. **30**(2): p. 345-7.

37. Ellis, R.J. and A.P. Minton, *Cell biology: join the crowd*. Nature, 2003. **425**(6953): p. 27-8.

38. Ellis, R.J., *Macromolecular crowding: obvious but underappreciated*. Trends Biochem Sci, 2001. **26**(10): p. 597-604.

39. Inomata, K., et al., *High-resolution multi-dimensional NMR spectroscopy of proteins in human cells*. Nature, 2009. **458**(7234): p. 106-9.

40. Miklos, A.C., et al., *Protein crowding tunes protein stability*. J Am Chem Soc, 2011. **133**(18): p. 7116-20.

41. Schlesinger, A.P., et al., *Macromolecular crowding fails to fold a globular protein in cells*. J Am Chem Soc, 2011. **133**(21): p. 8082-5.

42. Harada, R., Y. Sugita, and M. Feig, *Protein crowding affects hydration structure and dynamics*. J Am Chem Soc, 2012. **134**(10): p. 4842-9.

43. Minton, A.P. and J. Wilf, *Effect of macromolecular crowding upon the structure and function of an enzyme: glyceraldehyde-3-phosphate dehydrogenase*. Biochemistry, 1981. **20**(17): p. 4821-6.

44. Harada, R., et al., *Reduced native state stability in crowded cellular environment due to protein-protein interactions*. J Am Chem Soc, 2013. **135**(9): p. 3696-701.

45. Predeus, A.V., et al., *Conformational sampling of peptides in the presence of protein crowders from AA/CG-multiscale simulations*. J Phys Chem B, 2012. **116**(29): p. 8610-20.

46. Feig, M. and Y. Sugita, *Variable interactions between protein crowders and biomolecular solutes are important in understanding cellular crowding*. J Phys Chem B, 2012. **116**(1): p. 599-605.

47. McGuffee, S.R. and A.H. Elcock, *Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm*. PLoS Comput Biol, 2010. **6**(3): p. e1000694.

48. Yu, I., et al., *Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm*. Elife, 2016. **5**.

49. Ando, T. and J. Skolnick, *Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion*. Proc Natl Acad Sci U S A, 2010. **107**(43): p. 18457-62.

50. Nakano, S., D. Miyoshi, and N. Sugimoto, *Effects of molecular crowding on the structures, interactions, and functions of nucleic acids*. Chem Rev, 2014. **114**(5): p. 2733-58.

51. Nakano, S. and N. Sugimoto, *Model studies of the effects of intracellular crowding on nucleic acid interactions*. Molecular Biosystems, 2017. **13**(1): p. 32-41.

52. Xue, Y., et al., *Human telomeric DNA forms parallel-stranded intramolecular G-quadruplex in K+ solution under molecular crowding condition*. Journal of the American Chemical Society, 2007. **129**(36): p. 11185-11191.

53. Miyoshi, D., A. Nakao, and N. Sugimoto, *Molecular crowding regulates the structural switch of the DNA G-quadruplex*. Biochemistry, 2002. **41**(50): p. 15017-15024.

54. Heddi, B. and A.T. Phan, *Structure of Human Telomeric DNA in Crowded Solution*. Journal of the American Chemical Society, 2011. **133**(25): p. 9824-9833.

55. Kulkarni, M. and A. Mukherjee, *Understanding B-DNA to A-DNA transition in the right-handed DNA helix: Perspective from a local to global transition*. Prog Biophys Mol Biol, 2017. **128**: p. 63-73.

56. Jose, D. and D. Porschke, *Dynamics of the B-A transition of DNA double helices*. Nucleic Acids Res, 2004. **32**(7): p. 2251-8.

57. Cheatham, T.E., 3rd, et al., *A molecular level picture of the stabilization of A-DNA in mixed ethanol-water solutions*. Proc Natl Acad Sci U S A, 1997. **94**(18): p. 9626-30.

58. Cheatham, T.E., 3rd and P.A. Kollman, *Insight into the stabilization of A-DNA by specific ion association: spontaneous B-DNA to A-DNA transitions observed in molecular dynamics simulations of d[ACCCGCGGGT]2 in the presence of hexaamminecobalt(III)*. Structure, 1997. **5**(10): p. 1297-311.

59. Noy, A., et al., *Theoretical study of large conformational transitions in DNA: the B<-->A conformational change in water and ethanol/water*. Nucleic Acids Res, 2007. **35**(10): p. 3330-8.

60. Pastor, N., *The B- to A-DNA transition and the reorganization of solvent at the DNA surface*. Biophys J, 2005. **88**(5): p. 3262-75.

61. Gu, B., et al., *Solvent-induced DNA conformational transition*. Phys Rev Lett, 2008. **100**(8): p. 088104.

62. Arscott, P.G., et al., *DNA condensation by cobalt hexaammine (III) in alcohol-water mixtures: dielectric constant and other solvent effects*. Biopolymers, 1995. **36**(3): p. 345-64.

63. Livolant, F. and A. Leforestier, *Condensed phases of DNA: Structures and phase transitions*. Progress in Polymer Science, 1996. **21**(6): p. 1115-1164.

64. Bloomfield, V.A., *DNA condensation*. Current Opinion in Structural Biology, 1996. **6**(3): p. 334-341.

65. Schoen, I., H. Krammer, and D. Braun, *Hybridization kinetics is different inside cells*. Proc Natl Acad Sci U S A, 2009. **106**(51): p. 21649-54.

66. Lukacs, G.L., et al., *Size-dependent DNA mobility in cytoplasm and nucleus*. J Biol Chem, 2000. **275**(3): p. 1625-9.

67. Guigas, G. and M. Weiss, *Sampling the cell with anomalous diffusion - the discovery of slowness*. Biophys J, 2008. **94**(1): p. 90-4.

68. Golding, I. and E.C. Cox, *Physical nature of bacterial cytoplasm*. Phys Rev Lett, 2006. **96**(9): p. 098102.

69. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules*. Nat Struct Biol, 2002. **9**(9): p. 646-52.

70. Huang, J., et al., *CHARMM36m: an improved force field for folded and intrinsically disordered proteins*. Nat Methods, 2017. **14**(1): p. 71-73.

71. Best, R.B., et al., *Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles*. J Chem Theory Comput, 2012. **8**(9): p. 3257-3273.

72. Hart, K., et al., *Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium*. J Chem Theory Comput, 2012. **8**(1): p. 348-362.

73. Cornell, W.D., et al., *A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules*. Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.

74. Oostenbrink, C., et al., *A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6*. Journal of Computational Chemistry, 2004. **25**(13): p. 1656-1676.

75. Zgarbova, M., et al., *Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA*. J Chem Theory Comput, 2015. **11**(12): p. 5723-36.

76. Ivani, I., et al., *Parmbsc1: a refined force field for DNA simulations*. Nat Methods, 2016. **13**(1): p. 55-8.

77. Maffeo, C., et al., *Close encounters with DNA*. J Phys Condens Matter, 2014. **26**(41): p. 413101.

78. Perez, A., F.J. Luque, and M. Orozco, *Frontiers in molecular dynamics simulations of DNA*. Acc Chem Res, 2012. **45**(2): p. 196-205.

79. Cheatham, T.E., 3rd and P.A. Kollman, *Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution*. J Mol Biol, 1996. **259**(3): p. 434-44.

80. Feig, M. and B.M. Pettitt, *Sodium and chlorine ions as part of the DNA solvation shell*. Biophys J, 1999. **77**(4): p. 1769-81.

81. Feig, M. and B.M. Pettitt, *Modeling high-resolution hydration patterns in correlation with DNA sequence and conformation*. J Mol Biol, 1999. **286**(4): p. 1075-95.

82. Feig, M. and B.M. Pettitt, *A molecular simulation picture of DNA hydration around A- and B-DNA*. Biopolymers, 1998. **48**(4): p. 199-209.

83. Chen, J.H., C.L. Brooks, and J. Khandogin, *Recent advances in implicit solvent-based methods for biomolecular simulations*. Current Opinion in Structural Biology, 2008. **18**(2): p. 140-148.

84. Still, W.C., et al., *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics*. Journal of the American Chemical Society, 1990. **112**(16): p. 6127-6129.

85. Bashford, D. and D.A. Case, *Generalized born models of macromolecular solvation effects*. Annual Review of Physical Chemistry, 2000. **51**: p. 129-152.

86. Onufriev, A., *The Generalized Born Model: Its Foundation, Applications, and Limitations*. *Available online:* http://people.cs.vt.edu/~onufriev/PUBLICATIONS/gbreview.pdf

87. Lee, M.S., F.R. Salsbury, and C.L. Brooks, *Novel generalized Born methods*. Journal of Chemical Physics, 2002. **116**(24): p. 10606-10614.

88. Lee, M.S., et al., *New analytic approximation to the standard molecular volume definition and its application to generalized born calculations (vol 24, pg 1348, 2003)*. Journal of Computational Chemistry, 2003. **24**(14): p. 1821-1821.

89. Feig, M., W. Im, and C.L. Brooks, *Implicit solvation based on generalized Born theory in different dielectric environments*. Journal of Chemical Physics, 2004. **120**(2): p. 903-911.

90. Bernardi, R.C., M.C.R. Melo, and K. Schulten, *Enhanced sampling techniques in molecular dynamics simulations of biological systems*. Biochimica Et Biophysica Acta-General Subjects, 2015. **1850**(5): p. 872-877.

91. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding*. Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.

92. Ingolfsson, H.I., et al., *The power of coarse graining in biomolecular simulations*. Wiley Interdisciplinary Reviews-Computational Molecular Science, 2014. **4**(3): p. 225-248.

93. Chirico, G. and J. Langowski, *Kinetics of DNA Supercoiling Studied by Brownian Dynamics Simulation*. Biopolymers, 1994. **34**(3): p. 415-433.

94. Jian, H.M., T. Schlick, and A. Vologodskii, *Internal motion of supercoiled DNA: Brownian dynamics simulations of site juxtaposition*. Journal of Molecular Biology, 1998. **284**(2): p. 287-296.

95. Knotts, T.A., et al., *A coarse grain model for DNA*. Journal of Chemical Physics, 2007. **126**(8).

96. Savelyev, A. and G.A. Papoian, *Molecular Renormalization Group Coarse-Graining of Polymer Chains: Application to Double-Stranded DNA*. Biophysical Journal, 2009. **96**(10): p. 4044-4052.

97. Potoyan, D.A., A. Savelyev, and G.A. Papoian, *Recent successes in coarse-grained modeling of DNA*. Wiley Interdisciplinary Reviews-Computational Molecular Science, 2013. **3**(1): p. 69-83.

98. Sambriski, E.J., D.C. Schwartz, and J.J. de Pablo, *A mesoscale model of DNA and its renaturation*. Biophys J, 2009. **96**(5): p. 1675-90.

99. Hinckley, D.M., et al., *An experimentally-informed coarse-grained 3-Site-Per-Nucleotide model of DNA: structure, thermodynamics, and dynamics of hybridization*. J Chem Phys, 2013. **139**(14): p. 144903.

100. Vologodskii, A.V. and M.D. Frank-Kamenetskii, *Modeling supercoiled DNA*. Methods Enzymol, 1992. **211**: p. 467-80.

101. Kalhor, R., et al., *Genome architectures revealed by tethered chromosome conformation capture and population-based modeling*. Nature Biotechnology, 2012. **30**(1): p. 90-U139.

102. Ermak, D.L. and J.A. Mccammon, *Brownian Dynamics with Hydrodynamic Interactions*. Journal of Chemical Physics, 1978. **69**(4): p. 1352-1360.

103. Iniesta, A. and J.G. Delatorre, *A 2nd-Order Algorithm for the Simulation of the Brownian Dynamics of Macromolecular Models*. Journal of Chemical Physics, 1990. **92**(3): p. 2015-2018.

104. Vangunsteren, W.F. and H.J.C. Berendsen, *Algorithms for Brownian Dynamics*. Molecular Physics, 1982. **45**(3): p. 637-647.

105. Schlick, T., *T. Molecular Modeling and Simulation*. 2002, New York: Springer.

106. Adams, P.J., et al., *Optical properties of CsDNA films as a function of hydration*. J Biomol Struct Dyn, 1994. **11**(6): p. 1277-86.

107. Cheatham, T.E., 3rd, et al., *Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution*. J Biomol Struct Dyn, 1998. **16**(2): p. 265-80.

108. Cieplak, P., T.E. Cheatham, and P.A. Kollman, *Molecular dynamics simulations find that 3' phosphoramidate modified DNA duplexes undergo a B to A transition and normal DNA duplexes an A to B transition*. Journal of the American Chemical Society, 1997. **119**(29): p. 6722-6730.

109. Erfurth, S.C., P.J. Bond, and W.L. Peticolas, *Characterization of the A in equilibrium B transition of DNA in fibers and gels by laser Raman spectroscopy*. Biopolymers, 1975. **14**(6): p. 1245-57.

110. Fang, Y., T.S. Spisz, and J.H. Hoh, *Ethanol-induced structural transitions of DNA on mica*. Nucleic Acids Res, 1999. **27**(8): p. 1943-9.

111. Fuller, W., et al., *The Molecular Configuration of Deoxyribonucleic Acid. Iv. X-Ray Diffraction Study of the a Form*. J Mol Biol, 1965. **12**: p. 60-76.

112. Gao, Y.G., et al., *Influence of counter-ions on the crystal structures of DNA decamers: binding of [Co(NH3)6]3+ and Ba2+ to A-DNA*. Biophys J, 1995. **69**(2): p. 559-68.

113. Ivanov, V.I. and D. Krylov, *A-DNA in solution as studied by diverse approaches*. Methods Enzymol, 1992. **211**: p. 111-27.

114. Ivanov, V.I., et al., *The detection of B-form/A-form junction in a deoxyribonucleotide duplex*. Biophys J, 1996. **71**(6): p. 3344-9.

115. Jayaram, B., et al., *Free energy analysis of the conformational preferences of A and B forms of DNA in solution*. Journal of the American Chemical Society, 1998. **120**(41): p. 10629-10633.

116. Robinson, H. and A.H. Wang, *Neomycin, spermine and hexaamminecobalt (III) share common structural motifs in converting B- to A-DNA*. Nucleic Acids Res, 1996. **24**(4): p. 676-82.

117. Sprous, D., M.A. Young, and D.L. Beveridge, *Molecular dynamics studies of the conformational preferences of a DNA double helix in water and an ethanol/water mixture: Theoretical considerations of the A double left right arrow B transition*. Journal of Physical Chemistry B, 1998. **102**(23): p. 4658-4667.

118. Srinivasan, J., et al., *Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices*. Journal of the American Chemical Society, 1998. **120**(37): p. 9401-9409.

119. Vargason, J.M., K. Henderson, and P.S. Ho, *A crystallographic map of the transition from B-DNA to A-DNA*. Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7265-70.

120. Ellis, R.J., *Macromolecular crowding: an important but neglected aspect of the intracellular environment*. Curr Opin Struct Biol, 2001. **11**(1): p. 114-9.

121. Dill, K.A., *Theory for the folding and stability of globular proteins*. Biochemistry, 1985. **24**(6): p. 1501-9.

122. Elcock, A.H., *Models of macromolecular crowding effects and the need for quantitative comparisons with experiment*. Curr Opin Struct Biol, 2010. **20**(2): p. 196-206.

123. Hong, J. and L.M. Gierasch, *Macromolecular crowding remodels the energy landscape of a protein by favoring a more compact unfolded state*. J Am Chem Soc, 2010. **132**(30): p. 10445-52.

124. Onuchic, J.N. and P.G. Wolynes, *Theory of protein folding*. Curr Opin Struct Biol, 2004. **14**(1): p. 70-5.

125. Wang, Y., et al., *Macromolecular crowding and protein stability*. J Am Chem Soc, 2012. **134**(40): p. 16614-8.

126. Schutz, C.N. and A. Warshel, *What are the dielectric "constants" of proteins and how to validate electrostatic models?* Proteins, 2001. **44**(4): p. 400-17.

127. Gilson, M.K. and B.H. Honig, *The dielectric constant of a folded protein*. Biopolymers, 1986. **25**(11): p. 2097-119.

128. Antosiewicz, J., J.A. McCammon, and M.K. Gilson, *The determinants of pKas in proteins*. Biochemistry, 1996. **35**(24): p. 7819-33.

129. Dwyer, J.J., et al., *High apparent dielectric constants in the interior of a protein reflect water penetration*. Biophys J, 2000. **79**(3): p. 1610-20.

130. Warshel, A. and A. Papazyan, *Electrostatic effects in macromolecules: fundamental concepts and practical modeling*. Curr Opin Struct Biol, 1998. **8**(2): p. 211-7.

131. Akhadov, Y.Y., *Dielectric Properties of Binary Solutions*. 1981, Oxford, UK: Pergamon Press.

132. Despa, F., A. Fernandez, and R.S. Berry, *Dielectric modulation of biological water*. Phys Rev Lett, 2004. **93**(22): p. 228104.

133. Lee, M.S., et al., *New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations*. J Comput Chem, 2003. **24**(11): p. 1348-56.

134. Tanizaki, S., et al., *Conformational sampling of peptides in cellular environments*. Biophys J, 2008. **94**(3): p. 747-59.

135. Tjong, H. and H.X. Zhou, *Prediction of protein solubility from calculation of transfer free energy*. Biophys J, 2008. **95**(6): p. 2601-9.

136. Chocholousova, J. and M. Feig, *Implicit solvent simulations of DNA and DNA-protein complexes: agreement with explicit solvent vs experiment*. J Phys Chem B, 2006. **110**(34): p. 17240-51.

137. Ruscio, J.Z. and A. Onufriev, *A computational study of nucleosomal DNA flexibility*. Biophys J, 2006. **91**(11): p. 4121-32.

138. Tsui, V., et al., *NMR and molecular dynamics studies of the hydration of a zinc finger-DNA complex*. J Mol Biol, 2000. **302**(5): p. 1101-17.

139. Tsui, V. and D.A. Case, *Theory and applications of the generalized Born solvation model in macromolecular simulations*. Biopolymers, 2000. **56**(4): p. 275-91.

140. Drew, H.R., et al., *Structure of a B-DNA dodecamer: conformation and dynamics*. Proc Natl Acad Sci U S A, 1981. **78**(4): p. 2179-83.

141. Brooks, C.L., M. Berkowitz, and S.A. Adelman, *Generalized Langevin Theory for Many-Body Problems in Chemical-Dynamics - Gas-Surface Collisions, Vibrational-Energy Relaxation in Solids, and Recombination Reactions in Liquids*. Journal of Chemical Physics, 1980. **73**(9): p. 4353-4364.

142. Brooks, B.R., et al., *CHARMM: The Biomolecular Simulation Program*. Journal of Computational Chemistry, 2009. **30**(10): p. 1545-1614.

143. Hart, K., et al., *Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium*. Journal of Chemical Theory and Computation, 2012. **8**(1): p. 348-362.

144. Best, R.B., et al., *Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles*. Journal of Chemical Theory and Computation, 2012. **8**(9): p. 3257-3273.

145. Feig, M., J. Karanicolas, and C.L. Brooks, *MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology*. Journal of Molecular Graphics & Modelling, 2004. **22**(5): p. 377-395.

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

147. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics*. Journal of Molecular Graphics & Modelling, 1996. **14**(1): p. 33-38.

148. Schrödinger, L., *The PyMOL Molecular Graphics System 1.5.0.4*.

149. Malinina, L., et al., *Structure of the d(CGCCCGCGGGCG) dodecamer: a kinked A-DNA molecule showing some B-DNA features*. J Mol Biol, 1999. **285**(4): p. 1679-90.

150. Dickerson, R.E. and H.L. Ng, *DNA structure from A to B*. Proc Natl Acad Sci U S A, 2001. **98**(13): p. 6986-8.

151. Kollman, P.A., et al., *Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models*. Acc Chem Res, 2000. **33**(12): p. 889-97.

152. Brice, A.R. and B.N. Dominy, *Analyzing the robustness of the MM/PBSA free energy calculation method: application to DNA conformational transitions*. J Comput Chem, 2011. **32**(7): p. 1431-40.

153. Sitkoff, D., K.A. Sharp, and B. Honig, *Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models*. Journal of Physical Chemistry, 1994. **98**(7): p. 1978-1988.

154. Akerlof, G., *Dielectric constants of some organic solvent-water mixtures at various temperatures*. Journal of the American Chemical Society, 1932. **54**: p. 4125-4139.

155. Zimmerman, S.B. and B.H. Pheiffer, *Helical parameters of DNA do not change when DNA fibers are wetted: X-ray diffraction study*. Proc Natl Acad Sci U S A, 1979. **76**(6): p. 2703-7.

156. Zimmerman, S.B. and S.O. Trach, *Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli*. J Mol Biol, 1991. **222**(3): p. 599-620.

157. McPhie, P., Y.S. Ni, and A.P. Minton, *Macromolecular crowding stabilizes the molten globule form of apomyoglobin with respect to both cold and heat unfolding*. J Mol Biol, 2006. **361**(1): p. 7-10.

158. Dix, J.A. and A.S. Verkman, *Crowding effects on diffusion in solutions and cells*. Annu Rev Biophys, 2008. **37**: p. 247-63.

159. Minton, A.P., *Macromolecular crowding and molecular recognition*. J Mol Recognit, 1993. **6**(4): p. 211-4.

160. Zimmerman, S.B. and A.P. Minton, *Macromolecular crowding: biochemical, biophysical, and physiological consequences*. Annu Rev Biophys Biomol Struct, 1993. **22**: p. 27-65.

161. Rivas, G. and A.P. Minton, *Macromolecular Crowding In Vitro, In Vivo, and In Between*. Trends Biochem Sci, 2016. **41**(11): p. 970-981.

162. Feig, M., et al., *Crowding in Cellular Environments at an Atomistic Level from Computer Simulations*. J Phys Chem B, 2017. **121**(34): p. 8009-8025.

163. Zhou, H.X., *Protein folding and binding in confined spaces and in crowded solutions*. J Mol Recognit, 2004. **17**(5): p. 368-75.

164. Cheung, M.S., D. Klimov, and D. Thirumalai, *Molecular crowding enhances native state stability and refolding rates of globular proteins*. Proc Natl Acad Sci U S A, 2005. **102**(13): p. 4753-8.

165. Senske, M., et al., *Protein stabilization by macromolecular crowding through enthalpy rather than entropy*. J Am Chem Soc, 2014. **136**(25): p. 9036-41.

166. Wang, Y., C. Li, and G.J. Pielak, *Effects of proteins on protein diffusion*. J Am Chem Soc, 2010. **132**(27): p. 9392-7.

167. Yildirim, A., et al., *Conformational preferences of DNA in reduced dielectric environments*. J Phys Chem B, 2014. **118**(37): p. 10874-81.

168. Yoshikawa, K., et al., *Compaction of DNA Induced by Like-Charge Protein: Opposite Salt-Effect against the Polymer-Salt-Induced Condensation with Neutral Polymer*. Journal of Physical Chemistry Letters, 2010. **1**(12): p. 1763-1766.

169. Watson, J.D. and F.H. Crick, *The structure of DNA*. Cold Spring Harb Symp Quant Biol, 1953. **18**: p. 123-31.

170. Xu, Q.W., R.K. Shoemaker, and W.H. Braunlin, *Induction of B-a Transitions of Deoxyoligonucleotides by Multivalent Cations in Dilute Aqueous-Solution*. Biophysical Journal, 1993. **65**(3): p. 1039-1049.

171. Cheatham, T.E. and P.A. Kollman, *Insight into the stabilization of A-DNA by specific ion association: spontaneous B-DNA to A-DNA transitions observed in molecular dynamics simulations of d[ACCCGCGGGT](2) in the presence of hexaamminecobalt(III)*. Structure, 1997. **5**(10): p. 1297-1311.

172. Arscott, P.G., et al., *DNA Condensation by Cobalt Hexaammine(Iii) in Alcohol-Water Mixtures - Dielectric-Constant and Other Solvent Effects*. Biopolymers, 1995. **36**(3): p. 345-364.

173. Young, M.A. and D.L. Beveridge, *Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn*. J Mol Biol, 1998. **281**(4): p. 675-87.

174. Gu, B., et al., *Solvent-induced DNA conformational transition*. Physical Review Letters, 2008. **100**(8).

175. Brooks, B.R., et al., *CHARMM: the biomolecular simulation program*. J Comput Chem, 2009. **30**(10): p. 1545-614.

176. Gronenborn, A.M., et al., *A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G*. Science, 1991. **253**(5020): p. 657-61.

177. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water*. Journal of Chemical Physics, 1983. **79**(2): p. 926-935.

178. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes*. Journal of Computational Physics, 1977. **23**(3): p. 327-341.

179. Darden, T., D. York, and L. Pedersen, *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems*. Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.

180. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics*. J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.

181. Franklin, R.E. and R.G. Gosling, *Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate*. Nature, 1953. **172**(4369): p. 156-7.

182. Cheatham, T.E. and M.A. Young, *Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise*. Biopolymers, 2001. **56**(4): p. 232-256.

183. Korolev, N., et al., *On the competition between water, sodium ions, and spermine in binding to DNA: a molecular dynamics computer simulation study*. Biophys J, 2002. **82**(6): p. 2860-75.

184. Bonvin, A.M., *Localisation and dynamics of sodium counterions around DNA in solution from molecular dynamics simulation*. Eur Biophys J, 2000. **29**(1): p. 57-60.

185. Kalodimos, C.G., et al., *Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes*. Science, 2004. **305**(5682): p. 386-9.

186. Nekludova, L. and C.O. Pabo, *Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes*. Proc Natl Acad Sci U S A, 1994. **91**(15): p. 6948-52.

187. Shakked, Z., et al., *Determinants of repressor/operator recognition from the structure of the trp operator binding site*. Nature, 1994. **368**(6470): p. 469-73.

188. Guzikevich-Guerstein, G. and Z. Shakked, *A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein*. Nat Struct Biol, 1996. **3**(1): p. 32-7.

189. Olson, W.K., et al., *DNA sequence-dependent deformability deduced from protein-DNA crystal complexes*. Proc Natl Acad Sci U S A, 1998. **95**(19): p. 11163-8.

190. Lu, X.J., Z. Shakked, and W.K. Olson, *A-form conformational motifs in ligand-bound DNA structures*. J Mol Biol, 2000. **300**(4): p. 819-40.

191. Stagg, L., et al., *Molecular crowding enhances native structure and stability of alpha/beta protein flavodoxin*. Proc Natl Acad Sci U S A, 2007. **104**(48): p. 18976-81.

192. Minton, A.P., *Implications of macromolecular crowding for protein assembly*. Curr Opin Struct Biol, 2000. **10**(1): p. 34-9.

193. Gebala, M., et al., *Does Cation Size Affect Occupancy and Electrostatic Screening of the Nucleic Acid Ion Atmosphere?* J Am Chem Soc, 2016. **138**(34): p. 10925-34.

194. Gebala, M., et al., *Cation-Anion Interactions within the Nucleic Acid Ion Atmosphere Revealed by Ion Counting*. J Am Chem Soc, 2015. **137**(46): p. 14705-15.

195. Giambasu, G.M., et al., *Ion counting from explicit-solvent simulations and 3D-RISM*. Biophys J, 2014. **106**(4): p. 883-94.

196. Allred, B.E., M. Gebala, and D. Herschlag, *Determination of Ion Atmosphere Effects on the Nucleic Acid Electrostatic Potential and Ligand Association Using AH+.C Wobble Formation in Double-Stranded DNA*. J Am Chem Soc, 2017. **139**(22): p. 7540-7548.

197. Niki, H., Y. Yamaichi, and S. Hiraga, *Dynamic organization of chromosomal DNA in Escherichia coli*. Genes & Development, 2000. **14**(2): p. 212-223.

198. Fisher, J.K., et al., *Four-dimensional imaging of E. coli nucleoid organization and dynamics in living cells*. Cell, 2013. **153**(4): p. 882-895.

199. Nielsen, H.J., et al., *The Escherichia coli chromosome is organized with the left and right chromosome arms in separate cell halves*. Molecular Microbiology, 2006. **62**(2): p. 331-338.

200. Viollier, P.H., et al., *Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication*. PNAS, 2004. **101**(25): p. 9257-9262.

201. Wang, X., P. Montero Llopis, and D.Z. Rudner, *Bacillus subtilis chromosome organization oscillates between two distinct patterns*. PNAS, 2014. **111**(35): p. 12877-12882.

202. Wang, X.D., et al., *The two Escherichia coli chromosome arms locate to separate cell halves*. Genes & Development, 2006. **20**(13): p. 1727-1731.

203. Youngren, B., et al., *The multifork Escherichia coli chromosome is a self-duplicating and self-segregating thermodynamic ring polymer*. Genes & Development, 2014. **28**(1): p. 71-84.

204. Weng, X. and J. Xiao, *Spatial organization of transcription in bacterial cells*. Trends in Genetics, 2014. **30**(7): p. 287-297.

205. Lau, I.F., et al., *Spatial and temporal organization of replicating Escherichia coli chromosomes*. Molecular Microbiology, 2003. **49**(3): p. 731-743.

206. Fekete, R.A. and D.K. Chattoraj, *A cis-acting sequence involved in chromosome segregation in Escherichia coli*. Molecular Microbiology, 2005. **55**(1): p. 175-183.

207. Robinett, C.C., et al., *In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition*. Journal of Cell Biology, 1996. **135**(6): p. 1685-1700.

208. Teleman, A.A., et al., *Chromosome arrangement within a bacterium*. Current Biology, 1998. **8**(20): p. 1102-1109.

209. Webb, C.D., et al., *Bipolar localization of the replication origin regions of chromosomes in vegetative and sporulating cells of B-subtilis*. Cell, 1997. **88**(5): p. 667-674.

210. Cagliero, C., et al., *Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription*. Nucleic Acids Research, 2013. **41**(12): p. 6058-6071.

211. Le, T.B.K., et al., *High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome*. Science, 2013. **342**(6159): p. 731-734.

212. Marbouty, M., et al., *Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging*. Molecular Cell, 2015. **59**(4): p. 588-602.

213. Trussart, M., et al., *Defined chromosome structure in the genome-reduced bacterium Mycoplasma pneumoniae*. Nature Communications, 2017. **8**: p. 14665.

214. Umbarger, M.A., et al., *The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation*. Molecular Cell, 2011. **44**(2): p. 252-264.

215. Wang, X.D., et al., *Condensin promotes the juxtaposition of DNA flanking its loading site in Bacillus subtilis*. Genes & Development, 2015. **29**(15): p. 1661-1675.

216. Browning, D.F., D.C. Grainger, and S.J. Busby, *Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression*. Current Opinions in Microbiology, 2010. **13**(6): p. 773-780.

217. Dillon, S.C. and C.J. Dorman, *Bacterial nucleoid-associated proteins, nucleoid structure and gene expression*. Nature Reviews Microbiology, 2010. **8**(3): p. 185-195.

218. Wiggins, P.A., et al., *Strong intranucleoid interactions organize the Escherichia coli chromosome into a nucleoid filament*. PNAS, 2010. **107**(11): p. 4991-4995.

219. Wang, X. and D.Z. Rudner, *Spatial organization of bacterial chromosomes*. Current Opinions in Microbiology, 2014. **22**: p. 66-72.

220. Badrinarayanan, A., T.B. Le, and M.T. Laub, *Bacterial chromosome organization and segregation*. Annual Review of Cell and Developmental Biology, 2015. **31**: p. 171-199.

221. Dame, R.T., M.C. Noom, and G.J. Wuite, *Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation*. Nature, 2006. **444**(7117): p. 387-390.

222. Dame, R.T., C. Wyman, and N. Goosen, *H-NS mediated compaction of DNA visualised by atomic force microscopy*. Nucleic Acids Research, 2000. **28**(18): p. 3504-3510.

223. Higgins, C.F., et al., *A physiological role for DNA supercoiling in the osmotic regulation of gene expression in S. typhimurium and E. coli*. Cell, 1988. **52**(4): p. 569-584.

224. Xie, T., et al., *Spatial features for Escherichia coli genome organization*. BMC Genomics, 2015. **16**: p. 37.

225. Junier, I., O. Martin, and F. Kepes, *Spatial and topological organization of DNA chains induced by gene co-localization*. PLoS Computational Biology, 2010. **6**(2): p. e1000678.

226. Dily, F.L., F. Serra, and M.A. Marti-Renom, *3D modeling of chromatin structure: is there a way to integrate and reconcile single and population experimental data?* WIREs Computational Molecular Science 2017. **e1308**.

227. Duan, Z., et al., *A three-dimensional model of the yeast genome*. Nature, 2010. **465**(7296): p. 363-367.

228. Hu, M., et al., *Bayesian inference of spatial organizations of chromosomes*. PLoS Computational Biology, 2013. **9**(1): p. e1002893.

229. Zhang, Z., et al., *3D chromosome modeling with semi-definite programming and Hi-C data*. Journal of Computational Biology, 2013. **20**(11): p. 831-846.

230. Varoquaux, N., et al., *A statistical approach for inferring the 3D structure of the genome*. Bioinformatics, 2014. **30**(12): p. i26-i33.

231. Segal, M.R., et al., *Reproducibility of 3D chromatin configuration reconstructions*. Biostatistics, 2014. **15**(3): p. 442-456.

232. Lesne, A., et al., *3D genome reconstruction from chromosomal contacts*. Nature Methods, 2014. **11**(11): p. 1141-1143.

233. Dekker, J., *Mapping the 3D genome: Aiming for consilience*. Nature Reviews Molecular Cell Biology, 2016. **17**(12): p. 741-742.

234. Giorgetti, L. and E. Heard, *Closing the loop: 3C versus DNA FISH*. Genome Biology, 2016. **17**: p. 215.

235. Imakaev, M.V., G. Fudenberg, and L.A. Mirny, *Modeling chromosomes: Beyond pretty pictures*. FEBS Letters, 2015. **589**(20): p. 3031-3036.

236. Trussart, M., et al., *Assessing the limits of restraint-based 3D modeling of genomes and genomic domains*. Nucleic Acids Research, 2015. **43**(7): p. 3465-3477.

237. Williamson, I., et al., *Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization*. Genes & Development, 2014. **28**(24): p. 2778-2791.

238. Di Pierro, M., et al., *Transferable model for chromosome architecture*. PNAS, 2016. **113**(43): p. 12168-12173.

239. Zhang, B. and P.G. Wolynes, *Topology, structures, and energy landscapes of human chromosomes*. PNAS, 2015. **112**(19): p. 6062-6067.

240. Kalhor, R., et al., *Genome architectures revealed by tethered chromosome conformation capture and population-based modeling*. Nature Biotechnology, 2011. **30**(1): p. 90-98.

241. Tjong, H., et al., *Population-based 3D genome structure analysis reveals driving forces in spatial genome organization*. PNAS, 2016. **113**(12): p. E1663-E1672.

242. Hacker, W.C., S. Li, and A.H. Elcock, *Features of genomic organization in a nucleotide-resolution molecular model of the Escherichia coli chromosome*. Nucleic Acids Research, 2017. **45**: p. 7541-7554.

243. Grainger, D.C., et al., *Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome*. Nucleic Acids Research, 2006. **34**(16): p. 4642-4652.

244. Stracy, M., et al., *Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid*. PNAS, 2015. **112**(32): p. E4390-E4399.

245. Hong, S.H., et al., *Caulobacter chromosome in vivo configuration matches model predictions for a supercoiled polymer in a cell-like confinement*. PNAS, 2013. **110**(5): p. 1674-1679.

246. Wright, C.S., et al., *Intergenerational continuity of cell shape dynamics in Caulobacter crescentus*. Science Reports, 2015. **5**: p. 9155.

247. Robinow, C. and E. Kellenberger, *The bacterial nucleoid revisited*. Microbiological Reviews, 1994. **58**(2): p. 211-232.

248. Wang, X., P. Montero Llopis, and D.Z. Rudner, *Organization and segregation of bacterial chromosomes*. Nature Reviews Genetics, 2013. **14**(3): p. 191-203.

249. Bakshi, S., et al., *Time-dependent effects of transcription- and translation-halting drugs on the spatial distributions of the Escherichia coli chromosome and ribosomes*. Molecular Microbiology, 2014. **94**(4): p. 871-887.

250. Lee, S.F., et al., *Super-resolution imaging of the nucleoid-associated protein HU in Caulobacter crescentus*. Biophysical Journal, 2011. **100**(7): p. L31-L33.

251. Sanamrad, A., et al., *Single-particle tracking reveals that free ribosomal subunits are not excluded from the Escherichia coli nucleoid*. PNAS, 2014. **111**(31): p. 11413-11418.

252. Mondal, J., et al., *Entropy-based mechanism of ribosome-nucleoid segregation in E. coli cells*. Biophysical Journal, 2011. **100**(11): p. 2605-2613.

253. Bakshi, S., H. Choi, and J.C. Weisshaar, *The spatial biology of transcription and translation in rapidly growing Escherichia coli*. Frontiers in Microbiology, 2015. **6**: p. 636

254. Ortega, A., D. Amoros, and J. Garcia de la Torre, *Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models*. Biophysical Journal, 2011. **101**(4): p. 892-898.

255. Liu, Y., et al., *A model for chromosome organization during the cell cycle in live E. coli*. Science Reports, 2015. **5**: p. 17133.

256. Fang, G., et al., *Transcriptomic and phylogenetic analysis of a bacterial cell cycle reveals strong associations between gene co-expression and evolution*. BMC Genomics, 2013. **14**: p. 450.

257. Werner, J.N., et al., *Quantitative genome-scale analysis of protein localization in an asymmetric bacterium*. PNAS, 2009. **106**(19): p. 7858-7863.

258. Goley, E.D., A.A. Iniesta, and L. Shapiro, *Cell cycle regulation in Caulobacter: location, location, location*. Journal of Cell Science, 2007. **120**(Pt 20): p. 3501-3507.

259. Quardokus, E.M., N. Din, and Y.V. Brun, *Cell cycle and positional constraints on FtsZ localization and the initiation of cell division in Caulobacter crescentus*. Molecular Microbiology, 2001. **39**(4): p. 949-959.

260. Thanbichler, M. and L. Shapiro, *MipZ, a spatial regulator coordinating chromosome segregation with cell division in Caulobacter*. Cell, 2006. **126**(1): p. 147-162.

261. Mohl, D.A. and J.W. Gober, *Cell cycle-dependent polar localization of chromosome partitioning proteins in Caulobacter crescentus*. Cell, 1997. **88**(5): p. 675-684.

262. Mika, J.T. and B. Poolman, *Macromolecule diffusion and confinement in prokaryotic cells*. Current Opinions in Biotechnology, 2011. **22**(1): p. 117-126.

263. Dorman, C.J., *Genome architecture and global gene regulation in bacteria: making progress towards a unified model?* Nature Reviews Microbiology, 2013. **11**(5): p. 349-355.

264. Boles, T.C., J.H. White, and N.R. Cozzarelli, *Structure of Plectonemically Supercoiled DNA*. Journal of Molecular Biology, 1990. **213**(4): p. 931-951.

265. Ubbink, J. and T. Odijk, *Electrostatic-undulatory theory of plectonemically supercoiled DNA*. Biophysical Journal, 1999. **76**(5): p. 2502-2519.

266. Vologodskii, A.V. and N.R. Cozzarelli, *Conformational and Thermodynamic Properties of Supercoiled DNA*. Annual Review of Biophysics and Biomolecular Structure, 1994. **23**: p. 609-643.

267. Bliska, J.B. and N.R. Cozzarelli, *Use of Site-Specific Recombination as a Probe of DNA-Structure and Metabolism Invivo*. Journal of Molecular Biology, 1987. **194**(2): p. 205-218.

268. Erickson, H.P., *Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy*. Biological Procedures Online, 2009. **11**: p. 32-51.

269. Marko, J.F. and E.D. Siggia, *Statistical mechanics of supercoiled DNA*. Physical Review E, 1995. **52**(3): p. 2912-2938.

270. Luan, B. and A. Aksimentiev, *DNA Attraction in Monovalent and Divalent Electrolytes*. Journal of the American Chemical Society, 2008. **130**(47): p. 15754-15755.

271. Huang, J., T. Schlick, and A. Vologodskii, *Dynamics of site juxtaposition in supercoiled DNA*. PNAS, 2001. **98**(3): p. 968-973.

272. Smith, S.B., Y.J. Cui, and C. Bustamante, *Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules*. Science, 1996. **271**(5250): p. 795-799.

273. Wang, M.D., et al., *Stretching DNA with optical tweezers*. Biophysical Journal, 1997. **72**(3): p. 1335-1346.

274. Frank-Kamenetskii, M.D., et al., *Torsional and bending rigidity of the double helix from data on small DNA rings*. Journal of Biomolecular Structure and Dynamics, 1985. **2**(5): p. 1005-1012.

275. Hagerman, P.J., *Flexibility of DNA*. Annual Review of Biophysics and Biophysical Chemistry, 1988. **17**: p. 265-286.

276. Stigter, D., *Interactions of Highly Charged Colloidal Cylinders with Applications to Double-Stranded DNA*. Biopolymers, 1977. **16**(7): p. 1435-1448.

277. Vologodskii, A. and N. Cozzarelli, *Modeling of Long-Range Electrostatic Interactions in DNA*. Biopolymers, 1995. **35**(3): p. 289-296.

278. Yoo, J. and A. Aksimentiev, *The structure and intermolecular forces of DNA condensates*. Nucleic Acids Research, 2016. **44**(5): p. 2036-2046.

279. Beutler, T.C., et al., *Avoiding Singularities and Numerical Instabilities in Free-Energy Calculations Based on Molecular Simulations*. Chemical Physics Letters, 1994. **222**(6): p. 529-539.

280. Fischer, S. and M. Karplus, *Conjugate Peak Refinement - an Algorithm for Finding Reaction Paths and Accurate Transition-States in Systems with Many Degrees of Freedom*. Chemical Physics Letters, 1992. **194**(3): p. 252-261.

281. Voss, N.R. and M. Gerstein, *3V: cavity, channel and cleft volume calculator and extractor*. Nucleic Acids Research, 2010. **38**(Web Server issue): p. W555-W562.

282. Xu, S., M.V. Kamath, and D.W. Capson, *Selection of Partitions from a Hierarchy*. Pattern Recognition Letters, 1993. **14**(1): p. 7-15.

283. Klenin, K. and J. Langowski, *Computation of writhe in modeling of supercoiled DNA*. Biopolymers, 2000. **54**(5): p. 307-317.

284. Rangannan, V. and M. Bansal, *PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes*. BMC Research Notes, 2011. **4**: p. 257.

285. Mao, F., et al., *DOOR: a database for prokaryotic operons*. Nucleic Acids Research, 2009. **37**(Database issue): p. D459-463.

286. Uchiyama, I., et al., *MBGD update 2013: the microbial genome database for exploring the diversity of microbial world*. Nucleic Acids Research, 2013. **41**(Database issue): p. D631-D635.

287. Minton, A.P., *The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media*. J Biol Chem, 2001. **276**(14): p. 10577-80.

288. Weiss, M., et al., *Anomalous subdiffusion is a measure for cytoplasmic crowding in living cells*. Biophys J, 2004. **87**(5): p. 3518-24.

289. van den Berg, B., R.J. Ellis, and C.M. Dobson, *Effects of macromolecular crowding on protein folding and aggregation*. EMBO J, 1999. **18**(24): p. 6927-33.

290. Banks, D.S. and C. Fradin, *Anomalous diffusion of proteins due to molecular crowding*. Biophys J, 2005. **89**(5): p. 2960-71.

291. Chow, E. and J. Skolnick, *DNA Internal Motion Likely Accelerates Protein Target Search in a Packed Nucleoid*. Biophys J, 2017. **112**(11): p. 2261-2270.

292. Jung, J., et al., *GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations*. Wiley Interdiscip Rev Comput Mol Sci, 2015. **5**(4): p. 310-323.

293. Dorman, C.J., *Genome architecture and global gene regulation in bacteria: making progress towards a unified model?* Nat Rev Microbiol, 2013. **11**(5): p. 349-55.

294. Yeh, I.C. and G. Hummer, *Diffusion and electrophoretic mobility of single-stranded RNA from molecular dynamics simulations*. Biophys J, 2004. **86**(2): p. 681-9.

295. Saberi, S. and E. Emberly, *Chromosome driven spatial patterning of proteins in bacteria*. PLoS Comput Biol, 2010. **6**(11): p. e1000986.