Computer Simulation Study of Molecular Recognition in Model DNA Microarrays

Arthi Jayaraman, Carol K. Hall, and Jan Genzer
Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, North Carolina

ABSTRACT DNA microarrays have been widely adopted by the scientific community for a variety of applications. To improve the performance of microarrays there is a need for a fundamental understanding of the interplay between the various factors that affect microarray sensitivity and specificity. We use lattice Monte Carlo simulations to study the thermodynamics and kinetics of hybridization of single-stranded target genes in solution with complementary probe DNA molecules immobilized on a microarray surface. The target molecules in our system contain 48 segments and the probes tethered on a hard surface contain 8–24 segments. The segments on the probe and target are distinct and each segment represents a sequence of nucleotides (~11 nucleotides). Each probe segment interacts exclusively with its unique complementary target segment with a single hybridization energy; all other interactions are zero. We examine how the probe length, temperature, or hybridization energy, and the stretch along the target that the probe segments complement, affect the extent of hybridization. For systems containing single probe and single target molecules, we observe that as the probe length increases, the probability of binding all probe segments to the target decreases, implying that the specificity decreases. We observe that probes 12–16 segments (~132–176 nucleotides) long gave the highest specificity and sensitivity. This agrees with the experimental results obtained by another research group, who found an optimal probe length of 150 nucleotides. As the hybridization energy increases, the longer probes are able to bind all their segments to the target, thus improving their specificity. The hybridization kinetics reveals that the segments at the ends of the probe are most likely to start the hybridization. The segments toward the center of the probe remain bound to the target for a longer time than the segments at the ends of the probe.

INTRODUCTION

DNA microarrays have revolutionized the way biological research is done, enabling scientists to measure the expression patterns of thousands of genes in a single experiment. Microarrays are being used in a wide variety of applications (1–3):

1. To identify which genes are preferentially expressed by which cells,
2. To reconstruct the metabolic pathways for cell operation,
3. To identify which genes are differentially expressed in healthy versus diseased cells, enabling disease diagnosis and the development of drugs that can exploit this difference,
4. To discover therapeutic drugs tailored to the genetic profiles of patients, the ultimate goal being personalized medicine,
5. To screen for environmental toxins or pathogens based on changes in genetic profiles of exposed organisms, and
6. To facilitate legal identification.

A microarray is a small glass or nylon slide containing thousands of single-stranded genes or gene fragments immobilized on the surface in spots arranged in a grid, with one gene represented per spot (4). Fluorescently labeled single-strand target molecules in a sample solution exposed to the microarray surface bind specifically and hybridize to complementary probe molecules immobilized on the microarray surface. The molecular recognition of the target genes by the appropriate probe molecules is a consequence of the Watson-Crick basepairing rules. The four different nucleotides that make up a single-stranded DNA molecule—adenine (A), thymine (T), guanine (G), and cytosine (C), are compelled to bind (pair) to their basepair complements. Subsequent analysis of the pattern of fluorescence on the microarray surface allows scientists to identify the genes in the DNA sample solution and to determine their abundance.

The two main measures of microarray performance are sensitivity and specificity (5–11). Sensitivity refers to the hybridization signal/noise ratio. A high signal/noise ratio indicates high sensitivity and therefore, a more efficient detection of the genes under study. Specificity refers to the ability to discriminate between different nucleotides (8). The probes should be designed to discriminate between target and nontarget molecules differing by as little as a single nucleotide. The higher the specificity, the less likely is cross hybridization and generation of false positives. Current understanding of how to design microarrays for maximal sensitivity and specificity is limited, due in part to the shortage of publicly-available data on optimum design. Therefore there is a need for a fundamental understanding of the principles that govern the interplay between the various factors that affect microarray performance. Such knowledge is essential to fully exploit the incredible potential of microarrays.

Submitted March 29, 2006, and accepted for publication May 31, 2006.

Address reprint requests to C. K. Hall, E-mail: hall@eos.ncsu.edu.

0006-3495/06/09/2227/10 $2.00

doi: 10.1529/biophysj.106.086173

In recent years investigators have begun to examine the influence of various factors on microarray performance to optimize sensitivity and specificity (5–13). The factors that influence microarray sensitivity and specificity include: the choice of probe molecule sequence, length, and concentration; the target molecule sequence, length, and concentration; the probe and target nucleotide (G-C) compositions, the spacer length, and the temperature. Ramdas et al. (6) have evaluated experimentally the effect of oligonucleotide probe length and concentration on signal intensity (sensitivity) in microarrays. They observed that the signal intensity increases linearly with the length of the oligonucleotide. The signal intensity also increased as the probe concentration increased, although the effect of probe concentration on the signal intensity was minimal compared to the effect of probe length. Relogio et al. (9) have shown in vitro that while 60-nucleotide (60-mer) long oligonucleotide probes had 10-times the sensitivity of a 25-mer probe, they had much lower specificity than the 25-mer probes. Chou et al. (5) observed similar results, further suggesting that the addition of spacers could improve the signal intensity of short probes. Letowski et al. (8) have shown that probes with mismatches distributed over the entire length or at the center of the probe have higher specificity than probes with mismatches at the 3′ or 5′ ends. They also found that hybridizations done at temperatures 8–13°C below the T_{m} (melting temperature) of perfectly matching probes improved the specificity of the probe. Peterson et al. (7) have used surface plasmon resonance spectroscopy to study the effect of probe density on the kinetics of hybridization. They reported that at low probe densities, almost 100% of the probes hybridize and the kinetics of binding follows Langmuir-like behavior, whereas at high probe densities only 10% of the probes hybridize and the kinetics of binding is slow.

Investigators have also used computer simulations to gain insight into the structure and dynamics of DNA at the molecular level, to interpret experimental data, and to test analytical theories. Many different approaches have been taken in the modeling of DNA via simulations. The level of detail used in representing the geometry and energetics of DNA molecules depends on what aspects of DNA behavior one wishes to investigate. Atomic-resolution models provide the most realistic description of DNA geometry and energetics. The force fields are represented by empirical potentials that account for the intra- and intermolecular interactions between all of the atoms in the system (except hydrogen, in some cases). Atomistic simulations of DNA are generally performed using the traditional molecular dynamics (MD) method. Since the first MD simulations of nucleic acids (14,15) researchers have been able to reproduce the experimentally observed standard structures of single-stranded DNA (16,17), double-stranded DNA (18–21), and other motifs including anomalous structures (22–28). The dynamics of hybridization of DNA tethered to a surface and the effect of the surface on the conformation of the DNA has also been studied using all-atom MD simulations (29–31). Extensive reviews of the use of molecular dynamics for the simulation of a variety of nucleic acid systems can be found in the literature (19,20,32–39). It is important to note, however, that all-atom MD simulations usually have an already hybridized double-helix form of a DNA as an initial configuration and are used to study the stability of the double-helix DNA structure. Although all-atom MD simulations of DNA fragments can be performed on the nanosecond timescale within current computer capabilities, many of the physical and biological DNA processes of interest such as replication, transcription, and denaturation are observed at longer timescales. Also, simulations of large multichain systems of oligonucleotides with atomic detail are not feasible within the computational power currently available.

To simulate the behavior of DNA at longer timescales with current computational power, intermediate-resolution and low-resolution models have been developed (40–62). These models do not require many parameters yet provide a good general picture of DNA behavior. Most intermediate-resolution models are based on the bead-spring model. Each bead (united atom) can represent either a complete nucleotide unit or one of the three nucleic acid components: sugar, phosphate, and base. The focus of the study determines the level of coarse-graining in the prospective model. Garcia de la Torre and co-workers (40,41) have developed a model for short DNA chains in which each nucleotide is represented by one bead (united atom). The introduction of a series of stiff springs connecting neighboring beads on the same and complementary strands induces the system to adopt a helical conformation, allowing the calculation of helix dimensions and persistence length. Mergell et al. (42) introduced a generic model with basepairs represented as rigid ellipsoids and sugar-phosphate backbones represented as semirigid springs; this model was used to explore the local stacking and helical properties of DNA and the behavior under stretching. Drukker et al. (43,44) have developed a DNA model where each nucleotide is represented by two beads. One bead represents a backbone site (sugar plus phosphate) and the other the base. The model includes noncovalent and covalent interactions, bending and torsional angle contributions, and angle-dependent hydrogen bonding between bases. Next-nearest neighbor bonds along the backbone sites were introduced to produce stable duplexes. The model described DNA melting of a double-helix structure into single strands and is based on a previous two-dimensional representation of DNA (45) as a sequence of rigid bodies (base plus sugar) connected by flexible rods. Recently, Tepper and Voth (46) have developed a model for double-helix molecules in solution with explicit solvent and short-range interactions. Individual beads in the model do not represent specific groups of atoms but are evenly distributed to present a uniform distribution of interactions (hydrophobic/hydrophilic interactions, base-stacking interactions, and repulsion between adjacent phosphates along the backbone). The beads on the two strands are covalently linked. The twisting of this model into a double helix was examined as a function of the input parameters. The covalent linking between the basepairs on the two strands prevents application of this model to denaturation processes (i.e., single-strand separation) or hybridization (joining together of two single strands).

Low resolution (simple) models have been used to investigate macroscopic aspects and understand the basic physics behind the DNA behavior that stems from the polymeric nature of DNA molecules (polynucleotides). The best example of this class of models is the wormlike chain model (47–50), which represents the DNA molecule as a stiff polymer chain with variable persistence length; these models have been used to study DNA supercoiling and condensation (51–53). Another example is the elastic model, which represents the DNA molecule as a flexible rod or cylinder under the influence of ionic interactions (54–59). Lattice models have also been used to study denaturation of DNA (60,61) and the effect of stretching on nucleic acids (62).

All of the work discussed above provides valuable information on the structure and dynamics of DNA, but so far no one has used these models to specifically study hybridization of multiple probes and targets in DNA microarrays. The goal of our work is to use computer simulations to develop a comprehensive general understanding of the physical principles that govern the hybridization of target DNA molecules to probe DNA molecules in microarrays. We use Monte Carlo simulations of coarse-grained lattice-model DNA molecules on model microarray surfaces to uncover the basic physics underlying the hybridization process. The lattice model and the Monte Carlo simulation method give us the advantage of high computational speed. This in turn helps us to access the long timescales (approximately minutes) (63) involved in probe-target hybridization and makes the study of large system sizes feasible within current computational capabilities. Our work should culminate in a molecular level description of the hybridization process and a set of general guidelines for maximizing microarray sensitivity and specificity.

Our system consists of a single probe molecule tethered to a hard surface and a single target molecule. The probe and target molecules are modeled as self-avoiding chains on a cubic lattice. Each of the segments on the probe molecule recognizes and preferentially binds to its uniquely complementary segment on the target with an attractive interaction potential, ϵ. We examine how the hybridization of the probe and the target is affected by variation in probe length, hybridization strength ϵ, and the position of the complementary segments on the target.

Highlights of our results are the following. As the probe length decreases, the probability of binding all probe segments to complementary segments on the target increases. This in turn increases the ability of the probe to discriminate between perfectly complementary and partially complementary targets. This implies that shorter probes have higher specificity, which is in qualitative agreement with the experimental work done by Relogio et al. (9) and Chou et al. (5). In contrast, the longer probes have a higher probability than the shorter probes of binding to the target, which implies that longer probes have higher sensitivity. There is an optimum probe length (= 12 statistical segments) where we observe both good specificity and good sensitivity. Probes with segments that are complementary to the segments at either end of the target have higher specificity than probes with segments that are complementary to the central portion of the target. At strong hybridization strength ($\epsilon = 4$ kT), the probes tend to bind all of their segments to the target. At low hybridization strength ($\epsilon = 2$ kT), the probes tend to stay unbound or bind only a few segments to the target. At intermediate hybridization strength ($\epsilon = 3$ kT), the probes prefer to bind either short stretches (2–4 segments) or all of the segments to the target. Our study of the hybridization kinetics reveals that the segments at the ends of the probe are most likely to start the hybridization. The segments toward the center of the probe remain bound to the target for a longer time than the segments at the ends of the probe. The latter leads us to believe that the specificity of the probes is high if the mismatches in the target lie in the region complementary to the center portion of the probe, which is also observed experimentally by Letowski et al. (8).

The remainder of this article is organized as follows: Model and Method describes the molecular model and the simulation method. Results and Discussion describes our simulation results. A brief summary of our findings is provided in the Conclusion.

MODEL AND METHOD

We use lattice Monte Carlo simulation for our study since it is extremely fast and thought to faithfully mimic the large scale conformations of polymer chains. Our system consists of a single target molecule in solution and a single probe molecule tethered to a hard surface through a spacer. The probe and target are modeled as self-avoiding chains placed on a cubic lattice. The segments on the probe and target are distinct, i.e., instead of having four types of segments A, T, G, C, we have as many types of segments as there are segments along the probe. Each of the segments on the probe represents a sequence of nucleotides along the DNA single strand. We assume the dimension of each segment to be of the order of magnitude of the persistence length of a single-stranded DNA. The persistence length of a single-stranded DNA ranges from 0.8 nm to 5 nm depending on the ionic strength of the solvent (0.1–1 mM, respectively) (64). We assume the size of the segment to be 5 nm, which justifies modeling the DNA as a flexible chain. Since the rise per basepair for single-stranded DNA is 0.43 nm (64,65), each segment along the probe or target corresponds to ~11 nucleotides in a single-stranded DNA molecule. Each probe segment recognizes (preferentially attracts) its uniquely complementary segment on the target with an attractive interaction potential to mimic binding of complementary nucleotide pairs (A-T, G-C) on DNA. In other words, the i^{th} segment on the probe is complementary only to the j^{th} segment on the target, the $i + 1^{\text{th}}$ segment on the probe is complementary only to the $j + 1^{\text{th}}$ segment on the target, and so on. We refer to the attractive interaction potential between the complementary probe-target segments as the hybridization energy, ϵ. The interactions are only between segments that are nonbonded nearest-neighbors on the lattice. All other interactions in the system are zero. The dimensions of the simulation box in the x, y, and z directions are 48, 48, and 80, respectively. In the z direction, there is a hard surface at $z = 81$ and at $z = 0$. The target contains 48 segments. The probes contain 8–24 segments and are tethered to the surface at $z = 0$.

The initial configuration of the target is obtained by first placing the head-segment of the chain on a random location in the lattice. The second segment is placed on one of the six sites adjacent to the head-segment. The third segment is placed on a site next to the second segment, and this is repeated until the target chain is grown to the desired length. During this initialization process, if there is no vacant site for adding a segment onto one end of the chain, the segment is added to the other end of the chain. If adding the segment to either end fails due to the absence of a vacant site, the chain is moved in the box using reptation, kink jump, end moves, and crankshaft moves (66) until a vacancy is created. In the case of the probe, the head-segment is placed on a random location on the surface at $z = 1$ and the rest of the chain is grown in the same way as the target chain. We do not allow movement of the probe along the surface so the position of the head-probe segment is fixed. The other probe segments are moved in the box using a combination of kink jump, end moves, and crankshaft moves (66).

The simulation proceeds in three stages: initialization, equilibration, and production. In the initialization stage the system runs through 100,000 Monte Carlo steps (MC steps). In each MC step on average each segment along the target and probe is picked randomly and moved using a random combination of reptation, end moves, kink jump moves, and crankshaft moves (66). The moves are accepted or rejected based on the Metropolis algorithm (67). The initialization stage helps us avoid any bias that might arise due to the nature of the initial configuration of the chains. In the equilibration stage, the system goes through 8,000,000 MC steps, during which the standard chain moves are made to let the system equilibrate. In the production stage (an additional 5,000,000 MC steps) we obtain data on the property of interest after each 100 MC steps and calculate the block averages for every 100,000 MC steps. The equilibrium average for the desired property is the mean of all the block averages. We obtain equilibrium averages from 20 simulation trials; error bars are determined from the standard deviations. To quantify the extent of hybridization, we calculate the probability that a contiguous stretch of target segments binds to the complementary probe segments. The error bars throughout the article are within 11% of the value of the dependent variable.

RESULTS AND DISCUSSION

We study how the following factors affect the extent of hybridization of the probe to the target: 1), the probe length; 2), the hybridization energy between the segments on the probe and target, ϵ; and 3), the stretch along the target molecule that the probe is chosen to complement. In this study we consider probes of length 8, 12, 16, 20, and 24 segments and hybridization energies, ϵ, of 2 kT, 3 kT, and 4 kT. We also consider three types of probe based on the stretch along the target molecule that the probe is chosen to complement. Fig. 1 shows a cartoon of the three types of probes. The chain with black segments represents the target, and the chain with shaded segments represents the probe and the white segments represent the spacer. Table 1 contains the positions of the complementary segments on the target for each type of probe and every probe length. For example, for probe length = 8, in the end-type probe, the segments are complementary to segments 1–8 of the target; in the mid-type probe, the segments are complementary to segments 13–20 of the target; and in the center-type, the segments are complementary to segments 21–28 of the target.

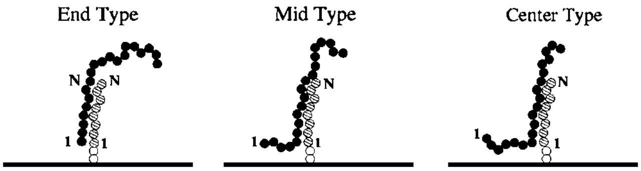

FIGURE 1 A schematic for the three types of probes for probe-length 8 and target of length 48 (target not drawn to scale). The segments of the probe are complementary to (*left panel*) segments 1–8 along the target (*End Type*); (*middle panel*) segments 13–20 along the target (*Mid Type*); and (*right panel*) segments 21–28 along the target (*Center Type*).

Effect of probe length

In Fig. 2 we plot the probability of binding a contiguous stretch of target segments to the probe versus number of contiguous bound target segments for probe-lengths 8, 12, 16, 20, and 24. The results shown in Fig. 2 are for a system with a single target that is 48-segments long, an end-type probe with a spacer four-segments long, and $\epsilon = 3$ kT. It is important to note that when we calculate the probability of binding a contiguous stretch of target segments, we do not break this stretch into smaller stretches and hence overcount the probability of having smaller stretches. For example, for probe-length 8, when all eight segments on the probe are bound, we only count that as one occurrence for an eight-segment-long contiguous stretch being bound, and not as eight occurrences of one segment being bound or as four occurrences of two segment stretches, etc.

For probe-length 8, the probability of binding all eight probe segments to their complementary target segments (long stretch) is higher than the probability of binding short stretches along the target. This implies that the probe prefers to bind either all of its contiguous segments or only a few contiguous segments to complementary segments on the target. When all of the probe segments bind to the complementary segments on the target, the free energy is lowered because the system gains considerable enthalpy and this is sufficient to overcome the considerable loss in entropy due to the configurational restraints associated with the binding of the segments. When only a few of the probe segments bind to complementary segments on the target, the free energy is lowered because the system gains a modest amount of enthalpy due to favorable interactions, and this is sufficient to overcome the modest loss of entropy due to the configurational constraints on the bound target segments.

TABLE 1 The position of the complementary segments on the target for each type of probe

Probe length	End type	Mid type	Center type
8	1–8	13–20	21–28
12	1–12	13–24	19–30
16	1–16	13–28	17–32
20	1–20	13–32	15–34
24	1–24	13–36	13–36

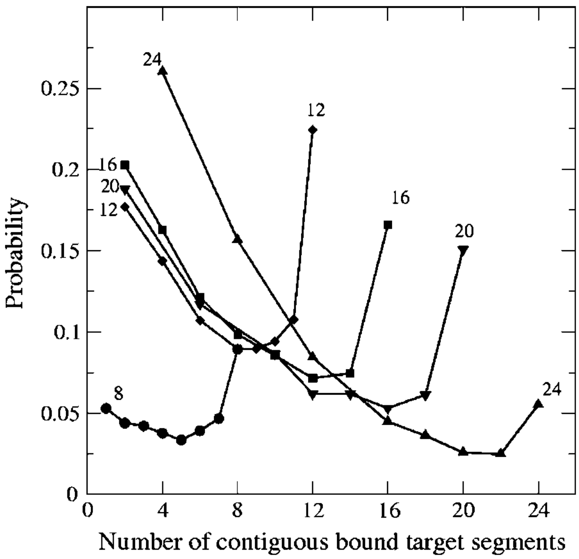

FIGURE 2 Effect of probe length on probability of binding a contiguous stretch of segments along target to the probe segments for end-type probes at $\epsilon = 3$ kT and spacer length = 4.

For probe-length 12 (*diamonds*) the probability of binding all 12 probe segments and the probability of binding small stretches two-segments in length are much higher than the probability of binding intermediate length stretches (~6–10 segments) along the target. The probability of binding long stretches is higher for probe-length 12 than for probe-length 8 because probe-length 8 has only eight segments to attract and bind the target while the probe of length 12 has 12 segments. We also observe that, as the probe length increases from 12 to 24, the probability of binding only few segments in the probe to the target increases, but the probability of binding all segments in the probe to the target decreases. This is because as the probe length increases, the enthalpic gain in binding all of the probe segments to the target is not high enough to overcome the entropic loss upon binding, the latter of which increases dramatically with chain length.

These results can be interpreted in terms of specificity and sensitivity. Shorter probes (except for probe-length 8) have a higher probability of binding all the probe segments to the target than longer probes. This means that shorter probes are better able to distinguish between matches and mismatches in the target and hence have higher specificity than longer probes. These results qualitatively agree with those obtained experimentally by Relogio et al. (9) and Chou et al. (5). Among all the probe lengths, probe-length 12 has the highest probability to bind all its segments to the target and therefore has the highest specificity.

As mentioned before, sensitivity is a measure of how well the microarray can detect rarely expressed genes. Therefore, the more often a fluorescently-labeled target is detected by the probe, the higher the fluorescence and hence the signal intensity (higher sensitivity). Sensitivity is also measured as the ratio of the number of targets detected to the total number of targets in solution. Since we have only a single target we cannot calculate the sensitivity. We can, however, make a good guess as to how the sensitivity varies for the different probe lengths by comparing the probability of the different probe lengths to bind the target. The higher the probability of a probe length to bind the target, the higher the sensitivity. We calculate the probability of binding the target as the ratio of number of data points when the probe binds the target by at least one segment to the total number of data points, averaged over 10 simulation trials. We have tabulated the probability of binding the target for the different probes lengths in Table 2. We see that probe-length 16 has the highest values of the probability, thus the highest sensitivity. In a future publication we will make a better assessment of the effect of probe length on the sensitivity by considering the case of multiple targets.

Therefore, we can conclude that probe-lengths 12 and 16 seem to give both good specificity and good sensitivity. Since our model assumes each segment to be ~11 nucleotides, our results suggest that probes molecules with 132–176 nucleotides will give good specificity and sensitivity. This 132–176 nucleotide range agrees with experimental results obtained by Chou et al. (5), who predict that 150 nucleotides is an optimal probe length for expression measurement.

Effect of position of complementary segments in the target

In Fig. 3 we plot the probability of binding all the probe segments to their complementary target segments versus probe length for end-type, mid-type, and center-type probes. The results shown in Fig. 3 are for a system with a single target 48-segments long, and spacer four-segments long at $\epsilon = 3$ kT.

For probe-length 8, the end-type probe has a higher probability than the mid-type and center-type probes, of binding all its segments to the target. This implies that the end-type probe has a higher specificity than the mid-type and center-type probes. The reason behind this is as follows. If the number of probe segments bound to the target is the same for all three types of probes, the enthalpy is equal in all three cases. Thus the free energy is affected only by the entropy term. This leads us to believe that the loss in configurational entropy upon hybridization increases from end type to center type. Thus the extent of hybridization is highest when the

TABLE 2 The probability that a probe of a given length binds to a target of length 48

Probe length	Probability
8	0.184
12	0.666
16	0.924
20	0.828
24	0.688

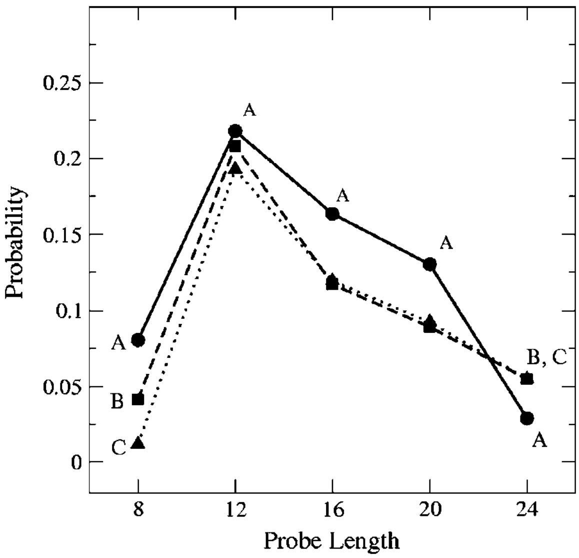

FIGURE 3 Comparison of probability of binding all the probe segments to the complementary target segments versus probe length for end-type (*A*), mid-type (*B*), and center-type probes (*C*).

probe is chosen to complement the end of the target. As the probe length increases from 12 to 16 to 20, the end-type probe continues to have a slightly higher probability than the mid-type and center-type probes for binding all its segments to the target. This is again due to the greater loss of configurational entropy for an end-type probe than a center-type probe. For probe-length 24, the mid-type and center-type probes have a higher probability than the end-type probe to bind all the probe segments to the target. In the case of probe-length 24, the mid-type and center-type probes are identical in terms of the target segments they are chosen to complement, as seen in the last row of Table 1, so we only need to compare end-type and mid-type probes. It is not clear why the center-type probes have a higher probability than the end-type probe to bind all the probe segments to the target (for probe-length 24) and we are not sure if this trend will continue for longer probes. We suspect that with increasing probe length (≥24), the loss of configurational entropy for mid-type (and center-type) probes decreases and becomes less than that of end-type probe. We will investigate this in a future publication.

Effect of hybridization energy ϵ

Fig. 4 shows the effect of varying the strength of the hybridization energy on the extent of hybridization for probe-lengths 8, 12, 16, 20, and 24. We plot the probability of binding all probe segments (Fig. 4 *a*) and short contiguous stretches of two probe segments (Fig. 4 *b*) to the complementary target segments versus varying length of end-type probes with spacers of length 4 at three hybridization energies 2 kT, 3 kT, and 4 kT. In our simulation, increasing hybridization energy, ϵ, can be interpreted as decreasing temperature because the reduced temperature T^* is equal to kT/ϵ.

For probe-length 8, as the interaction strength increases from 2 kT to 4 kT (equivalently T^* decreases from 1/2 to 1/4), the probability of binding all probe segments to their complementary segments on the target increases (Fig. 4 *a*). When $\epsilon = 2$ kT (high T^*), the enthalpic gain upon binding is small and cannot overcome the loss in entropy so the probability of binding all probe segments is low. When $\epsilon = 4$ kT (low T^*), the enthalpic gain upon binding is much higher than when $\epsilon = 3$ kT (intermediate T^*); this can overcome the loss in entropy upon binding, allowing the probe to readily bind all its segments to the target. The probability of binding a short contiguous stretch of two probe segments to the target also increases as the interaction strength increases from 2 kT to 4 kT for probe-length 8, but not by a significant amount (Fig. 4 *b*).

For probe-length 12 (Fig. 4 *a*), as the interaction strength increases from 2 kT to 4 kT, the probability of binding all probe segments to their complementary segments on the target increases. The reason for this is the same as that stated for probe-length 8. The probability of binding a short contiguous stretch of two probe segments for probe-length 12, unlike probe-length 8, is higher for 3 kT than for 4 kT or 2 kT (Fig. 4 *b*). This is because when $\epsilon = 4$ kT, the hybridization energy is so strong that the probe prefers to bind all its segments to the target to maximize the enthalpic gain rather

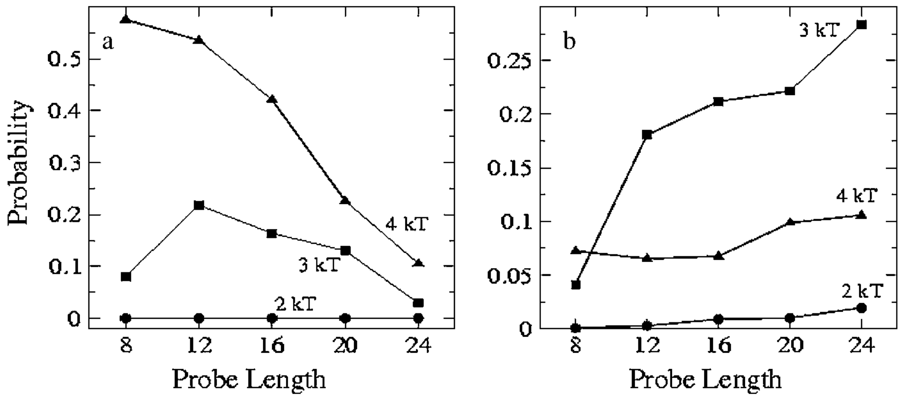

FIGURE 4 Effect of hybridization energy—2 kT (●), 3 kT (■), and 4 kT (▲)—on probability of binding (*a*) all probe segments and (*b*) short contiguous stretch of probe segments to the complementary target segments for end-type probes of varying length with spacers of length 4.

than to have only a few segments bind to the target. The same behavior is observed for probe-lengths 16, 20, and 24.

The effect of increasing ϵ on extent of hybridization can be compared to the effect of increasing the % G+C content on hybridization. It is known that a G-C basepair has three hydrogen bonds whereas an A-T pair has two hydrogen bonds; therefore, the higher the %G+C content in the probe, the higher the strength (ϵ) of binding to the complementary bases on the target. Therefore, our results suggest better specificity will be observed for probes with high ϵ (or high %G+C content). This is also seen experimentally by Letowski et al. (8), who have shown that even at higher temperature, probes with high %G+C content achieve better specificity than probes with low %G+C content. Furthermore, high ϵ (or high %G+C content) leads to the probe binding more readily to the target segments, which in turn leads to high sensitivity.

Effect of spacer length

Chou et al. (5) have shown that addition of spacers to short probes enhances the hybridization intensity by pushing the probe into the solution to improve the chances of target capture. Furthermore, they have shown that the effect of spacers is more prominent in the case of short oligonucleotide probes than in the case of long DNA probes. To study the effect of spacers, we plot in Fig. 5 the probability of finding a contiguous stretch of target segments bound to a probe of length 12 at $\epsilon = 3$ kT when the spacer length is varied from 0 segments to 24 segments. The probability of binding long stretches is the same for all the spacer lengths. The probability of binding short stretches is weakly dependent on the spacer length. In other words, the spacers do not have much of an effect on the extent of hybridization for probe-length 12 or for other probe lengths (not shown here for brevity). Although this contradicts the observation of Chou et al. (5), we suspect that this is because our system contains only a single probe. When there are multiple probes on the surface, there is a crowding effect and the probes try to get away from the neighboring probes by extending away from the surface. The spacers push the probes away from the surface, facilitating the binding of the target to the probe and improving the extent of hybridization. Due to the absence of multiple probes in our system, the spacers do not improve the extent of hybridization.

Kinetics of hybridization

To better understand the mechanism of probe-target hybridization, we analyze the kinetics of hybridization in our model. We begin by obtaining simulation data on which probe segment starts (nucleates) hybridization. Fig. 6 shows us the probability that each probe segment starts the hybridization for an end-type probe. In an end-type probe, the first and the N^{th} segments of the probe (see Fig. 1 for cartoon) are more likely to be nucleation sites for the hybridization process than the segments in the midportion of the probe at all probe lengths. These observations can be explained as follows. In the system containing end-type probes, we know that the first target segment binds to the first probe segment and the N^{th} probe segment binds to the N^{th} target segment and the two segments that have the most freedom of movement are the N^{th} probe segment and the first and last target segments. As a consequence, the N^{th} probe segment and the first probe segment tend to bind to their complementary segments before the other segments in the system do, thereby starting hybridization.

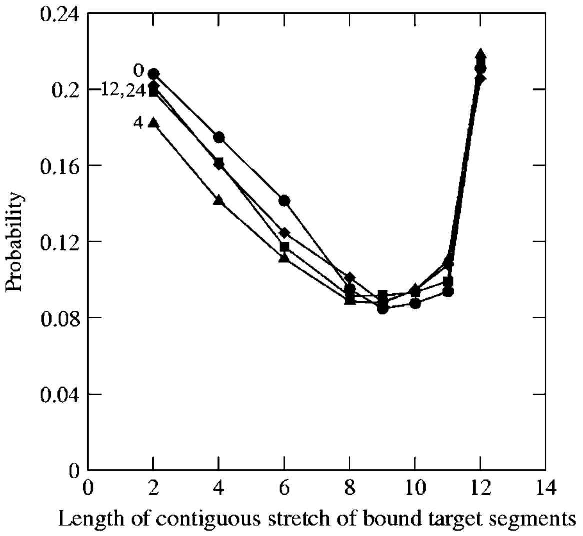

FIGURE 5 Effect of spacers of length 0 (●), 4 (▲), 12 (■), and 24 (♦) on probability of binding a contiguous stretch of segments along the target to the probe segments for end-type probe of length 12 segments.

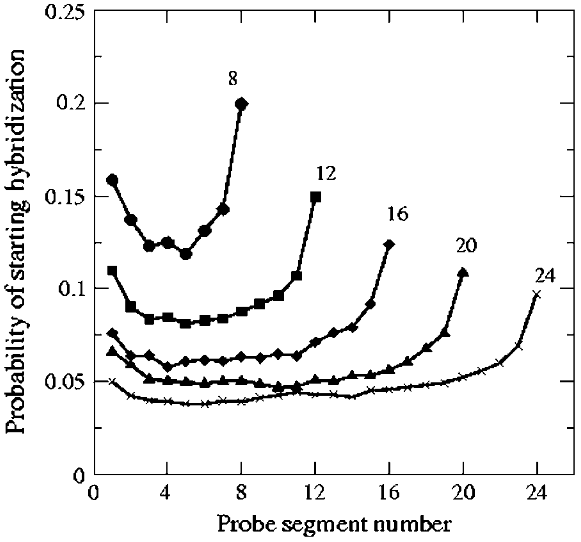

FIGURE 6 Probability of starting hybridization versus probe segment number for varying lengths of end-type probe, at $\epsilon = 3$ kT and spacer length = 4.

Once hybridization has started, the segments in the midportion of the probe are more likely to stay bound to their complements than the segments on the ends of the probe. This is seen in Fig. 7, which shows the fraction of simulation time during which a certain probe segment is bound for an end-type probe. This further suggests that since the probe segments in the midportion are hybridizing more often than the segments at the end, they are more likely to detect mismatches. Therefore end-type probes can detect mismatches in the midsection of the probe better than mismatches at the ends. This has also been observed experimentally by Letowski et al. (8), who found that the specificity of the probe was highest when the mismatches were in the center part of the probe.

CONCLUSION

We have used Monte Carlo simulations with a coarse-grained lattice DNA model to study the thermodynamics and the kinetics of hybridization of single-stranded target genes in solution with complementary probe DNA molecules immobilized on a microarray surface. The target molecules in our system contain 48 statistical segments and the probes tethered on a hard surface contain 8–24 segments. The segments on the probe and target are distinct, with each segment representing a short sequence of nucleotides (~11 nucleotides). Each segment along the probe interacts exclusively with its unique complementary segment on the target molecule with a single interaction (hybridization) energy; all other interactions are zero. We have examined the effect of probe length, hybridization energy (or equivalently temperature), and the position of the complementary segments in the target on the extent of hybridization.

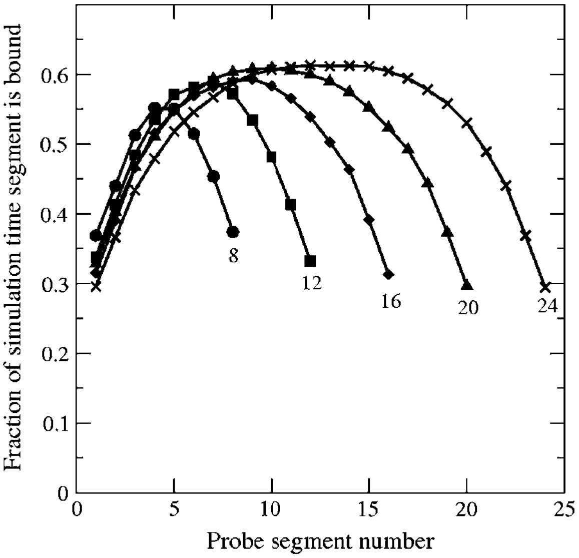

FIGURE 7 Fraction of simulation time that a segment is bound versus probe segment number for varying lengths of end-type probe, at $\epsilon = 3$ kT and spacer length = 4.

In this article we have shown that for systems containing a single probe molecule and a single target molecule, the probability of binding all probe segments to the target decreases as the probe length increases. The higher the probability of binding all probe segments, the higher is the specificity of the probe. Thus we expect shorter probes (12–16 segments) to be more specific than longer probes. Most of the experimental work qualitatively agrees with this trend that shorter probes have higher specificity (5,9). We cannot calculate the sensitivity since we are studying the hybridization of a single target, but we can make a good guess as to how the sensitivity varies for the different probe lengths by comparing the probability of the different probe lengths to bind the fluorescently-labeled target molecule. The higher the probability of a probe length to bind the target, the higher the sensitivity. We see that probes 12- and 16-segments long have higher sensitivity than the other probe lengths. Thus, our results suggest that probes 12- and 16-segments long (~132- and 176-nucleotides-long DNA probes) would give the highest sensitivity and specificity.

We have studied the effect of varying the temperature (T^*) on the extent of hybridization by examining the extent of hybridization at varying strengths of the hybridization energy (ϵ), since $T^* = kT/\epsilon$. As the hybridization energy increases, the longer probes are able to bind all their segments to the target. In other words, as the hybridization energy increases (or % G+C content increases), the specificity of the probe increases. The hybridization kinetics reveals that the segments at the ends of the probe are most likely to start the hybridization. The segments toward the center of the probe remain bound to the target for a longer time than the segments at the ends of the probe. Thus the probes can detect mismatches better if they are positioned toward the center of the probe. This has been observed by Letowski et al. (8), who have shown that when the mismatches are present at the 3′ and 5′-end, the probes were less specific than when the mismatches are present at the center.

It is important to keep the limitations of our model in mind. The model is simple and coarse-grained, and as such does not explicitly consider solvent-mediated interactions, electrostatic interactions, or atomistic details of the DNA molecule (bond angles, torsion angles, stacking interactions). It is important to note that our model is similar to the Poland-Scheraga model for DNA hybridization (68,69). In that model, hybridization between two complementary strands of DNA of equal length could occur only when bases with the same index along the strands bind. This is essentially a Gō-type model (70) which, based on observations for Gō-type protein models, means that the energy landscape will be relatively unfrustrated. We do not know to what extent the consequent low number of trapped complexes will affect the model's predictions for sensitivity and specificity. The behavior of the DNA probes and targets has been predicted based mainly on the chainlike nature of target and probe molecules and the interactions between the complementary segments of the probe and the target molecule. Furthermore, we have only shown the results for a single probe and a single target. As multiple probes and multiple targets are introduced into the system we expect to see a crowding effect with the targets binding simultaneously to multiple probes, leading to lowering of the specificity. We will present this work in a forthcoming publication.

We gratefully acknowledge useful discussions with Dr. Nancy Klauber-DeMore and Dr. Matthew Johnson.

This work was supported by the Office of Energy Research, Basic Sciences, Chemical Science Division of the U. S. Department of Energy under contract No. DE-FG05-91ER14181.

REFERENCES

1. Cummings, C. A., and D. A. Relman. 2000. Using DNA microarrays to study host-microbe interactions. *Emerg. Infect. Dis.* 6:513–525.
2. Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA.* 96:9212–9217.
3. Barrett, J. C., and E. S. Kawasaki. 2003. Microarrays: the use of oligonucleotides and CDNA for the analysis of gene expression. *Drug Discov. Today.* 8:134–141.
4. Amaratunga, D., and J. Cabrera. 2004. Exploration and Analysis of DNA Microarray and Protein Array Data. Wiley Series in Probability and Statistics. Wiley-Interscience, Wiley, New York.
5. Chou, C. C., C. H. Chen, T. T. Lee, and K. Peck. 2004. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res.* 32:e99.
6. Ramdas, L., D. E. Cogdell, J. Y. Jia, E. E. Taylor, V. R. Dunmire, L. Hu, S. R. Hamilton, and W. Zhang. 2004. Improving signal intensities for genes with low expression on oligonucleotide microarrays. *BMC Genomics.* 5:35.
7. Peterson, A. W., R. J. Heaton, and R. Georgiadis. 2001. The effect of surface probe density on DNA hybridization. *Nucleic Acids Res.* 29: 5163–5168.
8. Letowski, J., R. Brousseau, and L. Masson. 2004. Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide arrays. *J. Microbiol. Methods.* 57:269–278.
9. Relogio, A., C. Schwager, A. Richter, W. Ansorge, and J. Valcarcel. 2002. Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.* 30:e51.
10. Kane, M. D., T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50-mer) microarrays. *Nucleic Acids Res.* 28:4552–4557.
11. Urakawa, H., S. E. Fantroussi, H. Smidt, J. C. Smoot, E. H. Tribou, J. J. Kelly, and D. A. Stahl. 2003. Optimization of single basepair mismatch discrimination in oligonucleotide microarrays. *Appl. Environ. Microbiol.* 69:2848–2856.
12. Schultze, A., and J. Downward. 2001. Navigating gene expression using microarrays—a technology review. *Nat. Cell Biol.* 3:E190–E195.
13. Vainrub, A., and M. B. Pettitt. 2003. Sensitive quantitative nucleic acid detection using oligonucleotide microarrays. *J. Am. Chem. Soc.* 125:7798–7799.
14. Levitt, M. 1983. Computer simulation of DNA double-helix dynamics. *Cold Spring Harbor Symp. Quantum Biol.* 47:251–262.
15. Tidor, B., K. K. Irikura, B. R. Brooks, and M. Karplus. 1983. Dynamics of DNA oligomers. *J. Biomol. Struct. Dyn.* 1:231–252.
16. Isaksson, J., S. Acharya, J. Barman, P. Cheruku, and J. Chattopadhyaya. 2004. Single-stranded adenine-rich DNA and RNA retain structural characteristics of their respective double-stranded conformations and show directional differences in stacking pattern. *Biochemistry.* 43:15996–16010.
17. Sen, S., and L. Nilsson. 2001. MD simulations of homomorphous PNA, DNA and RNA single strands: characterization and comparison of conformations and dynamics. *J. Am. Chem. Soc.* 123: 7414–7422.
18. Seibel, G. L., U. C. Singh, and P. A. Kollman. 1985. A molecular dynamics simulation of double helical B-DNA including counterions and water. *Proc. Natl. Acad. Sci. USA.* 82:6537–6544.
19. Cheatham, T. E. I., and P. A. Kollman. 2000. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* 51:435–471.
20. Beveridge, D. L., and K. J. McConnell. 2000. Nucleic acids: theory and computer simulation. *Curr. Opin. Struct. Biol.* 10:182–196.
21. Feig, M., and M. B. Pettitt. 1997. Experiment vs. force fields: DNA conformation from molecular dynamics simulations. *J. Phys. Chem. B.* 101:7361–7363.
22. Shields, G. C., C. A. Laughton, and M. Orozco. 1997. Molecular dynamics simulations of the d(T.A.T) triple helix. *J. Am. Chem. Soc.* 119:7463–7469.
23. Luo, J., and T. C. Bruice. 1998. Nanosecond molecular dynamics of hybrid triplex and duplex of polycation deoxyribonucleic guanidine strands with complimentary DNA strand. *J. Am. Chem. Soc.* 120:1115–1123.
24. Weerasinghe, S., P. E. Smith, V. Mohan, Y. K. Cheng, and M. B. Pettitt. 1995. Nanosecond dynamics and structure of a model DNA triple-helix in saltwater solution. *J. Am. Chem. Soc.* 117:2147–2158.
25. Spackova, N., I. Berger, and J. Sponer. 1999. Nanosecond molecular dynamics simulations of parallel and antiparallel guanine quadruplex DNA molecules. *J. Am. Chem. Soc.* 121:5519–5534.
26. Spackova, N., I. Berger, and J. Sponer. 2001. Structural dynamics and cation interactions of DNA quadruplex molecules containing mixed guanine/cytosine quartets revealed by large scale MD simulations. *J. Am. Chem. Soc.* 123:3295–3307.
27. Spector, T. I., T. E. I. Cheatham, and P. A. Kollman. 1997. Unrestrained molecular dynamics of photodamaged DNA in aqueous solution. *J. Am. Chem. Soc.* 119:7095–7104.
28. Miaskiewicz, K., J. Miller, M. Cooney, and R. Osman. 1996. Computational simulations of DNA distortions by a *cis,syn*-cyclobutane thymine dimer lesion. *J. Am. Chem. Soc.* 118:9156–9163.
29. Hagan, M. F., and A. K. Chakraborty. 2004. Hybridization dynamics of surface immobilized DNA. *J. Chem. Phys.* 120:4958–4968.
30. Wong, K.-Y., and M. B. Pettitt. 2004. Orientation of DNA on a surface from simulation. *Biopolymers.* 73:570–578.
31. Wong, K.-Y., and M. B. Pettitt. 2001. A study of DNA tethered to surface by an all-atom molecular dynamics simulation. *Theor. Chem. Acc.* 106:233–235.
32. Olson, W. K. 1996. Simulating DNA at low resolution. *Curr. Opin. Struct. Biol.* 6:242–256.
33. Auffinger, P., and E. Westhof. 1998. Simulations of the molecular dynamics of nucleic acids. *Curr. Opin. Struct. Biol.* 8:227–236.
34. Lafontaine, I., and R. Lavery. 1999. Collective variable modelling of nucleic acids. *Curr. Opin. Struct. Biol.* 9:170–176.
35. Cheatham, T. E. I., and M. A. Young. 2001. Molecular dynamics simulations of nucleic acids: successes, limitations and promise. *Biopolymers.* 56:232–256.
36. Norberg, J., and L. Nilsson. 2002. Molecular dynamics applied to nucleic acids. *Acc. Chem. Res.* 35:465–472.
37. Giudice, E., and R. Lavery. 2002. Simulation of nucleic acids and their complexes. *Acc. Chem. Res.* 35:350–357.
38. Orozco, M., M. Perez, A. Noya, and F. J. Luque. 2003. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* 32: 350–364.

39. Cheatham, T. E. I. 2004. Simulation and modelling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* 14:360–367.

40. de la Torre, J. G., S. Navarro, and M. C. L. Martinez. 1994. Hydrodynamic properties of a double-helical model for DNA. *Biophys. J.* 66:1573–1579.

41. Huertas, M. L., S. Navarro, M. C. L. Martinez, and J. G. de la Torre. 1997. Simulation of the conformation and dynamics of a double-helical model for DNA. *Biophys. J.* 73:3142–3153.

42. Mergell, B., M. R. Ejtehadi, and R. Everaers. 2003. Modelling DNA structure, elasticity and deformations at the basepair level. *Phys. Rev. E.* 68:021911.

43. Drukker, K., and G. C. Schatz. 2000. A model for simulating dynamics of DNA denaturation. *J. Phys. Chem. B.* 104:6108–6111.

44. Drukker, K., G. Wu, and G. C. Schatz. 2001. Model simulations of DNA denaturation dynamics. *J. Chem. Phys.* 114:579–590.

45. Zhang, F., and M. A. Collins. 1995. Model simulations of DNA dynamics. *Phys. Rev. E.* 52:4217–4224.

46. Tepper, H. L., and G. A. Voth. 2005. A coarse-grained model for double-helix molecules in solution: spontaneous helix formation and equilibrium properties. *J. Chem. Phys.* 122:124906.

47. Marko, J. F., and E. D. Siggia. 1995. Stretching DNA. *Macromolecules.* 28:8759–8776.

48. Marko, J. F., and E. D. Siggia. 1995. Statistical mechanics of supercoiled DNA. *Phys. Rev. E.* 52:2912–2938.

49. Marko, J. F. 1998. DNA under high tension: overstretching, undertwisting and relaxation dynamics. *Phys. Rev. E.* 57:2134–2149.

50. Balaeff, A., A. Mahadevan, and K. Schulten. 1999. Elastic rod model of a DNA loop in the Lac operon. *Phys. Rev. Lett.* 83:4900–4903.

51. Schlick, T., and W. K. Olson. 1992. Supercoiled DNA energetics and dynamics by computer simulations. *J. Mol. Biol.* 223:1089–1119.

52. Chirico, G., and J. Langowski. 1994. Kinetics of DNA supercoiling studied by Brownian dynamics simulation. *Biopolymers.* 34:415–433.

53. Bouchiat, C., and M. Mezard. 1994. Elasticity model of a supercoiled DNA molecule. *Phys. Rev. Lett.* 80:1556–1559.

54. Abascal, J. L. F., and J. C. G. Montoro. 2001. Ionic distribution around simple B-DNA models. III. The effect of ionic charge. *J. Chem. Phys.* 114:4277–4284.

55. Montoro, J. C. G., and J. L. F. Abascal. 1995. Ionic distribution around simple DNA models. I. Cylindrically averaged properties. *J. Chem. Phys.* 103:8273–8284.

56. Montoro, J. C. G., and J. L. F. Abascal. 1998. Ionic distribution around simple DNA models. II. Deviations from cylindrical symmetry. *J. Chem. Phys.* 109:6200–6210.

57. Lyubartsev, A. P., and L. Nordenskiold. 1997. Monte Carlo simulation study of DNA polyelectrolyte properties in the presence of multivalent polyamine ions. *J. Phys. Chem.* 101:4335–4342.

58. Allahyarov, E., H. Lowen, and G. Gompper. 2003. Adsorption of monovalent and multivalent cations on DNA molecules. *Phys. Rev. E.* 68:061903.

59. Allahyarov, E., G. Gompper, and H. Lowen. 2004. Attraction between DNA molecules mediated by multivalent ions. *Phys. Rev. E.* 69:041904.

60. Carlon, E., E. Orlandini, and A. L. Stella. 2002. Roles of stiffness and excluded volume in DNA denaturation. *Phys. Rev. Lett.* 88: 198101.

61. Causo, M. S., C. Barbara, and P. Grasberger. 2000. Simple model for the DNA denaturation transition. *Phys. Rev. E.* 62:3958–3973.

62. Etchegoin, P., and R. C. Maher. 2003. A simple model for the mechanical stretching of (bio)polymers. *Physica A.* 323:551–560.

63. Fritz, J., E. B. Cooper, S. Gaudet, P. K. Sorger, and S. R. Manalis. 2002. Electronic detection of DNA by its intrinsic molecular charge. *Proc. Natl. Acad. Sci. USA.* 99:14142–14146.

64. Williams, M. C., J. R. Wenner, I. Rouzina, and V. A. Bloomfield. 2001. Effect of pH on the overstretching transition of double-stranded DNA: evidence of force-induced DNA melting. *Biophys. J.* 80:874–881.

65. Voet, D., and J. G. Voet. 1995. Biochemistry. John Wiley and Sons, Canada.

66. Verdier, P. H., and W. H. Stockmayer. 1962. Monte Carlo calculations on the dynamics of polymers in dilute solution. *J. Chem. Phys.* 36: 227–235.

67. Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087–1092.

68. Poland, D., and H. A. Scheraga. 1970. Theory of Helix-Coil Transition in Biopolymers. Academic Press, New York.

69. Garel, T., and H. Orland. 2004. Generalized Poland-Scheraga model for DNA hybridization. *Biopolymers.* 75:453–467.

70. Ueda, Y., H. Taketomi, and N. Gō. 1978. Studies on protein folding, unfolding and fluctuations by computer simulation. II. A three-dimensional lattice model of lysozyme. *Biopolymers.* 17:1531–1548.