Molecular Modelling and Simulation of Deoxyribonucleic Acids

Fei Zhang
Department of Computational Science, National University of Singapore
Republic of Singapore 119260. Email: feizhang@cz3.nus.sg

Michael A. Collins
Research School of Chemistry, Australian National University
ACT 2601, Canberra, Australia

Abstract Molecular Dynamics (MD) simulation and modelling of deoxyribonucleic acids (DNA) are discussed. A simplified molecular model for DNA is described. The model proceeds in the spirit of all-atom simulations but discards many unimportant degrees of freedom to preserve maximum simplicity. We use the reduced model to simulate several DNA sequences over about ten nanoseconds at different temperatures. It is shown that the model is able to reproduce some features of DNA dynamics. The helix structure of the DNA model is stable at low temperatures, but melts above a certain temperature. The melting transitions are found to be rather slow in time; it takes longer than ten nanoseconds to completely denature a DNA sequence of only 100 base pairs.

---

1. INTRODUCTION

In recent years, with the rapid increase of computer power and availability, Molecular Dynamics (MD) has become an increasingly important technique for simulating molecular-scale matter [Allen and Tildesley 1989]. The basis of the MD methods is to numerically integrate the coupled differential equations of motion for a system of particles (or atoms) interacting with each other according to a given force law. Since the MD simulation follows the detailed trajectory of each atom, many physical properties (thermodynamic, structural, and transport) of the materials can be calculated from the simulation data.

Biopolymers pose a unique challenge to MD simulation and modelling because they have very complicated internal structures and complex inter- and intramolecular interactions. In this paper we explore the development of a simple molecular model to study DNA dynamics. The reduced model, which uses only relatively few degrees of freedom, allows simulations of much larger DNA sequences over much longer timescales than is possible with all-atom approaches. It is shown that the present model can reproduce some but not all of the basic characteristics of the DNA dynamics. The model displays sharp thermal melting transitions for the DNA duplex, with melting temperature depending correctly on the base pair composition. In addition, we have demonstrated that the complete melting transition may take place only over rather long timescales (say, many tens of nanoseconds) even for a DNA of just 100 base pairs. We observe localized fluctuating modes in Hydrogen (H)-bond displacements in the MD simulations, though this model appears to exaggerate their magnitude. The inhomogeneity due to the difference in the strength of H-bonds between AT and GC pairs is found to have a strong effect on energy localization processes.

2. THE DNA MODEL

To model deoxyribonucleic acids (DNA), it is natural to use the *all-atom* approach [McCammon and Harvey 1987]. Unfortunately, the computational demand of all-atom MD is so large (there are about 60 atoms per base pair) that only short DNA fragments have been simulated over a nanosecond at most which is much shorter than the characteristic timescale of many physical and biological processes in DNA (e.g., replication, transcription, denaturation, and supercoiling. See Cantor et al. [1980]). Therefore, it is useful to develop reduced DNA models. If possible such models should contain the most important degrees of freedom and enough structural details to mimic reality, while allowing simulations of large sequences over long timescales. Moreover, an understanding of which reduced models best mimic reality is essential if very simple mathematical models are to be of use.

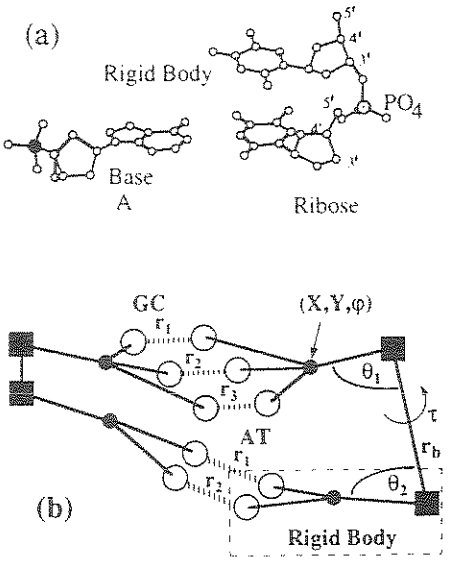

FIG. 1. (a) Schematic representation of a fragment of DNA double helix, showing a base pair and one adjoining nucleotide and phosphate backbone. All hydrogen atoms have been deleted for clarity. (b) Schematic representation of the simplified DNA model constructed from the all-atom model. The center of mass (black circle), hydrogen bonding sites (white circle), and backbone connections (black square) are indicated, with bond lengths, angles, and torsion angles defined.

The simplified DNA model is first built up with a double helix structure, whose equilibrium geometry is taken to be that for B-form DNA [Arnott and Hulins 1972]. As is well known, each strand of DNA is composed of a sequence of purine and pyrimidine bases or nucleotides, which are all flat molecules composed of one or two rings of nitrogen and carbon atoms [Saenger, 1984]. These molecules are quite stiff with respect to all molecular distortions. In particular, they have no low frequency conformational distortions so that they may be treated approximately as rigid bodies. Each base is attached to a five membered deoxyribose ring, which is connected to other such rings by a quite flexible backbone (See Fig.1a). The deoxyribose ring is not flat and has a number of preferred conformations. While transitions between these preferred conformations can occur at physiological temperatures, the ring is nevertheless significantly more rigid than the deoxyribose-PO4-deoxyribose backbone linkage. Thus, we ignore distortions of the ribose ring and fix it rigidly to each base in the standard conformation it takes in idealized B-DNA. The flexible backbone itself is treated as a structureless elastic rod. Each strand of this B-DNA model is then a sequence of rigid bodies (base-ribose) connected by flexible rods. The bases of each strand are hydrogen bonded to the complementary bases on the other strand, i.e., Adenine (A) to Thymine (T), and Guanine (G) to Cytosine(C). In the standard equilibrium B-DNA structure, the bases in a pair are nearly coplanar. In this first simplest model, we only allow these rigid base-ribose bodies to move in this XY plane. When projected onto this plane, the backbone connections at the ribose positions C3 and C5 are quite close to each other. Hence, for simplicity both the backbone connections are taken to apply at the one location. The distance between two adjacent planes is 3.38 angstrom. The motion of each rigid body can then be completely described in terms of the (X,Y) coordinates of the body's centre of mass, and an angle, φ, which determines the orientation of the body with respect to say the X axis. This model is depicted schematically in Figure 1(b).

The model potential energy has the form

$$V = \sum_n \sum_{j=1,2} (V_b + V_\theta + V_\tau) + \sum_n (V_{LJ} + V_H), \quad (1)$$

where the sum is over all base pairs (n) and over the two strands (j=1,2), and the backbone stretching and bending, the torsion, the nonbonded, and the hydrogen bonding interactions are

$$V_b = \frac{1}{2} K_b (r_b - r_b^{eq})^2, \quad (2)$$

$$V_\theta = \frac{1}{2} K_\theta [\cos(\theta_i) - \cos(\theta_i^{eq})]^2, \quad (3)$$

$$V_\tau = K_\tau [1 - \cos(\tau - \tau^{eq})], \quad (4)$$

$$V_{LJ} = 4\epsilon \left[(\frac{\sigma}{\bar{R}})^{12} - (\frac{\sigma}{\bar{R}})^6 \right], \quad (5)$$

$$V_H = V_0 \{\exp[-\alpha(r_i - r_i^{eq})] - 1\}^2 - \frac{1}{4} V_0 [1 + \tanh(\beta(r_i - r^*))]. \quad (6)$$

The model has been described in detail by Zhang and Collins [1995].

3. MD simulation results

Using the above model, we have simulated DNA chains of various lengths and base-pair compositions. In the following, the simulation results for three types of B-DNA fragments of 100 base pairs (100 bp) over a time of about 10 nanoseconds will be reported in detail. Simulations of shorter and longer sequences will also be briefly discussed. The three sequences are (i) Poly(A).Poly(T), (ii) Poly (G).Poly(C), and (iii) mixed content. Here, the notation Poly(.) represents a homopolymer consisting of one type of

base as indicated in the parentheses. In the fragment of 100 bp with mixed content some AT rich regions are deliberately introduced. In particular, the TATAAT sequence, known as the "Pribnow Box", is a highly conserved sequence appearing in many promoter sites in many related organisms [Kornberg 1974]. This sequence and the other AT rich regions are expected to be more amenable to base pair opening, as will be shown in the simulations.

The first feature of note in the dynamics of this model, is that the amplitude of the position fluctuations of different sites in the molecule increases from the inside to the outside of the double helix, and that the sites near the chain ends

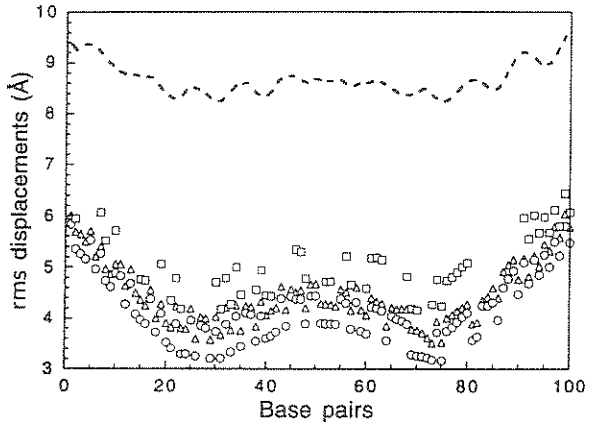

FIG. 2. Root-mean-square (rms) displacement of sites on the base pairs from their average positions for a DNA fragment of 100 bp with mixed content simulated over 10 nsec at temperature $T = 300$ K. The sites at the backbone connections (dashed line) move more than the H-bonding sites inside the helix (symbols). The first (triangles) and the third (squares) (for GC pairs) H-bonding sites move more than the second (circles). The chain ends move more than the middle.

move most (See Fig. 2). Moreover, the first and third hydrogen bonding sites in GC base pairs move more than the middle (second) site. These results are qualitatively in agreement with those reported in all-atom MD simulations of DNA . Thus we have a picture of the base-pair motion in which the base-ribose bodies move in and away from their partners while undergoing hindered rotation. The overall picture is not one in which the base and ribose swing about a hinge supplied by the (fixed) backbone; rather the reverse, where the base, ribose and backbone are hinged flexibly about the hydrogen bonds. This important qualitative aspect of the motion of the DNA was not taken into account in some previous attempts [Englander et al 1980] to derive simple nonlinear models to study the base pair planary rotations.

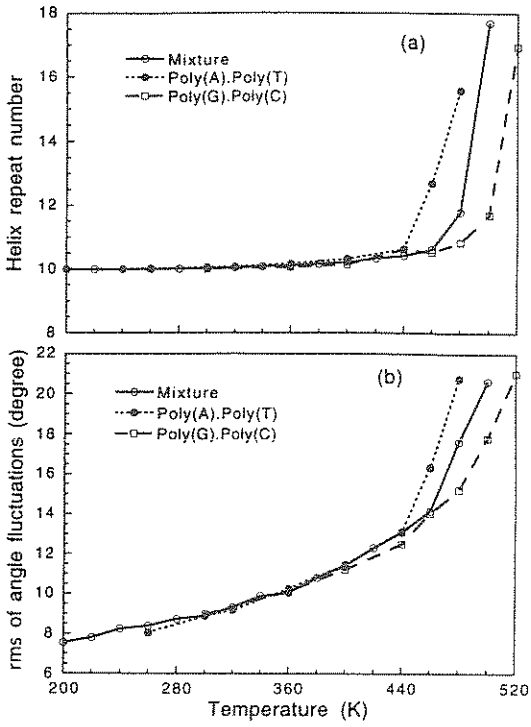

FIG. 3. (a) Helix repeat number for DNA segments of 100 base pairs and (b) rms of angle fluctuations (see the text).

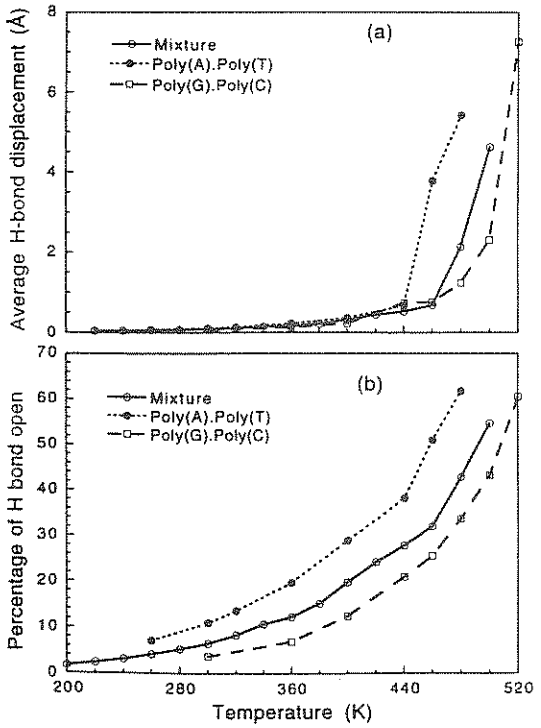

FIG. 4. (a) Average hydrogen-bond displacements and (b) percentage of H bond open, for simulations of DNA fragments of 100 bp over 10 nsec.

Another qualitative feature observed in all the simulations of this model, is that the double helical structure is stable at low temperatures. This can be clearly seen in Fig.3(a), which plots the "helix repeat number" against the temperature. Here, the "helix repeat number" is the average number of base pairs contained in a full turn. For the model B-DNA at the equilibrium geometry, the helix repeat number is exactly 10. A value higher than 10 indicates that the helix is "unwound" by comparison with the standard structure. The curves in Fig.3(a) stay close to 10 at low temperatures, but display a sharp increase at higher temperatures which correspond to melting of the DNA molecules. The fluctuations in the unwinding of DNA is given by the root-mean-square average of the angle between successive interbase center of mass vectors. See Fig.3 (b).

Hydrogen bond breaking is an important element in DNA thermal denaturation, replication and transcription. Here we measure the mean displacements of the H-bond lengths as shown in Fig 4(a), where the curves display similar behavior to that in Fig.3(a): A very slow increase of the H-bond displacement indicates the stability of the model DNA at low temperatures; and a rapid increase at temperatures above some transition point shows that the helix is actually melting. The percentage of H-bond "opening" has also been measured and presented in Fig. 4(b). As expected, base pairs in poly(A).poly(T) are the most amenable to opening, followed by the mixed and the poly(G).poly(C).

Figs. 3 and 4 clearly demonstrate that the present DNA model is able to show *sharp* melting transitions, and that the DNA fragments of different compositions have different melting temperatures. The GC homopolymer has the highest melting temperature (about 500K), the AT homopolymer has the lowest (460K), while the sequence with mixed content has intermediate melting temperature (480K). These characteristics are qualitatively in agreement with the results of thermal melting experiments for DNA Saenger [1984], although the melting temperatures are about 100 K too high, indicating that the simplified DNA model is too "stiff" overall.

Most interestingly, we find that the melting transitions require long timescales to develop, and that the larger the DNA sequence is, the longer the time is required for melting to occur. Note that we start each simulation from the ground state of the DNA. At low temperatures, the DNA appears to reach equilibrium in less than a nanosecond: the energy of the molecule, the helix repeat number and the H-bonds displacement (see Fig. 5), and the rms displacements of Figure 2, for example, do not ``drift" in value with time but merely fluctuate about a steady mean after less than one nanosecond. Thus the averages in Figs. 3 and 4, which are obtained from results of simulations over 10 nanoseconds, may represent the thermal equilibrium values at low temperatures. However, at temperatures above the "melting points", disorder in the model becomes apparent only after several nanoseconds of simulation (Fig.6).

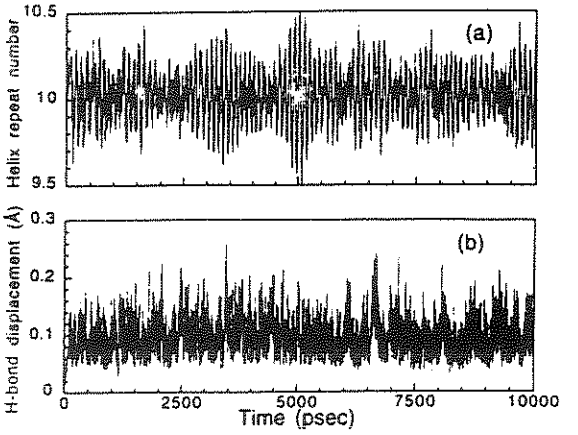

FIG. 5. (a) Helix repeat number and (b) average H-bond displacement, for 100 bp of mixed content at $T = 300$ K.

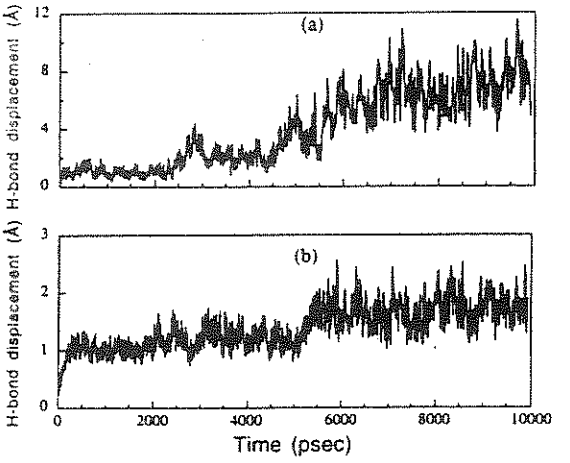

FIG. 6. Average H-bond displacement for DNA fragments with mixed content, showing the development of disorder at temperature $T = 480$ K: (a) 100 bp and (b) 400 bp. The larger the sequence, the longer the time scale required for melting to occur. Similar behavior is observed for the helix repeat number.

The large "steps" in Fig. 6 might be interpreted as arising from large amplitude distortions of the molecule or transitions of the molecule to high energy configurations which were not easily accessed at low temperature. These transitions might involve crossing high energy barriers and would therefore occur rarely even at high temperature. Thus, to reach thermal equilibrium in these simulations would require extremely long simulation times. For example, at temperatures above the melting point, ten-nanosecond simulations may only result in the melting of about 20 base pairs at both ends of the DNA helices (Fig.7). This suggests that the unwinding kinetics of DNA double helix is rather slow, as also indicated in experiments

[Cantor 1980]. Hence, we do not suggest that our simulation results at high temperatures reflect the true equilibrium values. Indeed, the "melting temperatures" indicated here should be taken as upper bounds, as melting may possibly occur at lower temperatures over much longer timescales than were simulated.

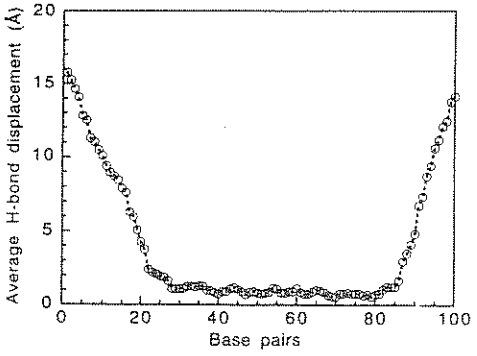

FIG. 7. Average H-bond displacement at $T = 480$ K, for simulation of a DNA fragment of 100 bp with mixed content over 10 nsec. The melting of the helix from both ends is seen.

We have also simulated B-DNA of shorter and longer sequences, and obtained similar results to those reported above. In particular, we have simulated sequences of up to 400 base pairs with mixed AT and GC content, and have found that the critical temperature for denaturation is about 480 K *coinciding* with that for a 100 bp sequence. However, the timescale for denaturation seems to be longer: above the melting temperature, one must wait at least several nanoseconds to see an apparent increase of disorder in the system (Fig.6). A short sequence such as a dodecamer retains its stable double helix structure below 350 K, but denatures at only slightly higher temperatures, due to the chain-end effects.

The opening of base pairs in DNA is an important process as it is a prerequisite for replication, transcription, and DNA-drug binding. See Gueron[1987], Ramstein and Lavery [1988]. In our simulations the hydrogen bond displacements are monitored to observe possible large amplitude excitations and their dependence on base pair composition. We find that at temperatures below the melting point, the model DNA exhibits some large-amplitude, though short-lived, fluctuations in the H-bonds and such excitations tend to repeat at nearby base-pairs. Most interestingly, compared to GC pairs there are more frequent large amplitude fluctuations in the H-bonds of AT pairs, and a sequences of several contiguous AT and TA base pairs are found to be more than usually prone to hydrogen bond breaking. This is understandable because a GC pair has one more H-bond than a AT pair.

4. Concluding remarks

In summary, we have investigated a relatively simple molecular model to study the dynamics of DNA molecules. The reduced model has allowed simulations of much larger DNA sequences over much longer timescales than is possible with all-atom approaches. It has been shown that the present model can reproduce some but not all of the basic characteristics of the DNA dynamics. The model displays sharp thermal melting transitions for the DNA duplex, with melting temperature depending correctly on the base pair composition. We have showed that the complete melting transition may take place only over rather long timescales even for a short segment of DNA double helix.

Overall, however, the simple model appears to be too "stiff", since the onset of melting appears to be significantly higher than experimentally observed. This may be primarily due to the restriction to two dimensional motion, in addition to the rather primitive description of the forces. Many very low frequency motions are excluded from the model, including relative motion of the strands along the helix axis, tilting and relative twisting of the bases. Three dimensional motion, which demands a more complete description of the forces between the stacked bases, ought to be essential in a better quantitative description of the DNA melting.

References

Allen, M.P., and D.J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 1989); Yonezawa F. ed., *Molecular dynamics simulations* (Springer, 1991).

Arnott, S., and D.W.L. Hulins, *Biochem. Biophys. Res. Comm. 47*, 1504,1972.

Cantor, C. R., and P.R. Schimmel, *Biophysical Chemistry* (W.H. Freeman and Company, 1980) p.1224.

Englander, S. M., et al., *Proc. Nat. Acad. Sci. USA 77*, 7222 , 1980.

Gueron, M., M. Kochoyan, and J.L. Leroy, *Nature, 328*, 89, 1987.

Kornberg, R., *DNA replication* (Freeman, San Francisco 1974) P. 242.

McCammon, J. A., and S. C. Harvey, *Dynamics of proteins and nucleic acids* (Cambridge University press, 1987).

Ramstein, J., and R. Lavery, *Proc. Natl. Acad. Sci. USA 85*, 7231 (1988).

Saenger, W., *Principles of Nucleic Acid Structure* (Springer-Verlag, 1984).

Zhang, F., and M.A. Collins, *Phys. Rev. E 52*, 4217, 1995.