Monte Carlo analysis of conformational transitions in superhelical DNA

Hongzhi Sun
Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York 10029

Mihaly Mezei
Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, New York 10029

Richard Fye
Sandia National Laboratory, Albuquerque, New Mexico 87185-5800

Craig J. Benham
Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York 10029

(Received 27 March 1995; accepted 16 August 1995)

Metropolis–Monte Carlo algorithms are developed to analyze the strand separation transition in circular superhelical DNA molecules. Moves that randomize the locations of unpaired regions are required in order to diminish correlations among the sampled states. This approach enables accurate simulations to be performed in reasonable computational times. Sufficient conditions to guarantee the formal correctness of the complete algorithm are proven to hold. The computation time required scales at most quadratically with molecular length, and is approximately independent of linking difference. Techniques are developed to estimate the sample size and other calculation parameters needed to achieve a specified accuracy. When the results of Monte Carlo calculations that use shuffling operations are compared with those from statistical mechanical calculations, excellent agreement is found. The Monte Carlo methodology makes possible calculations of transition behavior in cases where alternative approaches are intractable, such as in long molecules under circumstances where several runs of open base pairs occur simultaneously. It also allows the analysis of transitions in cases where the base pair separation energies vary in complex manners, such as through near-neighbor interactions, or the presence of modified bases, abasic sites, or bound molecules. © *1995 American Institute of Physics.*

I. INTRODUCTION

In vivo DNA is constrained into topological domains, within which superhelical stresses are modulated by topoisomerase enzymes. Many biological activities of DNA are regulated by the level of superhelicity imposed. Examples include the initiation of replication[1,2] and of transcription,[3,4] recombination,[5] and the uptake of homologous single strands.[6] Negative DNA superhelicity has long been known to destabilize the helix, inducing strand separations at specific locations.[7–9] Strand separation is an essential step in many superhelically modulated regulatory events. It is required for initiation of transcription and of replication, and also may be implicated in recombination, transposition, and other activities. The best characterized example involves *oriC*, the unique *E. coli* replication origin.[1] Experiments have shown that a specific location within this origin is susceptible to superhelical strand separation. If its base sequence is modified in a way that retains this susceptibility, *in vivo* origin activity is preserved. Sequence changes that degrade this susceptibility or move the site of separation, even by less than 100 bps, destroy *in vivo* origin function. No other attribute of this site or its base sequence affects activity.

Because superhelical strand separation plays essential roles in many biological functions of DNA, it is important to develop quantitatively precise methods to analyze it. The theoretical analysis of strand separation in superhelical DNA is complicated by the global nature of the constraint, and by the heteropolymeric character of the transition. Which duplex sites are destabilized depends in part on local sequence attributes, with separation energetically favored to occur at A+T-rich sites under normal physiological conditions. But superhelicity globally couples together the secondary structures of every base pair in the molecule. Transition at any one location alters the helicity there, which, by changing the distribution of superhelicity throughout the molecule, alters the level of stress experienced by every other base pair. The probability of transition at a particular site depends not just on its local sequence, but also on how transition there competes with all other possible transitions elsewhere in the molecule. This global coupling distinguishes superhelical strand separation from the standard Ising model in linear molecules, where the only coupling is between near neighbors.

The first approximate statistical mechanical analysis of this phenomenon was developed by Anshelevich *et al.*[10] There, a DNA molecule that actually was circular was regarded as being linear, with one inseparable base pair added at each end to decrease end effects. A partition function was calculated for that linear molecule by a standard recursion algorithm, and the results were modified by a renormalization step intended to account for the effects of superhelicity. This approach does not impose the true closed circular topological constraint on the superhelical DNA. Moreover, strand separated regions were regarded as being torsionally undeformable. Because the persistence length of DNA single strands is two orders of magnitude smaller than that of the *B*-form duplex,[11] large torsional deformations of separated regions require small amounts of energy. This can greatly affect the distribution of superhelicity and the extent of tran-

sition throughout the molecule. The utility of this method and the accuracy of its results are severely limited by the approximations made, particularly the assumption of undeformability. More recently, this approach has been improved by the imposition of self-consistency conditions on the renormalization step.[12]

A second approximate analysis focused on the extent of transition as a function of temperature.[13] Homopolymeric energetics were assumed, so every base pair required the same separation energy, and the statistical mechanical partition function was replaced by the largest term in the sum which expresses it. Although that analysis gives a reasonable explanation for the changes in the temperature dependence of transition between linear and superhelical DNAs, it does not provide a generally applicable formulation of the problem addressed here.

The most accurate approximate statistical mechanical method developed to date to analyze superhelical strand separation uses a different strategy.[14–16] An energy threshold is specified, and all states of strand separation are found whose free energies exceed that of the minimum energy state by no more than this threshold amount. The cumulative influence of the high energy states (those not satisfying the threshold condition) is estimated through a density of states calculation. From this data an approximate partition function is calculated, and approximate ensemble averages are determined. The accuracy of this approach is always high, and can be specified by appropriate placement of the energy threshold. The correct topological condition is imposed on the DNA, and the torsional deformability of separated regions is included. The results of calculations using this approach have been shown to agree exactly with experimental measurements of the extents and locations of superhelical strand separation in all molecules examined to date.[16] While this method is accurate and tractable for short sequences (less than 15 000 bps), it has two important limitations. First its numerical implementation is computationally tractable only when the states satisfying the energy threshold condition have small numbers r of open regions, typically $r<4$. In consequence, it cannot be applied accurately to long sequences in which numerous A+T-rich regions compete for transition. Also, the energetics of separation are assumed to be copolymeric, depending only on whether the base pair is AT or GC, and do not include near-neighbor effects. The limitation to copolymeric energetics precludes the analysis of effects on transition behavior of the presence of defects, such as apurinic sites, pyrimidine dimers, or chemically modified bases. The methods developed in this paper are tested for accuracy against calculations made using this technique, as these are known to be highly precise in the cases they can handle.[15,16]

This paper develops Monte Carlo simulation procedures to analyze strand separation in superhelical DNAs of specified sequence. Monte Carlo methods have been used to analyze the tertiary structure of superhelical DNA,[17–20] but they have not been applied to secondary structural transitions to date. Although this approach samples the equilibrium distribution instead of calculating it, it has several advantages over all existing alternative techniques. It can treat systems with complicated transition energetics, including near-neighbor effects and the presence of local inhomogeneities such as abasic sites, methylated or otherwise modified bases, or bound molecules. The topological state of the DNA is modeled correctly, and torsional deformations of the denatured regions are included. It can treat transitions at any temperature and in molecules of any size. There are no restrictions on the number of separated regions it can handle. The computation time required for a simulation scales at most quadratically with molecular length and in practice is nearly linear.

Development of efficient sampling methods requires the use of several types of fundamental moves, which are applied in a pattern. These are shown to satisfy the ergodicity and detailed balance conditions required to sample states according to the equilibrium distribution. By the careful use of a class of moves called shuffling operations, the convergence properties of the algorithm can be optimized. The development procedure used here extends the approach of Hastings,[21] who first applied Markov chain theory to the design of Monte Carlo simulation methods. A similar approach has been presented by Kandel *et al.*[22] in the case of cluster algorithms. We develop a more general approach for designing Monte Carlo sampling algorithms.

II. STRAND SEPARATION IN SUPERHELICAL DNA

Closed circularity fixes the linking number Lk of a DNA molecule. Its linking difference is $\theta=Lk-Lk_0$, where Lk_0 is the linking number of the relaxed state. When a molecule is negatively superhelical, $\theta<0$, the resulting stresses can destabilize the duplex, driving local strand separations.[7–9]

Consider a molecule containing N base pairs, supercoiled to a linking difference θ. A state of strand separation is determined by specifying the secondary structure of each base pair, so there are 2^N such states. Let $m_i=1$ if base pair i is separated and $m_i=0$ otherwise. For closure we specify $m_{N+1}=m_1$. A state of strand separation is determined by specifying the values of each m_i, $1\leqslant i\leqslant N$. The number r of runs of separation (i.e., open regions) in that state is

$$r=\sum_{i=1}^{N} m_{i+1}(1-m_i)$$

and the total number of separated base pairs is

$$n=\sum_{i=1}^{N} m_i .$$

The superhelical deformation θ is partitioned into three types of conformational changes, each of which requires free energy. Separation of the specified base pairs requires free energy

$$G_{\text{sep}}=ar+\sum_{i=1}^{N} m_i b_i . \tag{2.1}$$

Here a is the nucleation free energy needed to initiate an open region, and r is the number of open regions. Under normal physiological conditions, $a\cong 10.5$ kcal/mol,[16,23] with most of this free energy needed to disrupt the extra stacking

interaction required to open the first base pair in a run. Also, b_i is the free energy needed to separate base pair i, $1 \leqslant i \leqslant N$. A separation energy b_i is assigned to each base pair individually, and can depend on near-neighbor effects, the presence at specific sites of lesions which partially disrupt the duplex, or other factors.

Second, the two strands within a separated region can rotate around each other. If n separated base pairs are torsionally deformed at a rate of τ rad/bp, the required free energy is

$$G_{\text{tor}} = \tfrac{1}{2} C n \tau^2, \tag{2.2}$$

where C is the torsional stiffness associated to this deformation. Alternatively, one may allow different values of τ at each position in a separated region. In the sample calculations reported here, we assume every separated base pair experiences the same τ. This is done because these calculations are designed to test the accuracy of the presently developed Monte Carlo methods against statistical mechanical techniques in which the variability of τ was not considered.

Finally, the residual superhelicity remaining to stress the duplex is θ_r, the balance of θ not accommodated by either strand separation or interstrand twisting in the separated regions. This requires a free energy that has been shown to be quadratic in the deformation:

$$G_{\text{res}} = \tfrac{1}{2} K \theta_r^2, \tag{2.3}$$

where K is an experimentally measured constant.[16,23,24] If τ and θ_r are allowed to equilibrate in a state having n separated base pairs in r runs, then the free energy associated to that state is[15]

$$G(n, n_{\text{AT}}, r) = \frac{2\pi^2 CK}{4\pi^2 C + Kn} \left(\theta + \frac{n}{10.5} \right)^2 + ar + \sum_{i=1}^{N} m_i b_i . \tag{2.4}$$

The sample calculations reported below test the accuracy of the Monte Carlo procedures developed here by comparing their results with those from the previously developed statistical mechanical technique whose accuracy is known to be high.[15,16] For this reason we use the same expression for the free energy associated to a state that was used in that earlier work. Specifically, the separation energy is assigned either of two values, b_{AT} or b_{GC}, depending on the identity of the base pair involved, and τ is assumed to be constant and to equilibrate with θ_r. We reiterate, however, that the Monte Carlo procedures developed here can accommodate a wide range of more complex situations which the older procedure cannot handle.

All calculations reported here use the free energy parameter values appropriate to the experimental conditions of Kowalski *et al.*,[9,16] in which $T = 310$ °K and $[\text{Na}^+] = 0.01$ M. These are $b_{\text{AT}} = 0.258$ kcal/mol, $b_{\text{GC}} = 1.305$ kcal/mol, $C = 3.6$ kcal/rad^2, and $K = 2350RT/N$. The results found by the approximate statistical mechanical method using these values have been shown to agree precisely with experiment,[16] within the limits of experimental accuracy.

III. DEVELOPMENT OF MONTE CARLO METHODS

In this work we use the Metropolis Monte Carlo procedure, in which a fundamental move (also called a Monte Carlo trial) consists of the following two steps:

(I) Generate a candidate state j from the current state i according to a specified probabilistic rule.

(II) Use the energy difference between the two states to determine which to select as the new current state.

In step II we use the Metropolis criterion: We accept state j with probability $\exp[(G_i - G_j)/kT]$, if $G_j > G_i$, and with probability 1 if $G_j \leqslant G_i$. The procedure developed below uses a sequence of fundamental moves F_i performed in a specified pattern $(F_1, \ldots, F_m)$.

In order for this procedure to sample states in a distribution that converges to the equilibrium distribution in the limit of large sample size, it must satisfy two conditions. The rules employed must be capable of reaching every state (ergodicity), and the generation of states (step I) must be unbiased (detailed balance). These properties are best expressed using matrices.

The state generation step (step I) of a fundamental move F determines a square matrix M of dimension $W \times W$, where W is the number of states of the system. The possible current states of the system correspond to the rows of the matrix, and the candidate states to the columns. The element in the ith row and jth column of this matrix is p_{ij}, the probability that one will generate state j as candidate state given that the current state is i. The fundamental move F satisfies detailed balance if $p_{ij} = p_{ji}$ for all i, j so that M is symmetric.[21] Because some candidate state must be generated, the entries in each row sum to unity. This is the probability normalization condition.

Now, consider a simulation in which fundamental moves are performed in a pattern $(F_1, \ldots, F_m)$. For this simulation to satisfy detailed balance, each fundamental move must satisfy detailed balance so that the state generation matrices M_i, $i = 1, \ldots, m$, must all be symmetric. Ergodicity will be satisfied if every entry in the product generating matrix $R = M_1 M_2 \cdots M_m$ is positive. This means that the simulation is capable of reaching any final state from every initial state with positive probability in one repeat of the pattern. Because ergodicity can be guaranteed by a somewhat weaker condition, we call the positivity of R strong ergodicity.

Suppose a simulation consists of a pattern of fundamental moves of two types $(F_1, \ldots, F_m, S_1, \ldots, S_n)$, all of which satisfy detailed balance. If the subpattern $(F_1, \ldots, F_m)$ satisfies strong ergodicity, then the whole pattern will be (strongly) ergodic. This follows from the fact that if a $W \times W$ matrix A has every entry positive and another $W \times W$ matrix B has only non-negative entries with at least one positive entry in each row and column, then the product AB will have all positive entries. Moves of the F type are called basic operations and moves of S type are called shuffling operations.

Suppose a Monte Carlo algorithm has been constructed consisting of a pattern of fundamental moves satisfying strong ergodicity and detailed balance. Although this will guarantee that the algorithm samples the states in a way that

converges to the equilibrium distribution in the limit of large sample size, if successively sampled states are strongly correlated the rate of convergence will be very slow, often rendering the technique useless in practice. One recourse in such circumstances is to design supplementary shuffling operations that weaken these correlations, thereby producing rapid convergence and efficient simulation procedures. This approach is required to produce tractable algorithms to analyze superhelical strand separation, as shown below. In principle it can be applied to any Monte Carlo procedure in which convergence rates are slow due to the persistence of correlation among successively sampled states.

A. The standard Metropolis–Monte Carlo algorithm

Consider a circular DNA molecule containing N base pairs and supercoiled to a linking difference θ. A standard Monte Carlo algorithm consists of N fundamental moves executed in the pattern $(F_1,\ldots,F_N)$. Here F_i is the operation in which a candidate state is generated by flipping base pair i, either from closed to open or from open to closed, with probability $0<p<1$. We determine whether this candidate state is accepted by making a standard Metropolis decision. Performing this procedure, once per base pair proceeding along the entire length of the molecule, constitutes a standard Monte Carlo cycle (MCC). Because the states of any collection of base pairs can be changed in this process, the probability of passing from any initial state to any final state in one MCC is positive, so the procedure is strongly ergodic. Each move clearly satisfies detailed balance, because in the state generation step of F_i, the conformation of base pair i is flipped with the same probability p, regardless of its initial configuration. To generate a sample distribution using this approach, one sample state is chosen after λ MCCs, where λ is an adjustable parameter. The algorithm implementing this procedure is called MCA.

This procedure generates strongly correlated successively sampled states, even when λ is very large, as a sample calculation demonstrates. The molecule analyzed consists of a dimeric repeat of the sequence between positions 1301 and 3800 of the pBR322 DNA molecule. It contains two diametrically placed copies of the strongly destabilized site present at the β-lactamase gene $3'$ terminal region.[9] At linking difference $\theta=-24$ turns (well within the physiological range), the statistical mechanical analysis finds the probability of exactly one run of separation to be 0.993, with the separation equally likely to occur at either of the two destabilized regions. Figure 1(a) plots the equilibrium probability of separation of each base pair in this molecule, also called the transition profile, calculated using the statistical mechanical procedure. Figure 1(b) shows the profile calculated using the MCA algorithm. Here the initial state is an open loop comprising the base pairs between positions 400 and 500. Successive sample points are chosen after $\lambda=300$ MCCs, and the sample size is $U=20\,000$. We see that in the MCA calculation one site dominates over the other, despite their being identical. Which site is dominant depends on the initial state of the system.

The reason for this behavior is as follows. Because the initiation energy a is large ($a\cong 10.5$ kcal/mol), the probability of accepting a candidate state that differs at one site from the current state but has a larger number of runs is less than 10^{-7}. This means it is very unlikely that the secondary structure of a base pair will be changed if that base pair is interior to a region, be it open or closed. The only single base changes that have a significant chance of acceptance occur at junctions between open and closed regions. Moving between low energy states having the same number of open runs will be extremely improbable if the open regions in the two states are far from each other along the sequence. For the only feasible way to do this is by migration of an open region through single base pair openings and closings. Many such moves would be required, and if G+C-rich regions intervene, these moves will be individually improbable. Global coupling constrains the number of open base pairs, which further increases the difficulty of this movement. In consequence, successively sampled states remain highly correlated, even if very many Monte Carlo cycles are executed between samplings. In the example of Fig. 1(b), the separation is effectively trapped at one susceptible site and is unable to move to the other. These results show that the distribution found by standard Monte Carlo algorithm converges to the equilibrium distribution far too slowly to be useful in practice, despite satisfying ergodicity and detailed balance.

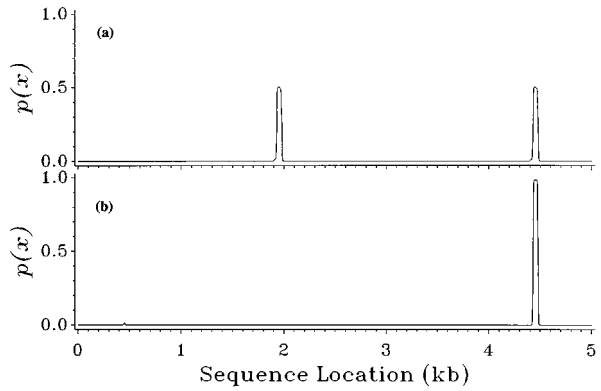

FIG. 1. A dimeric molecule is analyzed that is composed of two copies of a region of the pBR322 molecule, as described in the text. This molecule is supercoiled to a linking difference $\theta=-24$ turns. Part (a) gives the separation probability profile calculated using the approximate statistical mechanical procedure of Benham (Refs. 15 and 16), whose accuracy in evaluating equilibrium properties exceeds 99.9% in this case. One run of separation occurs, which is equally likely to be at either of the two strongly destabilized regions at positions 1900 to 2000 and 4400 to 4500, respectively. Part (b) shows the results of calculation using the MCA Monte Carlo algorithm with an initial state that is opened at positions 400–500. In this procedure the open region is trapped at the second susceptible site, and states in which the other region opens are not sampled.

B. Monte Carlo methods with shuffling operations

Although in principle increasing λ decreases correlations, in practice such large values of λ would be required by the standard Monte Carlo algorithm MCA as to make this approach unfeasible for our problem. To develop a practical algorithm, shuffling operations must be constructed that decorrelate successive sampled states. Because a standard MCC satisfies strong ergodicity and detailed balance, the performance of shuffling operations after a standard MCC

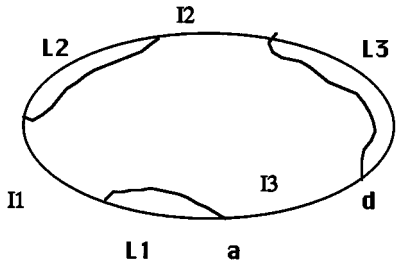

FIG. 2. The shift operation SH33 randomly picks one of the three loops shown, say, $L1$, and then randomly moves it within the closed region bounded by $L2$ on the left and $L3$ on the right. The lengths of all open regions are unchanged in this operation.

will yield a formally correct algorithm provided the shuffling operations satisfy detailed balance, as described above.

Four types of elementary moves have been designed, and four shuffling operations are constructed from them. These elementary moves decrease correlations among states by randomizing the positions and numbers of open regions without changing the total number of separated base pairs. The state generation steps of each type of elementary move are as follows. Rotation moves rotate all open loops a random amount around the circular molecule. This places the open regions at new positions, but keeps their lengths and separation distances fixed. Shift moves alter the relative positions of open regions, without changing their numbers or lengths. Squeeze moves redistribute the open base pairs among the open regions without altering either the total number of open base pairs or the number of open regions. Finally, exchange moves increase or decrease the number of open regions by either amalgamating open regions or subdividing a region. None of these elementary moves alter the total number of open base pairs. After a new state has been generated, a Metropolis decision is made regarding which state to accept. We describe the state generation step in each type of elementary move, and demonstrate that these satisfy detailed balance.

In the rotation move the positions of all open loops are rotated in unison a random amount around the molecule, keeping their lengths and the distances between them fixed. Since the number of possible rotations equals the total number of base pairs, setting the probability of each rotation to be $P=1/N$ assures detailed balance. This move, called ROTATION, can be applied to any state. Alternatively, its use can be restricted, for example to cases when $r=1$ or $r>7$.

Shift moves are applied only when there is more than one open region. Briefly, an open region is selected at random, and moved within the set of closed positions that abut it on either side. To describe this technique more precisely, consider the elementary move SH33, which is applied to states having three runs of separation, illustrated in Fig. 2. Denoting the length of open region i by L_i, the total number of open base pairs is $n=L_1+L_2+L_3$. The closed regions between adjacent open loops have lengths I_i, $i=1, 2, 3$. The order of open loops and closed intervals is shown in the figure. One can only move open loop L_1 within the closed regions I_1 and I_3 that about it. There are (I_3+I_1-1) ways this can be done without merging with the neighbor open regions. This includes the possibility that the segment L_1 keeps its original place. One can move either L_2 or L_3 in the same way within their respective intervals. There are (I_2+I_1-1) ways to move L_2, and (I_3+I_2-1) ways to move L_3, so the total number of possible moves of this type is $2(I_1+I_2+I_3)-3=2(N-n)-3$. This number depends only on the total number n of open base pairs in the three runs.

To perform the shift, one first calculates the probabilities P_1,P_2,P_3, of shifting open regions 1, 2, or 3, respectively, as

$$\begin{aligned} P_1&=(I_1+I_3-1)/(2(N-n)-3),\\ P_2&=(I_1+I_2-1)/(2(N-n)-3),\\ P_3&=(I_3+I_2-1)/(2(N-n)-3). \end{aligned} \tag{3.1}$$

To select which open loop is moved, one generates a random number *rand* in (0,1). If $rand<P_1$, then L_1 is chosen. If $P_1\leqslant rand<P_1+P_2$, then L_2 is selected, and otherwise L_3 is chosen. Suppose L_1 has been selected. To decide where L_1 is to be moved, a second random number *rand2* is generated. The position where L_1 is placed is determined by $\mathrm{INT}(rand2*(I_3+I_1-1))$, where $\mathrm{INT}(x)$ is the integer part of a number x. One places the first open base pair of L_1 in the position that is $\mathrm{INT}(rand2*(I_3+I_1-1))+1$ base pairs away from the end of its neighbor open region in the counterclockwise direction (L_3 in this case). Clearly this state generation step is symmetric: If one can generate state B from state A by a shift operation, one can also generate state A from state B by the same operation. Moreover, the probability of generating any accessible three-run state from any current three run state by this procedure is $1/(2(N-n)-3)$, so detailed balance is satisfied. The other elementary shift moves are constructed similarly. In the algorithms developed here, the shuffling operation SHIFT uses ROTATION to shuffle states having either one run or greater than seven runs, and SH22,...,SH77 to treat the other cases.

Elementary squeeze moves are applied to states having more than one run of separation. Their purpose is to change the distribution of open base pairs among the open runs without changing either the total number of open base pairs or the number of open runs. We randomly select an open run and subdivide it into two parts. One part is kept fixed, and the other is moved across the unique closed run that abuts it, and attached to the open region on the other side. This shrinks one open run, and simultaneously expands its neighbor by an equal amount, so the total number of open base pairs and the lengths of all closed regions remain constant. It can be shown that, from a given r run state with n open base pairs, one can generate $2(n-r)$ possible states by this procedure, independent of the details of the initial state. Moreover, the state generation step of the squeeze move is symmetric: If one can generate state B from state A by a squeeze operation, then one can generate state A from state B by the same operation. Equiprobable squeezing moves are constructed, using an approach analogous to that described above for the design of the shift moves. Being equiprobable, these clearly satisfy detailed balance. The shuffling operation SQUEEZE

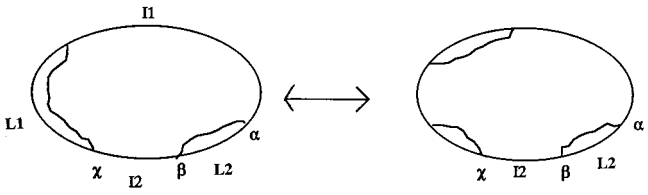

FIG. 3. The exchange move EX23 randomly selects an open region and a direction. Here a portion of $L1$ is moved into closed region $I1$. The selected loop $L1$ is randomly cut at an interior point, and the part proximal to $I1$ is moved into the $I1$ region. This produces the three run state shown on the right. EX32 performs the reverse operation, merging one of the three loops into one of its neighbors.

used in the algorithms developed below applies elementary squeeze moves when the number of open runs satisfies $2 \leqslant r \leqslant 7$, and uses ROTATION when there is either exactly one or more than seven separated runs.

Elementary exchange moves alter the number of separated regions without changing the total number of separated base pairs. This must be done in a way that satisfies detailed balance. As examples, we describe exchange between two run and three run states using Fig. 3. Suppose we start with a two-run state, the runs having lengths L_1 and L_2, respectively. The two-run to three-run move EX23 proceeds as follows. Randomly select an open region and divide it into two subregions. Keep one of these immobile, and move the other to a new position within its unique neighboring closed region. The resulting state has three open regions. The number of possible ways to divide run L_i and place one part within the neighbor closed region I_j is $N(L_i,I_j)=(L_i-1)(I_j-1)$. The length of the portion to be moved can vary from 1 to L_i-1. and there are I_j-1 ways to place that portion within the closed region of length I_j. Summation shows the total number of ways of going from any two-run state with n separated base pairs to any allowed three-run state is $(n-2)(N-n-2)$, which again depends only on n. To shuffle from the present two-run state to any available three-run state in a way that makes each three-run state equally likely to be chosen, we proceed as follows. First we calculate the probabilities $P_{L_iI_j}$:

$$P_{L_iI_j}=\frac{(L_i-1)(I_j-1)}{(n-2)(N-n-2)}. \tag{3.2}$$

To select which open region to cut and where to move the resulting open loop, one generates a random number *rand* in (0,1). If $rand<P_{L_1I_1}$ one cuts L_1 and moves the fragment into the I_1 region. If $P_{L_1I_1} \leqslant rand < P_{L_1I_1}+P_{L_1I_2}$ one cuts L_1 and moves the fragment into the I_2 region. If $P_{L_1I_1}+P_{L_1I_2} \leqslant rand < P_{L_1I_1}+P_{L_1I_2}+P_{L_2I_1}$ one cuts L_2 and moves the segment into the I_1 region. Otherwise, one cuts L_2 and moves the fragment into the I_2 region. Suppose we have decided to cut region L_i, and move the fragment into closed region I_j. Next, we must determine the cut location and the position to which the fragment is moved. We generate a second random number *rand2* in (0,1), and place the cut site after the base pair at position INT(*rand2*$^*(L_i-1))+1$ in L_i. Then we generate a third random number *rand3*, and start the newly created open run at the position INT(*rand3*$^*(I_j-1))+1$ base pairs into closed region I_j. All exchanges from any two-run state to each available three-run state are equiprobable with probability $1/(n-2)(N-n-2)$.

To satisfy detailed balance, each three-run state must have the same probability of exchanging back to any available two-run state. This operation is performed by EX32. Six two-run states can arise by shuffling from any given three-run state. These are the merges of L_i with either L_j or L_k, where i, j, k are all different. The merge of L_i with L_j is performed by moving the location of L_i until it abuts L_j so the two become contiguous. One first generates a random number *rand* in (0,1). If $rand \leqslant 6/(n-2)(N-n-2)$, one chooses either of the six possible merged two-run states with equal probability. The probability of shuffling from the given three-run state to any accessible two-run state is $1/(n-2)(N-n-2)$, so detailed balance for exchange moves is satisfied. If $rand>6/(n-2)(N-n-2)$, then shuffling to two-run states is not performed. In this situation, EX32 shuffles within three-run states using shift operations. Since this operation is performed only when three-run to two-run shuffles are forbidden, the probability of shifting between any two accessible three-run states in this case is

$$\frac{1-6/(N-n-2)(n-2)}{2(N-n)-3}.$$

In our algorithm we construct two categories of exchange shuffling operations, called INTERCHANGE1 and INTERCHANGE2. INTERCHANGE1 is composed of the elementary moves EX12, EX21, EX34, EX43, EX56, EX65 with ROTATION to treat states with run number higher than six. Similarly, INTERCHANGE2 is comprised of EX23, EX32, EX45, EX54, EX67, EX76, together with ROTATION to treat states having run number equal one or greater than seven. The reason for designing two exchange shuffling operations instead of one is to avoid complications in verifying detailed balance. [In each of the INTERCHANGE operations, i-run states can exchange with at most one other type, so the analysis presented above demonstrates detailed balance. It would be much more difficult to construct a provably correct algorithm if i-run states could exchange with both $(i-1)$-run states and $(i+1)$-run states in a single shuffling operation.]

In this implementation we confine shift, exchange, and squeeze steps to states having $r \leqslant 7$ only because states having larger run numbers occur very infrequently in the DNAs we analyze under Kowalski's experimental conditions.[9] For example, our calculations reported below show that phage λ DNA (48 502 bp) supercoiled to a superhelix density $\sigma=-0.055$ ($\theta=-254$ turns) has only a 1% chance of occurring in states having eight or more runs of separation. If needed, the elementary moves that also apply to states having larger run numbers can easily be constructed using the principles described here.

A shuffling Monte Carlo algorithm has been constructed that executes the following pattern. One standard Monte Carlo cycle MCC is followed by shuffling operations $(S_1S_2S_3S_4S_2S_3)^{\mu}$, where S_1 and S_4 are INTERCHANGE1 and INTERCHANGE2, respectively, S_2 is the SQUEEZE

operation and S_3 is the SHIFT operation. Sample calculations showed that S_3 could be eliminated from the pattern because its inclusion did not improve the convergence properties of the simulation. The number of shuffling operations performed after one MCC is $\nu=6\mu$ or $\nu=4\mu$ depending on whether or not S_3 is used in the simulation. This pattern is called a modified Monte Carlo cycle (MMCC). The detailed balance and strong ergodicity conditions sufficient for equilibrium sampling have been shown to hold. The simulation algorithm EXAMD is constructed by performing λ_s MMCCs before picking the next sample state.

The values chosen for ν and λ_s, the tunable parameters of this simulation, have important effects on convergence and simulation time. Design of an optimally efficient shuffling algorithm requires suitable choices of these parameters. Both values should be large enough so that the number of open base pairs and the numbers and locations of open regions can freely vary in the course of selecting the next sampled state. This means the total number of shuffling trials performed to sample one state, $\lambda_s\nu$, must be large enough so that the most infrequently selected shuffling move occurs. In practice, exchange moves have the smallest probabilities of realization [equal to $6/(n-2)(N-n-2)$ in EX32, for example]. In order to weaken correlations between successive sampled states, these types of exchanges must be realized occasionally. This requires a large number of shuffles per sample point. If we select $0.01\bar{n}N\leqslant\lambda_s\nu\leqslant\bar{n}N$, where $\bar{n}$ is the average number of open base pairs, then $\lambda_s\nu$ will be comparable to $(n-2)(N-n-2)$, so exchanges between states having different run numbers would be attempted several times. One can easily find an upper bound estimate for $\bar{n}$ using Eq. (2.4). When the superhelix density $\sigma=\theta/Lk_0$ ranges from -0.04 to -0.055 under the Kowalski[9] environmental conditions, one can show that $\bar{n}\leqslant 0.032N$. If we select ν and λ_s each to be large, with product $\lambda_s\nu\approx\bar{n}N$, then the time needed to generate the next sampled state grows at most quadratically with molecular weight. This shows that the EXAMD algorithm can treat long sequences efficiently.

The computation time t_{total} required by either MCA (MCC without shuffles) or EXAMD (MCC, with shuffles) can be expressed as:

$$t_{\text{total}}\propto[t(N)+t(\nu)]\lambda_s U. \tag{3.3}$$

Here U is the sample size, $t(N)$ is the time needed for one complete standard MCC, which is proportional to N, and $t(\nu)$ is the time needed to perform ν shuffling operations, which is proportional to ν. In MCA, $t(\nu)=\nu=0$ and $\lambda_s=\lambda$, so the simulation time is proportional to $t(N)\lambda U$. In this case λ must be extremely large to weaken strong correlations among the sampled states. Sample calculations on the dimeric molecule whose results are reported in Fig. 1(b) above show virtually permanent correlations, extending over 20 000 sampled points, even when $\lambda=300$. The introduction of the shuffling trials in EXAMD reduces λ to a much smaller value λ_s. In practice, one shuffle requires approximately the same time as the trial of one base pair in an MCC, so the shuffling time $t(\nu)$ of the EXAMD algorithm is much smaller than the time $t(N)$ needed to perform one MCC. Thus, a dramatic increase in efficiency results from the introduction of shuffling operations. The convergence rate is greatly increased because correlations among sampled states are rapidly degraded by shuffling, resulting in a highly efficient algorithm. These improvements are illustrated by sample calculation on the dimeric test molecule using the EXAMD algorithm with $\lambda_s=50$, $\nu=1600$ and 20 000 sample points. Figure 4 shows that EXAMD accurately depicts the transition behavior in this test case. The absolute deviation between these results and the equilibrium distribution is less than 0.008 at every position. The simulation time for this analysis using EXAMD was 8 h, approximately half that required by MCA.

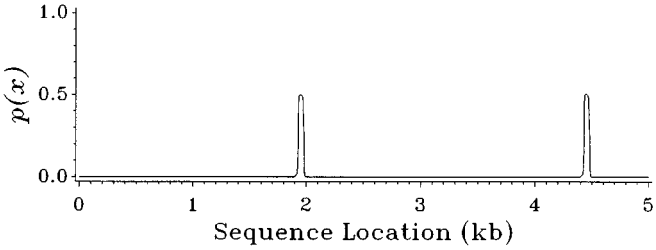

FIG. 4. The EXAMD method is used to calculate the transition profile of the dimeric molecule treated in Fig. 1. The probability of separation of each base pair in the molecule, calculated in this way, deviates from its equilibrium value by no more than ± 0.008. The results do not depend on initial conditions. The simulation time used here is about half that used by MCA to calculate the results shown in Fig. 1(b).

C. A modification of the algorithm

To improve the overall efficiency of the simulation procedure one must reduce the computation time $t(N)$ of the unit MCC. One approach modifies the standard MCC by treating base pairs in blocks rather than individually. This approach is easiest to implement when the transition energies are copolymeric. Although it also can be used in cases having more complicated energetics, such as near-neighbor effects, it becomes more cumbersome to implement and the time it saves becomes smaller.

In a standard MCC, a change in the state of a base pair interior to a region (open or closed) will increase the number of open runs, hence has a very small chance of being accepted. In consequence, the only trials with a significant chance of success in a standard MCC are performed at boundaries of open loops. Consider a closed region whose interior contains n_{AT} AT base pairs and n_{GC} GC base pairs. The probability that all of these base pairs remain closed after one MCC is

$$P_{cbp}=[1-p\ \exp(-\Delta G(\text{AT})/RT)]^{n_{\text{AT}}} \times[1-p\ \exp(-\Delta G(\text{GC})/RT)]^{n_{\text{GC}}}. \tag{3.4}$$

Here p is the probability used in the state generation step of the standard MCC, and $\Delta G(\text{AT})$ [resp. $\Delta G(\text{GC})$] is the total free energy cost of opening one AT (resp. GC) pair in this region. This cost is very large, about 10–12 kcal/mol, because a new run is initialized, so $P_{cbp}\cong 1$. Accordingly, we may modify the standard MCC in the following way. Trials performed at sites interior to or at the boundaries of open

loops will be done in the standard sequential order. Interiors of closed regions are treated by calculating P_{cbp} from the base composition of the region, then generating a random number *rand* in (0,1). If $rand<P_{cbp}$, no change is made in the region involved. Standard trials are commenced at the end of this region and continued until the next modified trial can be performed. If $rand\geqslant P_{cbp}$, then at least one base pair in this interior region will be opened. In this case, the probability that the first open pair is an AT is

$$p_{\mathrm{AT}}=\frac{n_{\mathrm{AT}}\exp[-\Delta G(\mathrm{AT})/RT]}{n_{\mathrm{AT}}\exp[-\Delta G(\mathrm{AT})/RT]+n_{\mathrm{GC}}\exp[-\Delta G(\mathrm{GC})/RT]}. \tag{3.5}$$

We choose a second random number *rand2* in (0,1). If $rand2<p_{\mathrm{AT}}$ then we will open an AT pair in this region. Otherwise we open a GC pair. If the opening base pair is AT, then we choose one of the n_{AT} AT pairs with equal probability, and similarly for opening a GC pair. After performing this opening, we use the standard trials to treat the base pairs next to this position in order until the next interior of a closed region is encountered. One circuit of the molecule performed in this way is called an approximate Monte Carlo cycle (AMCC). This approach cannot be shown to satisfy detailed balance, so any algorithm constructed in this way must be regarded as approximate. However, the practical differences between this approach and a formally exact one are slight because the probability of opening a base pair interior to a closed region is very low.

The algorithm APPMD has been constructed using the pattern of one AMCC, followed by shuffling operations $(S_1S_2S_3S_4S_2S_3)^{\mu}$. The succession of standard and modified trials in the AMCC traverses the molecule quickly. In practice less than 2% of the base pairs are open in a state under normal physiological conditions, so successive trials of individual base pairs are needed at a small fraction of sites. This results in a substantial savings of computer time without sacrificing significant accuracy, as the sample calculations reported below show.

D. Estimates of sample size and other parameters

An estimate of the minimum sample size U_0 needed to achieve a given level of accuracy in a Monte Carlo simulation can be made using the Chebyshev[25] and Kolmogorov[26] inequalities. The variations of accuracy with sample size that are achieved in practice will be described in the Sec. IV.

If $X_1,X_2,\ldots$ are random variables, then the Chebyshev inequality states that

$$P\left\{\left|\frac{X_1+\cdots+X_U}{U}-E\left(\frac{X_1+\cdots+X_U}{U}\right)\right|\geqslant\epsilon\right\} \leqslant\frac{\mathrm{Var}((X_1+\cdots+X_U)/U)}{\epsilon^2}. \tag{3.6}$$

If $X_1,X_2,\ldots$ are assumed independent and identically distributed (denoted by i.i.d.), this becomes Kolmogorov's inequality:

$$P\left\{\left|\frac{X_1+\cdots+X_U}{U}-E(X_1)\right|\geqslant\epsilon\right\}\leqslant\frac{c^2}{U\epsilon^2}, \tag{3.7}$$

where $c^2=\mathrm{Var}(X_1)$. In our problem the sampled states may not be exactly i.i.d. but approach this condition if λ_s (or λ) is sufficiently large that successive points are effectively uncorrelated. If our sampled states are regarded as i.i.d., then several useful estimates can be obtained from this inequality.

Let $B_s(i)$ be the random variable whose value is 1 if base pair s is open in state i, and 0 otherwise. If the equilibrium probability of separation of the base pair s is P_s^B, then inequality (3.7) gives

$$P\left\{\left|\frac{\sum_{i=1}^{u}B_s(i)}{U}-P_s^B\right|\geqslant\epsilon\right\}\leqslant\frac{P_s^B(1-P_s^B)}{U\epsilon^2}. \tag{3.8}$$

Here $\mathrm{Var}(B_s)=P_s^B(1-P_s^B)\leqslant 1/4$. Setting $\epsilon=0.02$ and using the maximum possible variance $\mathrm{Var}(B_s)=1/4$, this inequality states that a simulation having $U=20\,000$ i.i.d. sampled states will have at most a 3.13% chance that its error in estimating P_s^B for any particular base pair exceeds 0.02.

Let r be the random variable corresponding to the number of open runs, so $r(i)=k$, if in state i the number of open runs is k. Now the Kolmogorov inequality states

$$P\left\{\left|\frac{\sum_{i=1}^{U}r(i)}{U}-\bar{r}\right|\geqslant\epsilon\right\}\leqslant\frac{(r_{\max}-r_{\min})^2/4}{U\epsilon^2}, \tag{3.9}$$

where $\bar{r}$ is the average number of open runs. Here $(r_{\max}-r_{\min})^2/4\geqslant\mathrm{Var}(r)$, and $r_{\max}$ and $r_{\min}$ are the maximum and minimum numbers of open runs appearing in the sampled states. In practical simulations we find that $(r_{\max}-r_{\min})^2/4\leqslant 4$. If 20 000 states are sampled, then the ensemble average number of runs will be estimated correctly within ± 0.05 approximately 92% of the time under the i.i.d. assumption.

The same method can be used to estimate the accuracy of calculations of the average number of separated base pairs. One finds an effective maximum number of separated base pairs $n_{\max}$ where the probability of states having more than this number of open base pairs is essentially 0. Similarly, one finds the minimum number $n_{\min}$. Let $\bar{n}$ be the average number of open base pairs and n be the random variable corresponding to the number of open base pairs, i.e., $n(i)=k$, if in state i the number of open base pairs is k. Then we have:

$$P\left\{\left|\frac{\sum_{i=1}^{U}n_i}{U}-\bar{n}\right|\geqslant\epsilon\right\}\leqslant\frac{(n_{\max}-n_{\min})^2/4}{U\epsilon^2}. \tag{3.10}$$

The choices $\epsilon=1$ and $n_{\max}-n_{\min}<80$, are reasonable for a real simulation. This formula shows that when $U=20\,000$ an i.i.d. simulation will estimate the expected number of open base pairs correctly to ± 1 bp 92% of the time.

These evaluations assume the largest possible variance, hence provide worst case estimates of the deviations under the i.i.d. assumption. They indicate that a sample size of $U_0=20\,000$ is reasonable for present purposes. It is not so large as to require very long simulation times, and it suffices for reasonable accuracy.

IV. RESULTS

We have developed three Monte Carlo algorithms for analyzing strand separation transitions in circular superheli-

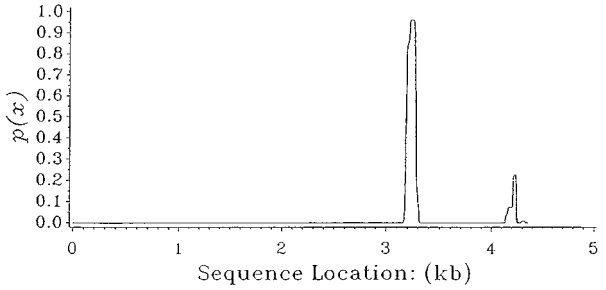

FIG. 5. The transition probablity profile calculated for pBR322 DNA by the approximate statistical mechanical procedure is shown. The calculation assumes linking difference $\theta=-30$ turns, at $[Na^+]=0.01$ M and $T=37°C$. Two regions of high separation tendency are observed both theoretically and experimentally.

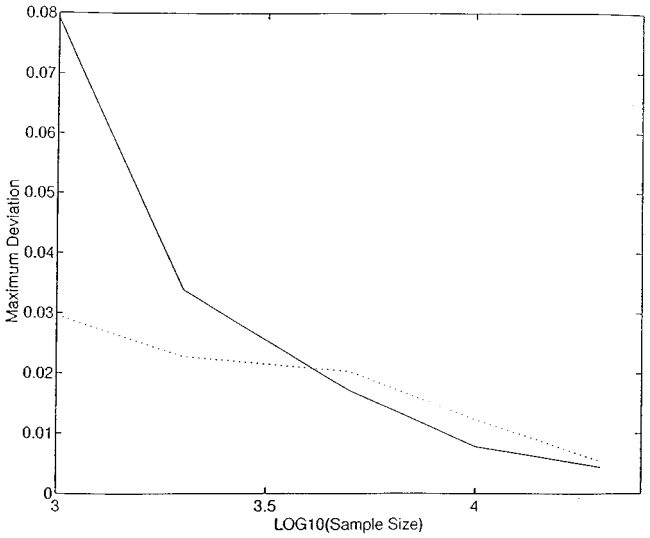

FIG. 6. The absolute maximum deviation between the statistical mechanical and Monte Carlo separation probability in the entire pBR322 DNA sequence, $D=\max_{1\leqslant i\leqslant N}|d(i)|$, is plotted against sample size. The solid line gives the maximum deviation of the APPMD algorithm, and the dotted line gives that of the EXAMD algorithm. The sample sizes are 10^3, 2×10^3, 5×10^3, 10^4, and 2×10^4.

cal DNAs. The standard Monte Carlo algorithm MCA performs Monte Carlo cycles without shuffling. EXAMD augments the standard MCC with shuffling operations. APPMD uses approximate cycles AMCC that have been modified by block estimates for opening of closed regions, together with shuffling operations. The MCA technique has been shown above to suffer extreme convergence problems that preclude its practical utility. In this section we present the results of sample calculations performed using the EXAMD and APPMD algorithms. These results are compared for accuracy with those from approximate statistical mechanical calculations whose precision is known to be very high.[15,16]

A. Tests of accuracy, convergence rate, and computational speed

The first collection of sample calculations were designed to evaluate the accuracy, convergence properties and relative speeds of the EXAMD and APPMD algorithms. For this purpose strand separation in the pBR322 DNA sequence was analyzed at linking difference $\theta=-30$ turns. We performed sample calculations using each of the algorithms with sample sizes $U=1000$, 2000, 5000, 10 000, and 20 000 states. In the EXAMD procedure, we selected the sample states after performing $\lambda_s=50$ modified Monte Carlo cycles (MMCC), with $\nu=1600$ shuffling operations performed after each MCC. In APPMD, values of $\nu=240$ and $\lambda_s=150$ were used. The free energy parameters appropriate for Kowalski's experimental conditions were used.[9,16] To facilitate precise comparison with the statistical mechanical results, the separation energy assumes only two values, b_{AT} and b_{GC}, depending on the identity of the base pair involved.

Figure 5 shows the probability profile calculated by the statistical mechanical technique of Benham[15,16] under these circumstances. That calculation is at least 99.9% accurate in all calculated ensemble averages. Two regions of the pBR322 sequence are shown to be destabilized by stress. Region $R1$ lies between positions 3100 and 3350, while region $R2$ occurs between positions 4100 and 4300. These results agree closely with those from experiments.[9,16]

The probability profiles computed by the Monte Carlo simulation methods both show transition to be confined to the same two regions $R1$ and $R2$. To analyze the accuracy of our algorithms, we subtract the probability $p_{SM}(i)$ of separation of base pair i, calculated using the statistical mechanical method, from its value $p_{MC}(i)$ found by each of the Monte Carlo procedures. This determines the deviation $d(i)=p_{MC}(i)-p_{SM}(i)$ of the Monte Carlo results from the analytical probability profile at each position i. Figure 6 compares the absolute maximum deviations $D=\max_{1\leqslant i\leqslant N}|d(i)|$ of the probability profiles obtained by the EXAMD and APPMD procedures from that obtained by statistical mechanics. The results are reported as functions of sample size. These results show that the two procedures have similar stability and accuracy for sample sizes $U>2000$. Values of other parameters calculated in these simulations are shown in Table I. In all cases the results for EXAMD and APPMD are comparable in accuracy.

It is interesting to compare these simulation results with the Kolmogorov estimates found in last section. Inequality (3.8) shows that the fluctuation in the separation probability P_s^B of the base pair at position s will be smaller than 0.02 with probability

$$p^f(s)\geqslant 1-\frac{P_s^B(1-P_s^B)}{8}, \tag{4.1}$$

assuming i.i.d. samples and $U=20\,000$ sampled states. If the random variables B_s and B_q are independent when s and q are different sites, then the probability P that every site on the pBR322 DNA molecule deviates from its exact value by less than 0.02 is

$$P=\prod_{s=1}^{4363}\left(1-\frac{P_s^B(1-P_s^B)}{8}\right). \tag{4.2}$$

Since P_s^B is known from statistical mechanical calculations, we find $P=0.0392$ in this case. If we exclude those sites where the probability of separation is smaller than 0.03, then

TABLE I. Statistical quantities vs sample sizes. The first column lists the sample size values U for which the simulations are done. Every row is the ensemble average value obtained corresponding to the U indicated in the first column.

EXAMD $\theta=-30$	Open base pairs	Open AT pairs	Open GC pairs	$\langle G\rangle$
$U=1000$	99.206	75.239	23.967	120.375
$U=2000$	99.310	75.329	23.981	120.353
$U=5000$	99.389	75.574	23.815	120.507
$U=10\,000$	99.025	75.150	23.863	120.358
$U=20\,000$	99.069	75.252	23.817	120.413
APPMD $\theta=-30$	Open base pairs	Open AT pairs	Open GC pairs	$\langle G\rangle$
$U=1000$	98.444	74.275	24.169	120.082
$U=2000$	99.292	75.446	23.846	120.439
$U=5000$	98.975	75.147	23.828	120.398
$U=10\,000$	98.934	75.207	23.727	120.434
$U=20\,000$	98.994	75.273	23.721	120.452
Statmech mthd[a]	99.037	75.084	23.953	120.4185

[a]Statistical mechanics approach (Ref. 15). Initial run: 2200–2541. Seed $=-5$.

we find $P=0.0422$. This shows that a simulation with i.i.d. sampling and sample size $U=20\,000$ will have a 96% chance of finding at least one base pair whose deviation from the exact result is larger than 0.02. In our sample calculations the maximum deviations found by EXAMD and APPMD were both less than 0.01 when $U=20\,000$. This suggests that the convergence of both algorithms is comparable to that which would occur if successively sampled points were i.i.d., hence entirely uncorrelated.

To test this claim we performed the APPMD simulation four times using different initial states. Three of these simulations used the sample size $U=20\,000$ and the fourth used $U=32\,000$. No deviations beyond 0.02 were found in any of these simulations. From the worst case Kolmogorov estimate we find that 86% of i.i.d. simulations having $U=32\,000$ would exhibit deviations exceeding 0.02. The chances of four independent simulations all having maximum deviation less than 0.02 is 1.05×10^{-5}. Recall that the accuracy estimates [expressions (3.9) and (3.10)] made from the Kolmogorov inequality used an overestimate of the variance, hence may underestimate the convergence rates correspondingly. We analyzed EXAMD in a similar way, performing six simulations using different starting conditions. In all cases the sample size $U=20\,000$ was chosen. Two of these simulations had maximum deviations smaller than 0.02. The above analysis suggests that the probability of such an occurrence under the i.i.d. assumption is about 0.020. These results indicate that the Monte Carlo sampling procedures that include shuffling operations converge at rates that are comparable to, and possibly even better than, those that would occur under strictly independent sampling. Without shuffling, the sampled states will be confined near local energy minima for long times, so successively sampled states will remain strongly positively correlated indefinitely. But shuffling trials facilitate moves from one local minimum to another, making them comparable in ease to what would occur under independent sampling.

These results show that the EXAMD and APPMD algorithms, both with shuffling trials, give results that converge more rapidly and sample the equilibrium distribution much more effectively than does the standard algorithm MCA, which lacks shuffling. However, APPMD executes significantly faster than does EXAMD. Simulations on a DEC 3000/800 computer for sample size $U=20\,000$ points required 9 h for EXAMD, but only 2.5 h for APPMD. This speedup was achieved without a significant loss of accuracy.

B. Scaling with molecular length

Consider two molecules of different lengths supercoiled to the same superhelix density. On average, the longer molecule usually will have more open base pairs and more open runs, other factors being equal. To see why this occurs, note that the difference in separation energy between AT and GC base pairs under the assumed conditions is approximately 1 kcal/mol, while the energy required to open a run of separation is 10 kcal/mol. Now, consider states having n separated base pairs. Suppose the energetically most favored r-run state contains $n_{\mathrm{AT}}(r)$ AT base pairs. An $r+1$-run state containing n separated base pairs must have A+T richness at least 11 base pairs greater than $n_{\mathrm{AT}}(r)$ to be more energetically favored, because the cost of initiating one more run must be offset by the savings due to the increased A+T richness. This can happen only when the expected number of open base pairs is large. For example, if the AT-richest one run state with $n=60$ open base pairs has $n_{\mathrm{AT}}(1)=50$, then no multiple-run state with the same n can be energetically favored, even if entirely comprised of AT base pairs. States having small numbers of runs of separation (i.e., $r\leqslant2$) are favored when the expected number of separated base pairs is small (roughly $\leqslant100$ bp). For short molecules ($N\leqslant5000$ bp) this occurs throughout the range of physiological linking differences. For long molecules, however, the expected number of runs of separation grows with linking difference.

The complexity of the approximate statistical mechanical technique of Benham[15] increases rapidly with run number. In practice, calculations where states with four or more runs occur are not feasible using this method. The only alternative approach presently available is Monte Carlo simulation.

To assess how the performance of the Monte Carlo algorithm scales with molecular length, we analyzed the phage λ DNA molecule containing 48 502 base pairs. Simulations were performed for various linking numbers using the APPMD algorithm with $\nu=1600$ and $\lambda_s=160$. Each simulation found $U=22\,500$ sampled states. Other physical parameters were the same as in the analysis of pBR322 described above.

Table II shows the distribution of states with r open runs for phage λ DNA at several linking differences θ. When $\theta=-254$, corresponding to the physiological superhelical density $\sigma=-0.055$, six-run states are the most populated. (In pBR322 DNA at this superhelical density the probability of states with more than one run is less than 0.25.) An accurate analysis of this transition is not possible using the approximate statistical mechanical method, due to the large number of runs.

TABLE II. Open run fractions. The first column lists all linking difference values for which the simulations are done. Every row is the results obtained corresponding to the linking difference indicated in the first column.

θ	1-run	2-run	3-run	4-run	5-run	6-run	7-run	8-run
−177	0.295	0.653	0.051	0.001				
−187	0.053	0.705	0.229	0.013				
−197	0.001	0.444	0.472	0.080	0.003			
−207		0.151	0.591	0.237	0.020	0.001		
−217		0.022	0.380	0.492	0.100	0.006		
−227		0.001	0.128	0.564	0.273	0.033	0.002	
−237			0.020	0.355	0.491	0.125	0.008	
−247			0.007	0.104	0.521	0.321	0.046	0.001
−254			0.007	0.041	0.370	0.465	0.102	0.015

Table III shows several average values calculated for phage λ DNA at various linking differences θ using the APPMD algorithm. These include the average numbers of open base pairs, open AT pairs, open GC pairs, and the average free energy $\langle G \rangle$. All these quantities increase approximately linearly with $|\theta|$.

The CPU time required by the DEC 3000/800 computer to perform one simulation on phage λ DNA using APPMD ranged from 19.5 to 24 h, with about a half hour increase for each change of −10 in linking difference. Analogous calculations on pBR322 ($N=4363$ bp) at the same superhelical density required approximately 2.5 h. Thus, in practice the execution time of the APPMD algorithm scales approximately linearly with molecular length.

To test the accuracy of the APPMD algorithm for long molecules, we compared its computed probability profile for phage λ with that calculated by the approximate statistical mechanical method at a small superhelix density where that method retains its accuracy. We choose $\theta=-177$ turns, at which the probability of states with $r \geqslant 4$ runs is 0.001. The maximum deviation D between the profiles calculated by these two methods is $D<0.018$, comparable to that for pBR322 DNA with similar sample size. Thus, the number of sampled states required to achieve a given level of accuracy in a Monte Carlo simulation with shuffling operations is effectively independent of molecular length.

We compared the two Monte Carlo algorithms APPMD and EXAMD by performing simulations on phage λ DNA at linking difference $\theta=-177$. In EXAMD we set $\lambda_s=40$ and $\nu=6400$, while in APPMD we fixed $\lambda_s=160$ and $\nu=1600$. These choices made the total number $\lambda_s \nu$ of shuffling operations performed before picking each sampled state the same for both algorithms. In each case we computed the maximum deviation $D=\max_{1 \leqslant i \leqslant N}|d(i)|$ from the probability profile calculated by the statistical mechanical algorithm. The maximum deviation for EXAMD is $D_E \leqslant 0.026$, and for APPMD it is $D_A \leqslant 0.018$. The deviation between EXAMD and APPMD never exceeds 0.015. Our results show that the accuracies attained by EXAMD and APPMD are comparable for all calculated quantities. The values of $\langle G \rangle$ calculated by these procedures agree within 0.2%, while the expected numbers of separated base pairs agree to better than 2%. Thus, the imprecision caused by the formal failure of the APPMD algorithm to satisfy detailed balance is not significant. However, APPMD executes much faster. When $\theta=-177$, the simulation times were 19.5 h for APPMD and 52 h for EXAMD.

To test the stability of the APPMD algorithm, we made two simulations from different initial states at linking difference $\theta=-247$. The result shows that the two values of $\langle G \rangle$ calculated agree to within 0.04%, and the maximum deviation between the two probability profiles is 0.024. The fact that both APPMD and EXAMD rapidly converge to the equilibrium distribution, as shown by the calculations on pBR322 DNA, also demonstrates their numerical stability.

TABLE III. Statistical quantities vs linking differences. The first column lists all linking difference values for which the simulations are done. Every row is the ensemble average values obtained corresponding to the linking difference indicated in the first column.

θ	Open base pairs	Open AT pairs	Open GC pairs	$\langle G \rangle$
−177	129.24	106.14	23.10	470.36
−187	191.97	154.18	37.79	520.75
−197	258.86	204.98	53.88	571.38
−207	327.65	257.36	70.29	622.39
−217	398.34	311.39	86.95	673.90
−227	468.87	365.04	103.83	725.27
−237	539.91	419.06	120.84	776.75
−247	610.94	473.08	137.86	828.46
−254	660.64	510.67	149.97	864.60

V. DISCUSSION

The Metropolis–Monte Carlo procedures developed here provide a new method for calculating equilibrium properties of the strand separation transition in superhelical DNA. A formally exact method is developed that contains specialized shuffling operations to increase convergence speed. This method is shown to satisfy the ergodicity and detailed balance conditions required for formally correct sampling of the equilibrium. Its convergence properties are shown to be comparable to those achieved with i.i.d. sampling. An alternative approach using composite steps (single base pair tests combined with block region tests) also was developed, which executes significantly faster. Although this APPMD algorithm does not satisfy detailed balance, the accuracy of its results is comparable to that achieved by the formally exact EXAMD procedure. However, we note that APPMD is only efficient when copolymeric transition energetics are used.

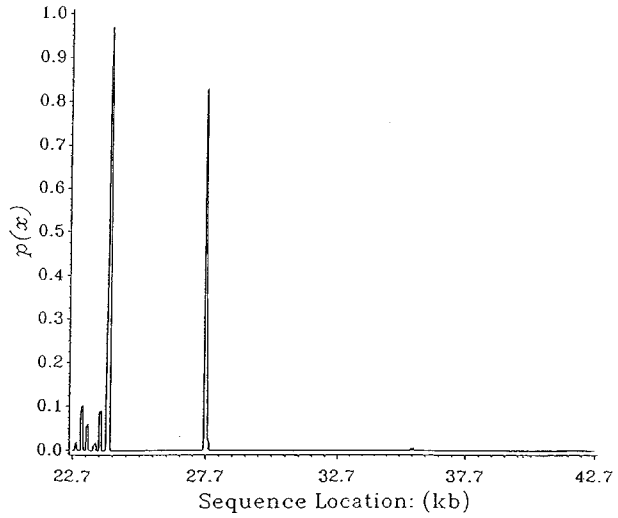

FIG. 7. The transition profile of phage λ DNA calculated at linking difference $\theta=-187$ turns is shown. The part of the sequence that is not plotted showed no destabilization in this calculation.

The results of these Monte Carlo procedures agree closely with those from statistical mechanical calculations, whose accuracy can be made as high as desired by setting a threshold appropriately.[15] The accuracy of these simulation procedures is shown to be approximately as good as what could be expected if the i.i.d. condition held.

The Monte Carlo approach developed here does not have the limitations of alternative methods. The sample size needed to achieve a prescribed accuracy can be estimated in advance, and calculations having that accuracy can be performed for long DNAs at any reasonable linking difference. The execution time grows at most quadratically with molecular length, and in practice behaves approximately linearly. Execution time increases slowly with imposed linking difference, regardless of the number of runs of separation involved.

Although copolymeric transition energetics were used here to facilitate comparison with the results from the statistical mechanical method, the exact Monte Carlo procedure permits calculations with transition energetics having any complexity. Thus, one can include near-neighbor effects and structural modifications such as base methylation, lesion formation, ligand binding, or other alterations that affect transition energetics. Also, calculations of the transition properties of molecules at high temperatures can be perform. These effects cannot be included in the approximate statistical mechanical procedure as currently structured. Calculations analyzing transitions in these situations will be presented elsewhere.

The present Monte Carlo method does have one significant drawback when compared to the approximate statistical mechanical procedure. Using the latter technique one can calculate the incremental free energy needed to separate any base pair in the sequence,[27] thereby finding sites that are partly destabilized by imposed stress. These are sites where superhelicity significantly reduces the energy required for separation, although not enough to induce their opening with significant probability. Such sites may be biologically important, as they may constitute targets for the activities of other molecules. This destabilization energy cannot be accurately calculated using the Monte Carlo method because states in which such sites are separated have a low probability of being sampled.

The results of a Monte Carlo analysis of strand separation in phage λ DNA (48 502 bp) are shown in Fig. 7. That calculation illustrates the ability of this method to treat long DNA sequences. This opens the possibility of analyzing entire sequences the size of eucaryotic topological domains, a feat that is not feasible using the approximate method.

A complete theoretical analysis of superhelical DNA structure must include deformations of tertiary structure as well as the alterations of secondary structure treated here. Monte Carlo statistical sampling methods already have been proposed to treat superhelical tertiary structure.[17–20,28] A central reason for developing Monte Carlo methods to treat secondary structure transitions is because this is the other required step toward handling the complete problem. Once Monte Carlo sampling techniques have been developed separately for the secondary and the tertiary structural aspects of superhelical DNA conformation, one can amalgamate them into a unified technique to analyze superhelical DNA structure in its full generality. This will be the focus of future work.

ACKNOWLEDGMENTS

H. Sun thanks Dr. I. P. Sugar for helpful discussions. This work was supported in part by grants to C. J. Benham from the National Institutes of Health and the National Science Foundation. It will be presented by H. Sun in partial fulfillment of the requirements for the doctoral degree in the Department of Biomathematical Sciences of the Mount Sinai School of Medicine.

[1] D. Kowalski and M. J. Eddy, EMBO J. **8**, 4335 (1989).
[2] M. Mattern and R. Painter, Biochim, Biophys. Acta **563**, 293 (1979).
[3] G. Pruss and K. Drlica, Cell **56**, 521 (1989).
[4] H. Weintraub, P. Cheng, and K. Conrad, Cell **46**, 115 (1986).
[5] E. Richet, P. Abcarian, and H. Nash, Cell **46**, 1011 (1986).
[6] K. L. Beattie, R. C. Wiegand, and C. M. Radding, J. Mol. Biol. **116**, 783 (1977).
[7] J. Vinograd, J. Lebowitz, and R. Watson, J. Mol. Biol. **33**, 173 (1968).
[8] W. Dean and J. Lebowitz, Nature **231**, 5 (1971).
[9] D. Kowalski, D. Natale, and M. J. Eddy, Proc. Natl. Acad. Sci. USA **85**, 9464 (1988).
[10] V. V. Anshelevich, A. V. Vologodskii, A. V. Lukashin, and M. D. Frank-Kameneskii, Biopolymers **18**, 2733 (1979).
[11] V. Bloomfield, D. Crothers, and I. Tinoco, *Physical Chemistry of Nucleic Acid* (Harper & Row, New York, 1974), pp. 258–260.
[12] S. Katsura, F. Makishima, and H. Nishimura, J. Biomol. Struct. Dynam. **10**, 639 (1993).
[13] S. Sen and R. Majumdar, Biopolymers **27**, 1479 (1988).
[14] C. J. Benham, Proc. Natl. Acad. Sci. USA **76**, 3870 (1979).
[15] C. J. Benham, J. Chem. Phys. **92**, 6294 (1990).
[16] C. J. Benham, J. Mol. Biol. **225**, 835 (1992).
[17] S. D. Levene and D. M. Crothers, J. Mol. Biol. **189**, 73 (1986).
[18] K. V. Klenin, A. V. Vologodskii, V. V. Anshelevich, A. M. Dykhne, and M. D. Frank-Kamenetskii, J. Mol. Biol. **217**, 413 (1991).
[19] V. B. Zhurkin, N. B. Ulyanov, A. A. Gorin, and R. L. Jernigan, Proc. Natl. Acad. Sci. USA **88**, 7046 (1991).
[20] M. O. Fenley, W. K. Olson, I. Tobias, and G. S. Manning, Biophys. Chem. **50**, 255 (1994).

[21] W. K. Hastings, Biometrika **57**, 97 (1970).
[22] D. Kandel, R. Ben-Av, and E. Domany, Phys. Rev. Lett. **65**, 941 (1990).
[23] W. R. Bauer and C. J. Benham, J. Mol. Biol. **234**, 1184 (1993).
[24] D. E. Pulleyblank, M. Shure, D. Tang, J. Vinograd, and H. Vosberg, Proc. Natl. Acad. Sci. USA **72**, 4280 (1975).
[25] M. Eisen, *Introduction to Mathematical Probability Theory* (Prentice-Hall, Englewood Cliffs, NJ, 1969), p. 24.
[26] P. A. P. Moran, *An Introduction to Probability Theory* (Clarendon, Oxford, 1968), p. 358.
[27] C. J. Benham, Proc. Natl. Acad. Sci. USA **90**, 2999 (1993).
[28] A. V. Vologodskii, S. D. Levene, K. V. Klenin, M. D. Frank-Kamenetskii, and N. R. Cozzarelli, J. Mol. Biol. **227**, 1224 (1992).