SHUFFLING METROPOLIS ALGORITHMS FOR MONTE CARLO SIMULATIONS AND THEIR APPLICATION TO BIOLOGICAL SYSTEMS: THE STRAND SEPARATION TRANSITION IN SUPERHELICAL DNA AND MEMBRANE PHASE TRANSITIONS

by

Hongzhi Sun

A dissertation submitted to the Graduate Faculty in Biomedical Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1995

This manuscript has been read and accepted for the Graduate Faculty in Biomedical Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

9/28/95
Date

Craig J. Benham
Chair of Examining Committee

9/28/95
Date

Executive Officer

Sylvan Wallenstein

Istvan Sugar

Mihaly Mezei

Supervisory Committee

The City University of New York

Abstract

Shuffling Metropolis Algorithms for Monte Carlo Simulations and Their Application to Biological Systems: The Strand Separation Transition in Superhelical DNA and Membrane Phase Transitions

by

Hongzhi Sun

Advisor: Craig J. Benham, Ph.D.

Metropolis-Monte Carlo algorithms are developed to analyze the strand separation transition in circular superhelical DNA molecules and phase transitions in biological membranes. In both cases shuffling operations are introduced to the simulation algorithms in order to diminish correlations among the sampled states, and thereby speed convergence. The theoretical basis for shuffling Monte Carlo algorithms is developed first. Sufficient conditions to guarantee the formal correctness of these algorithms are proved to hold.

To treat the DNA problem, moves that randomize the locations of unpaired regions are required. 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 in DNA containing modified bases, abasic sites, or bound molecules.

Since ergodicity is not a required property for shuffling operations, it is easy to construct these operations according to the specific of the system. The design, application and efficiency of the shuffling operations in Monte Carlo simulations are also demonstrated on an Ising model of the phase transition of a one-component phospholipid membrane. The results of the simple two-state membrane model agree with the calorimetric data. According to the simulation the gel-to-liquid crystalline transition of dipalmitoyl-phosphatidylcholine multilamellar vesicles (DPPC MLV) is a second-order phase transition, which is close to the critical point.

ACKNOWLEDGMENTS

I am greatly indebted to my thesis advisor Dr. Craig J. Benham for his encouragement, advice, patience and support during my Ph. D. research work. I am also indebted to many individuals who helped in my thesis work. Those whom I wish especially to thank include Drs. Istvan Sugar, Mihaly Mezei and Sylvan Wallenstein. It is a pleasure to acknowledge the assistance of my wife Qun Han, who helped take care of a number of details.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

Chapter 1

THE THEORETICAL BASIS FOR SHUFFLING MONTE CARLO ALGORITHMS

(1.1)Introduction

Metropolis-Monte Carlo methods are widely used simulation techniques in studies of biomolecular systems. Instead of calculating properties of the thermodynamic equilibrium distribution using statistical mechanical methods, Monte Carlo simulation samples this distribution. Metropolis, *et al.* (1953) first introduced this simulation technique, applying it to a problem whose complete statistical mechanical analysis was not feasible. Since then, it has become increasingly useful in many scientific and engineering fields as the continuing increase in computer speed brings more problems into range.

In the early 70's Hastings (1970) first gave a mathematical formulation of general Monte Carlo simulation methods using Markov chain theory. This formulation produced several important advances. Using this approach, Monte Carlo algorithms can be proven to be formally correct by demonstrating that simple conditions are satisfied (described later). This puts the simulation method on a solid mathematical foundation, and makes the estimation of the convergence speed a Monte Carlo simulation algorithm an important problem. These general conditions guaranteeing the correctness of a simulation procedure can be satisfied in many specific ways, each of which can be developed into a Monte Carlo algorithm. This fact has

stimulated the search for the optimal Monte Carlo method for a given system, the so called "smart Monte Carlo method". Peskun (1981) has given some guidelines for optimization of algorithms. Other approaches that start from physical consideration of the system under study include force-biased Monte Carlo (Rao, *et al.* 1979), energy-scaled displacement Monte Carlo (Goldman, 1983; Mezei, *et al.* 1987), configurational bias Monte Carlo (de Pablo, *et al.* 1992), and cluster Monte Carlo algorithms (Swendsen & Wang 1986; Wolff 1989; Kandel, 1990;). All these algorithms can yield efficient convergence speeds for specific problems.

Another way of designing Monte Carlo algorithms is to combine different types of moves, in either an ordered way or a random way. This method is similar to shuffling cards with different shuffling methods. This "shuffling" approach to designing Monte Carlo algorithms has been used in many lattice systems (Kolinski, et al.,1987). The advantage of this method is that it can combine within the algorithm many types of moves which potentially can accelerate the convergence of the simulation. In spite of its potentially wide applicability, this method has not been formulated in a rigorous mathematical way previously. In this chapter we develop such a formulation.

We determine conditions which guarantee the formal correctness of shuffling Monte Carlo algorithms. The rigorous proof of the correctness of this type of algorithm uses the Markov chain approach. In later chapters shuffling Monte Carlo simulation algorithms are designed to treat two important biological systems: strand separation transitions in supercoiled DNA and phase

transitions in biological membranes. The ability of the resulting shuffling algorithms to efficiently sample the equilibrium distribution in these cases illustrates the power of this approach.

(1.2) Hastings' formulation using Markov chains

Let i=1, 2, ..., W enumerate the configurations of the system. Systems having small number of configurations presumably can be treated by direct enumeration, so do not require Monte Carlo methods. In all systems considered here W is very large. The equilibrium probability distribution over the configurations can be described by a row vector $\pi=(\pi_1,\pi_2,...,\pi_W)$, where the ith element π_i of the vector is the probability of finding the system in configuration i, so $\pi_i > 0$ for all i. It should be noted that the value of π_i is frequently unknown, but the ratio of π_i/π_j can be determined usually for any pair of states or configurations (i,j). A direct calculation of π_i would require evaluation of the statistical mechanical partition function. This may be impossible if W is very large, or the system has other complicating factors. Suppose we wish to evaluate the ensemble average value

$$< f > = \sum_{i=1}^{W} f(i)\pi_i , \qquad (1)$$

where f is a function defined on the configurations. An approximate value for < f > can be found by sampling configurations generated by a Markov chain on the states 1, 2, ..., W.

Let $M=\{q_{ij}\}$ be the state generation matrix of an arbitrary ergodic Markov chain on the configurations. Here q_{ij} is the conditional probability of generating configuration j given current configuration i. Ergodicity means that there is a positive probability of going from configuration i to configuration j in some finite number of transitions, for each pair of configurations i and j. Then one can determine a second matrix $P=\{p_{ij}\}$ from the matrix M such that π satisfies the condition

$$\lim_{n\to\infty}\pi^0 P^n=\pi, \tag{2}$$

where π^0 is any initial distribution of the system. Thus, every component of π^0 is non-negative and the sum of all the components is 1. The element p_{ij} of transition matrix P is defined as

$$p_{ij}=q_{ij}\alpha_{ij} \quad (i\neq j), \quad p_{ii}=1-\sum_{j\neq i}p_{ij}, \tag{3}$$

where $0\leq\alpha_{ij}\leq 1$ is chosen to satisfy the reversibility restriction

$$\pi_j q_{ji}\alpha_{ji}=\pi_i q_{ij}\alpha_{ij}, \tag{4}$$

and at least for one i we have

$$p_{ii}>0.$$

The condition $p_{ii}>0$ is easy to satisfy in most practical situations. Condition (4) implies $\pi_j p_{ji}=\pi_i p_{ij}$, which corresponds to the detailed balance condition in equilibrium statistical mechanics.

In order to perform a Monte Carlo simulation for the transition matrix P the following two steps are repeated:

Step I Candidate state generation

If the system is in configuration i, the next candidate configuration j is generated by using the distribution given by the ith row of M.

Step II Decision making

The system is changed into configuration j with probability α_{ij} and stays in configuration i with probability $1-\alpha_{ij}$.

One repeat of Step I and Step II together gives one trial T. During a Monte Carlo simulation the trials are repeated many times. If the initial probability distribution is π^0, n repeats of the trial will bring the distribution to $\pi^0 P^n$. According to Eq. 2, $\pi^0 P^n \approx \pi$ when n is sufficiently large. If the simulation starts from an arbitrary initial configuration i and after every nth trial a new configuration is sampled, then the collection of sampled configurations will approximate the equilibrium distribution π when n is chosen large enough. In this way, the statistical quantity $< f >$ is estimated by

$$<\mathrm{f}> \cong \frac{\sum_{k=1}^{U} f(k)}{U}, \qquad (5)$$

where k indexes the sampled configurations and U is the sample size.

Two simple and widely used methods for determining α_{ij} are the following. When M is a *symmetric* matrix

$$\alpha_{ij} = \begin{cases} 1 & (\pi_j / \pi_i \geq 1) \\ \pi_j / \pi_i & (\pi_j / \pi_i < 1) \end{cases} \qquad (6)$$

or

$$\alpha_{ij} = \pi_j / (\pi_i + \pi_j). \qquad (7)$$

Eq. 6 is the Metropolis method (Metropolis, *et al.* 1953) while Eq. 7 is its Barker variant (Barker, 1965). The Metropolis method can be applied only when M is symmetric. Otherwise, detailed balance will be violated (see Eq. 4). In other words, the symmetry of the matrix M is equivalent to the detailed balance condition if the Metropolis method is applied.

A very important problem that occurs when this sampling method is used relates to the speed of convergence. How large must n be to guarantee that $\pi^0 P^n$ is sufficiently near π ? This is very difficult to answer in many situations. Only a qualitative answer has been given (Moran, 1968). The transition matrix P constructed above has eigenvalues

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_W|. \qquad (8)$$

Let $\pi^o(n) = \pi^o P^n = (\pi_1^o(n), \ldots, \pi_W^o(n))$. Then according to Moran (1968) $|\pi_k^o(n) - \pi_k| < b|\lambda_2|^n$ for some number b. However, there is no general way to estimate the value of λ_2 (Diaconis, 1993) because the matrix P has dimension $W \times W$, which commonly is huge in problems of interest.

(1.3) A generalized formulation of Markov chain Monte Carlo simulation

When performing a Monte Carlo simulation according to a scheme constructed as described in section (1.2), all the random moves in the candidate state generation step are determined by the matrix M. In other words, one performs the same type of trial at each step. A natural generalization of the method is to use different types of trials within a Monte Carlo simulation.

If the transition matrices $M^{(1)}$, $M^{(2)}$, ..., $M^{(r)}$ are used in a Monte Carlo algorithm, then one can construct $p_{ij}^{(1)}, p_{ij}^{(2)}, ..., p_{ij}^{(r)}$ and $\alpha_{ij}^{(1)}, \alpha_{ij}^{(2)}, ..., \alpha_{ij}^{(r)}$ from $q_{ij}^{(1)}, q_{ij}^{(2)}, ..., q_{ij}^{(r)}$ respectively, according to Eq. 3, such that Eq. 4 is satisfied for each index k (detailed balance). Here $q_{ij}^{(k)}$ is the (i,j)th entry of matrix $M^{(k)}$, and $i,j=1, 2,...,W$. The generalized version of the Markov chain Monte Carlo algorithm is formulated in the following way:

Step I' Candidate state generation

If the system is in configuration i, the next candidate configuration j is generated by probabilistic rules. These rules are represented by probability matrices $M^{(1)}$, $M^{(2)}$, ..., $M^{(r)}$. If the kth rule is applied, then the candidate state j is generated by using the distribution given by the ith row of $M^{(k)}$.

Step II' Decision making

The system will change into configuration j with probability $\alpha_{ij}^{(k)}$ and will still stay in configuration i with probability $1-\alpha_{ij}^{(k)}$ if the kth rule, the matrix $M^{(k)}$, is applied.

Performing Step I' then Step II' constitutes one trial. Different types of trials can be performed in a simulation. An algorithm performing a Monte Carlo simulation can be developed by repeating different types of trials according to a deterministic pattern $T_{(\nu_1)}T_{(\nu_2)}...T_{(\nu_h)}$. The index ν_i in this pattern indicates that the trial $T_{(\nu_i)}$ is constructed from the candidate generating rule $M^{(\nu_i)}$. The application of the candidate generating rules according to this pattern leads to a generating matrix $\hat{M}$

$$\hat{M}=M^{(\nu_1)}M^{(\nu_2)}...M^{(\nu_h)}. \tag{9}$$

A sufficient condition for formal correctness of the algorithm (*i.e.*, convergence to the equilibrium distribution) is the following *ergodicity property* : The system can go from any initial state to any other state by applying $\hat{M}$ a finite number of times (the proof is shown in section (1.4) below).

It is clear that performing Step II' modifies the matrices that were used in Step I', producing matrices $P^{(1)}$, $P^{(2)}$,..., $P^{(r)}$, which are the mathematical representations of the trials $T_{(1)}$, $T_{(2)}$,..., $T_{(r)}$, respectively. The orderly application of the trials according to the deterministic pattern $T_{(\nu_1)}T_{(\nu_2)}...T_{(\nu_h)}$ leads to a resultant transition probability matrix $\hat{P}$:

$$\hat{P}=P^{(\nu_1)}P^{(\nu_2)}...P^{(\nu_h)}, \tag{10}$$

which is the equivalent of the transition matrix P given in section (1.2).

Metropolis Monte Carlo algorithms are used to analyze many statistical thermodynamic systems. In these systems the equilibrium distribution $\pi=(\pi_1,\pi_2,...\pi_W)$ will depend on the type of contacts between the system and its surrounding and also on the nature of the system's configurations. For example, if there is only thermal contact and each quantum mechanical state of the system corresponds to one configuration, then

$$\pi_i = e^{-E_i/kt} / \sum_{j=1}^{W} e^{-E_j/kt}, \qquad (11a)$$

where E_i is the energy of the i-th quantum state, t is the equilibrium temperature of the system, k is the Boltzmann constant, and W is the number of quantum states. For statistical mechanical purposes the description of the system is frequently less detailed. In this case the quantum states of the system are grouped together in such a way that the g_i quantum states of the i-th configuration all have approximately the same energy E_i, and then

$$\pi_i = g_i e^{-E_i/kt} / \sum_{j=1}^{W} g_j e^{-E_j/kt}, \qquad (11b)$$

where g_i is the degeneracy of the i-th configuration and W is the total number of configurations.

To use the Metropolis criterion in Step I', each $M^{(k)}$ is required to be a symmetric matrix. When performing Step II', the system will move into state j from current state i with probability $\min\{1,\pi_j/\pi_i\}$. Otherwise, the system will remain in configuration i. This is equivalent to saying that each α_{ij} in Eq. 3 is to be chosen as $\min\{1,\pi_j/\pi_i\}$. The other condition in Eq. 3, $p_{ii}>0$ for at least one configuration i, is satisfied automatically, as is shown below. For each $M^{(q)}$, the corresponding $P^{(q)}$ in this special case has its (i,j)th element as follows:

$$P_{ij}^{(q)} = M_{ij}^{(q)} \frac{\pi_j}{\pi_i} \quad if \quad \pi_i > \pi_j, \quad for \quad i \neq j$$

$$P_{ij}^{(q)} = M_{ij}^{(q)} \quad if \quad \pi_i \leq \pi_j, \quad for \quad i \neq j \qquad \text{for } q=1,2,\ldots,r$$

$$P_{ii}^{(q)} = 1 - \sum_{j \neq i}^{W} P_{ij}^{(q)}.$$

If $\pi_i = \pi_j$ for all i, j, then every state is equally probable in the equilibrium distribution, and a Monte Carlo simulation is unnecessary. Otherwise, let i be the configuration with $\pi_i = \max\{\pi_j, 1 \leq j \leq W\}$. Then $M_{ij}^{(q)} > P_{ij}^{(q)}$ for at least one j because $\pi_i > \pi_j$ for at least one j. Thus,

$$1 = \sum_{j=1}^{W} M_{ij}^{(q)} \geq \sum_{j \neq i}^{W} M_{ij}^{(q)} > \sum_{j \neq i}^{W} P_{ij}^{(q)}.$$

Therefore, $P_{ii}^{(q)}>0$. The (i,i)th element in $\hat{P}$, denoted by $\hat{P}_{ii}$, is also positive because $\hat{P}_{ii} \geq P_{ii}^{(v_1)} P_{ii}^{(v_2)} \ldots P_{ii}^{(v_h)} > 0$.

(1.4) Validity of the generalized version of the Markov chain Monte Carlo algorithm.

To demonstrate the validity of this approach it suffices to show that, for any initial distribution π^0 of the system,

$$\lim_{n\to\infty}\pi^0(\hat{P})^n=\pi. \tag{12}$$

This is the equivalent of Eq. 2 of section (1.2).

Proposition 1: If B is the equilibrium distribution matrix (each row equals the row vector π of the equilibrium distribution), then Eq. 12 is equivalent to

$$\lim_{n\to\infty}(\hat{P})^n=B. \tag{13}$$

Proof

(a) Eq. 12 implies Eq. 13.

Let e_i be the unit row vector of dimension W, whose ith component is 1 and the rest are zeros. Let the initial distribution be $\pi^0= e_i$. Substituting e_i into Eq. 12 we get

$$\lim_{n\to\infty}e_i(\hat{P})^n=\lim_{n\to\infty}(\hat{P}_{i1}(n),\hat{P}_{i2}(n),...,\hat{P}_{iW}(n))=\pi \tag{14}$$

where $\hat{P}_{ij}(n)$ is the (i,j)th element of $(\hat{P})^n$ matrix. According to this expression the ith row vector of the limit matrix $\lim_{n\to\infty}(\hat{P})^n$ exists and equals π. As i is arbitrary, this shows that every row of the limit

matrix exists and is equal to π. Thus the limit matrix exists and is the equilibrium distribution matrix B.

(b) Eq. 13 implies Eq. 12

We have, for all e_i's

$$e_i B = e_i \lim_{n\to\infty}(\hat{P})^n = \lim_{n\to\infty} e_i(\hat{P})^n. \tag{15}$$

Any distribution π^0 can be expressed in the following way

$$\pi^0 = \sum_{j=1}^{W} \pi_j^0 e_j \tag{16}$$

where $\pi_j^0 \geq 0$ and $\sum_{j=1}^{W} \pi_j^0 = 1$. Then the following is true

$$\lim_{n\to\infty} \pi^0(\hat{P})^n = \sum_{j=1}^{W} \pi_j^0 e_j \lim_{n\to\infty}(\hat{P})^n = \sum_{j=1}^{W} \pi_j^0 e_j B = \sum_{j=1}^{W} \pi_j^0 \pi = \pi. \tag{17}$$

QED

The following propositions will prove Eq. 13.

Proposition 2: If M is a transition matrix (ergodicity is not required) and if P is obtained according to Eqs. 3-4 then $BP=PB=B$.

Proof

Let us consider the (i,j)th element of the product matrix BP.

$$\begin{aligned}(BP)_{ij} &= \sum_{m=1}^{W} \pi_m p_{mj} = \pi_j p_{jj} + \sum_{m\neq j} \pi_m p_{mj} \\ &= \pi_j(1 - \sum_{m\neq j} p_{jm}) + \sum_{m\neq j} \pi_m p_{mj} \\ &= \pi_j + \sum_{m\neq j} (\pi_m p_{mj} - \pi_j p_{jm}) = \pi_j = B_{ij}. \end{aligned} \tag{18}$$

Thus we proved that $BP=B$. On the other hand it is clear that $PB=B$ because the sum of the elements in each row of P is 1 and the column entries of B matrix are identical. By combining the above two results we get $BP=PB=B$.

QED

It follows that the transition matrices $P^{(k)}$ obtained by modifying the configuration generating matrices $M^{(k)}$ commute with the equilibrium distribution matrix B :

$$B P^{(k)}=P^{(k)}B=B. \tag{19}$$

By using this relationship for every k one can get a similar relationship for the resultant transition probability matrix, $\hat{P}$:

$$B \hat{P}=\hat{P}B=B. \tag{20}$$

Proposition 3: There is a positive integer l such that for $m \geq l$ $\hat{P}^m$ is positive, *i.e.*, every matrix element is positive.

Proof

According to Eq. 3 there is at least one diagonal element $\hat{p}_{ii}$ of matrix $\hat{P}$ is positive. Now, if $\hat{M}$ in Eq. 9 satisfies the *ergodicity property*, then $\hat{P}$ is also ergodic, and therefore there exists an integer $l>0$ such that $\hat{P}^l$ is positive (Feller, 1966).

QED.

Proposition 4: If A is a positive probability transition matrix then the following limit exits:

$$\lim_{n\to\infty}(A)^n = C \tag{21}$$

and the rows of the limit matrix C are distributions and are identical with each other (Moran, 1968).

Now we can prove Eq. 13. According to Proposition 3, $(\hat{P})^m$ is a positive matrix for every $m \geq l$ and thus Proposition 4 is applicable for $A = (\hat{P})^m$, which results in

$$\lim_{n\to\infty}(\hat{P})^n = C. \tag{22}$$

Let us multiply Eq. 22 by the equilibrium distribution matrix, and apply repeatedly the relationship of Proposition 2 ($B\hat{P} = \hat{P}B = B$):

$$BC = B\lim_{n\to\infty}(\hat{P})^n = \lim_{n\to\infty} B(\hat{P})^n = \lim_{n\to\infty} B(\hat{P}) = \lim_{n\to\infty} B = B. \tag{23}$$

This shows that $B = BC$. On the other hand, $BC = C$ is also true, because all elements in a column of C are identical, and B is a probability transition matrix where the sum of every row is 1. According to these two relationships, we get $C = B$ and from Eq. 22

$$\lim_{n\to\infty}(\hat{P})^n = B. \tag{24}$$

QED

(1.5) Introducing shuffling trials into a Monte Carlo algorithm

From a mathematical point of view, shuffling trials are a type of generalized trial as described in section (1.3). We name some types of operations as shuffling operations based on physical reasons.

Shuffling operations are stochastic operations which, when included in a pattern $T_{(1)}$, $T_{(2)}$, ..., $T_{(N)}$ of trials, increases the convergence rate of Monte Carlo algorithm to the equilibrium distribution of the system.

In conventional algorithms the trials usually allow *local changes* in the system which are physically likely within a short time. Unfortunately, when these trials are used the sampled configurations frequently remain strongly correlated over long segments of the Markov chain. That is, the system is trapped into metastable states for long periods of the simulation. In consequence, the equilibrium distribution of the system is not attainable within a feasible computational time. The probability of long, strongly correlated segments in the Markov chain can be reduced by using trials which allow global, non-physical, changes in the system. The type of global changes appropriate in a given system must be specifically designed for that system in order to efficiently sample the whole space of configurations.

Shuffling trials are characterized by candidate state generation matrices, $S^{(1)}$, $S^{(2)}$,... of Markov chains on the configurations. However, there is no *ergodicity* requirement imposed on each $S^{(k)}$ or on the combination of them (for reasons described below). Following Eqs. 3-4, one can construct $\alpha_{ij}(S^{(k)})$ and the corresponding transition matrices $R^{(k)}=\{r_{ij}^{(k)}\}$ for each index k. Here a candidate configuration j, generated from current configuration i by $S^{(k)}$, is accepted with probability $\alpha_{ij}(S^{(k)})$. The probability that the system still stays in configuration i is $1-\alpha_{ij}(S^{(k)})$. The application of several shuffling trials

according to a deterministic pattern $\mathbf{u}=(u_1,u_2,...u_g)$ leads to the definition of the transition probability matrix $\hat{R}$

$$\hat{R}=R^{(u_1)}R^{(u_2)}...R^{(u_g)}. \tag{25}$$

This $\hat{R}$ differs from $\hat{P}$ in that $\hat{S}=S^{(u_1)}S^{(u_2)}...S^{(u_g)}$ need not satisfy the *ergodicity* requirement.

One can prove that a Monte Carlo simulation interrupted regularly by shuffling trials is a formally correct simulation (*i.e.*, leads to the equilibrium distribution of the system). The proof is based on propositions given in the last section and on the following two propositions.

Proposition 5: One has

$$BR^{(k)}=R^{(k)}B=B \tag{26}$$

and

$$B\hat{R}=\hat{R}B=B. \tag{27}$$

The proof is the same as in proposition 2.

Proposition 6: If A is a $W\times W$ positive matrix then $AR^{(k)}$ is positive for every k=1,2,..., and $A\hat{R}$ is also positive.

Proof:

Let us assume that the (i,j)th element of the product matrix is zero and thus the product matrix is not positive

$$(AR^{(k)})_{ij} = \sum_{m=1}^{W} A_{im} r_{mj}^{(k)} = 0. \qquad (28)$$

Since every $A_{im}>0$, the above equation can be satisfied only if all elements in the jth column of $R^{(k)}$ are zeros. From the construction of $R^{(k)}$, it follows that all elements of the jth row of $R^{(k)}$ must be zero too. However, this contradicts the definition of probability transition matrix, *i.e.*, $\sum_{m=1}^{W} r_{jm}^{(k)} = 1$. Thus our initial assumption is incorrect, *i.e.*, the product matrix $AR^{(k)}$ must be positive. Similarly, $A\hat{R}$ is positive.

QED

Now we can prove the correctness of a simulation where the Monte Carlo algorithm is interrupted by one shuffling trial $R^{(k)}$ (or by a series of shuffling trials $\hat{R}$) after every mth trial pattern. Or more precisely, we have to prove that the transition probability matrix $(\hat{P})^m R^{(k)}$ satisfies the condition of the correct simulation: $\lim_{n\to\infty}[(\hat{P})^m R^{(k)}]^n = B$, where m is chosen such that $(\hat{P})^m$ is a positive matrix.

According to Proposition 2 $(\hat{P})^m$ is a positive transition matrix if $m \geq l$. Thus the product of $(\hat{P})^m$ and a shuffling matrix $R^{(i)}$ also is a positive matrix by proposition 6. Since $(\hat{P})^m R^{(i)}$ is positive, Proposition 4 is applicable

$$\lim_{n\to\infty}[(\hat{P})^m R^{(k)}]^n = D, \qquad (29)$$

where the rows in the limit matrix D are distributions and identical with each other. Let us multiply Eq. 29 with the equilibrium distribution matrix, and apply the Commutativity properties ($B(\hat{P})^m = B$

and $BR^{(k)}=B$) repeatedly. Similar to the derivation of Eqs. 23-24, we get

$$\lim_{n\to\infty}[(\hat{P})^m R^{(k)}]^n = B. \tag{30}$$

QED

When multiple shuffling trials are used according to a pattern, it is clear that

$$\lim_{n\to\infty}[(\hat{P})^m \hat{R}]^n = B. \tag{31}$$

At this stage, the shuffling algorithm can be generalized further. In a shuffling algorithm the trial pattern can be written as $T_{(1)}T_{(2)}...T_{(n)}U_{(1)}U_{(2)}...U_{(g)}$, where $T_{(i)}$ is a common trial and $U_{(j)}$ is a shuffling trial. The resultant transition probability matrix corresponding to this trial pattern can be denoted by $\hat{H}^E$. Correctness of the algorithm implies that matrix $\hat{H}^E$ is positive. A random change of the trial order (for both T and U) in the trial pattern is called a random permutation of the pattern.

Proposition 7: A shuffling Monte Carlo algorithm in which random permutations of a trial pattern are performed (instead of repeating the trial pattern $T_{(1)}T_{(2)}...T_{(n)}U_{(1)}U_{(2)}...U_{(g)}$) is a correct algorithm if the probability of the occurrence of trial pattern $T_{(1)}T_{(2)}...T_{(n)}U_{(1)}U_{(2)}...U_{(g)}$ is positive.

Proof

To simplify the proof of this proposition, suppose n=1 and g=1 in the trial pattern. In this Monte Carlo algorithm trial pattern TU

and UT will be performed with probability p and 1-p, respectively, where 0<p<1. Let the resultant transition matrices of TU and UT be $\hat{H}^E$ and $\hat{Q}^E$, respectively. In this algorithm the transition matrix $\hat{H}^E$ occurs with probability p and transition matrix $\hat{Q}^E$ occurs with probability 1-p, i.e., the resultant transition matrix in the algorithm is $\hat{V}^E = p\hat{H}^E + (1-p)\hat{Q}^E$. Since $\hat{H}^E$ is a positive matrix and every entry in $\hat{Q}^E$ is non-negative, $\hat{V}^E$ is a positive matrix too. The correctness of the algorithm follows from the positivity of $\hat{V}^E$ matrix and from the fact that both $\hat{H}^E$ and $\hat{Q}^E$ satisfy detailed balance.

QED

Shuffling algorithms are special cases of Metropolis Monte Carlo simulations. As stated at the end of section (1.3), each $M^{(q)}$ must be *symmetric*. Similarly, each $S^{(k)}$ must be *symmetric* too. There is no ergodicity requirement imposed on the $S^{(k)}$, because this is already satisfied by the non-shuffling moves. Whether a new candidate state is accepted depends on the Metropolis energy criterion in the cases considered below.

CHAPTER 2

MONTE CARLO ANALYSIS OF STRAND SEPARATION TRANSITIONS IN SUPERHELICAL DNAS

(2.1) Background

(2.1.1) General formulation

In vivo DNA is constrained into topological domains, within which molecular stresses are regulated by imposed superhelicity. Whereas much of the DNA in prokaryotic cells is negatively supercoiled, in eucaryotes only DNA in actively transcribing chromatin is supercoiled. Modulation of DNA superhelicity affects many biological events, including the initiation of replication (Kowalski & Eddy, 1989; Mattern & Painter, 1979) and of transcription (Pruss & Dilica, 1989; Weintraub *et al.*, 1986), recombination (Richet *et al.*, 1986), and the uptake of homologous single strands (Beattie *et al.*, 1978).

Negative DNA superhelicity has long been known to destabilize the helix (Vinograd, *et al.*, 1968; Dean & Lebowitz, 1971; Kowalski, *et al.*, 1988). Duplex unwinding occurs in many superhelically modulated regulatory events. Strand separation is required for the initiation of transcription and of replication, and also may be implicated in recombination, transposition, and other events. It is essential to develop a quantitative understanding of superhelical duplex destabilization because of the importance of strand separation in diverse superhelically mediated biological activities.

The theoretical analysis of strand separation in superhelical DNA molecules 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. However, superhelicity globally couples together the secondary structures of every site in a circular DNA molecule. Transition at any one site alters its helicity, which changes the distribution of superhelicity throughout the molecule, and thereby affects the level of torsional stress experienced by all other sites. Hence 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 on 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.

A formally exact statistical mechanical analysis of superhelical duplex destabilization requires calculation of the governing partition function Z (Benham, 1990),

$$Z = \sum_{i \in W} \exp\left\{ -G(i) \Big/ kt \right\} . \qquad (1)$$

Here W is the set of all states (*i.e.*, configurations) of strand separation, i is one such state, and $G(i)$ is its associated free energy. (This is a free energy because each individual state of separation is itself an average over microstates having different solvent

structures, base orientations, etc.) The equilibrium probability of state i is

$$p_i = \exp(-G(i)/Rt)/Z. \qquad (2)$$

Because the number of states increases exponentially with sequence length, exact calculations enumerating all states are feasible only for short sequences.

An approximate statistical mechanical method to calculate properties of the superhelical strand separation transition in circular DNA has been developed previously (Benham, 1979; 1990; 1992). This approach specifies an energy threshold, and explicitly finds all states whose free energy exceeds 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 results of calculations using this method have been shown to be in close quantitative agreement with experimental measurements of the extents and locations of superhelical strand separation in all molecules examined to date (Benham, 1992). While this method is accurate and fast for short sequences (less than 15,000 bps), it has some limitations. The number of states explicitly included, and hence also the time required for a calculation, increases approximately exponentially with the threshold. The combinatorics of the strand separated states dictates that this approach is computationally tractable only in

situations where the low energy states, *i.e.*, those satisfying the energy threshold condition, have small numbers $r < 4$ of runs (contiguous open regions) of separation. In consequence, this method cannot be applied accurately to long sequences in which numerous A+T-rich regions compete for transition. Also, the energetics of base pair separation can only be specified as a function of base pair identity, AT or GC. More detailed energetics, such as near-neighbor effects, cannot be incorporated into this approximate statistical mechanical approach as it is presently structured.

Another approximate statistical mechanical method has been developed by Anshelevich *et al.* (1979). That approach calculates a partition function for a linear molecule by a recursion algorithm and then approximately accounts for the superhelical constraint by an energy renormalization technique. Clearly, this approach does not impose the actual closed circular topological constraint. A DNA molecule is regarded as being linear, with one inseparable base pair added at each end to decrease end effects. This condition differs significantly from the global coupling experienced by circular molecules. Moreover, the strand separated regions are regarded as being torsionally undeformable. In fact, the persistence length of DNA single strands is two orders of magnitude smaller than that of the B-form duplex (Bloomfield *et al.*, 1974). So residual superhelical stresses can torsionally deform the separated regions. The assumption of undeformability, (not made in the other approximate statistical mechanical analysis or in the presently developed Monte Carlo procedure), severely limits the utility of this method and the accuracy of its results. More recently, this approach has been

improved by the imposition of self-consistency conditions on the renormalization step (Katsura *et al.*, 1993).

This chapter develops Monte Carlo simulation procedures to analyze heteropolymeric strand separation transitions in superhelical DNAs of specified sequence. Although Monte Carlo methods have been used to analyze deformations of tertiary structures in superhelical DNAs (Levene & Crothers, 1986; Klenin *et al.*, 1991; Zhurkin *et al.*, 1991), they have not been applied to secondary structure transitions to date. In this approach the configurations available to the system are sampled with frequencies that approximate those of the equilibrium distribution. This technique has several advantages over existing alternatives. It can treat systems with complicated transition energetics, including such factors as near neighbor modulation of the transition energetics, the presence of abasic sites, modified bases or other lesions. The topological state of the DNA is modeled exactly, and torsional deformations of the denatured regions are allowed. The computation time required for a simulation increases at most quadratically with molecular length. The Monte Carlo methods can handle states in which many regions are separated, as occur in long DNA sequences.

(2.1.2) Strand separation in supercoiled DNA

Closed circularity fixes the linking number Lk of a DNA molecule. The linking difference associated to this closed circular DNA is $\theta = Lk - Lk_0$, where Lk_0 is the linking number of the lowest energy, relaxed state. When a molecule is negatively supercoiled, *i.e.*,

$\theta<0$, the resulting stresses can destabilize the duplex, inducing local strand separations to occur (Vinograd *et al.*, 1968; Dean & Lebowitz, 1971; Kowalski *et al.*, 1988). To model this phenomenon, 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 possible states in this analysis. We set $m_i=1$ if base pair i is separated in a given state, $m_i=0$ otherwise. For closure we specify $m_{N+1}=m_1$. Then the number r of runs of separation (*i.e.*, open regions) 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 experienced by the molecule is partitioned into three types of conformational changes, each of which requires free energy. First, separation of the specified base pairs requires free energy

$$G_{sep} = ar + \sum_{i=1}^{N} m_i b_i . \tag{3}$$

Here a is the free energy needed to initiate a run of transition. This arises primarily from the extra stacking interaction that must be disrupted when a run of strand separation is initiated. Under normal physiological conditions, $a \cong 10.5$ kcal/mol (Benham, 1992; Bauer & Benham, 1993). Also, b_i is the free energy needed to separate base pair i, $1 \le i \le N$. This free energy associated with each base pair may be assigned individually. It can include effects such as near neighbor interactions or the presence of lesions, such as pyrimidine dimers or apurinic sites, which partially disrupt the duplex. Second, the two

strands within a separated region can rotate around each other. If n separated base pairs are torsionally deformed to a helicity τ (rad/bp), the required free energy is

$$G_{tor}=\frac{1}{2}Cn\tau^2. \tag{4}$$

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

$$G_{res}=\frac{1}{2}K\theta_r^2, \tag{5}$$

where K is an experimentally measured constant. 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 (Benham, 1990)

$$G(n,n_{AT},r)=\frac{2\pi^2CK}{4\pi^2C+Kn}(\theta+\frac{n}{10.5})^2+ar+\sum_{i=1}^{N}m_ib_i \tag{6}$$

The sample calculations reported below are designed to test the accuracy of the Monte Carlo procedures by comparison with the results from the approximate statistical mechanical methods of Benham (1992), For this reason an initial formulation uses the same expression for the free energy associated to a state as was used in

that earlier work. Thus, the separation energy was assigned either of two values, b_{AT} or b_{GC}, depending on the identity of the base pair involved. We reiterate, however, that the Monte Carlo procedures developed here can easily accommodate a wide range of possibly more complex energy laws.

All calculations performed here use energy parameter values found to hold under the experimental conditions of Kowalski et al. (1988), in which t=310° K and [Na^+]=0.01M. These values are: b_{AT}=0.258 kcal/mol, b_{GC}=1.305 kcal/mol, C=3.6 kcal base pair/rad^2 and $K=2350\ Rt/N$. The results found by the approximate statistical mechanical method using these values have been shown to agree closely with experiment (Benham, 1992).

(2.2) Methods

(2.2.1) A standard Metropolis-Monte Carlo algorithm

First we develop a standard Metropolis-Monte Carlo simulation algorithm, MCA, without shuffling, to treat strand separation in supercoiled DNA. Consider a circular DNA molecule containing N base pairs and supercoiled to a linking difference θ. There are 2^N states of strand separation available to this system, as described above. Denote the current state by A with free energy G(A). Changing the secondary structure of one base pair in A (either from closed to open or from open to closed) with probability $1>p>0$ produces another, possibly different, state B, whose energy is G(B). To determine if we accept B, we apply the Metropolis criterion: If G(B)-G(A) is positive, a random number *rand* in the open interval (0,1) is generated and a

comparison between $rand$ and $d = \exp((G(A)-G(B))/Rt)$ is made. If $rand > d$, B is rejected so that the current state remains A. In all other cases, B is chosen. Next, this current state is taken as the new starting point, and the process is repeated. This procedure, performed once per base pair proceeding along the entire length of the molecule, is called one 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. To generate a sample distribution using this approach, one sample state is chosen after λ MCCs, where λ is an adjustable parameter.

To prove the formal correctness of this procedure, we show that the detailed balance and the ergodicity conditions listed in section (1.3) are satisfied. This algorithm has matrix $\hat{M}=M_1...M_N$, where N is the total number of base pairs, and M_i corresponds to the step where base pair i is probabilistically altered. Since $\hat{M}$ is constructed by flipping each base pair with the same probability p, proceeding along the entire molecule, the entries in $\hat{M}$ can be computed by regarding this process as an N-fold Bernoulli trial in which 2^N states are available. Any two distinct states C_1 and C_2 will differ at m positions, $1 \leq m \leq N$. One may generate C_2 from C_1 by a series of N elementary steps, in which one changes the state of separation at each of these m positions and keeps all others the same. The probability of this occurring is $p^m(1-p)^{N-m} > 0$. This shows that the entry $p(C_2|C_1)$ in $\hat{M}$ is positive, so ergodicity condition is satisfied. The detailed balance condition is clear since $p(C_1|C_2)$ also equals $p^m(1-p)^{N-m}$.

Strong correlations still can be found in the standard algorithm MCA mainly due to two facts. First, because the initiation energy a is large ($a \cong 10$-12 kcal/mol), the probability of accepting a candidate state differing at one site from the current state but having a larger number of runs, is less than 10^{-7}. This makes it 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. Second, transitions between low energy states having the same number of open runs also will be difficult by the standard method if the unpaired regions in the two states are far from each other along the sequence. It is extremely unlikely to move between these states by migration of the open region through single base pair openings and closings. This is particularly true if G+C-rich regions intervene. Global coupling limits the total number of open base pairs, which further increases the difficulty of this movement. These observations indicate why the correlation between states will remain strong in this standard Monte Carlo simulation algorithm MCA, even with very large values of λ. In principle, increasing λ will decrease correlations. However, such large values of λ are needed to generate sufficiently weakly correlated sampled states by this procedure as to be unfeasible for our problem. This has been demonstrated by the calculations reported in the **Results** in section (2.3). Shuffling operations are required to destroy the strong correlations between successive sampled states.

(2.2.2) Metropolis Monte Carlo methods with shuffling operations

We use shuffling trials to reduce the strong correlations occurring in the MCA sampling process. The shuffling trials chosen in this algorithm randomize the locations of separated regions without changing the total number of separated base pairs. According to section (1.3) it is clear that adding shuffling trials after a standard MCC yields a formally correct algorithm. We use $S^{(k)}$ to indicate a shuffling generating matrix corresponding to step I'. The matrix $R^{(k)}$ corresponding to a shuffling trial constructed by using method similar to Eq. 3 of section (1.2) is called a shuffling selection matrix.

The shuffling generation matrices are constructed from elementary operations. These consist of a shift class consisting of ROTATION, SH22,....., SH77, a squeezing class comprised of SQ22, SQ33,..... , SQ77, and an interchange class containing EX12, EX21, EX23, EX32, ..., EX67 and EX76.

In the ROTATION operation the positions of all open loops are rotated in unison by a random amount around the molecule, as illustrated in Figure 1. This moves the open regions to new positions, but keeps their lengths and the distances between them fixed. Since the number of possible rotations equals the number of total base pairs, it is clear that if state i can rotate to state j, then state j also can rotate to state i. Setting the probability of each rotation equal to $P = 1/N$, assures detailed balance because the Metropolis energy criterion will be applied in decision making step. The ROTATION operation can be applied to any state. Alternatively, its use can be restricted, for example to cases when $r = 1$ or $r > 7$.

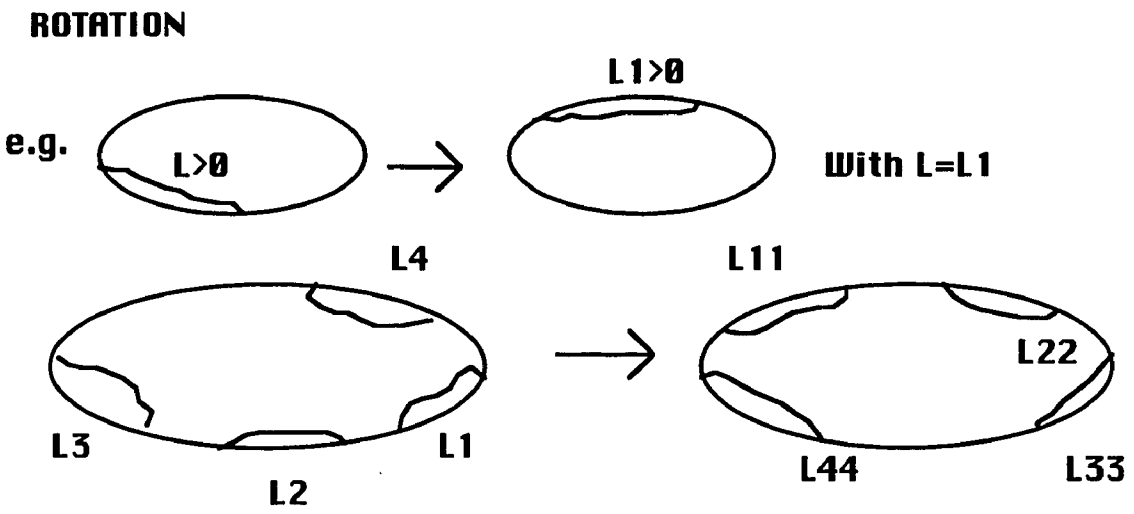

Figure 1

Fig. 1 The upper left picture describes a state having one open region, denoted by L. The rotation operation randomly moves this loop to another place on the molecule, indicated as L1. Here L and L1 have equal lengths. The lower left picture shows a state having four open regions, L1, L2, L3 and L4. The rotation operation randomly moves these four loops to new positions, while keeping their separation distances and lengths fixed.

Shift operations change the relative positions of open regions without altering their numbers or lengths. They only apply when there are $r>1$ open regions. A particular open region is selected, and moved within the set of closed positions that abut it on either side. To describe this class of operations we consider SH33, which shifts among states having three runs of separation, as illustrated in Figure 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 intervals between adjacent loops

have lengths I_i, i=1, 2, 3, and the order of open loops and closed intervals is shown in the figure.

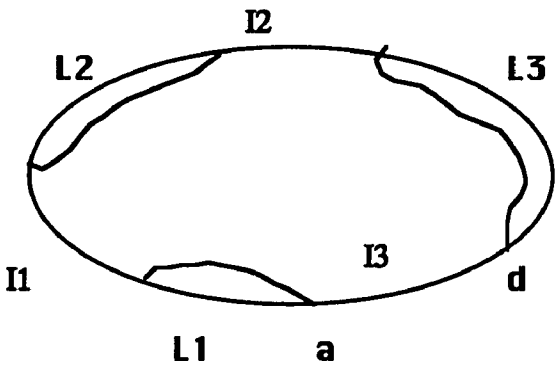

Figure 2

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.

One can only move open loop L_1 within the region containing it and bounded by L_3 and L_2. There are (I_3+I_1-1) ways this can be done without merging with other 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 a shift operation, one first calculates the probabilities P_1, P_2, P_3, of shuffling open regions 1, 2, or 3, respectively, as

$$P_1 = (I_1 + I_3 - 1) / (2 (N - n) - 3),$$
$$P_2 = (I_1 + I_2 - 1) / (2 (N - n) - 3), \qquad (7)$$
$$P_3 = (I_3 + I_2 - 1) / (2 (N - n) - 3).$$

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 $P_1 \le rand < P_1+P_2$, 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 INT($rand2\,(I_3+I_1-1)$), where 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 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). 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 shift operations for other different number of runs are constructed analogously.

In the algorithms developed here, ROTATION is used to shuffle states having either one run or greater than seven runs, and SH22, ..., SH77 are used to treat the other cases. A matrix corresponding to this shuffling operation is constructed, which is called SHIFTALL in our program.

Squeeze operations are applied to states having more than one run of separation. Their purpose is to change the distribution of open base pairs among the runs without changing either the total number of open base pairs or the number of runs. We randomly decrease the size of one of the open runs, and simultaneously increase the size of a neighboring open run 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. Equiprobable squeezing operations are constructed, using an approach analogous to that described above for the design of the SH33. The shuffling generation matrix SQUEEZEALL consists of SQ22, ..., SQ77 and ROTATION, which again is used to treat states with one run or more than seven runs.

It is more difficult to design balanced operations that change the number of open runs without altering the total number of open base pairs. This is the purpose of the exchange operations. As examples, we describe EX23 and EX32 using Figure 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 neighbor closed region. (Note that there is only one neighboring closed region to the subregion being moved.) The resulting state has three open regions. Let $N(L_i,I_j)$ be the number of possible ways that one can divide run L_i into two and place one part within the neighbor closed region I_j. Then $N(L_i,I_j) = (L_i-1)(I_j-1)$,

because the length of the portion to be moved can vary from 1 to L_i-1, and for any one of these fragments there are I_j-1 ways to place it within the I_j region. Summation shows the total number of ways of going from a two-run state to any allowed three-run state to be $(n-2)(N-n-2)$. This number again depends only on the number of open base pairs in the initial two-run state.

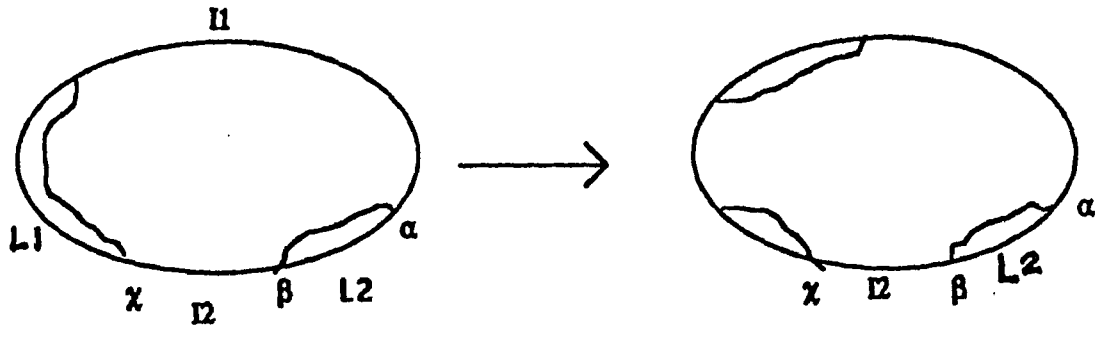

Figure 3

Fig. 3 The exchange move. EX23 randomly selects an open region and a direction, say open region L1, a portion of which will be moved into closed region I1. The selected loop 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 in the right side picture. EX32 performs the reverse operation, merging one of the three loops into one of its neighbor loops.

To make an equiprobable shuffling from the present two-run state to any available three-run state, one first calculates the probabilities $P_{L_iI_j}$:

$$P_{L_iI_j} = \frac{(L_i - 1)(I_j - 1)}{(n-2)(N-n-2)}. \qquad (8)$$

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 $rand < P_{L_1I_1}$, then one cuts L_1 and moves the fragment into the I_1 region. If $P_{L_1I_1} \leq rand < P_{L_1I_1} + P_{L_1I_2}$, then one cuts L_1 and moves the fragment into the I_2 region. If $P_{L_1I_1} + P_{L_1I_2} \leq rand < P_{L_1I_1} + P_{L_1I_2} + P_{L_2I_1}$, then 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.

Next, one must determine the exact cut location and the position to which the movable fragment is moved. Suppose open region L_i has been chosen to be cut. One generates a second random number *rand2* in (0,1), and chooses the cut site to be after the base pair at position INT(*rand2*(L_i-1))+1 in L_i. Suppose the open subregion generated is to be moved into the closed region I_j. A random number *rand3* is generated, and the newly created open run is started at the position INT(*rand3*(I_j-1))+1 base pairs into I_j. The resulting shuffle operation from any two-run state to each available three-run state is equiprobable with probability $1/(n\text{-}2)(N\text{-}n\text{-}2)$.

To satisfy the detailed balance condition, 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. In this

operation, one first generates a *rand* in (0,1). If *rand* is smaller than $6/(n\text{-}2)(N\text{-}n\text{-}2)$, then one chooses one of the 6 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 $1/(n\text{-}2)(N\text{-}n\text{-}2)$, so detailed balance for exchange moves is satisfied. If *rand* is greater than $6/(n\text{-}2)(N\text{-}n\text{-}2)$, then shuffling to two-run states is not performed. In this situation, EX32 shuffles within three-run states, as required by the probability normalization condition implied in the definition of transition matrix. Correct shuffles within three-run states having n open base pairs are required to be independent of any details of a particular three-run state. Otherwise detailed balance is extremely difficult to satisfy. One option is to use SH33. Since this operation on three-run states is allowed only when three-run to two-run shuffles are forbidden, the probability of shifting between three-run states in this case becomes

$$\frac{1-6/(N-n-2)(n-2)}{2(N-n)-3}, \tag{9}$$

which guarantees probability normalization.

In our algorithm we construct two shuffling generation matrices that use exchange operations, called INTERCHANGE1 and INTERCHANGE2. INTERCHANGE1 is composed of EX12, EX21, EX34, EXC43, EX56, EX65 with ROTATION to treat states with run number higher than six. Similarly, EX23, EX32, EX45, EX54, EX67, EX76, together with ROTATION to treat states having run number equal one or greater than seven, comprise the other matrix INTERCHANGE2. The reason for designing two interchange shuffling operations

instead of one is to avoid complications in verifying detailed balance. The generating matrices for INTERCHANGE1 and INTERCHANGE2 can be decomposed into diagonal blocks, allowing an easier checking of detailed balance.

In the algorithm developed here we chose to confine shift, exchange and squeeze operations to states having run number $r \leq 7$ only because states having larger run number occur very infrequently in DNAs of reasonable length at physiological temperature. For example, our calculations show that phage λ DNA (48,502 bp) supercoiled to a superhelix density σ=-0.055 (θ=-254 turns) has only a 1% chance of occurring in states having eight or more runs of separation. If needed, shuffling trials that apply to states having larger run numbers can be constructed using the principles described here.

A modified Monte Carlo cycle (MMCC) consists of one standard MCC followed by a series of shuffling trials. The resulting simulation algorithm is named EXAMD. The tunable parameters of this algorithm are v, the number of shuffling trials performed after each standard MCC, and λ_s, the number of MMCCs performed before picking the next sample state. The values chosen for these parameters have important effects on simulation time.

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

$$t_{total} \propto [t(N)+t(v)] \times \lambda_s \times U. \qquad (10)$$

Here U is the sample size, $t(N)$ is the time needed for one complete standard MCC, which is proportional to N, and $t(v)$ is the time needed for finishing v shuffles, which is proportional to v. In MCA, $t(v)=v=0$, $\lambda_s = \lambda$ and the simulation time is proportional to $t(N)\times\lambda\times U$. In this case λ must be very large to avoid strong correlations among the sampled states. Sample calculations have shown that correlations remain even when λ=1000 (data not shown). The introduction of the shuffling trials in EXAMD reduces λ to the much smaller value λ_s. In practice, one shuffle requires approximately the same time as the trial of one base pair in an MCC. The shuffling time $t(v)$ is much smaller than the total MCC time $t(N)$ because the standard MCC tests each base pair in succession, and N is usually large, typically N>2000. A dramatic increase in efficiency arises from the introduction of shuffling operations. The convergence speed is greatly increased because correlations among sampled states are rapidly degraded by shuffling. This results in a much more efficient algorithm. Sample calculations illustrating these improvements are described in the **Results** section (2.3) below.

Design of an optimally efficient shuffling algorithm requires correct choices of the sampling parameters λ_s and v. Selection of their values is guided by the following consideration. The total number of shuffling trials performed to select one sample state is $\lambda_s v$. This quantity must be large enough so that the most infrequently selected operation occurs. These are exchange moves, which change the number of runs, where small probabilities occur (viz. $6/(n-2)(N-n-2)$ for EX32). In order to weaken correlations between sample states, these types of exchanges must be selected

occasionally. This requires a larger number $\lambda_s v$ of shuffles per sample point. If we select $0.01\bar{n}N \leq \lambda_s v \leq \bar{n} \times N$, where $\bar{n}$ is the average number of open base pairs. Then $\lambda_s v$ will be comparable to $(n\text{-}2)(N\text{-}n\text{-}2)$, so exchanges of states with different run numbers can happen easily. Here both v and λ_s should be chosen to be large enough so that both the number of open base pairs and the numbers and locations of open regions can freely vary in the selection of the next sampled state.

One can easily find an upper bound estimate for $\bar{n}$ using equation (6). If all base pairs in the sequence are AT's, then the lowest energy state will have one run of separation. The number of open base pairs found by minimizing the energy is

$$n = \left(\frac{\sqrt{2\pi^2 C}(4\pi^2 C - 2350Rt\sigma)}{\sqrt{2\pi^2 C + 110.25 b_{AT}}} - 4\pi^2 C \right) \frac{N}{2350Rt} \qquad (11)$$

Where $\sigma = \theta / LK_o$ ranges from -0.04 to -0.055 under normal physiological conditions and $LK_o = N/10.5$. In this range equation (11) shows that $0.018N \leq n \leq 0.032N$. This n will be an upper bound for $\bar{n}$ on the true sequence. In real calculation $\bar{n} \leq 0.018N$. If we select v and λ_s each to be large, with product $\lambda_s v \leq 0.018N^2$, then the total simulation time grows at most quadratically with molecular weight, for fixed sample size U. This shows that the method can treat long sequences efficiently.

(2.2.3) Further improvements to the algorithm

Methods that reduce the computation time $t(N)$ of the unit MCC will improve the overall efficiency of the resulting simulation procedure. One approach modifies the standard MCC by treating the base pairs in certain regions in blocks rather than individually. This modification is easiest to implement when the transition energetics are heteropolymeric so b has two values, b_{AT} and b_{GC}. Although it also can be implemented in cases having more complicated transition energetics, it becomes more cumbersome and the time savings decrease.

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

$$P_{cbp} = [1 - p\exp(-\Delta G(AT)/Rt)]^{n_{AT}}[1 - p\exp(-\Delta G(GC)/Rt)]^{n_{GC}} \qquad (12)$$

Here p is the probability used in the construction of the standard MCC, and $\Delta G(AT)$ (resp. $\Delta G(GC)$) is the free energy cost if an AT (resp. GC) pair is opened in this interior region. This cost is very large, about 10-12 kcal/mol, because a new run is initialized, so $P_{cbp} \cong 1$. Accordingly, we modify the standard MCC algorithm in the following way: Trials performed at sites in open loops and at their boundaries will be done in the usual base pair-by- base pair way. For interiors of closed regions the trial is modified as follows. We produce a

random number $rand$ in (0, 1). If $rand<P_{cbp}$, then 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 $rand\geq P_{cbp}$, 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

$$p_{AT} = \frac{n_{AT}\exp(-\Delta G(AT)/Rt)}{n_{AT}\exp(-\Delta G(AT)/Rt)+n_{GC}\exp(-\Delta G(GC)/Rt)} \qquad (13)$$

We choose a second random number $rand2$ in (0,1). If $rand2<p_{AT}$ then we will open an AT pair in this region and 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 an opening GC pair. After performing this opening, we use the standard trials to treat the base pairs next to this first open base pair in clockwise order until the next interior of a closed region is encountered. In this procedure the base pairs on either side of the newly opened site are treated in different ways. Those in the counterclockwise direction have been considered in aggregate, while those in the clockwise direction are treated individually. This approach cannot be shown to satisfy detailed balance. In consequence, the algorithm developed 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 an interior base pair in this way is very low.

The algorithm APPMD performs unit MCCs in this way, with shuffling operations performed after each MCC. This succession of

standard and modified trials 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 calculation time without sacrificing significant accuracy.

(2.2.4) Estimates of sample size and other parameters

An estimate of the minimum sample size U_o needed to achieve a given level of accuracy in a Monte Carlo simulation can be made using the Chebyshev (Eisen, 1969) and Kolmogorov inequalities (Moran, 1968). The estimates of U_o found here should be regarded as suggesting a lower bound on the actual sample size needed. The accuracy achieved in practical calculations having different sample sizes will be described in the **Results** section (2.3).

Let X_1, X_2, ... be random variables. Then the Chebyshev inequality states that

$$P\left\{\left|\frac{X_1+.....+X_U}{U}-E(\frac{X_1+.....+X_U}{U})\right|\geq\varepsilon\right\}\leq\frac{Var(\frac{X_1+.....+X_U}{U})}{\varepsilon^2} \qquad (14)$$

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

$$P\left\{\left|\frac{X_1+.....+X_U}{U}-E(X_1)\right|\geq\varepsilon\right\}\leq\frac{Var(X_1)}{U\varepsilon^2}, \qquad (15)$$

In our problem the sampled states are not exactly i.i.d. but approach this condition if λ_s (or λ) is large enough so successive points are nearly uncorrelated. If our sampled states are regarded as i.i.d., then several useful estimates can be obtained from this inequality.

The transition profile is the collection of ensemble average probabilities of transition of every base pair in the sequence. Let $B_s(i)$ be the random variable whose value is 1 if the s-th base pair is open in state i, and 0 otherwise. If the exact equilibrium probability of separation of the s-th base pair is P_s^B, then inequality (15) gives:

$$P\left\{\left|\frac{\sum_{i=1}^{U} B_s(i)}{U} - P_s^B\right| \geq \varepsilon\right\} \leq \frac{P_s^B(1-P_s^B)}{U\varepsilon^2} \qquad (16)$$

Here $Var(B_s)=P_s^B(1-P_s^B) \leq 1/4$. We take $\varepsilon=0.02$ and use the maximum possible variance $Var(B_s)=1/4$. Then a simulation having $U=20{,}000$ i.i.d. sampled states will have a 3.13% chance that its error in estimating P_s^B for any particular base pair exceeds 0.02.

The accuracy of the calculated average number of open runs can be estimated in a similar manner. Let r be the random variable corresponding to the number of open runs, i.e., $r(i)=k$, if in state i the number of open runs is k. We have the following formula

$$P\left\{\left|\frac{\sum_{i=1}^{U} r(i)}{U} - \bar{r}\right| \geq \varepsilon\right\} \leq \frac{(r_{max}-r_{min})^2/4}{U\varepsilon^2} \qquad (17)$$

where $\bar{r}$ is the exact value of the average number of open runs, $(r_{max} - r_{min})^2/4 \geq Var(r)$ (see appendix), 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 $(r_{max} - r_{min})^2/4 \leq 4$. If a bound on the accuracy of ε=0.05 is regarded as adequate and 20,000 states are sampled, then the ensemble average number of runs will be estimated correctly within ±0.05 approximately 92% of the time.

The same method can be used to estimate 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 base pair separation 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:

$$P\left\{\left|\frac{\sum_{i=1}^{U} n_i}{U} - \bar{n}\right| \geq \varepsilon\right\} \leq \frac{(n_{max} - n_{min})^2/4}{U\varepsilon^2} \qquad (18)$$

The choices ε=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 results indicate that a sample size of U_o=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.

(2.3) Results

In this section we present the results of several sample calculations, which implement the three algorithms, i.e., MCA (standard Monte Carlo cycles MCC without shuffling), EXAMD (standard MCC with shuffling), and APPMD (MCC modified by block estimates for opening of closed regions with shuffling). Their results are compared for accuracy with those from statistical mechanical calculations of known precision (Benham, 1990; 1992).

The first collection of sample calculations was designed to evaluate the accuracy, convergence properties and relative speeds of the algorithms developed above. These calculations are performed on the pBR322 DNA sequence at linking difference θ=-30 turns under low salt conditions, $[Na^+]$=0.01M, at 37^{O}C. These are the conditions under which Kowalski experimentally determined the locations and extents of strand separation in this molecule, from which data the values of the governing energy parameters have been determined (Kowalski. et al., 1988; Benham, 1992). First we examine the influence of sample size on accuracy. We perform sample calculations using each of the three algorithms at linking difference θ=-30 with sample sizes U=1000, 2000, 5000, 10000 and 20000 sampled states. In MCA, each sampled state is picked after λ=200 MCCs. In EXAMD λ_s=50 is used with ν=1600 shuffling operations performed after each MCC. These choices are considerably larger than required for accurate analysis of such a short molecule. They were selected to make the simulation time per sampled state approximately equal for these two methods. In APPMD, ν =240 and λ_s=150 were used.

Figure 4 shows the probability profile calculated by the statistical-mechanical technique of Benham (1990, 1992) under these circumstances. In that calculation the threshold was chosen to yield accuracy exceeding 99.9% 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 is between positions 4100 and 4300. These results agree closely with those from experimental determinations (Kowalski *et al.*, 1988; Benham, 1992).

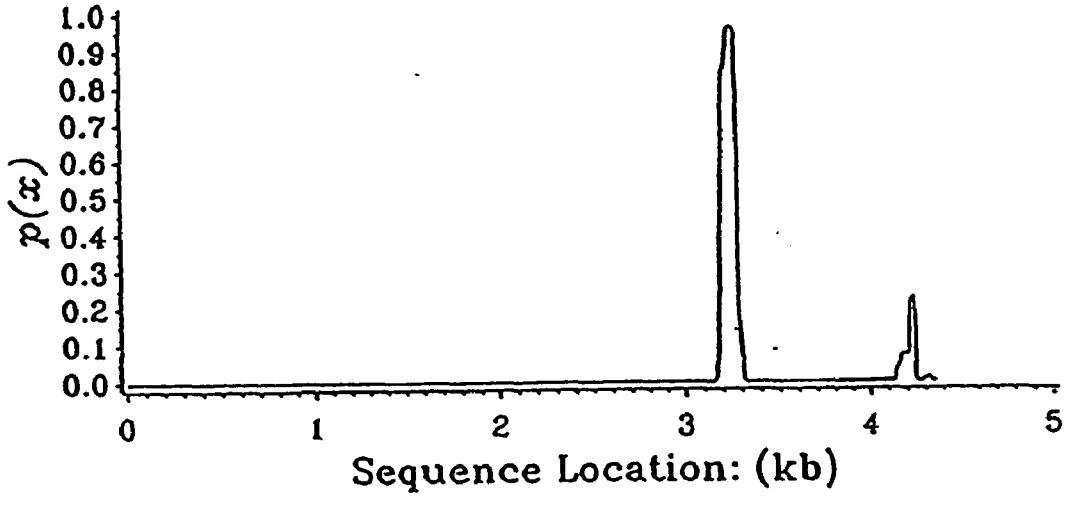

Figure 4

Fig. 4 The transition probability profile calculated for pBR322 DNA by the approximate statistical mechanical procedure is shown. The calculation assumes linking difference θ=-30 turns, at [Na^+]=0.01M and T=37°C. Two regions of high separation tendency are observed.

The probability profiles computed by Monte Carlo simulation also show transitions to be confined to these two regions. To analyze the accuracy of our simulation algorithms, we subtract the probability profile obtained using statistical mechanics from those found by each of the three Monte Carlo procedures. This determines the deviation $d(i)=p_{MC}(i)-p_{SM}(i)$ of the Monte Carlo results from the exact probability profile. Figure 5 shows these deviations as functions of position for sample size $U=20{,}000$ over the regions where separation occurs. The solid line gives the results from EXAMD which uses shuffling, while the dotted line is obtained using MCA, which does not. The maximum deviation of EXAMD is approximately an order of magnitude less than that of MCA. Successive sampled states found by the MCA procedure retain significant correlations, even when $\lambda=200$ MCCs separate them. Clearly, the use of shuffling operations significantly improves the accuracy of the results.

Next, Figure 6 plots the maximum deviation $D=\max|d(i)|$ in each of the two sensitive regions R1 and R2 between the statistical mechanical and the MCA and EXAMD simulation results for sample sizes $U=1000$, 2000, 5000, 10000, and 20000. These comparisons again show that simulation results become much more accurate when shuffling operations are applied. Moreover, the accuracy of the EXAMD algorithm, as measured by its maximum deviation from the statistical mechanical results in each of the two regions of significant separation, improves monotonically with sample size. The MCA algorithm lacking shuffling operations shows much larger and more case-specific variations with sample size.

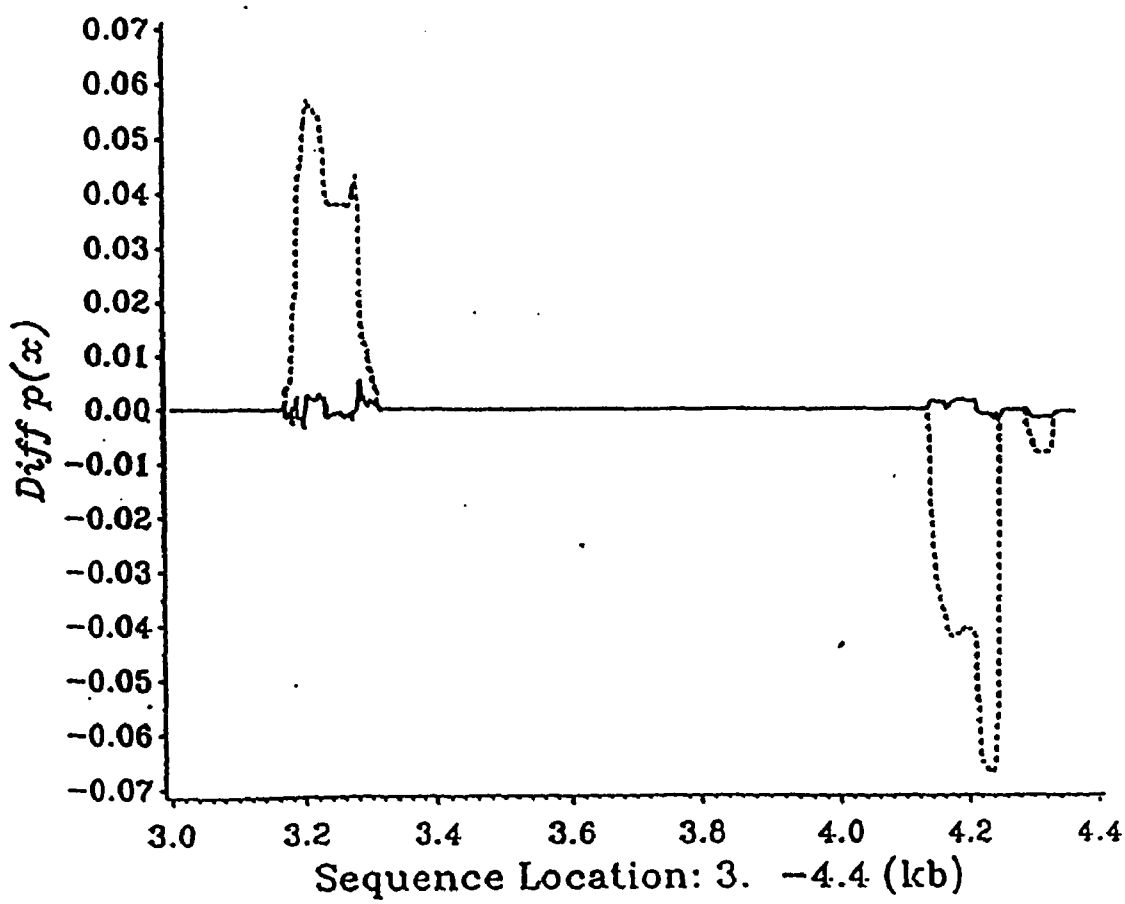

Figure 5

Fig. 5 This figure plots the deviation $d(x)=pMC(x)-pSM(x)$, which is the difference between the values of the separation probability for base pair x calculated by a Monte Carlo procedure $pMC(x)$, and the values $pSM(x)$ calculated by statistical mechanics. The X coordinate is the base pair position. The Y coordinate is the deviation of the probability profile calculated by two Monte Carlo methods from those obtained in Fig. 4. The dotted line is obtained from MCA (without shuffling), and the solid line is obtained from EXAMD (with shuffling). The introduction of shuffling operations decreases the maximum deviations by approximately one order of magnitude. In both cases the sample size is 20000.

We also note that the EXAMD and APPMD algorithms do not need relaxation time. The current state becomes essentially independent of the choice of initial state after very few sample points are taken.

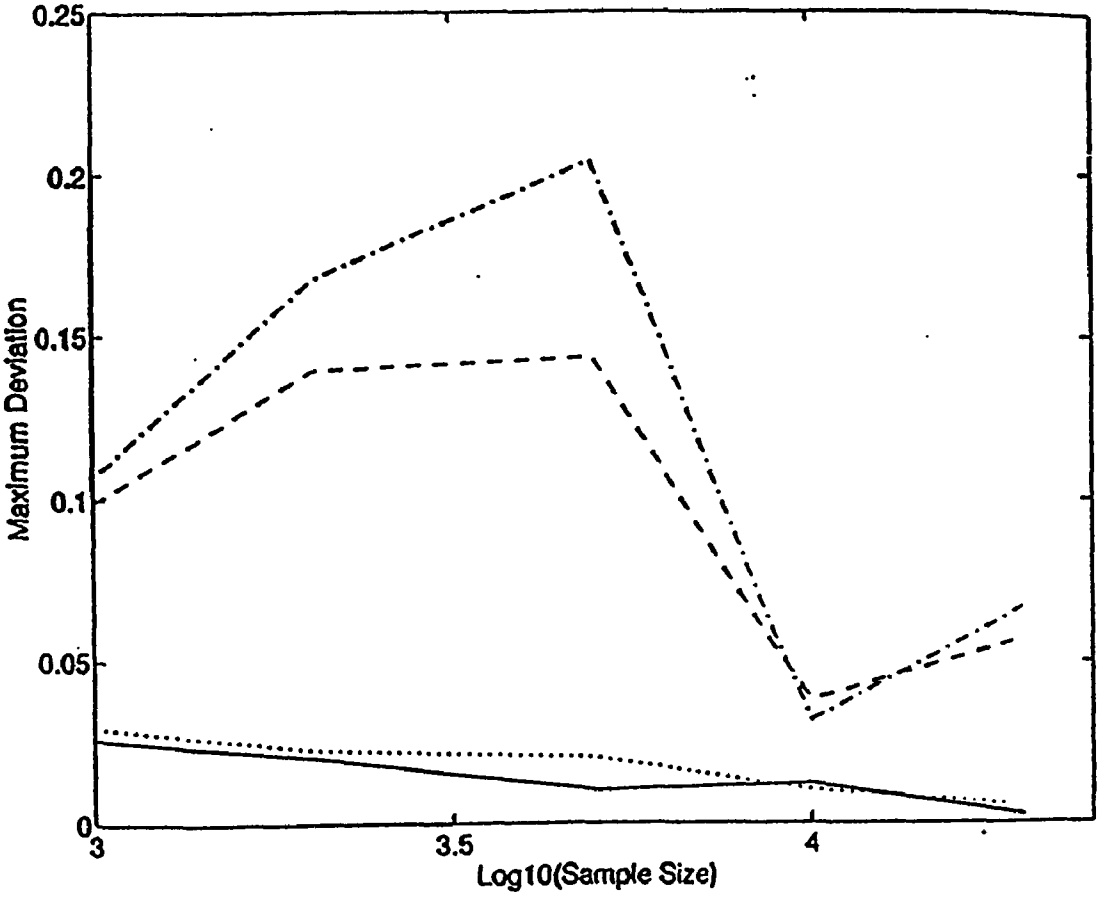

Figure 6

Fig. 6 The maximum deviation $D=\max|d(i)|$ in the two sensitive regions that are given in Fig. 4 are shown as functions of sample size. The line marked '-·-' gives the maximum deviations of the MCA results in region 1 (base pairs from 3100 to 3350); The line marked '--' gives deviations of the MCA in region 2 (base pairs from 4100 to 4300). The dotted line shows the maximum deviation of EXAMD in region 1, while the solid line gives the maximum deviations of EXAMD in region 2. The sample sizes are 10^3, 2×10^3, 5×10^3, 10^4, and 2×10^4.

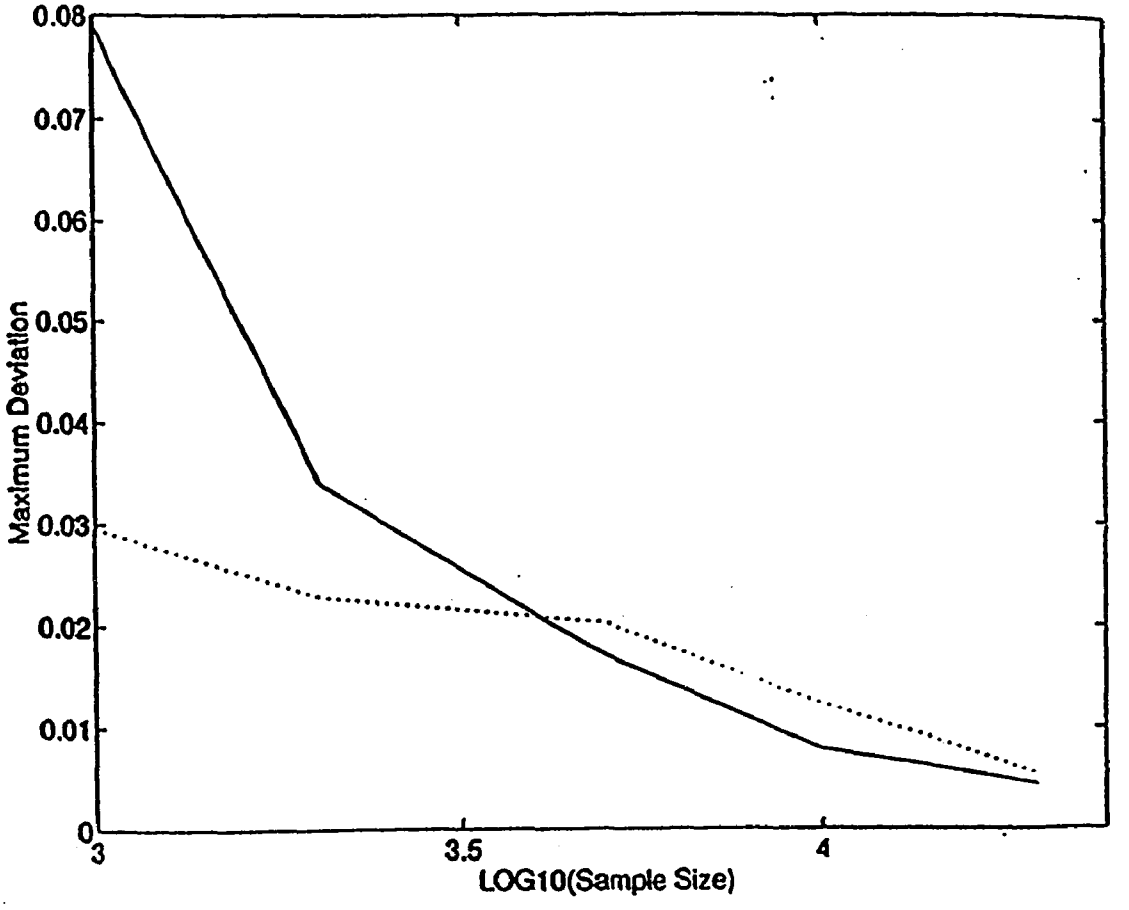

Figure 7

Fig. 7 The maximum absolute deviation between the statistical mechanical and Monte Carlo separation probability in the entire pBR322 DNA sequence, $D=\max_{1\le i\le N}|d(i)|$, is plotted against sample size. Solid line gives the maximum deviation of the APPMD algorithm, 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.

In contrast, MCA requires a long relaxation time. In our test calculation, even after collecting 1000 sample points by the MCA procedure the state retains some dependence on initial conditions (Data not shown). This shows that generating approximately

uncorrelated sample states using the MCA algorithm requires λ>1000.

In Figure 7 we compared the absolute maximum deviations $D=\max_{1\le i\le N}|d(i)|$ over the entire pBR322 molecule of the probability profiles obtained by EXAMD and APPMD from that obtained by statistical mechanics. The results are reported as functions of sample size. These results show that the two procedures have almost the same stability and accuracy for sample sizes exceeding 2000. However, APPMD executed significantly faster than EXAMD. The simulation times on a DEC 3000/800 computer for sample size U=20000 points was 9 hrs for EXAMD, and only 2.5 hrs for APPMD.

Values of other parameters calculated in these simulations are shown in Table 1. The quantity $<G>$ is the average free energy of the sampled states. In all cases the results for EXAMD and APPMD are comparable in accuracy. Achieving similar accuracy using MCA procedure requires calculations that are so long as to be unfeasible.

It is interesting to compare these simulation results with the Kolmogorov estimate found in last section. Inequality (16) states that fluctuations in the separation probability P_s^B for the base pair at position s, assuming sampling points are i.i.d. and sample size 20000, obey

$$P\left\{\left|\frac{\sum_{i=1}^{20000} B_s(i)}{20000} - P_s^B\right| \ge 0.02\right\} \le \frac{P_s^B(1-P_s^B)}{8} \qquad (19)$$

This means that at each site s the fluctuation in the probability P_s^B will be smaller than 0.02 with probability

$$p^f(s)=1-\frac{P_s^B(1-P_s^B)}{8}. \qquad (20)$$

Table 1
Statistical quantities vs. sample sizes

EXAMD θ=-30	open base pairs	open AT pairs	open GC pairs	$<G>$
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=10000	99.025	75.150	23.863	120.358
U=20000	99.069	75.252	23.817	120.413
APPMD θ=-30	open base pairs	open AT pairs	open GC pairs	$<G>$
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=10000	98.934	75.207	23.727	120.434
U=20000	98.994	75.273	23.721	120.452
statmech mthd*	99.037	75.084	23.953	120.4185

The first column lists the sample size values U for which the simulations are done. Each row contains the ensemble average values obtained corresponding to the U indicated in the first column.
*Statistical mechanics approach (Benham, 1990).
initial run:2200-2541 Seed=-5.

The i.i.d. condition assumes that the random variables B_s and B_q are independent when s and q are different sites. Hence the probability

P that every site on the pBR322 DNA molecule deviates from its exact value by less than 0.02 is:

$$P=\prod_{s=1}^{4363}(1-\frac{P_s^B(1-P_s^B)}{8}) \qquad (21)$$

Since we know P_s^B 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 the resulting probability is P=0.0422. This shows that a simulation with i.i.d. sampling and sample size 20000 will have a 96% chance of finding at least one base pair whose deviation from the exact result is larger than 0.02. On average 96 out of 100 such simulations will have some base pair where the error in evaluating the probability of separation exceeds 0.02.

To test this claim we performed the APPMD simulation four times using different initial conditions. Three of these simulations used the sample size U=20000, and the fourth used U=32000 sample points. No deviations beyond 0.02 were found in any of these simulations. When U is 32000, equation (21) gives P=0.133. If we exclude those base pairs whose probability of separating according to the statistical mechanical analysis is less than 0.03, then P=0.139, so 86% of simulations with this sample size would find deviations exceeding 0.02. The chances of four independent simulations all having maximum deviation less than 0.02 is 1.05×10^{-5}. We performed a similar analysis of EXAMD, in which six simulations were performed with different initial conditions. In all cases the sample size was U=20000. Two of these simulations had maximum

deviations smaller than 0.02. The above analysis suggested that the probability of such an occurrence under the i.i.d. assumption is 0.020. These results indicate that our Monte Carlo sampling procedure is comparable to or better than what would occur under strictly independent sampling. To understand this, one must examine the way in which the i.i.d. assumption is violated in our simulations. When successive sample points are not i.i.d., expected fluctuations are analyzed using inequality (14), which states that for each base pair s if sample size is 20000 and ε=0.02, then

$$P\left\{\left|\frac{\sum_{i=1}^{20000} B_s(i)}{20000} - P_s^B\right| \geq 0.02\right\} \leq \frac{Var\left(\frac{B_s(1)+.....+B_s(20000)}{20000}\right)}{4\times 10^{-4}} \qquad (22)$$

If the quantity on the right of this inequality is smaller than $\frac{P_s^B(1-P_s^B)}{8}$ in (19), then a larger value of P arises in equation (21), as happens in shuffling simulations. This will occur when successive sampled states are negatively correlated. This shows that the introduction of shuffling may change correlations of successive sampled states from strongly positive to slightly negative. Without shuffling, the sampled states will be confined near local energy minima for long times, so successively sampled states are strongly positively correlated. But shuffling trials may facilitate moves from one local minimum to another, making them even easier than they would be in the case of independent sampling. This would induce negative correlations between successively sampled states, making their quality even better than would occur in the i.i.d. case.

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 MCA, which lacks shuffling. Also, the APPMD algorithm using the modified trials executes significantly faster than does EXAMD, with comparable accuracy.

Next, we examine the influence of molecular length on these calculations. For this purpose we analyzed the phage λ DNA molecule containing 48,502 base pairs using APPMD algorithm. We choose v=1600 and λ_s=160. The sample size in each simulation is U=22500 sampled points. Simulations are performed for various linking numbers. Other physical parameters are the same as in the analysis of pBR322 described above.

Consider two molecules of different lengths supercoiled to the same superhelix density. The longer molecules usually will have a larger number of open base pairs and also a larger average number of open runs. 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 kcals. Now, consider states having n separated base pairs. Suppose the energetically most favored r-run state contains $n_{AT}(r)$ AT base pairs. If there is an r+1-run state also containing n separated base pairs, whose A+T-richness is at least 11 base pairs greater than this, it will have lower energy because the cost of initiating one more run is more than offset by the savings due to the increased A+T-richness. Thus, states having small numbers of runs of separation are favored when the expected number of

separated base pairs is small (roughly ≤ 100 bp). For short molecules ($N \leq 5{,}000$ 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. Table 2 shows how the fraction of open runs in the phage λ DNA changes with linking difference θ. This shows that six-run states are most populated when θ=-254, which corresponds to a superhelical density of σ=-0.055. In pBR322 DNA at this superhelical density, the probability of states with more than one run is less than 0.25.

Table 2
Open run fractions

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

The first column lists all linking difference values for which the simulations were done. Each row contains the results obtained corresponding to the linking difference indicated in the first column.

Table 3 shows several average values calculated for phage λ DNA using the APPMD algorithm. These include the average numbers of open base pairs, open AT pairs, open GC pairs, and the average free energy $<G>$, calculated at various values of linking differences θ. All these quantities increase approximately linearly as θ decreases.

Table 3
Statistical quantities vs. linking differences

θ	open base pairs	open AT pairs	open GC pairs	$<G>$
-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

The first column lists all linking difference values for which the simulations were done. Each row contains the ensemble average values obtained corresponding to the linking difference indicated in the first column.

The time required to perform these simulations increases only slightly with the magnitude of the imposed linking difference θ. The CPU time used by the DEC 3000/800 computer to do these calculations using APPMD ranged from 19.5 hrs to 24 hrs, with about a half hour increase for each change of -10 in linking difference. This compares favorably with the approximate statistical mechanical analysis, where the number of states satisfying a fixed threshold condition, and hence the computation time, increases rapidly as the molecule becomes more negatively supercoiled.

To test the accuracy of the APPMD algorithm, we compared its computed probability profile for phage λ with that calculated by the approximate statistical mechanical method. When θ=-177 turns, the maximum deviation between these profiles is less than .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 (with shuffling operations) is effectively independent of the molecular length.

We note that the difficulty of the exact statistical mechanical analysis increases very rapidly with run number. In practice, calculations where states with four or more runs occur are not feasible using this technique. Therefore, accurate comparisons between the statistical mechanical and the Monte Carlo techniques are possible only at linking differences where low energy states having four or more runs of separation do not occur. The Monte Carlo approach is the only known feasible way to calculate separation probabilities under conditions where large numbers of runs occur, i.e., for long DNA molecules that are substantially supercoiled.

To make a comparison between APPMD and EXAMD, we also performed simulations on phage λ DNA at θ =-177 and -187, respectively. In EXAMD we set λ_s=40 and ν=6400. In APPMD we set λ_s=160 and ν=1600. The total number $\lambda_s\nu$ of shuffling operations performed before picking each sampled state is the same in these two algorithms. In each case we computed the maximum deviation $D=\max_{1\le i\le N}|d(i)|$ from the probability profile calculated by the statistical mechanical algorithm. At θ=-177, the maximum deviation for EXAMD is D $\le$0.026, and for APPMD it is D$\le$0.018. When θ=-187, D$\le$0.049 for

EXAMD and D≤0.059 for APPMD, while the difference between EXAMD and APPMD never exceeds 0.015. Both EXAMD and APPMD begin to deviate from the statistical mechanics results around this linking difference. Table 2 shows that at θ=-187, four-run states have fractional probability 0.013. The statistical mechanics algorithm ignores states having run number greater than three, causing it to lose accuracy in this range.

These results on phage λ DNA again show that the accuracies attained by EXAMD and APPMD are comparable for all calculated quantities. The values of $<G>$ calculated by these procedures agree within 0.2%, while the expected numbers of separated base pairs agree to better than 2%. However, APPMD is significantly faster. When θ=-177, the simulation times were 19.5 hrs for APPMD and 52 hrs for EXAMD. When θ=-187, these simulation times were 20 hrs and 55 hrs, respectively. All calculations were performed on a DEC 3000/800 computer.

To test the stability of the APPMD algorithm, we made two simulations from different initial states at linking difference θ=-247. The result shows that the two values of $<G>$ 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 converge to the equilibrium distribution, as shown by the calculations on pBR322 DNA, also demonstrated their numerical stability.

(2.4) Discussion

The Monte Carlo procedure developed here provides a new method for calculating equilibrium properties of the superhelical DNA strand separation transition that does not have the limitations of alternative methods. The sample size needed to achieve a prescribed accuracy can be estimated in advance by the procedures given here. Calculations having that accuracy can be performed for long DNAs at any reasonable linking difference. The size of the calculation can be estimated to grow at most quadratically with molecular length. In practice, our calculations show that the CPU time required for a simulation of sample size U=20,000 increases approximately linearly with molecular length, and very slowly with imposed linking difference, regardless of the number of runs of separation involved.

The results of the Monte Carlo simulations that use shuffling operations agree very closely with those from the statistical mechanical procedure, whose accuracy can be made as high as desired by setting the threshold appropriately. The accuracy of these simulation procedures is at least as good as, and sometimes even better than, could be expected if the i.i.d. condition held.

The energetics of separation of each base pair in the sequence can be individually specified in the Monte Carlo procedure. Thus, one can include near-neighbor effects and structural modifications such as methylation, lesion formation, ligand binding, or other alterations that could affect transition energetics. These cannot be included in the approximate statistical mechanical procedure as currently structured. Calculations modeling 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 that technique one can calculate the incremental free energy needed to separate any base pair in the sequence (Benham, 1993). This finds sites that are partly destabilized by imposed stress, so superhelicity significantly reduces the energy of separation, though not enough to induce 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 Figure 8. This 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 (Klenin *et al.*, 1991; Levene & Crothers, 1986; Zhurkin *et al.*, 1991; Vologodskii *et al.*, 1992). A central reason for developing Monte Carlo methods to treat secondary structure transitions is that this is a required step in 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 attempt to amalgamate them into a unified technique to analyze superhelical DNA structure in its full generality.

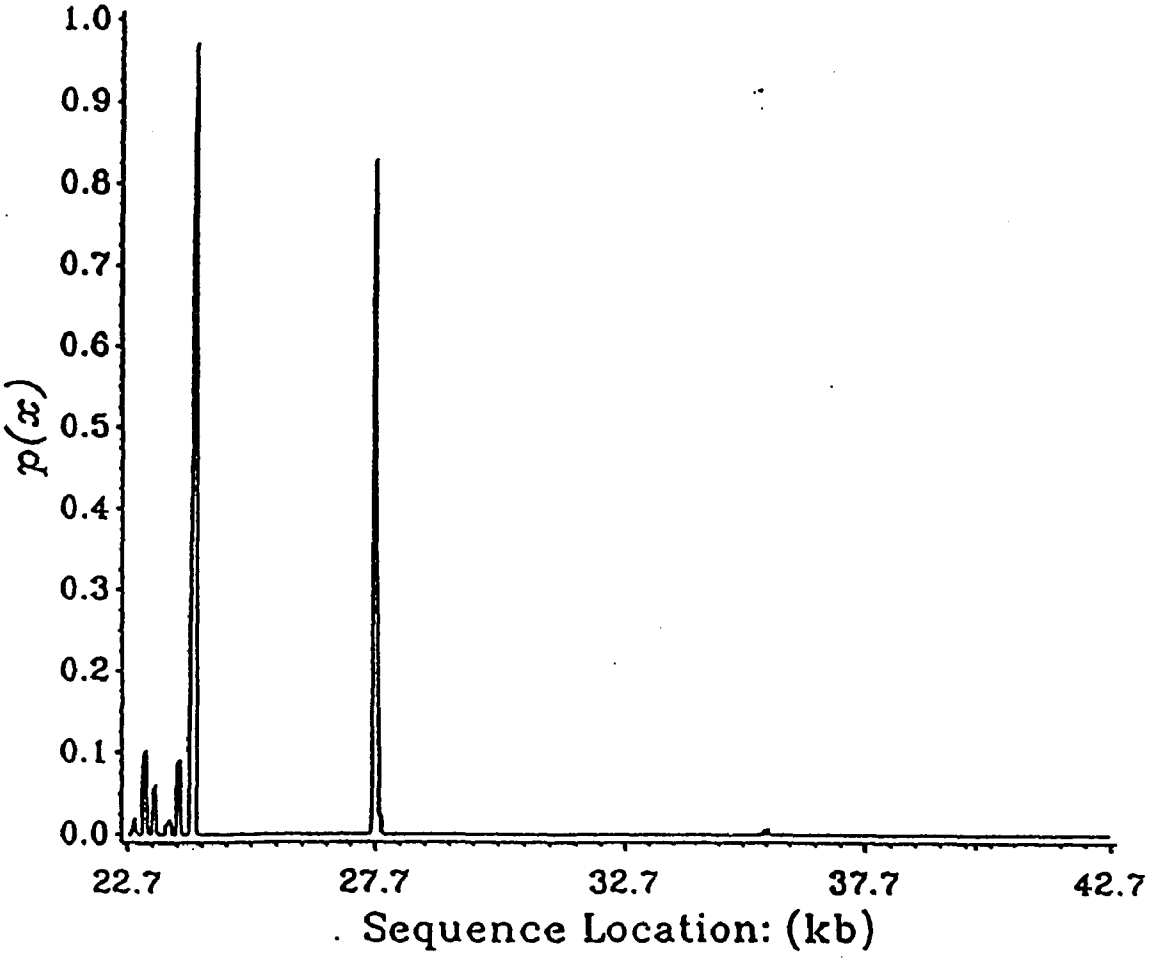

Figure 8

Fig. 8 Transition profile of phage λ DNA, as calculated at linking difference θ=-187 turns. The part of the sequence that is not plotted showed no destabilization in this sample calculation.

CHAPTER 3

SHUFFLING METROPOLIS-MONTE CARLO SIMULATIONS OF THE PHASE TRANSITION IN PHOSPHOLIPID MEMBRANES

(3.1)Two-state Ising model of membrane phase transitions

The gel-liquid crystalline phase transitions of DPPC MLV (dipalmitoyl-phosphatidylcholine multilamellar vesicles) are sharp transitions (the half width is about 0.07° C) with no hysteresis. The phase transition enthalpy is 8.5-8.7 kcal/mol (Albon & Sturtevant, 1978; Biltonen, 1990), and the volume change is 0.037ml/g (Albon & Sturtevant, 1978; Peskun, 1981).

According to the wide angle x-ray diffraction measurements (Sugar, *et al.*, 1994) on DPPC membranes, the acyl chains of the phospholipid molecules are located on the points of a triangular lattice in both phases, however the lattice spacing is larger in liquid crystalline phase. Laser Raman (Yellin and Levin, 1977) and deuterium-NMR (Seelig and Niederberger, 1974) spectroscopy on DPPC membranes revealed that the acyl chains are in all-trans conformation in gel phase, although one rotational isomer may appear sometimes at the chain end. In the liquid crystalline phase, however, there are about 4-5 rotational isomers/chain with changing location and orientation along the chain. The intrachain energy increases with the number of rotational isomers by 500 cal/mol (Person and Pimentel, 1953). The Van der Waals interaction between the acyl chains is a short-range interaction which decreases with the 5th power of the interchain separation (Nagle and Wilkinson, 1978).

In our theoretical analysis one monolayer of the lipid bilayer is modeled using a triangular lattice. Each lattice point represents an acyl chain. A lattice point exists in either gel (g) or liquid crystalline (l) state. V_g and V_l are the volume of a chain in the gel and in the liquid crystalline state, respectively. The intrachain energies in the gel state, E_g, and in the liquid crystalline state, E_l, are assumed to be constant and independent of the location and orientation of the rotational isomers. It follows that the energy level E_l is highly degenerate. According to the calorimetric data (Albon and Sturtevant, 1978) the ratio of the degeneracy of the two energy level E_l and E_g is estimated to be $W_l/W_g \approx 1020$.

Only nearest neighbor interactions between the lattice points are considered. E_{gg}, E_{gl} and E_{ll} are the interaction energies between a pair of nearest neighbor chains of gel-gel, gel-liquid crystalline and liquid crystalline-liquid crystalline states, respectively. The interaction energies are assumed to be unaffected by the location and orientation of the rotational isomers in the interacting chains, and thus the interaction energies also are degenerate. The degeneracies of E_{gg}, E_{gl} and E_{ll} are W_{gg}, W_{gl} and W_{ll}, respectively.

Depending on the state of each lattice point the system can be in different configurations. Let us consider a configuration C where N_g lattice points are in the gel state and N_l lattice points are in the liquid crystalline state, and N_{gg}, N_{gl} and N_{ll} are the number pairs of nearest neighbor lattice points in the gel-gel, gel-liquid crystalline and liquid crystalline-liquid crystalline states, respectively. The energy E_C and volume V_C of this configuration are

$$E_C = E_g N_g + E_l N_l + E_{gg} N_{gg} + E_{gl} N_{gl} + E_{ll} N_{ll}.$$
$$V_C = V_g N_g + V_l N_l. \qquad (1)$$

The number of microstates belonging to this configuration, W_C, is calculated from the degeneracies of the lattice states and of the nearest neighbor interactions

$$W_C = W_g^{N_g} W_l^{N_l} W_{gg}^{N_{gg}} W_{gl}^{N_{gl}} W_{ll}^{N_{ll}}. \qquad (2)$$

The lattice is in thermal and mechanical contacts with its surroundings at temperature t and pressure p. The partition function of the system is

$$Q(N,t,p) = \sum_{\{\phi\}} \exp[-(E_\phi + pV_\phi)/kt] = \sum_{\{C\}} W_C \exp[-(E_\phi + pV_\phi)/kt], \qquad (3)$$

where $N(=N_g+N_l)$ is the total number of lattice points in the system and k is Boltzmann constant. The first summation in Eq. 3 is taken for every microstate and the second is for every configuration.

In the case of periodic boundary conditions the following two relationships exist:

$$N_{gg} = 3N_g - (N_{gl}/2) \qquad (4)$$

and

$$N_{ll} = 3N_l - (N_{gl}/2). \qquad (5)$$

During the Monte Carlo simulation different lattice configurations are selected randomly as candidate states of the

system. Let us assume that the lattice is in configuration C and the candidate state is in configuration C'. The trial state is accepted if

$$R < \frac{\pi_{C'}}{\pi_C} = \frac{W_{C'}}{W_C} \exp[-(E_{C'} - E_C)/kt] = \exp[-(\chi' - \chi)/kt], \qquad (6)$$

where symbols with prime refer to configuration C'. Here R is a random number generated between 0 and 1, π_C is the equilibrium probability of configuration C, and the χ function is defined by

$$\chi = E_C + pV_C - kt \ln W_C. \qquad (7)$$

After substituting Eqs. 1-2 into Eq. 7 and then eliminating N_{gg} and N_{ll} by means of Eqs. 4-5, one gets χ as a function of N_{gl} and N_l,

$$\chi = [\Delta H - t\Delta S]N_l + wN_{gl}, \qquad (8)$$

with $w = [w_E - tw_S]$. In Eq. 8 the following notations are used,

$$\Delta H = E_l + pV_l + 3E_{ll} - (E_g + pV_g + 3E_{gg}) \qquad (9)$$

and

$$\Delta S = k \ln W_l - k \ln W_g, \qquad (10)$$

where w_E is the energy part of the cooperativity parameter $w(=w_E - t w_S)$, given by

$$w_E = E_{gl} - (E_{gg} + E_{ll})/2, \qquad (11)$$

while w_S is the entropic part of the cooperativity parameter,

$$w_S = k[\ln W_{gl} - (\ln W_{gg} + \ln W_{ll})/2]. \qquad (12)$$

In Eq. 9 ΔH is the transition enthalpy (the enthalpy change per site when the system changes from the completely gel to the completely liquid crystalline phase). For DPPC ΔH=4.25-4.35 kcal/mol/chain (Albon and Sturtevant, 1978).

The value of ΔS in Eq. 10 can be calculated from ΔH and the phase transition temperature t_m. At the phase transition temperature the probability of gel phase membrane is equal to that of the liquid crystalline phase membrane, i.e.,

$$\pi(N_g = N, N_{gl} = 0) = \pi(N_l = N, N_{gl} = 0),$$

or (13)

$$\exp[-\chi(N_g = N, N_{gl} = 0)/kt_m] = \exp[-\chi(N_l = N, N_{gl} = 0)/kt_m].$$

After substituting Eq. 8 into the equation above we get the following relationship:

$$\Delta S = \Delta H / t_m \qquad (14)$$

and by using this relation ΔS=13.85 cal/deg/chain for DPPC MLV is obtained.

(3.2) Methods

(3.2.1) The Common Glauber method

In the Glauber method (Glauber, 1963) a Monte Carlo cycle (MCC) is defined such that the number of opportunities for lattice points to change state is equal to the number of points in the lattice, N. The points within a lattice are picked randomly, with probability $1/N$. The algorithm, according to the Markov chain theory, can be mathematically represented by a $2^N \times 2^N$ candidate state generation matrix M, where 2^N is the number of configurations in the lattice. During candidate state generation there is direct access to N different configurations from the current state, thus every row of the matrix M contains N elements, each equal to $1/N$, while the remaining $2^N - N$ elements are zeros. The random picking of lattice points ensures the symmetry of this matrix. Any configuration of the membrane can be reached from any other configuration within N trials because any two configurations differ at no more than N lattice points. Since the transition matrix M satisfies the symmetry and ergodicity conditions, the transition probability matrix P constructed according to Eqs. 3-4 of chapter 1 generates a formally correct simulation. However, while simulating the membrane phase transition we get long, strongly correlated sequences of configurations when $\chi' - \chi > 3kt$ during the initiation of a new phase region. (Here k=2 cal/mol/deg). According to Eq. 8 the change in χ is $\Delta\chi = zw$ (z=6), when a liquid crystalline state is initiated within a gel phase region at t_m, and thus the long, correlated sequences appear when $w \geq 3kt_m/z = kt_m/2$. In this case the convergence to the equilibrium distribution becomes very slow, and the equilibrium distribution is not attainable within a feasible computer time.

(3.2.2) The Glauber method with shuffling

Convergence of the Glauber method can be accelerated by introducing the following simple shuffling trial into the algorithm: Change all gel-state molecules in the membrane to liquid-state molecules and at the same time change all liquid-state molecules to gel-state. This global, non-physical operation promotes phase changes without energy-intensive local initiation of the new phase. The operation is truly shuffling because the respective shuffling generation matrix S is a symmetric, stochastic $2^N \times 2^N$ matrix. In each row of S there is only one non-zero element, which is 1. The transition probability matrix of the shuffle trial, R, is constructed from S according to Eqs. 3-4 of section (1.2). In the actual simulation the Glauber method is interrupted by shuffling after every 17th MCC (There is no special reason for choosing 17. In fact, any integer ≥ 2 works). The respective transition probability matrix $P^{17N}R$ generates a correct simulation because P^{17N} is a positive matrix. The proof is as follows.

Proposition 8: P^{2N} is a positive matrix, where P is the transition probability matrix of the common Glauber algorithm, and N is the total number of the lattice points.

Lemma 1.

If there is a configuration i such that $p_{ii}>0$, then for any configuration j of the membrane the (i,j)th element, $p_{ij}(N)$, of the product matrix P^N is positive.

Proof

Configuration i of the membrane can be changed into any chosen configuration j by changing the state of t∈[1,N] lattice points, *i.e.*, the system is ergodic $p_{i,l_1}p_{l_1,l_2}\cdots p_{l_{t-1},j}>0$, where l_1, l_2,...,l_{t-1} are the intermediate $t-1$ configurations between state i and j, and thus the following inequality holds for the (i,j)th element of the product matrix P^t

$$p_{ij}(t) \geq p_{i,l_1}p_{l_1,l_2}\cdots p_{l_{t-1},j}>0 \tag{15}$$

Therefore, $p_{ij}(N) \geq p_{ii}^{N-t} p_{i,l_1}p_{l_1,l_2}\cdots p_{l_{t-1},j}>0$.

Lemma 2

If there is a configuration i such that $p_{ii}>0$, then P^{2N} is a positive matrix.

Proof

From *Lemma 1* we have $p_{ij}(N)>0$ for any j. This means one can select a chain of N-1 intermediate states (l_1, l_2,...,l_{N-1}) leading from state i to j such that

$$p_{ij}(N) \geq p_{i,l_1}p_{l_1,l_2}\cdots p_{l_{N-1},j}>0 \tag{16}$$

According to the construction of P, $p_{ik}>0$ means $p_{ki}>0$ too. Thus one can return from state j to state i on the same path with a positive probability, i.e.,

$$p_{ji}(N) \geq p_{j,l_{N-1}}p_{l_{N-1},l_{N-2}}\cdots p_{l_1,i}>0. \tag{17}$$

This shows that $p_{jl}(N)$ is also positive. Finally, any element $p_{kl}(2N)$ in P^{2N} satisfies the following inequality:

$$p_{kl}(2N) \geq p_{kl}(N)p_{ll}(N)>0 \qquad (18)$$

Thus P^{2N} is positive.

The assumption in this proof that there is a state i such that $p_{ii}>0$ follows when i is such that $\pi_i = \max\{\pi_j, 1 \leq j \leq 2^N\}$.

(3.2.3) Example of $N=4$

In the simplest case the membrane is a 2×2 lattice ($N=4$) and the number of configurations is 16. By labeling the configurations with 4-digit binary numbers one can give a natural order to the configurations. For example, 0000(=0) means that every molecule is in gel state, 0001(=1) means that the first three molecules are in gel state while the last one is in liquid state, etc.. The 16x16 matrix M has the following form:

$$M = \begin{bmatrix} A & I/4 & I/4 & 0 \\ I/4 & A & 0 & I/4 \\ I/4 & 0 & A & I/4 \\ 0 & I/4 & I/4 & A \end{bmatrix},$$

where $I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $A = \begin{bmatrix} 0 & 1/4 & 1/4 & 0 \\ 1/4 & 0 & 0 & 1/4 \\ 1/4 & 0 & 0 & 1/4 \\ 0 & 1/4 & 1/4 & 0 \end{bmatrix}$ (19)

Here the rows and columns of M matrix are indexed according to the code numbers of the configurations. The 16×16 shuffling generating matrix S has a much simpler structure than the M matrix. Every element is 1 in the antidiagonal of S matrix, while the other elements are zeros.

(3.3) Results and discussions

(3.3.1) Equilibrium and Hysteresis

Monte Carlo simulations of the gel-to-liquid crystalline phase transition of lipid membranes have been performed by means of two different protocols: (1) the common Glauber method, and (2) the Glauber method interrupted after every 17th Monte Carlo cycle by a shuffling trial. In the simulations the following parameters of DPPC MLV phase transition were utilized: phase transition enthalpy ΔH = 4350 cal/mol chain; phase transition entropy ΔS=14 cal/deg/mol chain. The lattice size was N=65x65. The cooperativity parameters were $w_E = 0$ cal/mol chain, w_S=-1.3 cal/deg/mol chain. These cooperativity parameters are not directly measurable. At these w_E and w_S values the change of χ during the initiation of a new phase at t_m is $3.9kt$. This change is so large that in the case of the common Glauber method one can expect very slow convergence to the equilibrium distribution at the phase transition temperature.

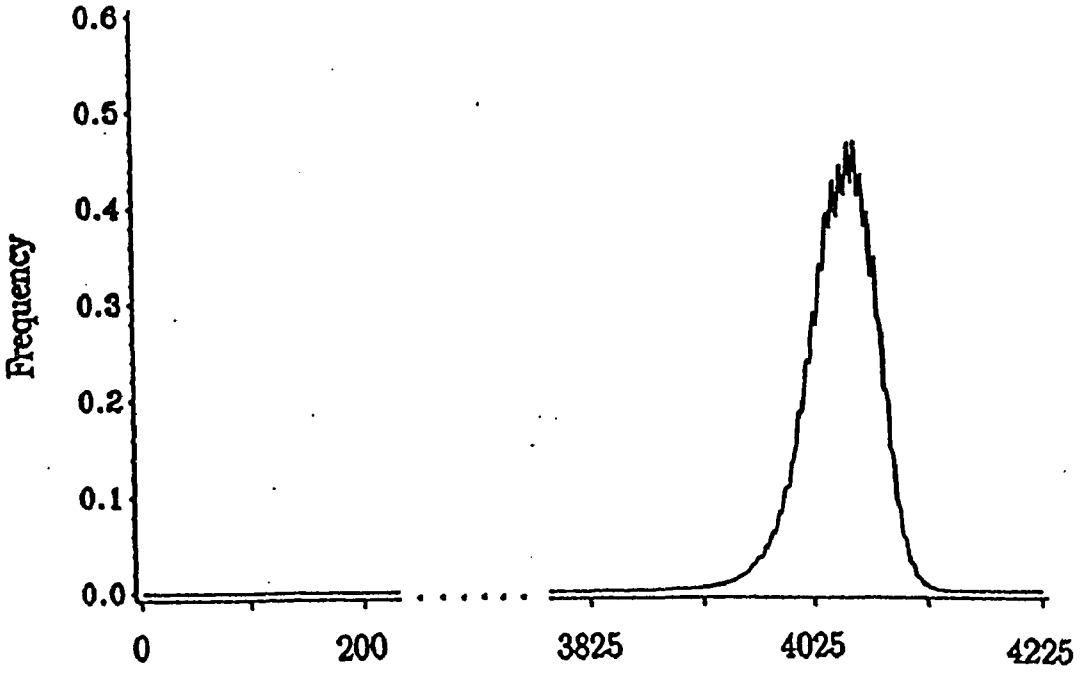

Figure 9a

Fig. 9 Frequency distributions of the number of liquid state acyl chains, N_l. a) Common Glauber method. Initial state: all-liquid crystalline. Temperature: $t=t_m$-1.48(K). Parameters of the simulations: $w_E=0$ cal/mol chain, w_S=-1.3 cal/deg/mol chain, ΔH=4350 cal/mol chain, ΔS=14 cal/deg/mol chain. To construct each frequency distribution the data were grouped into 128 classes.

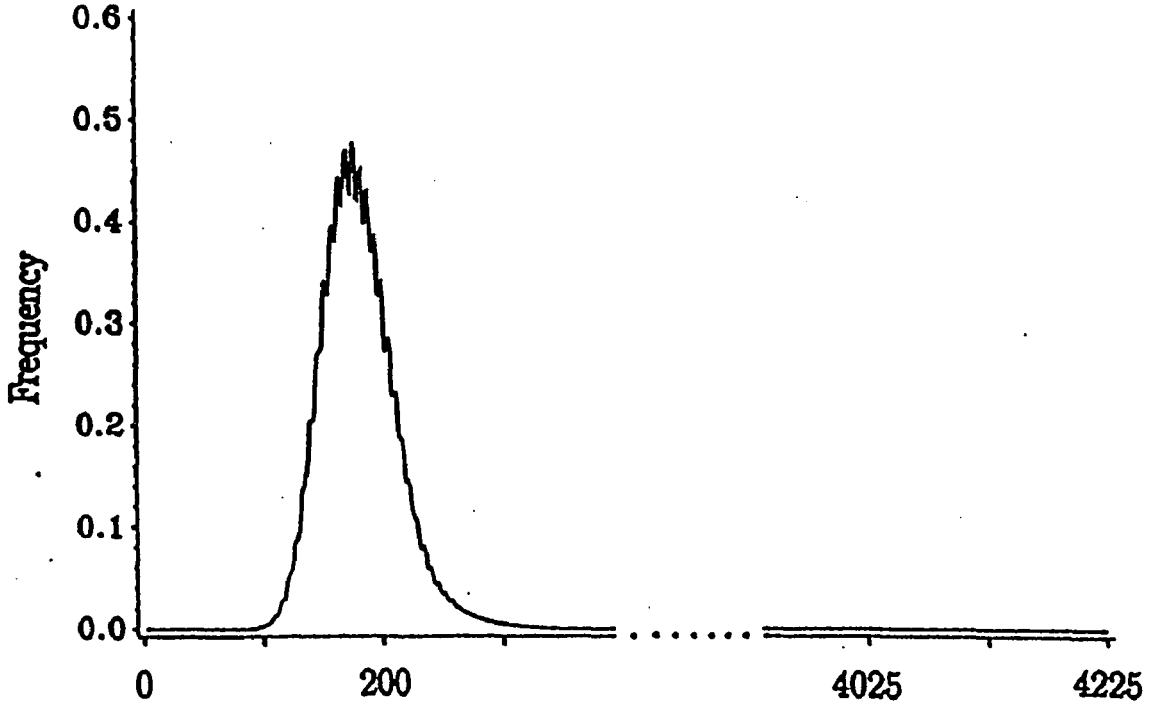

Figure 9b

Fig. 9 Frequency distributions of the number of liquid state acyl chains, N_l. b) Common Glauber method. Initial state: all-gel phase. Temperature: $t=t_m$-1.48(K). Parameters of the simulations: $w_E=0$ cal/mol chain, w_S=-1.3 cal/deg/mol chain, ΔH=4350 cal/mol chain, ΔS=14 cal/deg/mol chain. To construct each frequency distribution the data were grouped into 128 classes.

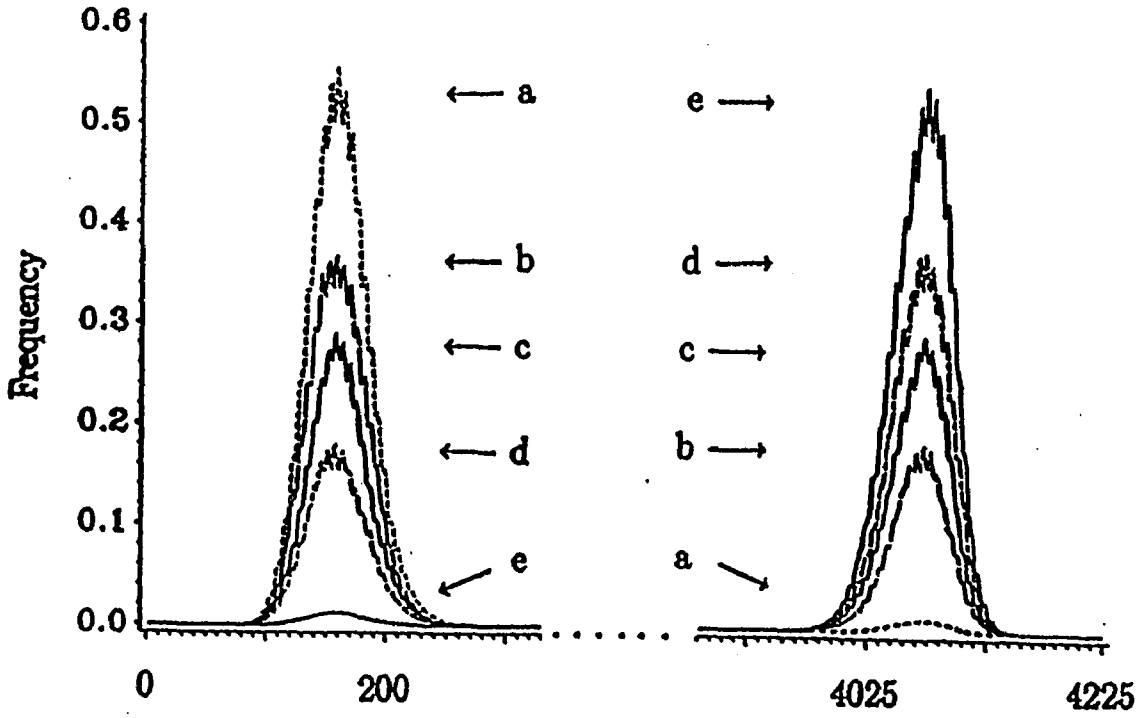

Figure 9c

Fig. 9 Frequency distributions of the number of liquid state acyl chains, N_l. c) Glauber method combined with shuffling trial after every 17th Monte Carlo cycle. The biphasic distribution curves obtained at five different temperatures are letter-coded at the tip of each peak (a: $t=t_m$-0.05 K; b: $t=t_m$-0.01 K; c: $t=t_m$ K; d: $t=t_m$+0.01 K; e: $t=t_m$+0.05 K). Similar distribution curves were obtained regardless of the initial state. Parameters of the simulations: $w_E=0$ cal/mol chain, w_S=-1.3 cal/deg/mol chain, ΔH=4350 cal/mol chain, ΔS=14 cal/deg/mol chain. To construct each frequency distribution the data were grouped into 128 classes.

Simulations were performed at eleven different temperatures near and at the midpoint of the transition, starting from either an all-gel state or an all-liquid crystalline state. After an equilibration period of 1000 Monte Carlo cycles the snapshots of 10^5 consecutive cycles were analyzed in order to get a frequency distribution of N_l.

Frequency distributions in Figs.9a and 9b belong to all-liquid crystalline and all-gel initial states, without using shuffling trials, respectively. Figs. 9a,b show that, near the midpoint temperature of the transition, the common Glauber method results in different distributions for different initial conditions, while the combined use of the Glauber method and shuffling trials leads to the same distribution, independent of the initial configurations as shown in Fig. 9c. Thus in the case of the common Glauber method the equilibrium distribution was not attained after 10^5 MC cycles.

The convergence to the equilibrium was monitored by means of the average of N_l:

$$\overline{N}_l(NC) = [\sum_{i=1}^{NC} N_l(i)]/NC. \qquad (20)$$

Where NC is the number of MC cycles. In Fig. 10 $\overline{N}_l$ is plotted against NC. Starting the simulations from all-gel state, with shuffling the equilibrium value ($<N_l>$=0.69 at $t=t_m$+0.01K) was closely approached by $\overline{N}_l$ after $\approx$ 700 MC cycles. However, without shuffling, $\overline{N}_l$ remained close to zero at $t=t_m$+0.01K even after 10^5 MC cycles.

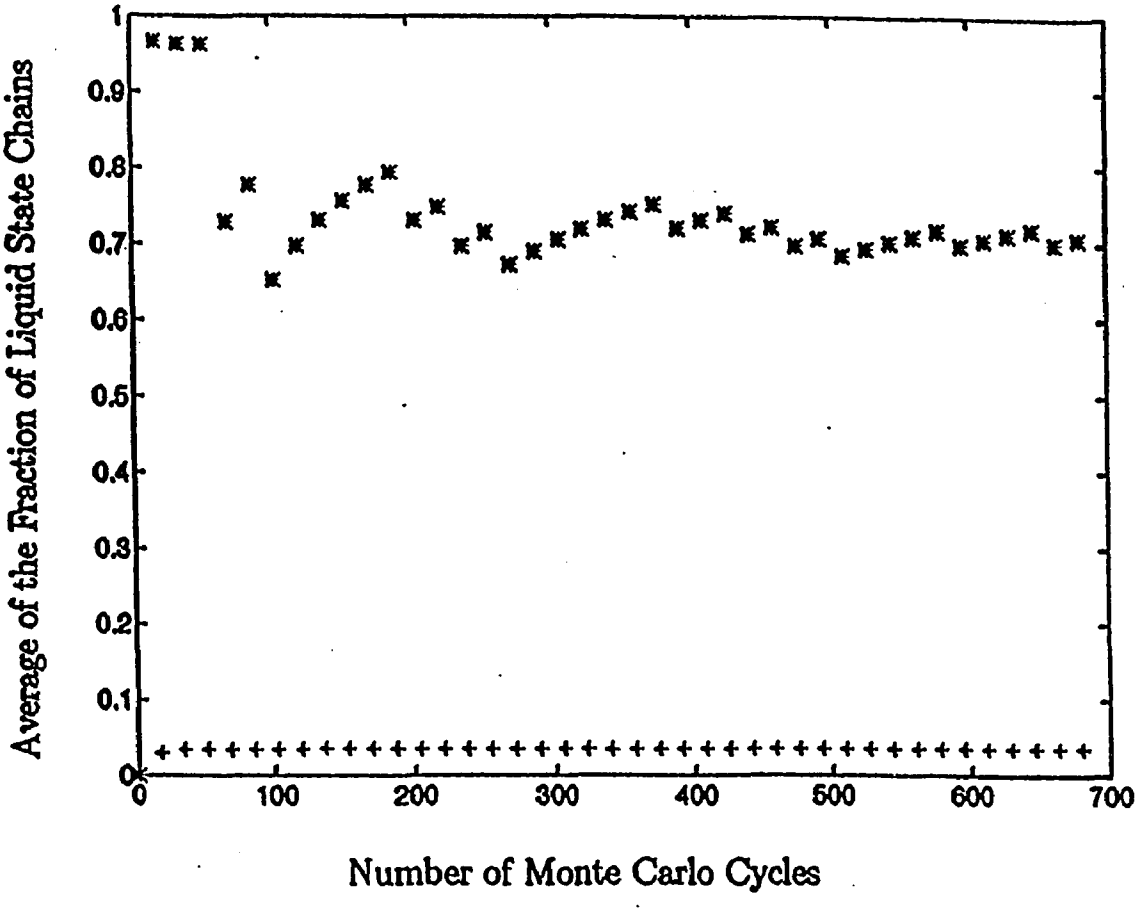

Figure 10

Fig. 10 Convergence to the equilibrium value of N_l. The average of the number of liquid state chains $\overline{N_l}$ (section 3.3.1 Eq. 20) is plotted against the number of MC cycles. Initial state: all-gel. Temperature: $t=t_m+0.01$ K. (+): Standard Glauber method. (*): Glauber method combined with shuffling trial after every 17th Monte Carlo cycle. The other parameters are given in the legend to Fig. 9.

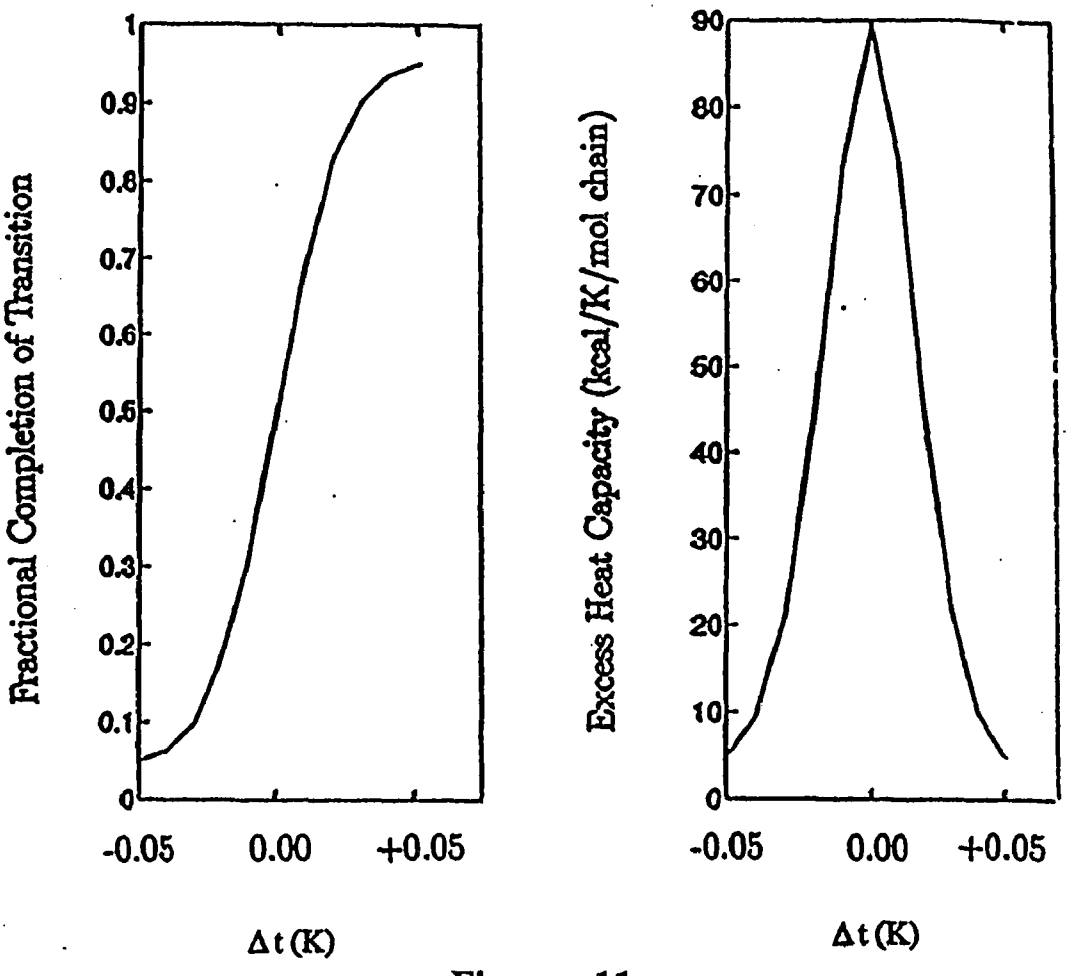

Figure 11

Fig. 11 Fractional completion and the molar excess heat capacity of the transition. a) Fractional completion, $\theta=\langle N_l\rangle/N$ vs. $\Delta t = t-t_m$). b) The molar excess heat capacity vs. Δt. The molar excess heat capacity c_p is calculated from the equilibrium fluctuations of N_l and N_{lg} by

$$c_p=[(\Delta H)^2\langle(N_l-\langle N_l\rangle)^2\rangle+w_E^2\langle(N_{lg}-\langle N_{lg}\rangle)^2\rangle]/NRt^2$$

The average number of chains in liquid crystalline state $\langle N_l\rangle$ and the fluctuations $\langle(N_l-\langle N_l\rangle)^2\rangle$, $\langle(N_{lg}-\langle N_{lg}\rangle)^2\rangle$ are determined from the respective frequency distribution. The parameters of the simulations are given in the legend to Fig. 9.

In the case of the common Glauber method the initial phase of the system was very persistent. Out of 15 simulations at $t=t_m+1.48K$ and $NC=10^5$ the system jumped only once from the initial gel phase to liquid crystalline phase. Thus, simulating the phase transition without shuffling, a hysteresis greater than $2\times1.48K$ can be expected at $NC=10^5$.

In Fig. 11 the fractional completion of the phase transition $\theta(=<N_l>/N)$ is plotted against the temperature, where the equilibrium value of the number of liquid state lattice points is determined from the respective frequency distribution (Fig. 9c).

The results in Figs. 9 - 11 demonstrate that by introducing simple shuffling trials into the conventional Glauber method one can obtain the equilibrium distributions of the system within a feasible computation time. This method makes possible the simulation of very sharp phase transitions without hysteresis, as occurs in the phase transitions of one-component phospholipid membranes.

(3.3.2) Phase Transition of DPPC MLV

The gel-to-liquid crystalline phase transition of DPPC MLV has been thoroughly studied, both experimentally and theoretically during the last two decades. According to the calorimetric experiments of Albon and Sturtevant (1978) the midpoint of the transition and the phase transition enthalpy are $t_m=314.55K$ and $\Delta H\approx4250$ cal/mol chain, respectively. Using the thermodynamic parameters above (t_m and ΔH) the phase transition was simulated by means of our two-state membrane model. The cooperativity

parameters, w_E was varied until a good fit was obtained between the experimental and simulated excess heat capacity values at t_m, while w_S =0 was chosen. This choice is partially justified by the observed asymmetry of the excess heat capacity curve of DPPC MLV. The opposite choice, simulation with w_E =0 and $w_S \neq 0$, would lead to a symmetric excess heat capacity curve.

By studying the effects of the lattice size on the results of the simulations it was noticed that, at a constant value of the cooperativity parameter w_E, the excess heat capacity at t_m increases with increasing lattice size. In order to get the experimental excess heat capacity at t_m, the value of the cooperativity parameter was adjusted at every lattice size (see Table 4).

Table 4
Parameters of the simulations at different lattice sizes.

lattice size	w_E cal/mol/chain	$c_p(t_m)$ kcal/deg/mol	NC
65X65	345.06	66.94	10^5
100X100	341.10	67.68 ± 0.5	$4x10^5$
150X150	339.80	67.06 ± 2.4	$5x10^5$

Thus with increasing lattice size the fitted value of the cooperativity parameter slightly decreases and tends to level off after a lattice size of 100x100. Despite this small change in the cooperativity parameter the respective frequency distributions of N_l/N (see Fig. 12) show qualitative differences with lattice size, changing from bimodal to unimodal curves.

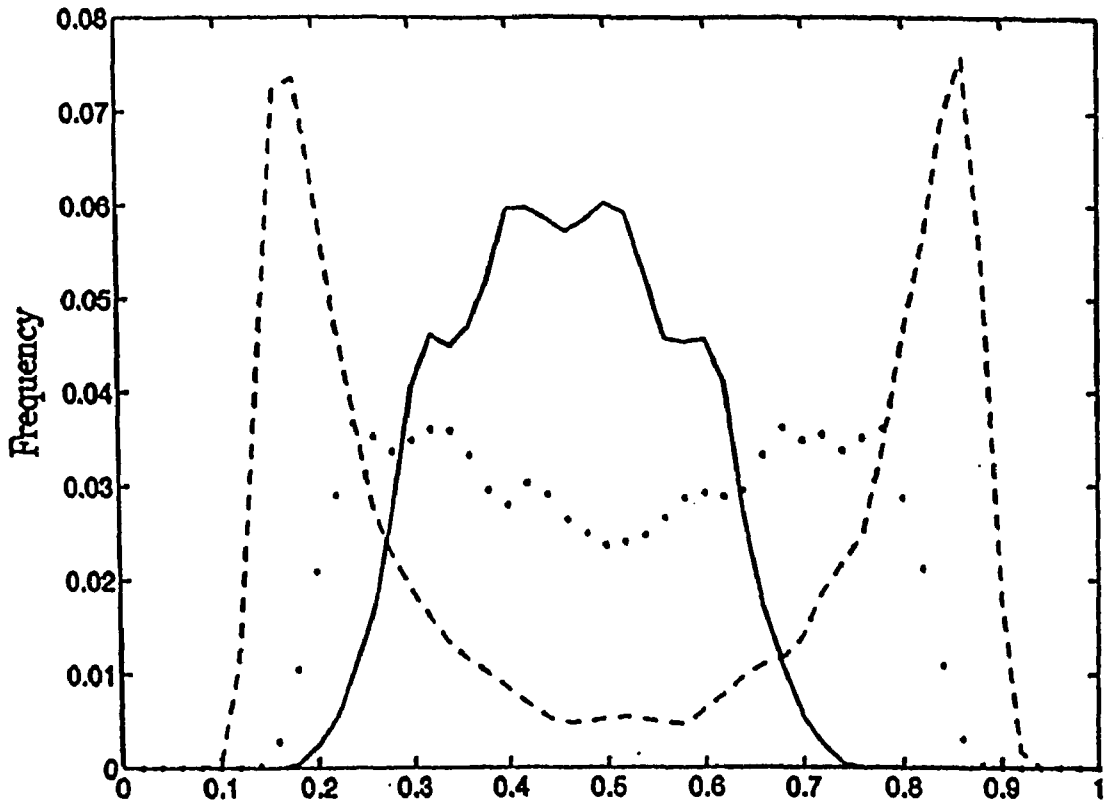

Figure 12

Fig. 12 Frequency distribution of N_l/N at the midpoint temperature of the phase transition at different lattice sizes. Dashed line: lattice size N=65x65, dotted line: N=100x100, solid line: N=150x150. Common parameters of the simulations are: ΔH=4250 cal/mol chain, t_m=314.55 K, w_g=0 cal/deg/mol chain, while the lattice size dependent w_E and NC values are listed in Table 4. To construct each frequency distribution the data were grouped into 50 classes.

The simulated phase transitions can be classified by means of the frequency distribution of N_l at the midpoint of the transition. In the case of first-order phase transitions, the distribution functions of the fluctuating extensive variables are inhomogeneous, whereas, in

the second-order phase transition case they are unimodal (Hill, 1985; Sugar, 1987). Note that these criteria for the classification are more general than those based on the continuity of the transition curves, and they are applicable for both small and large systems.

According to this classification the frequency distributions in Fig. 12 represent first-order phase transitions at 65x65 and 100x100 lattice sizes, and second-order phase transition at 150x150 lattice sizes. This shows that one can expect second-order phase transitions at even larger lattice sizes. Thus according to the simple two-state membrane model the main transition of DPPC MLV is a second-order phase transition, which, however, is very close to the critical point.

The main phase transition of DPPC MLV has been considered by many to be first-order transition (Albon & Sturtevant, 1978; Doniach, 1978; Mouritsen, 1991). Exception to this view have been taken by others (Kanehisa & Tsong, 1978; Freire & Biltonen, 1978; Mitaku, et al., 1983) who have interpreted heat capacity data in terms of coexistence of gel and liquid crystalline clusters over the temperature range of the transition. More recently, Biltonen (1990) has argued on the basis of experimental thermodynamic data that the transition is not first order. Corvera *et al.* (1993) have applied finite size scaling theory to the membrane model developed by Pink (1980). They also concluded that, for the usual set of parameters employed, no first-order phase transition exists. This conclusion is in agreement with our result.

Recently, Sugar et al. (1994) simulated the gel-to-liquid crystalline phase transition of DPPC small unilamellar vesicles (SUV) by the same two-state membrane model with $w_S=0$, $w_E=282.4$

cal/mol chain. There is a 57 cal/mol chain difference between the values of the cooperativity parameters of MLV and SUV, probably reflecting the packing differences between DPPC MLV and DPPC SUV membranes (Chrzeszczyk, et al., 1977; Ruocco & Shipley, 1982). DPPC SUV also shows a second-order phase transition, but it is farther from the critical point. At the phase transition temperature on average half of the membrane is in gel and the other half is in liquid crystalline phase, and N_l fluctuates around the average by 4% for SUV and 17% for MLV.

(3.3.3) Excess Heat Capacity Curve of DPPC MLV

The excess heat capacity was calculated at different temperatures, while the parameters of the simulations were kept constant. The results of the simulations are shown in Fig. 13. Diagonal crosses and open circles belong to the simulation parameters listed in the first and the third row of Table 4, respectively. There is a qualitative difference between the experimental and calculated excess heat capacity curves. The calculated transition is narrower than the experimental transition curve at $t<t_m$ and tends to be broader at $t>t_m$ (Albon & Sturtevant, 1978) (solid line in Fig. 13). Recently, however, Biltonen (1990) repeated this experiment using a scanning rate of 0.1K/h, which is an order of magnitude slower than that used by Albon and Sturtevant (1978). In this case the observed asymmetry of the peak agreed with the calculated one (dash-dotted line in Fig. 13).

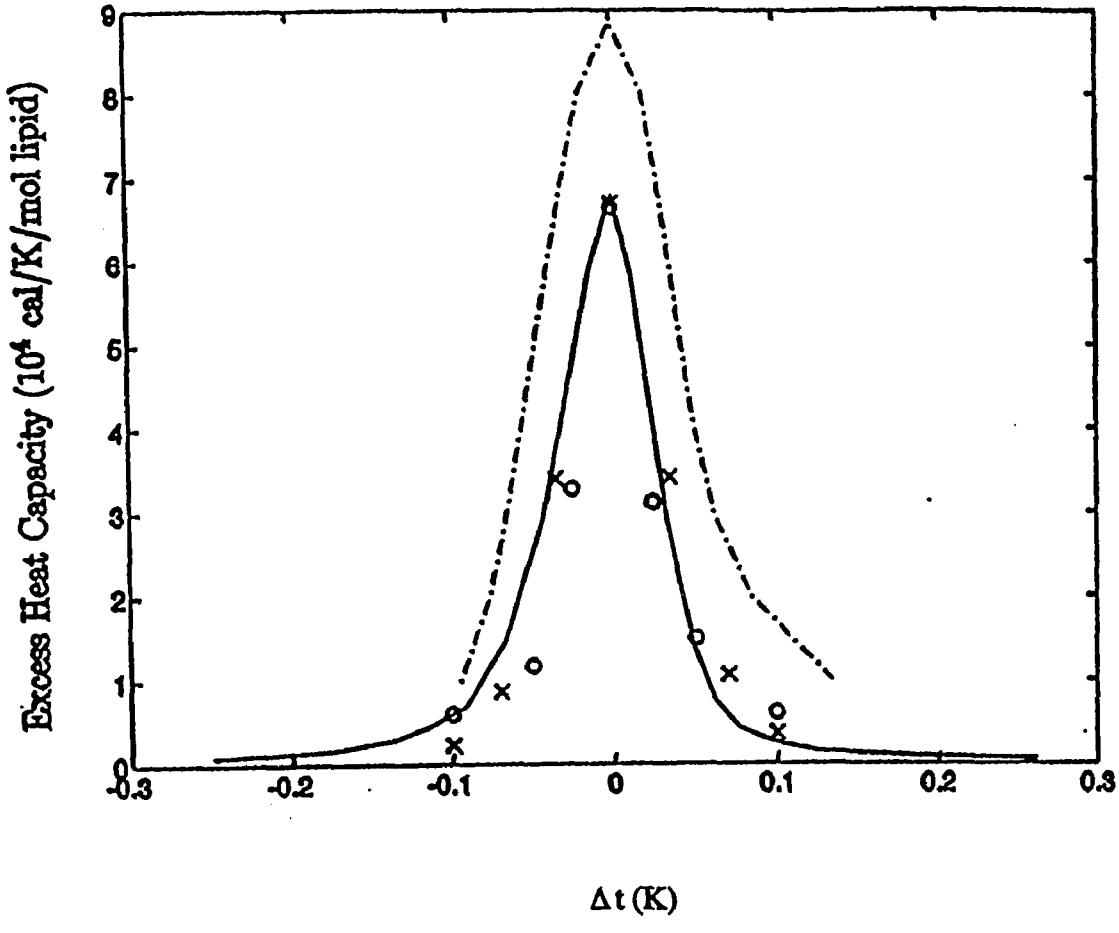

Figure 13

Fig. 13 Excess heat capacity curve of DPPC MLV. Solid line: experimental result (Albon and Sturtevant, 1978). Dashed line: experimental result (Biltonen, 1990). Data were calculated by means of the two-state membrane model and using the following sets of parameters: ΔH=4250 cal/mol chain, t_m=314.55 K, w_S=0 cal/deg/mol chain. x: w_E=345.06 cal/mol chain with lattice size 65X65; o: w_E=339.80 cal/mol chain with lattice size 150X150. $\Delta t=t-t_m$.

Unfortunately, this curve cannot be used for a quantitative comparison with the simulated data. Due to the lack of a careful calibration at slow scanning rate (personal communication with Mr.

K. Thompson, laboratory assistant to Dr. Biltonen at Univ. of Virginia, Charlottesville) the transition enthalpy obtainable from the published excess heat capacity curve is about 1 kcal/mol lipid larger than the commonly accepted 8.5-8.7 kcal/mol lipid (Albon &.Sturtevant, 1978) transition enthalpy of DPPC MLV main phase transition.

APPENDIX

Proposition: Let n be greater than 2 and ξ be a random variable, ξ = 1, 2, 3, ..., n with probabilities p_1, p_2, p_3, ..., p_n, respectively. Then $D\xi = \mathrm{Var}(\xi) = E(\xi - E\xi)^2$ reaches maximum when $p_1 = p_n = 1/2$ and the rest p_i , i=2, ..., n-1, are all zeroes. The maximum of $D\xi$ is $\max D\xi = (n-1)^2/4$.

Proof: $D\xi = p_1 + 2^2 p_2 + 3^2 p_3 + ... + n^2 p_n - (p_1 + 2p_2 + 3p_3 + ... + np_n)^2$

(1°) If $p_1 \le p_n$, then let η be another random variable such that η = 1, 2, 3, ..., n with probabilities $p_1 + p_2$, 0, p_3, p_4, ..., p_n, respectively. $D\eta$ can be expressed as $D\eta = (p_1+p_2) + 3^2 p_3 + ... + n^2 p_n - ((p_1+p_2) + 3p_3 + ... + np_n)^2$.

$$D\eta - D\xi = -3p_2 + (p_1 + 2p_2 + 3p_3 + ... + np_n)^2 - ((p_1+p_2) + 3p_3 + ... + np_n)^2$$

$$= -3p_2 + (2p_1 + 3p_2 + 2\times 3p_3 + 2\times 4p_4 + ... + 2\times np_n)\times p_2$$

$$= p_2 \times (2p_1 + 3p_2 + 2\times 3p_3 + 2\times 4p_4 + ... + 2\times np_n - 3)$$

$$\ge p_2 \times (3p_1 + 3p_2 + 2\times 3p_3 + 2\times 4p_4 + ... + (2n-1)p_n - 3)$$

$(\because p_n \ge p_1)$

$$\ge p_2 \times (3p_3 + ... + (2n-4)p_n) \ge 0.$$

$(\because p_1 + p_2 + ... + p_n = 1)$

$$\therefore D\eta \ge D\xi$$

Now, in the same way one can show that if ξ = 1, 2, 3, ..., n with probabilities p_1, 0, p_3, ..., p_n, respectively, ($p_1 + 0 + p_3 ..... + p_n = 1$ at this time), and $p_1 \le p_n$, then η = 1, 2, 3, ..., n with probabilities $p_1 + p_3$, 0, 0, p_4, ..., p_n, respectively, still satisfies $D\eta \ge D\xi$. This process can continue until at some stage one finds that $p_1 > p_n$.

(2°) If $p_1 > p_n$, then let η = 1, 2, 3, ..., n with probabilities p_1, p_2, p_3, p_4, ..., p_{n-2},0,$p_{n-1}+p_n$, respectively, one can find the following inequality:

$$D\eta - D\xi =$$

$$(2n-1)p_{n-1}\text{-}$$

$$p_{n-1}\left(2p_1+2\times 2p_2+2\times 3p_3+.....+2\times(n-2)p_{n-2}+(2n-1)p_{n-1}+2\times np_n\right)$$

$$=$$

$$p_{n-1}(2n-1-(2p_1+2\times 2p_2+2\times 3p_3+.....+2\times(n-2)p_{n-2}+(2n-1)p_{n-1}+2\times np_n))$$

$$\geq p_{n-1}(2n-1-3p_1-2\times 2p_2-2\times 3p_3-.....-2(n-2)p_{n-2}-(2n-1)p_{n-1}-(2n-1)p_n)$$

This is because $p_1 > p_n$. Finally, one has the following obvious inequality: $D\eta - D\xi \geq p_{n-1}(2n-1-(2n-1)(p_1+p_2+.....+p_n))=0$. Again, in the same way one can show that this process can continue until at some stage one finds that $p_1 \leq p_n$.

Repeating these two procedures one can construct a random variable $\gamma(\xi)$ which depends on ξ such that $\gamma(\xi)$ takes values 1 and n with some probabilities q_1 and q_n, respectively, and $q_1 + q_n = 1$. $\gamma(\xi)$ satisfies $D\gamma(\xi)\geq D\xi$. It is easy to show that these γ's have $(n-1)^2/4$ as the upper bound for $D\gamma$ and when $q_1 = q_n$ =1/2, the upper bound can be reached.

QED

Corollary: Let n be greater than 2 and ξ be a random variable, ξ =m,m+1, ..., n with probabilities p_m, p_{m+1}, ..., p_n, respectively. Then $D\xi = E(\xi - E\xi)^2$ reaches maximum when $p_m = p_n$ =1/2 and the rest p_i, i=m+1, ..., n-1, are all zeroes. The maximum of $D\xi$ is $\max D\xi = (n-m)^2/4$.

BIBLIOGRAPHY

Albon, N.; Sturtevant, J. M. **(1978)** Nature of the gel to liquid crystal transition of synthetic phosphatidylcholines. *Proc. Natl. Acad. Sci. USA* , **75,** 2258-2260

Anshelevich, V. V., Vologodskii, A. V., Lukashin, A. V. & Frank-Kameneskii, M.D. **(1979).** Statistical-mechanical treatment of violations of the double helix in supercoiled DNA *Biopolymers.* **18,** 2733-2744.

Barker, A.A. **(1965).** Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Aust. J. Phys.* **18,** 119-33.

Bauer, W.R. & Benham, C.J. **(1993)** The free energy, enthalpy and entropy of native and of partially denatured closed circular DNA. *J. Mol. Biol.* **234,** 1184-1196

Beattie, K.L., Wiegand, R.C. & Radding, C.M. **(1977).** Uptake of homologous single-stranded fragments by superhelical DNA. *J. Mol. Biol.* **116,** 783-803.

Benham, C. J. **(1979).** Torsional stress and local denaturation in supercoiled DNA. *Proc. Nat. Acad. Sci. U.S.A.* **76,** 3870-3874.

Benham, C. J. **(1990).** Theoretical analysis of heteropolymeric transitions in superhelical DNA molecules of specified sequence. *J. Chem. Phys.* **92,** 6294-6305.

Benham, C. J. **(1992)** Energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.* **225,** 835-847.

Benham, C. J. **(1993).** Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci. *Proc. Nat. Acad.* **90,** 2999-3003

Biltonen, R.L. **(1990)** A statistical-thermodynamic view of cooperative structural changes in phospholipid bilayer membranes: their potential role in biological function. *J. Chem. Thermodyn.* **22,** 1-19.

Bloomfield, V., Crothers, D. & Tinoco, I. **(1974)** *Physical Chemistry of Nucleic Acid.* Harper & Row, New York NY pp.258-260

Chrzeszczyk, A.; Wishnia, A.; Springer, C.S. **(1977)** The intrinsic structural asymmetry of highly curved phospholipid bilayer membranes. *Biochim. Biophys. Acta* **470,** 161-169.

Chung, K.L. **(1976)** *Markov Process with Stationary Transition Probability;* Springer-Verlag, Heidelberg

Dean, W. & Lebowitz, J. **(1971)** Partial alteration of secondary structure in native superhelical DNA. *Nature* **231**, 5-8.

de Pablo, J.J., Laso, M. & Suter, U.W. **(1992)** Simulation of polyethylene above and below the melting point. *J. Chem. Phys.* **96**, 2394.

Diaconis, P. & Saloff-Coste, L. **(1993)** Comparison theorems for reversible Markov chains. *The Annals of Applied Probability.* **3**, 696-730

Doniach, S. **(1978)** Thermodynamic fluctuations in phospholipid bilayers. *J. Chem. Phys.* **68,** 4912-4916.

Eisen, M **(1969)** *Introduction to Mathematical Probability Theory.* pp.24, Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

Feller, W. **(1966)** *An Introduction to Probability Theory and Its Applications.* Vol. 1, 3rd edition. John Wiley and Sons, Inc. New York London Sydney. Chap. XV. Especially, pp.426 problems 21 and 22.

Flory, P.J. **(1969)** *Statistical Mechanics of Chain Molecule.* John Willy & Sons, Interscience Publishers, New York. Chap. V.

Glauber, R.J. **(1963)** Time-dependent statistics of the Ising model. *J. Math. Phys.* **4,** 294-307

Hastings, W. K. **(1970).** Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 1, 97-109.

Hill, T.L. **(1985)** *Cooperativity Theory in Biochemistry;* Springer-Verlag, Section 1

Hill, T.L. **(1963)** *Thermodynamics of Small Systems;* Benjamin: New York.

Hill, T.L. **(1986)** *An Introduction to Statistical Thermodynamics;* Dover Publications: New York.

Kamenetskii, M. D. **(1991).** Computer simulation of DNA supercoiling. *J.Mol. Biol.* **217**, 413-419.

Kandel, D., Ben-Av, R. & Domany, E. **(1990).** Cluster dynamics for fully frustrated system. *Phys. Rev. Lett.* **65**, 941-944.

Katsura, S., Makishima, F. & Nishimura, H. **(1993)** Statistical mechanical approach for predicting the transition to non-B DNA structures in supercoiled DNA. *J. Biomol. Str. Dyn.* **10**, 639-656.

Kawasaki, K. **(1972)** *Phase Transitions and Critical Phenomena;* Academic Press: London. **Vol.2**, p. 443.

Keizer, J. **(1972)** *On the Solutions and the Steady States of a Master Equation;* Plenum Press, New York

Klenin, K. V., Vologodskii, A. V., Anshelevich, V. V., Dykhne, A. M. &

Frank-Kowalski, D., Natale, D. & Eddy, M. **(1988).** Stable DNA unwinding, not breathing, accounts for single-strand-specific nuclease hypersensitivity of specific A+T-rich sequences. *Proc.*

Nat. Acad. Sci. U.S.A. **85**, 9464-9468.

Kolinski, A., Skolinik, J. & Yaris, R. **(1987)** Monte Carlo studies on equilibrium globular protein folding. I. Homopolymeric lattice methods of beta-barrel protein. Biopolymers *26*, 937-962

Kowalski, D. & Eddy, M. J. **(1989)**. The DNA unwinding element: a novel, *cis*-acting component that facilitates opening of the *Escherichia coli* replication origin. *EMBO J.* **8**, 4335-4344.

Levene, S. D. & Crothers, D. M. **(1986)**. Topological distributions and the torsional rigidity of DNA: A Monte Carlo study of DNA circles. *J. Mol. Biol.* **189**, 73-83.

Mattern, M. & Painter, R. **(1979)**. Dependence of mammalian DNA replication on DNA supercoiling. *Biochim, Biophys. Acta* **563**, 293-305.

Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.N.; Teller, E. **(1953)** Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087-1092.

Mezei, M., Bencsath, K.A., Goldman, S. & Singh, S. **(1987)** The detailed balance energy-scaled displacement Monte Carlo algorithm. *Molecular Simulation* **Vol.1**, 87-93

Mizuuchi, K., Gellert, M. & Nash, H. **(1978)**. Involvement of supertwisted DNA in integrative recombination of bacteriophage lambda. *J. Mol. Biol.* **121**, 375-392.

Moran, P. A. P. **(1968)** *An Introduction to Probability Theory.* pp.358,Clarendon Press. Oxford.

Mouritsen, O.G. **(1991)** Theoretical models of phospholipid phase transitions. *Chem. Phys. Lipids.* **57**, 179-194.

Nagle, F.J. and Wilkinson **(1978)** Lecithin bilayers density

measurements and molecular interactions. *Biophys. J.* **23**, 159-175

Person, W.B. and Pimentel, G.C. **(1953)** Thermodynamic properties and the characteristic CH_2 frequencies of n-paraffins. *J. Am. Chem. Soc.* **75**, 532-538

Peskun, P.H. **(1981)** Guidelines for choosing the transition matrix in Monte Carlo methods using Markov chains. *J. Comp Phys.* **40,** 327-344

Pink, D.A.; Green, T.J.; Chapmann, D. **(1980)** Raman scattering in bilayers of saturated phosphatidylcholines. Experiment- and theory. *Biochemistry.* **19**, 349-356

Pruss, G. & Drlica, K. **(1989)** DNA supercoiling and prokaryotic transcription. *Cell* **56**, 521-523.

Rao, M., Pangali, C. & Berne, B.J. **(1979)** On the force bias Monte Carlo simulation of water: Methodology, optimization and comparison with molecular dynamics. *Mol. Phys.*, **37**, 1773

Richet, E., Abcarian, P. & Nash, H. **(1986).** The interaction of recombination proteins with supercoiled DNA: Definimg the role of supercoiling in lambda integrative recombination. *Cell* **46**, 1011-1021.

Ruocco, M.J.; Shipley, G.G. **(1982)** Characterization of the subtransition of hydrated dipalmitoyl phosphatidylcholine bilayers. Kinetics, hydration and structural studies. *Biochim. Biophys. Acta* **691**, 309- 320.

Seelig, J and Niederberger, W **(1974)** Deuterium-labeled lipids as structural probes in liquid crystalline bilayers. A deuterium magnetic resonance study. *J. Am. Chem. Soc.* **96**, 2069-2072

Sugar, I.P. **(1987)** Cooperativity and classification of phase transitions. Application to one- and two-component phospholipid membranes. *J. Phys. Chem.* **91,** 95-101.

Sugar, I.P.; Biltonen, R.L.; Mitchard, N. **(1994)** Monte Carlo simulations of membranes: Phase transition of small unilamellar dipalmitoylphatidylcholine vesicles. *Methods in Enzymology.* **240,** 569-593.

Swendsen, R.H.; Wang, J.S. **(1986)** Nonuniversal critical dynamics in Monte Carlo simulation. *Phys. Rev. Lett.* **58**, 86-88.

Vologodskii, A. V., Levene, S. D., Klenin, K. V., Frank-Kamenetskii, M.& Cozzarelli, N. R. **(1992)**. Conformational and thermodynamic properties of supercoiled DNA. *J. Mol. Biol.* **227**, 1224-1243.

Vinograd, J., Lebowitz, J. & Watson, R. **(1968)**. Early and late helix-coil transitions in closed circular DNAs. *J.Mol.Biol.* **33**, 173-197.

Weintraub, H., Cheng, P. & Conrad, K. **(1986)**. Expression of transfected DNA depends on DNA topology. *Cell* **46**, 115-122.

Wolff, U. **(1989)** Collective Monte Carlo updating for spin systems. *Phys. Rev. Lett.* **62,** 361-364.

Yellin, N and Levin **(1977)** Hydrocarbon chain trans-gauche isomerization in phospholipid bilayer gel assemblies. *Biochemistry* **16**, 642-647

Zhurkin, V. B., Ulyanov, N. B., Gorin, A. A. & Jernigan, R. L. **(1991)**. Static and statistical bending of DNA evaluated by Monte Carlo simulations. *Proc. Nat. Acad. Sci. U.S.A.* **88**, 7046-7050.