A simple physical description of DNA dynamics: quasi-harmonic analysis as a route to the configurational entropy

2007 J. Phys.: Condens. Matter 19 076103

(http://iopscience.iop.org/0953-8984/19/7/076103)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 076103 (14pp) doi:10.1088/0953-8984/19/7/076103

A simple physical description of DNA dynamics: quasi-harmonic analysis as a route to the configurational entropy

S A Harris[1,3] **and C A Laughton**[2]

[1] School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, UK
[2] Centre for Biomolecular Sciences, University of Nottingham, Nottingham NG7 2RD, UK

E-mail: s.a.harris@leeds.ac.uk

Received 2 December 2005, in final form 20 November 2006
Published 23 January 2007
Online at stacks.iop.org/JPhysCM/19/076103

Abstract

It has become increasingly apparent that the dynamic as well as the structural properties of biological macromolecules are important to their function. However, information concerning molecular flexibility can be difficult to obtain experimentally at the atomic level. Computer modelling techniques such as molecular dynamics (MD) have therefore proved invaluable in advancing our understanding of biomolecular flexibility. This paper describes how a combination of atomistic MD simulations and quasi-harmonic analysis can be used to describe the dynamics of duplex DNA, with a particular emphasis on methods for calculating differences in configurational entropies. We demonstrate that DNA possesses remarkably simple mechanical properties relative to globular proteins, making it an ideal system for exploring biomolecular flexibility in general. Our results also highlight the importance of solvent viscosity in determining the dynamic behaviour of DNA in aqueous solution.

M Supplementary data are available from stacks.iop.org/JPhysCM/19/076103

(Some figures in this article are in colour only in the electronic version)

1. Introduction

A revolution in the understanding of biochemical interactions came from an appreciation of the relationship between the structure of biomolecules and their function. Molecular recognition processes in which, for example, a DNA binding protein recognizes one DNA sequence over another are central to biological communication and control within the cell. Discrimination at the molecular level is achieved through thermodynamics; the protein will bind to the sequence which gives the most favourable change in free energy. Detailed structural information relating

[3] Author to whom any correspondence should be addressed.

to macromolecules and their complexes (obtained primarily from x-ray crystallography) has shown that molecular interactions and information transfer in biology occur through molecular shape and chemical complementarity since these govern the thermodynamic changes which occur on complexation. The DNA double helix is an ideal example of the ability of biology to use chemical structure to encode biological information. More recently, both experimental (e.g. NMR, neutron scattering and isothermal calorimetry) and theoretical methods (such as molecular dynamics simulation) have shown that the dynamic properties of these molecules can be just as important to their function. This is unsurprising, as free energy changes contain an entropic as well as an enthalpic contribution. Biological macromolecules possess the physical properties of soft condensed matter, and therefore they spend a non-negligible amount of time away from their time averaged structure due to thermal agitation [1]. Consequently, dynamics and entropy must also play a role in any biological process which relies on shape recognition. The importance of entropic changes in regulating the interaction of proteins and DNA with other molecules is highlighted by several experimental studies of so-called 'dynamic allostery', where the binding of a single substrate affects the affinity of the biomolecule for a second ligand through dynamic changes alone [2, 3]. There have been a few successful attempts to describe such processes theoretically in proteins by estimating the change in entropy on binding the first and second ligand using coarse grained models [4–6]. For DNA, we were able to use a combination of atomistic MD and the theoretical analysis described in this paper to show that binding a single substrate predisposes the dynamics (rather than structure) of the biomolecule to accommodating the second [2]. These theoretical models provide a mechanism for dynamic allostery by proposing that most of the overall unfavourable entropy change in forming the 2:1 ligand–protein or ligand–DNA complex occurs when the first ligand binds, leading to a highly cooperative interaction between binding events [2–6]. This striking example of the biological importance of entropic effects shows the need for accurate and reliable theoretical methods to describe and quantify biomolecular dynamics.

Biological macromolecules are large and chemically inhomogeneous. Their dynamic behaviour is highly coupled and spans timescales from femtoseconds (high-frequency local bond vibrations) to milliseconds (protein folding). This complexity has made it difficult to obtain a theoretical description of thermal fluctuations in biomolecular systems. Consequently, entropic changes have proven difficult to calculate and have been frequently overlooked. The current study describes how insight into DNA dynamics can be obtained by a combination of classical molecular dynamics (MD) simulations and quasi-harmonic analysis. These methods have now been successfully applied to understanding the thermodynamics of a number of nucleic acid interactions including sequence selective drug–DNA association [2], DNA–RNA complexes [7], and protein–DNA complexes [8], and to interpreting the results of single molecule nanomanipulation experiments which stretch and twist DNA in the laboratory [9, 10]. All of these studies have illustrated the importance of entropic effects. Furthermore, several databases collating MD trajectories for biomolecular systems have recently become available online [11, 12]. These have employed similar analysis techniques to proteins and DNA in order to summarize the increasingly large data sets generated as computational facilities continue to improve. The success of the techniques is due to the surprising simplicity of the dynamic behaviour of DNA over the nanosecond timescales currently available to classical MD. However, the philosophy underlying the analysis is non-trivial and will be discussed in some detail. In particular, there is some controversy in the literature concerning the calculation of configurational entropies from MD simulations which will be addressed [13, 14]. In contrast to the other descriptions of quasi-harmonic analysis available in the literature [15], this paper will focus on the underlying physics of DNA dynamics and comment on the limitations of the method. The hope is that this will lead to better physical descriptions of biomolecular dynamics

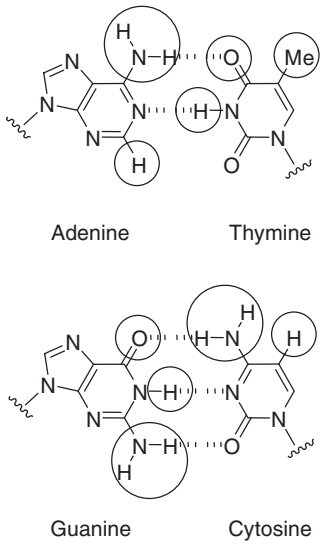


Figure 1. The DNA base pairs. Circled atoms are removed from the MD trajectory prior to analysis so that the two data sets are equivalent.

which will in turn advance our understanding of the role of thermal fluctuations and entropy in biology.

2. Methods and analysis

2.1. Molecular dynamics simulations of GC and AT DNA sequences

All classical MD simulations are fully atomistic and surround the DNA in an explicit solvent layer of 10 Å in each orthogonal direction. The simulations are performed using the AMBER8 suite of programs [16], which provides the current state of the art representation of nucleic acids within the field of biomolecular simulation. The two DNA sequences $d(GC)_{30}$ and $d(AT)_{30}$ were built using the NUCGEN module. Sufficient K^+ counterions were added to electrically neutralize the system, and the DNA was solvated with a rectangular box of TIP3P water molecules. The system was equilibrated using a standardized multi-step protocol [17]. Periodic boundary conditions in conjunction with the PME algorithm were used to properly reproduce long-range electrostatic interactions that are so important to the stability of duplex DNA. MD was performed over 15 ns for each 30mer sequence at NTP with an integration timestep of 2 fs and SHAKE to restrain all bonds to hydrogen. Solute coordinates were saved every 1 ps (the positions of water molecules and counterions are discarded); only the final 10 ns of the trajectory are used in the analysis.

To quantitatively compare the dynamic behaviour of two systems, it is necessary for them to contain an equal number of atoms. The 30mer sequences used in this study were designed so that removal of certain key atoms from the trajectory after MD will leave most of the system intact but produce two equivalent data sets, as shown in figure 1.

2.2. The nature of DNA dynamics revealed by quasi-harmonic analysis

Over the very shortest (femtosecond) timescales observable by classical simulation, DNA dynamics is dominated by very high-frequency bond vibrations as atoms oscillate about

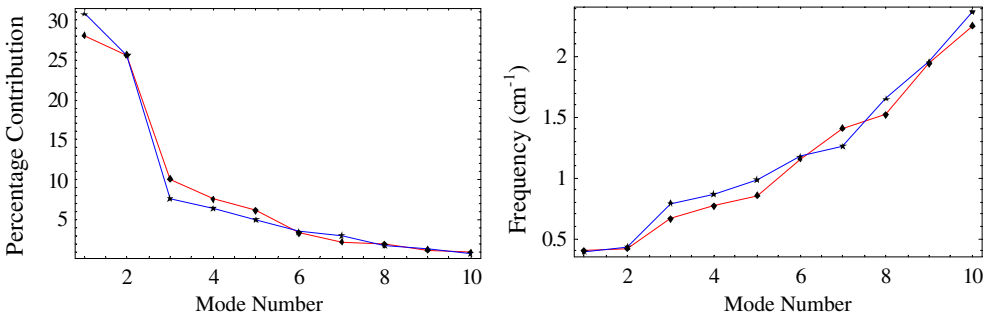


Figure 2. Percentage contribution (left) and frequencies (right) of the top ten eigenvalues for $d(AT)_{30}$ diamonds (red) and $d(CG)_{30}$ stars (blue).

their individual equilibrium position. However, the diffusion of groups of atoms away from their equilibrium positions gives the molecule some fluid-like characteristics over longer (picosecond) timescales. Over the longest (nanosecond) timescales currently accessible by atomistic MD, the helix undergoes large-scale global oscillatory motions which are reminiscent of quasi-harmonic normal modes. As these motions involve the largest displacements, they dominate the range of molecular conformations generated by thermal agitation and are therefore of most relevance to biology. The aim of quasi-harmonic analysis (which is frequently called principal component analysis, or PCA) is to extract such motions from the trajectory statistically. It is thereby possible to describe the complex set of molecular structures produced by an MD simulation in terms of a small number of time-dependent components or modes which can be considered individually. Quasi-harmonic analysis thereby provides a simplified account of the molecule's dynamic behaviour by focusing on the subspace in which most atomic motion occurs; this space is often termed the 'essential subspace'. Therefore, as well as providing useful physical information, it is often used as a technique for data compression [11, 12, 15].

2.3. The method of quasi-harmonic analysis

The data set consists of a series of DNA configurations generated by MD simulation. Each of these is comprised of the $3N$ Cartesian coordinates describing the positions of each atom within the structure, where $N = 1618$ for both of the systems considered in this study. Global translation and rotation of the molecule within the simulation box is removed prior to the analysis by least squares fitting to a reference structure. The first step of quasi-harmonic analysis is to construct the $3N \times 3N$ covariance matrix which maps correlations in atomic position over the trajectory. For a trajectory containing M coordinate sets, the elements of this covariance matrix are given by

$$C_{p,q} = \frac{1}{M} \sum_{m=1}^{M} \left(X_{m,p} - \langle X_p \rangle\right) \left(X_{m,p} - \langle X_p \rangle\right) \tag{1}$$

where $\langle X_p \rangle$ is the mean position of atom p during the simulation. Numerical diagonalization of this covariance matrix provides a set of $3N$ statistically independent eigenvectors or modes and their corresponding eigenvalues. The modes are ordered in terms of decreasing eigenvalue, since the modes with the largest eigenvalues have made the most significant contribution to the dynamics. Figure 2 (left) shows the percentage contribution of the top ten eigenvalues extracted from each of the two trajectories. The sharp drop in eigenvalue with mode number (from ∼30% for mode 1 down to under 1% for mode 10) suggests that only a few components

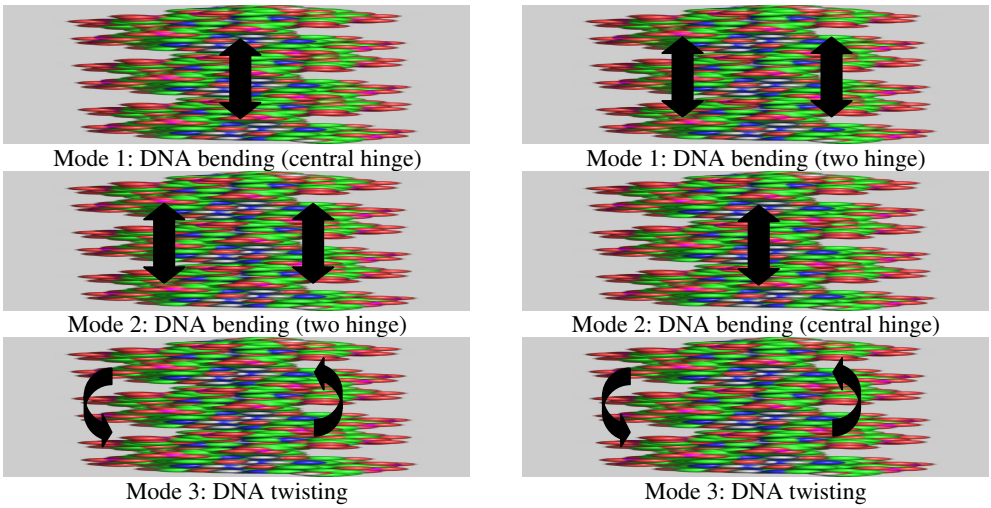


Figure 3. The first three modes of oscillation extracted from 10 ns trajectories of $d(AT)_{30}$ (left) and $d(GC)_{30}$ (right) using quasi-harmonic analysis.

contribute significantly to the trajectory. The frequencies (figure 2, right) calculated from the mass weighted covariance matrix compare favourably with the results obtained using normal mode analysis for shorter (18mer) DNA sequences [18]. However, these frequencies must be interpreted with caution. The eigenvectors and eigenvalues obtained depend both on the length of the simulation and the position of the sampling window used in the analysis, as will be discussed in section 2.5.

2.4. Physical interpretation of the eigenvectors and eigenvalues

Quasi-harmonic analysis identifies a set of independent structural changes in the molecule and ranks them in order of decreasing size. It is instructive to characterize these structural changes by producing animations of the first few modes as represented schematically in figure 3 (see also supplementary information available at stacks.iop.org/JPhysCM/19/076103).

The alternating sequences $d(AT)_{30}$ and $d(GC)_{30}$ chosen for this study both display the simple modes of oscillation (bending, twisting) expected for a uniform helical polymer. Such motions are important biologically as they describe the manner in which the biomolecule is particularly flexible. Many proteins which bind selectively to DNA distort the helix as part of the recognition process. Clearly, the duplex must be flexible or the interaction could not proceed. Thermodynamically, this implies a prohibitively large enthalpic penalty. The response of a given DNA sequence to an applied structural change will depend on how 'sympathetic' this distortion is to the natural flexibility of the molecule, which is described by the modes extracted through quasi-harmonic analysis. Although we have found in general that the dynamics of DNA is remarkably simple, more complex sequences have been found to possess highly specific 'dynamic motifs' (for example, TpA steps introduce a flexible 'hinge joint' into the duplex) which assist DNA binding proteins to discriminate between one binding site and another [19–21]. These variations in duplex flexibility are less well characterized than sequence-dependent structural differences which can be measured experimentally with x-ray crystallography and NMR. This importance of sequence-dependent flexibility at the single base pair level is discussed in more detail in section 2.6.

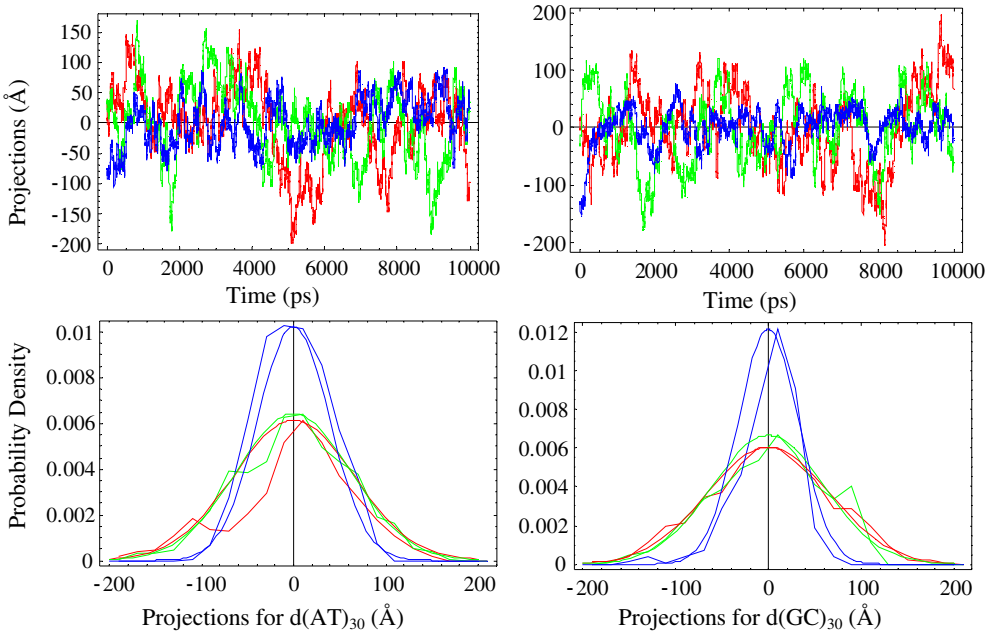


Figure 4. Projections and probability distributions for the first (red), second (green) and third (blue) quasi-harmonic modes for $d(AT)_{30}$ (left) and $d(GC)_{30}$ (right). The frequency spectra are weighted towards higher frequencies and have narrower probability distributions as the mode number increases. Gaussian distributions with equivalent variances and means are shown for comparison.

The modes extracted by quasi-harmonic analysis involve bending and twisting, as would be expected for a helical rod. Therefore, it may be more appropriate to perform the analysis using helical rather than Cartesian coordinates, although the algorithm would be considerably more complex to implement. The aim of quasi-harmonic analysis is to identify key independent motions within an MD trajectory, and this is only possible using an appropriate basis set. For example, a linear combination of two eigenvectors is necessary to describe a simple circular orbit in Cartesian coordinate space, even though intuitively there is only one motion. To date, we have assumed that the displacements are small enough for the small-angle approximation to apply so that a Cartesian representation is sufficient. However, it may be necessary to revisit this question in the future.

2.5. Projections and time dependence of the modes

The time dependences of the first three modes of oscillation are shown for $d(AT)_{30}$ (left) and $d(GC)_{30}$ (right) in figure 4. These are extracted by projecting the relevant eigenvector back onto each successive structure in the original trajectory.

All of the first three eigenvectors describe motions which are oscillatory about some mean structure. There is no net drift in any of the projections during the course of the simulation to suggest a continuous change in the overall conformation of the molecule. In this manner, quasi-harmonic analysis provides a measure of convergence and equilibration of an MD trajectory [15]. The probability densities shown in figure 4 are all very close to the Gaussian distribution expected for a harmonic oscillator in a heat bath, although there is clearly

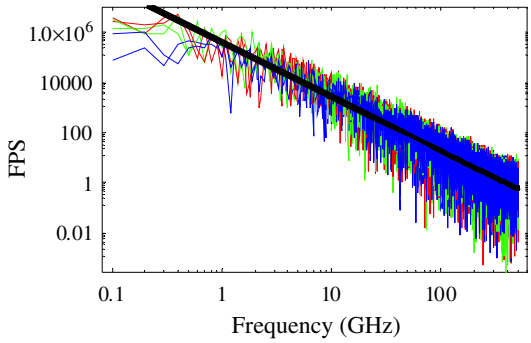


Figure 5. The frequency power spectrum (FPS) of the first (red), second (green) and third modes of $d(AT)_{30}$ and $d(GC)_{30}$. Very similar behaviour is observed for all three modes of both systems. Superimposed is a straight line fit to the data (black) with a gradient of -2.12 ± 0.07.

some undersampling of the very lowest modes. The duplex is therefore well described by a simple coupled oscillator model. However, the time dependence is not perfectly sinusoidal, but contains a complex spectrum of frequencies. Figure 5 shows the frequency power spectrum (FPS) calculated for the first three modes of both sequences. The FPS was calculated using the Mathematica program [22] from the Fourier transforms of the projections shown in figure 4.

The absence of peaks in the FPSs shown in figure 5 indicates that these low-frequency oscillations are heavily overdamped by frequent collisions between the DNA and molecules in the surrounding solvent. At frequencies above 1 GHz, the FPSs all obey an inverse square power law decay (ω^{-2}) with increasing frequency. Using a simple harmonic oscillator model coupled to a stochastic Langevin noise term $F(t)$ to describe the behaviour of each mode gives

$$m\ddot{x} + \varsigma\dot{x} + \kappa x = F(t) \tag{2}$$

where κ is the spring constant associated with the mode and ς is the frictional relaxation time. Equation (2) is equivalent to the damped elastic cylinder model originally applied to short DNA fragments by Lankas *et al* [23]. In an overdamped system, viscosity dominates inertia so that the effective mass of the modes can be set to zero. The power spectrum can then be shown to be [24]

$$\mathrm{FPS} = \frac{2kT\varsigma}{\kappa^2 + \varsigma^2\omega^2}. \tag{3}$$

Equation (3) implies that the FPS decays as ω^{-2} for high frequencies, in agreement with the data shown in figure 5 for the most prominent modes of $d(AT)_{30}$ and $d(GC)_{30}$. An estimate of the relaxation coefficients ς from the MD simulation is unlikely to be reliable as there is clearly insufficient sampling to quantify any of the low-frequency behaviour of the modes over the 10 ns timescale of the simulation. Nevertheless, the data do show deviation away from ω^{-2} decay for frequencies between 0.1 and 1 GHz, suggesting that the relaxation time should lie somewhere between 1 and 10 ns. This has serious implications for the sampling times required to quantify the relaxation kinetics of overdamped biomolecules using MD. The data shown in figure 5 imply that it would be desirable to sample frequencies as low as 0.01 GHz for a DNA 30mer, corresponding to MD simulations of 100 ns in length. This would be prohibitively computationally expensive for systems of this size. What the data clearly show, however, is the importance of interactions with the solvent in the dynamic behaviour of DNA.

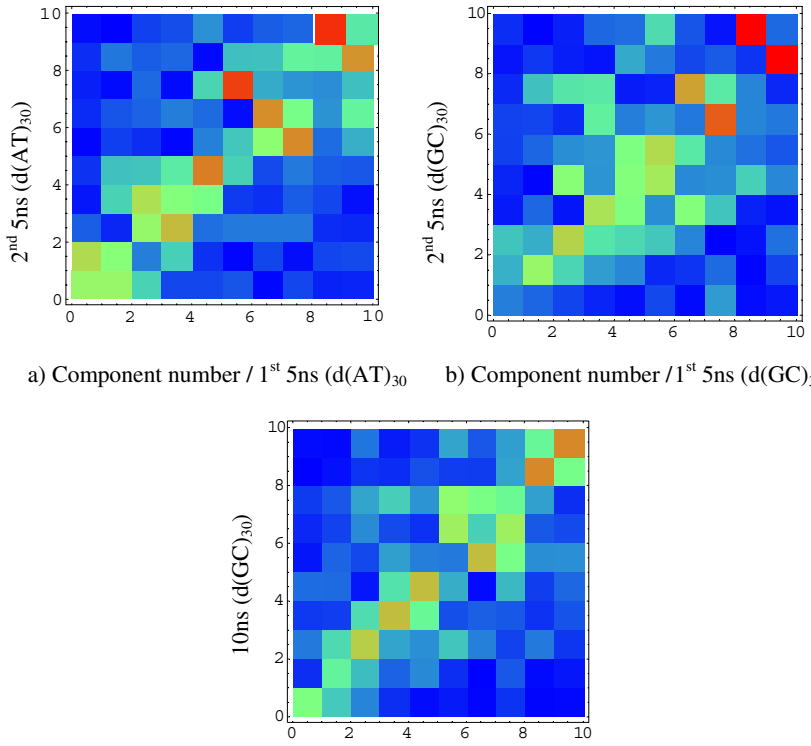
Proteins and nucleic acids are notoriously sensitive to changes in their environment [25, 26], and the role of water in determining the thermodynamics of biological interactions is well known [27]. However, the biological importance of friction and solvent damping at the atomic level has received less attention. Simulations in which the solvent is represented implicitly using generalized Born methods display very different dynamic behaviour in which many of the high-frequency modes are missing [28]. Although such models are thought to provide a realistic description of the dielectric properties of water (and provide thermodynamic quantities that are in agreement with explicitly solvated calculations), the viscous nature of the solvent is poorly represented by this treatment. However, recent neutron scattering experiments performed in H_2O and in D_2O for comparison have shown that solvent viscosity and damping can indeed affect biomolecular kinetics [29]. Intuitively, it seems likely that biological processes which require large conformational changes or translocations (such as gene transcription or replication) must be heavily influenced by solvent interactions and friction, although such effects remain poorly understood [30].

2.6. Time evolution of the eigenvectors

The eigenvectors define a set of structural changes that represent the modes of flexibility of a given DNA sequence. However, it is difficult to estimate elastic constants (such as the bending and twisting modulus) from quasi-harmonic analysis. Firstly, it is not always possible to characterize the motion associated with a given eigenvector precisely. Furthermore, the eigenvectors show a marked dependence on the length and position of the sampling window used in the analysis. This becomes clear whenever eigenvectors from different portions of a single simulation trajectory are compared. The MD trajectories for $d(AT)_{30}$ and $d(GC)_{30}$ were both split into two 5 ns segments and quasi-harmonic analysis was performed independently on each half. The correlation between the eigenvector sets obtained for the first and second halves of the simulations was then measured by calculating the dot products between the top ten modes, as shown in figure 6. Similarly, the eigenvectors obtained for $d(AT)_{30}$ were compared with those for $d(GC)_{30}$ using the full 10 ns trajectory.

Although the eigenvectors are clearly time dependent, the matrices show significant overlaps clustered around the diagonals. As a simulation progresses, modes can be promoted or suppressed and they may even recouple. The oscillatory motions in DNA are not expected to be perfectly independent as the trajectory contains anharmonic contributions from non-bonded terms in the potential function describing the macromolecule and the solvent. Some mixing of modes over longer timescales is therefore to be expected.

Although the general mechanical properties of the helix do remain qualitatively the same (even the eigenvectors obtained from the two different simulation trajectories $d(AT)_{30}$ and $d(GC)_{30}$ are remarkably similar, as previously reported [15]), the time dependences of the eigenvectors and eigenvalues extracted using quasi-harmonic analysis are sufficient to make these quantities an unreliable *quantitative* measure of the modes of flexibility of the duplex. Experimental measurements of the flexibility of DNA at the single base pair level are also difficult to obtain; however, a statistical analysis of the crystal structures of protein–DNA complexes provides results consistent with MD studies [19, 20, 23]. The macroscopic bending and twisting flexibility of DNA has been measured using a number of complementary techniques; in particular, single molecule nanomanipulation experiments have made it possible to stretch or twist long DNA molecules in the laboratory. Although these experiments have provided a wealth of information on the macroscopic elastic properties of duplex DNA, they provide very different information than is required to understand sequence-selective ligand–DNA or protein–DNA recognition. Theoretical studies which have attempted to use MD



a) Component number / 1st 5ns ($d(AT)_{30}$ b) Component number / 1st 5ns ($d(GC)_{30}$)

c) Component number / 10ns ($d(AT)_{30}$)

Figure 6. Dot product matrices showing the correlation between eigenvectors obtained from subdividing an MD trajectory into shorter sections ((a), (b)). The eigenvectors from the 10 ns simulations of $d(AT)_{30}$ and $d(GC)_{30}$ are also compared (c). All significant correlations are clustered around the diagonals. A high correlation between eigenvectors (>0.7) is shown in red; dissimilar modes (<0.4) are shown in blue.

simulations of short sequences to infer global elastic properties for comparison have shown that there are serious conceptual difficulties in relating local and global flexibilities for highly anisotropic molecules such as nucleic acids [7, 23]. For instance, there is the possibility that the duplex may be locally flexible while appearing to be relatively rigid over macroscopic length scales if motions in different regions cancel, as discussed in detail by Noy *et al*. While the eigenvectors do provide useful qualitative structural information about duplex flexibility, we now demonstrate that the Schlitter entropy provides a more reproducible method to quantitatively compare dynamic behaviour. Furthermore, since the Schlitter entropy hopes to represent a true thermodynamic quantity, the method does provide information that can be compared with experiment.

2.7. Calculating the configurational entropy

The free energy change driving biomolecular association contains both a (static) enthalpic contribution and a (dynamic) entropic contribution. The configurational entropy provides a quantitative measure of the relative flexibilities of two or more systems. Entropies are notoriously difficult to calculate using computer simulation. The need to explore all of the conformational space available to a complex molecule makes these calculations highly

computationally demanding. Biomolecules can also be very soft. Many proteins consist of disordered loop regions held together by stiffer secondary structural elements such as α-helices and β-sheets. Oscillatory and diffusive motions are highly coupled within these systems, making their dynamics difficult to describe. However, the results of quasi-harmonic analysis indicate that duplex DNA is rather well described by a simple coupled oscillator model. Evaluating the entropy of such a classical oscillator within the canonical ensemble suggests that the entropy difference between simulations A and B can be obtained trivially from the eigenvalues (λ_i) from quasi-harmonic analysis using [31]

$$\Delta S_{B\to A} = \frac{1}{2}k\sum_{i=1}^{3N-6}\ln\frac{\lambda(B)_i}{\lambda(A)_i}. \tag{4}$$

Although theoretically correct, in practice equation (4) does not provide a suitable method for calculating entropy differences from MD simulation. The very smallest eigenvalues obtained from quasi-harmonic analysis are frequently non-physical and do not describe true degrees of freedom, but are just there to maintain the completeness of the data set. Although these inaccuracies are small, they can contribute significantly to the entropy when equation (4) is used, since the logarithm becomes large and negative for eigenvalues which are very much less than unity. One possible solution to this problem is to define an arbitrary cut-off beyond which eigenvalues are considered negligible; however, this often introduces artefacts into the results. A more reliable solution is to ensure that numerical noise cannot make a significant contribution. The very smallest eigenvalues represent frequencies which lie outside the classical regime. It is not surprising that these contribute spuriously to the entropy as only a discrete set of states is available to a quantum oscillator, whereas a continuum of states is allowed classically. Any method which correctly suppresses these unphysically high frequencies whilst still reproducing the true, classical degrees of freedom should provide a more reliable method for calculating the entropy than equation (4). An obvious solution is to use the formula for a quantum mechanical (QM) oscillator, as suggested by Andricioaei and Karplus [14]:

$$S_{\mathrm{qm}} = k\sum_{i=1}^{3N-6}\left[\frac{\alpha_i}{\mathrm{e}^{\alpha_i}-1} - \ln\left(1-\mathrm{e}^{-\alpha_i}\right)\right]. \tag{5}$$

Alternatively, the Schlitter equation provides an approximation to the QM entropy in the classical limit [13]:

$$S_{\mathrm{Sc}} = \frac{1}{2}k\sum_{i}^{3N-6}\ln\left(1+\frac{e^2}{\alpha_i^2}\right) \qquad \text{where } \alpha = \frac{\hbar\omega}{kT}. \tag{6}$$

The entropy can also be calculated directly from the eigenvalues of the mass weighted covariance matrix (γ_i) using the equipartition theorem:

$$S_{\mathrm{Sc}} = \frac{1}{2}k\sum_{i}^{3N-6}\ln\left(1+\frac{kTe^2}{\hbar^2}\gamma_i\right). \tag{7}$$

The Schlitter equation reduces to equation (4) in the classical limit (when α is small), but also tends to zero for infinitely high frequencies, as required. For a simple justification of the formula, note that the first-order Taylor expansion of equation (6) is equivalent to the first-order expansion for the quantum harmonic oscillator. The Schlitter equation is rather counterintuitive as differences in classical entropy do not usually depend on the quantum of action $\hbar$; however, it is required because the shortest timescales explored by classical MD lie very close to (and to some extent overlap with) the QM regime. Formally, it is more

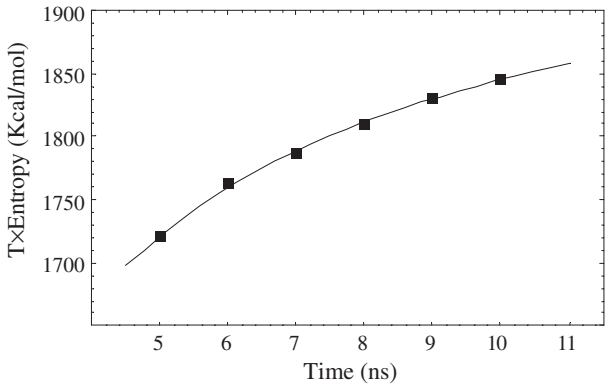


Figure 7. Configurational entropy as a function of the length of the sampling window. Fitting equation (6) gives numerical values for S_∞, A and n of 2053, 978 and 0.67, respectively.

appropriate to use the Schlitter equation to calculate the entropy from classical MD than the fully QM expression (although others have expressed the opposite opinion). Unfortunately, a classical MD simulation cannot provide an accurate representation of the time evolution of atomic positions in the semi-classical regime, so treating these degrees of freedom particularly rigorously is somewhat dishonest. Clearly, both treatments are identical in the classical limit where molecular mechanics operates, and provide reassuringly similar results.

2.8. The convergence of the configurational entropy

The configurational entropy of the system is a logarithmic measure of the volume of conformational space accessible to a molecule at a given temperature. For simulations of finite length, this volume is restricted not only by the constraints of molecule structure, but also by undersampling. Therefore, the entropy calculated from an MD trajectory (using either the Schlitter or the QM approach) will contain a hidden time dependence that must be eliminated before the measurement is representative of the whole thermodynamic ensemble. The Schlitter entropy calculated for $d(AT)_{30}$ as a function of the length of the simulation is shown in figure 7; almost identical results were obtained for $d(GC)_{30}$.

For a system containing N atoms, it is formally necessary to include at least $3N - 6$ configurations in the quasi-harmonic analysis to ensure completeness of the data set. The Schlitter entropy was calculated over the shortest allowable timescale of 5 ns (corresponding to 5000 configurations for $N = 1618$) in 1 ns increments up until the full 10 ns obtained by MD. The apparent flexibility of the duplex is strongly dependent on the length of the sampling window, and has not converged over the 10 ns simulation. However, in contrast to the eigenvectors obtained from quasi-harmonic analysis, the Schlitter entropy is insensitive to the position of the sampling window. For example, the Schlitter entropies calculated for the first and second 5 ns halves of the trajectory agree to within 1% of the total. This indicates that the dynamics of the duplex is determined by an underlying energy landscape which is smooth over MD timescales. In contrast, proteins are expected to have a rough energy landscape containing many multiple minima [32]. Nevertheless, the Schlitter entropy has also proven useful for quantifying dynamic changes in globular proteins, despite this additional complexity [8, 33–35]. The two contrasting cases of rough and smooth energy landscapes are compared in figure 8. One important consequence of a smooth energy landscape is that the Schlitter entropy can be determined accurately for a sampling window of a given length

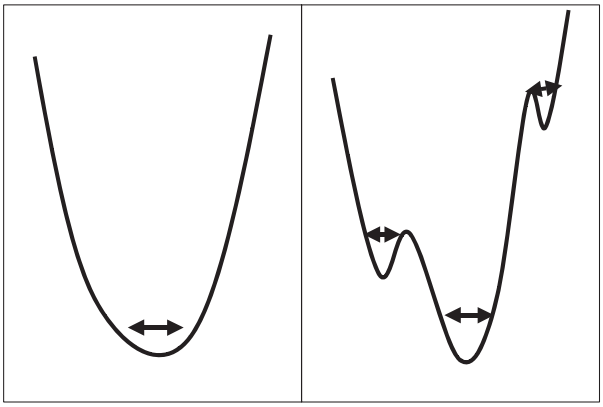


Figure 8. The relationship between the shape of configuration space and the predictability of the Schlitter entropy. If the molecule moves over a smooth energy surface (left) then the entropy will be insensitive to the position of the sampling window, as is the case for duplex DNA. However, if the surface contains many multiple minima (right) then the entropy will change as different regions are explored. The extrapolation will be less reliable for the discontinuous surface.

(hence the scatter in the data points shown in figure 7 is negligibly small and no error bars are shown). It is therefore possible to fit a function (such as equation (6)) to the curve shown in figure 7 with some confidence and predict the entropy for infinite simulation times by extrapolation [2].

$$S = S_\infty - \frac{A}{t^n}. \tag{8}$$

Although the constant S_∞ is interpreted as the configurational entropy, as yet there is no equivalent physical interpretation for the fitting constants A or n. Intriguingly, we have always obtained a value for n that is close to $2/3$. We hope to develop a theoretical model which provides some physical insight both into the fitting parameters and the functional form of equation (8), since it might then be possible to establish upper and lower bounds for the errors in the entropy predicted by extrapolation. Currently, this is difficult to estimate. In all cases when we have applied this method (e.g. to measure the decrease in entropy caused by binding a drug in the minor groove), we have been able to detect a sufficiently large change in the dynamics that we could measure the direction of the entropy change with confidence, even if the errors in the magnitude are not known with certainty [2, 9].

The convergence of the Schlitter entropy provides some interesting insight into the rate at which the 30mer duplex samples conformational space. A fit to equation (8) implies that the 10 ns simulation samples nearly 90% of the dynamic behaviour of the duplex. However, a further 30 ns is required to capture 95% of the dynamics, and over 350 ns is necessary to get to within 1%. Surprisingly, 50% of the total flexibility of the molecule can be observed in only 1 ns, but improving the sampling to over 95% is prohibitively computationally expensive. This is consistent with the observation that MD does provide useful information about the structure and dynamics of biomolecules despite the limitations in timescale. However, special simulation methods (e.g. steered MD) have to be employed to observe rare events, and relaxation timescales for systems of this size are currently too long to be quantified using state of the art simulation techniques.

3. Conclusions

Quasi-harmonic analysis provides a combination of qualitative and quantitative information about the dynamic behaviour of DNA in MD simulations. In general, DNA duplexes are found to undergo global bending and twisting oscillations over nanosecond timescales which are heavily overdamped due to the high viscosity of the aqueous solvent. The eigenvectors and eigenvalues from quasi-harmonic analysis can also provide useful insight into the structural changes associated with the major modes of flexibility of a given sequence. For example, in a previous study of sequence selective drug–DNA recognition we compared eigenvectors in the presence and absence of the ligand to illustrate the similarity between the major modes of flexibility of this sequence and the change in conformation required to accommodate the drug [21]. However, the eigenvectors and eigenvalues should only be interpreted semi-quantitatively since they will vary as the simulation progresses. Furthermore, the relationship between local and global flexibilities of anisotropic systems such as DNA remains poorly understood theoretically, and there is currently no reliable method for determining the global persistence length from microscopic measurements.

In contrast, the Schlitter entropy provides a more robust measure of the flexibility of a DNA sequence for an MD trajectory of a given length. The entropy for infinite simulation times can then be estimated by extrapolation. Most of the biologically important consequences of molecular flexibility at atomic length scales are due to its contribution to the thermodynamics of receptor–ligand recognition. It is notoriously difficult to calculate the entropic contribution to the overall free energy change from MD (whereas the enthalpic term is obtained trivially from the molecular mechanics forcefield). The Schlitter equation combined with quasi-harmonic analysis provides a useful method for estimating changes in configurational entropy in duplex DNA, and can therefore contribute to our understanding of the thermodynamics of biomolecular interactions. The method has always provided a good agreement with the experimental observations for the systems that we have studied to date. However, these calculations only required that the *sign* of the entropy change is correct. No direct comparison between the magnitude of the configurational entropy changes obtained through quasi-harmonic analysis and experiment has yet been possible to our knowledge. Clearly, further thermocalorimetric studies on DNA–ligand interactions to validate these results would be highly desirable.

The DNA helix behaves essentially as a semi-flexible polymer; consequently its dynamic behaviour can be described using a simple quasi-harmonic approximation. Unfortunately, this approach may not be as appropriate for studying protein dynamics. Proteins contain a mixture of highly flexible and relatively stiff structural elements coupled together into a folded globular structure. Therefore, an equivalent description of protein dynamics would require a model in which the motion of diffusive loop regions could be coupled to the damped oscillatory motion of the stiffer elements (analysis of a single alpha-helix would be qualitatively equivalent to an analysis of a DNA oligomer). The simple behaviour of DNA makes it an ideal model system for the development of new methods to describe and quantify biomolecular dynamics. Future calculations will hope to gain more insight into the physics that governs the convergence of the Schlitter entropy and rate of exploration of conformational space. Furthermore, the importance of water viscosity in biomolecular dynamics, as highlighted by this study, warrants further investigation by both theory and experiment.

Acknowledgments

We would like to thank Simon Woods, Richard Graham, Bhavin Khatri and Tom McLeish for useful discussions, and Geoff Wells for reading the manuscript. We are also grateful to the Royal Society for providing funds to purchase the computer resources used in this study.

References

[1] Cooper A 1984 *Prog. Biophys. Mol. Biol.* **44** 181–214
[2] Harris S A, Gavathiotis E, Searle M S, Orozco M and Laughton C A 2001 *J. Am. Chem. Soc.* **123** 12658–63
[3] Popovych N, Sun S, Ebright R H and Kalodimos C G 2006 *Nat. Struct. Biol.* **13** 831–8
[4] Cooper A and Dryden D T 1984 *Eur. Biophys. J.* **11** 103–9
[5] Hawkins R J and McLeish T C B 2006 *Biophys. J.* **91** 2055–62
[6] Hawkins R J and McLeish T C B 2004 *Phys. Rev. Lett.* **93** 098104
[7] Noy A, Pérez A, Lankas F, Luque J F and Orozco M 2004 *J. Mol. Biol.* **343** 627–38
[8] Dixit S B, Andrews D Q and Beveridge D L 2005 *Biophys. J.* **88** 3147–57
[9] Harris S A, Sands Z and Laughton C A 2005 *Biophys. J.* **88** 1684–91
[10] Wereszynski J and Andricioaei I 2006 *Proc. Natl Acad. Sci. USA* **103** 16200–5
[11] Beveridge D L *et al* 2004 *Biophys. J.* **87** 3799–813
[12] Tai K *et al* 2004 *Org. Biomol. Chem.* **2** 3219–21
[13] Schlitter J 1993 *Chem. Phys. Lett.* **215** 617–21
[14] Andricioaei I and Karplus M 2001 *J. Chem. Phys.* **115** 6289–92
[15] Pérez A, Blas J R, Rueda M, López-Bes J M, de la Cruz X and Orozco M 2005 *J. Chem. Theor. Comput.* **1** 790–800
[16] Case D A *et al* 2004 *AMBER 8* (San Francisco, CA: University of California)
[17] Shields G C, Laughton C A and Orozco M 1997 *J. Am. Chem. Soc.* **119** 7463–9
[18] Matsumoto A and Go N 1999 *J. Chem. Phys.* **110** 11070–5
[19] Olson W K, Gorin A A, Liu X J, Hock L M and Zhurkin V B 1998 *Proc. Natl Acad. Sci. USA* **95** 11163–8
[20] Matsumoto A and Olson W K 2002 *Biophys. J.* **83** 22–41
[21] Bostock-Smith C E, Harris S A, Laughton C A and Searle M S 2001 *Nucl. Acids. Res.* **29** 693–702
[22] Wolfram Research Inc. 2003 *Mathematica, Version 5.0* Champaign, Illinois
[23] Lankaš F, Šponer J, Hobza P and Langowski J 2000 *J. Mol. Biol.* **299** 695–709
[24] Khatri B S, Kawakami M, Byrne K, Smith D A and McLeish T C B 2007 *Biophys. J.* doi:10.1529/biophysj.106.097709
[25] Fuller W, Forsyth T and Mahendrasingam A 2004 *Phil. Trans. R. Soc.* B **359** 1237–48
[26] Neidle S 2002 *Nucleic Acid Structure and Recognition* (Oxford: Oxford University Press)
[27] Finney J L 2004 *Phil. Trans. R. Soc.* B **359** 1145–65
[28] Sands Z A and Laughton C A 2004 *J. Phys. Chem.* B **108** 10113–9
[29] Halle B 2004 *Phil. Trans. R. Soc.* B **359** 1207–24
[30] Tehei M, Madern D, Pfister C and Zaccai G 2001 *Proc. Natl Acad. Sci. USA* **98** 14356–61
[31] Karplus M and Kushick J 1981 *Macromolecules* **14** 325–32
[32] Kitao A, Hayward S and Gō N 1998 *Proteins* **33** 496–517
[33] Schäfer H, Daura X, Mark A E and van Gunsteren W F 2001 *Proteins Struct. Funct. Genet.* **43** 45–56
[34] Schäfer H, Smith L J, Mark A E and van Gunsteren W F 2001 *Proteins Struct. Funct. Genet.* **46** 215–24
[35] Hsu S D, Peter C, van Gunsteren W F and Bonvin A M J J 2005 *Biophys. J.* **88** 15–24