PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Large-scale molecular dynamics simulation of DNA: implementation and validation of the AMBER98 force field in LAMMPS

Christina Grindon, Sarah Harris, Tom Evans, Keir Novik, Peter Coveney and Charles Laughton

Phil. Trans. R. Soc. Lond. A 2004 **362**, doi: 10.1098/rsta.2004.1381, published 15 July 2004

10.1098/rsta.2004.1381

THE ROYAL SOCIETY

Large-scale molecular dynamics simulation of DNA: implementation and validation of the AMBER98 force field in LAMMPS

BY CHRISTINA GRINDON[1], SARAH HARRIS[1], TOM EVANS[2], KEIR NOVIK[2], PETER COVENEY[2] AND CHARLES LAUGHTON[1]

[1]*School of Pharmaceutical Sciences, University of Nottingham, Nottingham NG7 2RD, UK* (charles.laughton@nottingham.ac.uk)
[2]*Centre for Computational Science, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK* (p.v.coveney@ucl.ac.uk)

Published online 5 May 2004

Molecular modelling played a central role in the discovery of the structure of DNA by Watson and Crick. Today, such modelling is done on computers: the more powerful these computers are, the more detailed and extensive can be the study of the dynamics of such biological macromolecules. To fully harness the power of modern massively parallel computers, however, we need to develop and deploy algorithms which can exploit the structure of such hardware. The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) is a scalable molecular dynamics code including long-range Coulomb interactions, which has been specifically designed to function efficiently on parallel platforms. Here we describe the implementation of the AMBER98 force field in LAMMPS and its validation for molecular dynamics investigations of DNA structure and flexibility against the benchmark of results obtained with the long-established code AMBER6 (Assisted Model Building with Energy Refinement, version 6). Extended molecular dynamics simulations on the hydrated DNA dodecamer d(CTTTTGCAAAAG)$_2$, which has previously been the subject of extensive dynamical analysis using AMBER6, show that it is possible to obtain excellent agreement in terms of static, dynamic and thermodynamic parameters between AMBER6 and LAMMPS. In comparison with AMBER6, LAMMPS shows greatly improved scalability in massively parallel environments, opening up the possibility of efficient simulations of order-of-magnitude larger systems and/or for order-of-magnitude greater simulation times.

Keywords: molecular dynamics; high-performance computing; DNA; LAMMPS; AMBER

1. Introduction

Molecular modelling was at the heart of the discovery of the structure of DNA by Watson & Crick (1953). They famously used rudimentary mechanical models that

One contribution of 16 to a Theme 'The mechanics of DNA'.

were assembled in order to produce a three-dimensional molecular structure consistent with the observed X-ray diffraction pattern of DNA crystals. Today, of course, molecular modelling has moved on considerably, owing in very large measure to the dramatic impact of the computer on scientific research. At the most basic level, computer graphics affords a virtual environment for constructing and visualizing the most complicated three-dimensional structures of macromolecules, including proteins and nucleic acids. Beyond their purely qualitative value, classically based molecular modelling methods enable us to evaluate low-lying energy states of complex molecules using energy minimization and Monte Carlo techniques, while molecular dynamics (MD) provides detailed information about the motions of these molecules. The size and time-scales accessible by such methods are circumscribed in large measure by the speed of contemporary computers, which increases relentlessly with each passing year. Computational speed has been enhanced by the advent of parallel computing over the past decade but, as is also true of other less dramatic developments in computing, the structure of the parallel algorithms deployed also needs to be optimized to extract maximum benefit from today's massively parallel supercomputers. This paper is concerned with addressing these issues for the molecular modelling of DNA. We shall be specifically concerned with MD simulations of DNA on parallel supercomputers, a subject in its infancy today.

MD simulations are increasingly proving their worth for the fuller understanding of biomolecular structure and dynamics (Karplus & McCammon 2002). In the search for ever-more realistic and observationally useful simulations, the pressure has been on the practitioner to simulate increasingly complex systems (e.g. proteins in membranes) over increasing lengths of time. The rationale behind the drive for more complex systems is straightforward: the behaviour of a biomolecule *in vivo* is highly dependent on its environment, and the treatment of this in anything less than atomistic detail generally produces results that are unacceptable for biological purposes, although implicit models of solvation (Delgado-Buscalioni & Coveney 2003; Hawkins *et al.* 1996; Weiser *et al.* 1999) are now showing some promise. The desire for ever more extended simulations is driven by a number of factors. Firstly, MD simulations are often used as a form of structure optimization, where initially constructed models are allowed to relax through MD until some sort of equilibrated state is achieved. The complexity of many systems makes this a process impossible to achieve through deterministic energy minimization alone (due to the multiple minimum problem) but it is often very slow by MD. Secondly, MD simulations are perhaps the only method available for the study of dynamical processes at the atomic level of detail, the example *par excellence* being protein folding (Duan & Kollman 2001; Fersht & Daggett 2002; Wu *et al.* 2002). With the computational facilities currently available, simulation time-scales for systems of just average complexity (e.g. a small peptide in solution) of over *ca.* 100 ns can be regarded as heroic (Duan & Kollman 1998), yet many interesting and biologically relevant processes take place on the micro- or millisecond time-scale (Yakushevich 1998). Thirdly, even if the time-evolution of a system is not of interest, MD simulations are often used as an efficient method of generating a thermodynamically relevant ensemble of structures for a biomolecule from which thermodynamic parameters may be calculated (Beveridge & DiCapua 1989; Kollman 1993; Kollman *et al.* 2000). The accuracy of these calculations depends critically on the conformational sampling from MD and, for large systems with impor-

tant low-frequency modes of conformational flexibility, simulations of at least many nanoseconds may be required (cf. the protein folding problem itself).

Conformational flexibility has long been recognized as a hallmark of DNA structure, and a vital consideration for a full understanding of DNA function and recognition. Many of the important conformational changes that DNA can undergo take place on the micro- to millisecond time-scale: for example, the 'breathing' of the bases, in which the normal pattern of Watson–Crick hydrogen bonds is temporarily disrupted as a base swings out of the helix and is exposed to solvent, or perhaps is recognized by a DNA-binding protein. Currently, we cannot simulate this spontaneous process at the atomic level because of the time-scale issues discussed above. We must either bias the simulations and 'force' the breathing event (Varnai & Lavery 2002), or study the dynamics of non-natural bases, such as difluorotoluene, which breathe on a much faster time-scale (Cubero *et al.* 1999), or move to non-atomistic representations where much longer time-scales are computationally feasible (Wattis *et al.* 2001). But even for the quite minor conformational adjustments that accompany, for example, the binding of ligands into the minor groove of B-form duplex DNA, a proper understanding of DNA dynamics is vital. This is because it is now becoming increasingly clear that the most important factors that drive recognition are not always enthalpic (e.g. related to an understanding of specific interactions made between the drug and the DNA) but are often entropic (Haq *et al.* 1999). And while it has been assumed that the major contribution to the entropy changes that accompany drug–DNA recognition comes from solvent reorganization, Harris *et al.* (2001) and others have recently shown that changes in the configurational entropy of the DNA can also be vitally important. While estimates of the enthalpic components of a recognition process may readily be made from the examination of static (e.g. X-ray crystallographic) structures by molecular modelling methods, and estimates of solvation effects may also be made from such information, they provide no clues at all as to any configurational entropic factors. For this we must have information about the dynamical behaviour of the DNA and the ligand, and how it changes between the bound and unbound state. As an example of this, we have shown by combining nuclear magnetic resonance (NMR) structure-determination methods with extended MD computer simulations that there are important changes to the dynamics and flexibility of the DNA decamer duplex d(GGTAATTACC)$_2$ when it binds a molecule of Hoechst 33258 in the minor groove of the central A tract (underlined) (Bostock-Smith *et al.* 2001). The binding reduces the ability of the duplex to bend at the normally very flexible TA steps, which we have shown is related to associated changes in minor groove width which 'clamp' the ligand in position. More recently, we have taken advantage of the development of methods to quantify configurational entropy changes to provide a full thermodynamic analysis of how Hoechst 33258 binds to the DNA duplex d(CTTTTGCAAAAG)$_2$ (Harris *et al.* 2001). NMR titration experiments had shown that this DNA duplex binds two molecules of Hoechst, one to each A_4/T_4 tract (underlined), in a highly cooperative manner such that no 1:1 drug/DNA complex could ever be detected (Gavathiotis *et al.* 2000). Through a series of extended MD simulations, we were able to show that this cooperativity was due to changes in the configurational entropy of the system. Binding of the first drug molecule caused major stiffening in the DNA structure, reducing its configurational entropy considerably. Binding of the second drug molecule could then take place with little further stiffening effect, and so was favoured. The reliability of cal-

culations of this type is critically dependent on the possibility of obtaining extended MD trajectories. We find that, as the simulation length increases, so does the calculated configurational entropy, as the DNA samples new areas of conformational space. The increase is not linear, but tends to a limit, and while we appear to be able to estimate that limit quite reasonably by curve fitting, the process requires considerable extrapolation beyond currently accessible simulation time-scales, and so is open to dispute. Improvements to studies of this type clearly require order-of-magnitude increases in simulation times, but since current simulations of this type typically already consume months of central-processor-unit (CPU) time, this is not a realistic option for day-to-day studies.

Parallel processing provides the most obvious approach to reducing time-to-solution for such calculations, but most common general-purpose MD algorithms are not well suited to parallel computing, let alone massively parallel processing. Indeed, the most popular MD life science packages (Brooks *et al.* 1983; Case *et al.* 1999) are long-established codes which were not originally designed with parallel implementation in mind. The overwhelming majority of such 'legacy' codes has been parallelized in the most direct fashion using the so-called 'replicated data' paradigm, which assigns the data on all atoms in an MD simulation to all N processors on the parallel computer. Such codes scale very poorly as both the size of the model and the number of processors are increased. For this reason, even on modern supercomputers, these codes are unable to exploit such unprecedented computing power to the full; indeed, the codes are rarely deployed on more than a very small number of processors, a situation which highlights the importance of 'smart' algorithms in harnessing maximum benefit from modern parallel computers. An alternative paradigm, which makes use of 'spatial domain decomposition', distributes the computation over spatially disjoint domains in the system which are handled by separate processors; thus, for N atoms and P processors, each processor carries data on N/P atoms, and is hence scalable to much larger systems, with far less stringent memory limitations. It should be clear that spatial domain decomposition in particular puts heavy demands on efficient interprocessor communication, especially for problems in which Coulomb interactions are dominant, as their long range guarantees that atoms on other processors influence the behaviour of those on each local processor. For this reason, one can expect the best performance to be delivered only on tightly coupled parallel machines, rather than on weakly coupled clusters, which have recently been gaining in popularity on grounds of cost. The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) (Plimpton & Hendrickson 1996) is one such recently developed, highly scalable MD code which implements spatial domain decomposition and is hence suitable for deployment on massively parallel supercomputers.

If such new codes are to gain widespread acceptance, it is vital that they are shown to produce results that agree with those obtained using the 'legacy' codes. However, there is a problem with this. Practical issues regarding differences in time-stepping, temperature coupling and treatment of long-range electrostatics, as well as the intrinsically chaotic nature of MD (which means that phase-space trajectories can be expected to diverge exponentially fast regardless of how close the initial conditions), mean that simulations run using LAMMPS or a similar code can never be expected to be identical to those run using the AMBER code, despite the use of the same force field. In this paper we describe how we established and implemented a series of benchmark comparisons that can be made between the results of

two simulations produced using two different computer architectures and/or software codes, to provide stringent and quantitative measurements of similarity despite this problem. We chose to apply the method to the study of the dynamics of the DNA dodecamer d(CTTTTGCAAAAG)$_2$ and its recognition by the minor-groove binding ligand Hoechst 33258 described above (Harris *et al.* 2001). The reasons for choosing this system included the volume of dynamics data already available, and the range of static, dynamic and thermodynamic parameters that had been obtained and in many cases correlated with experimental data. These original simulations were performed using the AMBER6 (Case *et al.* 1999) suite of MD simulation programs and the associated AMBER98 force field (Cornell *et al.* 1995), which has been extensively validated for the simulation of both DNA and protein systems (Cheatham *et al.* 1999). Indeed, it is *the* preferred potential parametrization for describing Watson–Crick hydrogen bonding in DNA. The simulations were performed on a single-processor machine (SGI Origin 200). Here we describe the porting of the AMBER98 force field to LAMMPS, and the optimization of simulation parameters for the same DNA system. The simulations were run on a massively parallel machine (Cray T3E). An approach based on principal component analysis (PCA) (Amadei *et al.* 1993; Cubero *et al.* 1999; Sherer *et al.* 1999; Wlodek *et al.* 1997) has been used as a sensitive test to ascertain the extent to which not only the energetics, but also the dynamics, of this system are consistent between the original AMBER and new LAMMPS simulations. As we shall show in the present paper, through application of the benchmarking methods and careful adjustment of simulation parameters the dynamical behaviour (and derived thermodynamic parameters) of a DNA system in LAMMPS simulations can be brought into good agreement with that obtained using AMBER, with the important advantage of greatly improved computational efficiency on tightly coupled massively parallel processors.

2. Methods

(*a*) *Simulations*

All simulations in this study were run on up to 64 processors of an 816-processor Cray T3E-1200E supercomputer. Full details of the AMBER simulation protocols have been given elsewhere (Harris *et al.* 2001). LAMMPS MD simulations were conducted using exactly the same periodically solvated DNA system. Briefly, this consists of the DNA dodecamer d(CTTTTGCAAAAG)$_2$, 22 sodium counterions to establish electrical neutrality, and 1748 TIP3P model (transferable intermolecular potential, three-point) water molecules (a total of 6028 atoms). The initial configuration of the system was taken from the NMR data (Gavathiotis *et al.* 2000) and energy minimized using AMBER. Periodic boundary conditions were applied (initial box dimensions *ca.* 32 Å × 36 Å × 52 Å) to a canonical (NVT) ensemble. Electrostatic interactions were calculated by the particle–particle, particle–mesh (PPPM) (Hockney & Eastwood 1988) method using a grid order of 5 and a cut off of 9 Å on the direct sum. A Lennard-Jones cut-off of 9 Å was used for non-bonded interactions. A dielectric constant of 1 was used, atomic positions were dumped every picosecond and the neighbour list was updated every 15 steps. AMBER simulations typically use the holonomic constraints algorithm SHAKE (Ryckaert *et al.* 1977) to constrain all bonds, permitting a 2 fs integration time-step. SHAKE is also implemented in LAMMPS, but

as an alternative it also permits multiple time-stepping (rRespa, or reversible reference system propagation algorithm (Plimpton *et al.* 1997; Tuckerman *et al.* 1991)) in which computationally expensive terms (in particular, the non-bonded interactions) are evaluated less often than those requiring a small time-step (the bonded interactions). Strictly speaking, the TIP3P water model is designed for use with Shake, not rRespa. However, we have published a study of non-rigid TIP3P which establishes that, if anything, its properties are better than those of standard (Boek *et al.* 1996). Unlike rRespa, the use of Shake is generally observed to be an impediment to efficient simulations on parallel architectures, so to test this we investigated the use of Lammps with both Shake and rRespa. Before the 'production' phase of the simulations, all systems were thermalized and equilibrated. The requirements for this process vary according to the quality of the initially constructed configuration, and details of the temperature coupling method. While thermal equilibration typically takes 100 ps or so, in our experience conformational equilibrium may take up to 1 ns to be established for such DNA systems, and can only be confirmed retrospectively through techniques such as PCA (see below). The protocol for Lammps MD studies using Shake consisted of a 10 ps run in which the temperature of the system was raised from 0 to 300 K, then 1 ns of dynamics at $T = 300$ K to ensure equilibration. All 2 ns production runs started with the configuration and velocities from this 1 ns checkpoint. rRespa simulations required a slower warming run to avoid problems with neighbour list updates: 200 ps warming from 0 to 100 K, 200 ps warming from 100 to 200K, 200 ps warming from 200 to 300 K, then 400 ps equilibration at $T = 300$ K. Again, all 2 ns production runs started with the coordinates and velocities from this 1 ns time-point.

(*b*) *Analysis methods*

Atomic Cartesian coordinate fluctuations and time-averaged structures were calculated using the Amber utility program *ptraj*. Visualization of trajectories was performed using the graphics software package VMD (Humphrey *et al.* 1996). PCAs were performed according to the methods previously described using in-house programs (Cubero *et al.* 1999; Sherer *et al.* 1999). The comparison of trajectories via the calculation of eigenvector overlaps was performed according to Hess (2000). Individual snapshots from the Lammps simulations, stripped of solvent and ions, were input to the Amber program *sander* in order to calculate energies with the Generalized Born/Surface Area implicit solvation model (Hawkins *et al.* 1996; Weiser *et al.* 1999), as previously described for the Amber simulations (Harris *et al.* 2001). Configurational entropies were also calculated from the mass-weighted eigenvectors from the PCA (above) according to the method of Schlitter (1993) as previously described (Harris *et al.* 2001). Entropies calculated by this method are sensitive to the length of the MD simulations: in essence, longer simulations tend to lead to a fuller exploration of conformational space by the molecule and so to a higher calculated configurational entropy. However, we find that the dependence of the calculated entropy $S(t)$ on the simulation length t can be fitted very well to a function of the form

$$S(t) = S_\infty - \frac{A}{t^n}.$$

The rationale for a function of this form comes from the analysis of a quite general time-series model with a stationary temporal covariance structure (A. Dryden 2003,

personal communication). According to this model, if the configuration at each point depends solely on the previous configuration (a Markov model), the exponent n is 1. Longer-range dependencies reduce n. For these studies, and the previous studies with which these results are compared, we found a value for n of 0.67 was appropriate. From the fitting procedure the entropy for a simulation of infinite length, S_∞ (in addition to the other parameter of the fit, A), can be estimated.

3. Results and discussion

(*a*) *Porting of the force field*

Since the internal architecture of the LAMMPS code is quite different from that of AMBER, the first test was to ensure that, given the same force field and the same configuration of the macromolecule, LAMMPS and AMBER calculated the same molecular mechanics energy for the system. Single-point energy calculations performed using AMBER and LAMMPS on configurations of the solvated DNA system showed excellent agreement (table 1). Bond, angle and dihedral energies are identical, while there are very small differences in the calculated van der Waals and electrostatic energies. However, since these are sums over a very large number of individual interactions, such differences were not regarded as significant. While such an analysis confirms that LAMMPS and AMBER calculate essentially identical energies given identical input structures, it does not show that the calculated forces are identical. The analysis of MD simulations is obviously the most sensitive and practically important way of testing this.

(*b*) *Analysis of simulations on the DNA alone*

Extended simulations were carried out using LAMMPS on the free (unliganded) duplex to compare the results obtained using rRESPA and SHAKE with those previously obtained using AMBER (with SHAKE). Each simulation was run on 64 processors of the Cray T3E and analysis was carried out on 2 ns equilibrated portions of the trajectories.

The similarity between the results obtained in each case and the benchmark AMBER simulation was assessed using four measures. The first was the root-mean-square Cartesian coordinate deviation (RMSD) between the time-averaged structure obtained using LAMMPS and that from AMBER. This gives a static structural measure of similarity. The second was a comparison of the average energy of the solute in each LAMMPS simulation compared with that obtained using AMBER. For this, each snapshot from the 2 ns trajectories was stripped of its waters and counterions, and used as the input for a single-point energy calculation in AMBER, using the Generalized Born with Surface Area correction (GB/SA) implicit solvation model to provide a solvation term. The reason for doing this, rather than directly comparing energy terms as described above, was that it permitted a direct comparison with previous published AMBER data for this DNA sequence. Thirdly, the configurational entropy of each LAMMPS trajectory was calculated, and compared with that obtained using AMBER. This gives a general but very sensitive measure of how the dynamics of the systems compare, and is particularly important if the ultimate desire is to calculate thermodynamic quantities from such simulations (see below). Fourthly, the dynamics of each simulation was investigated in more detail through the comparison

Table 1. *Static energy analysis, using both* Amber*6 and* Lammps*, of a representative structure of the DNA duplex* d(CTTTTGCAAAAG)$_2$

energy type	Amber6 (kcal mol^{-1})	Lammps (kcal mol^{-1})
bond	0.0239	0.0239
angle	399.8833	399.8833
dihedral	438.7989	438.7989
total van der Waals	2 532.0560	2 532.2143
total electrostatic	−27 167.3825	−27 167.0726
total energy	−23 796.6204	−23 796.1522

Table 2. *Benchmark comparisons for* Lammps *MD simulations of* d(CTTTTGCAAAAG)$_2$ *against previous* Amber*6 results*

system	RMSD (Å)	average energy (bond energy removed) (kcal mol^{-1})	T^* entropy (TS) (kcal mol^{-1})	PCA overlap
Amber6	n/a	−4397.95	693.67	n/a
Lammps/rRespa	0.58	−4443.38	727.13	0.6315
Lammps/Shake	0.59	−4435.32	732.07	0.6796

of principal components. PCA of a trajectory produces a description of the essential quasi-harmonic modes of deformation of the structure (eigenvectors). Motion along each eigenvector describes the way in which the structure samples configurational space in terms of a set of orthogonal modes. If two separate simulations are dynamically equivalent, they must sample this space in the same way. This can be achieved if they have the same eigenvectors, or they are linear combinations of each other. The first 10 eigenvectors from all simulations capture the majority of the important modes of flexibility of the system and so a comparison of these is deemed sufficient. To quantify this comparison, we used the eigenvector overlap measure of Hess (2000), calculated as the normalized sum of the dot products between each eigenvector from the Lammps simulation with each eigenvector from the reference Amber simulation.

The results obtained are shown in table 2. Clearly, using Lammps with either Shake or rRespa, the time-averaged structure of the DNA is very similar to that obtained using Amber (see figure 1). In our experience RMSD values of this magnitude (less than 0.6 Å) are frequently observed between independent simulations of the same system using the same MD code. However, the average energy (recalculated using the GB/SA method in Amber) of the two Lammps simulations is *ca.* 65 kcal mol^{-1} more negative than that of the reference simulation, and their entropy (TS at 300 K) is *ca.* 36 kcal mol^{-1} more positive. The PCA overlap values are good: we typically observe that separate independent simulations of the same system using the same MD code lead to trajectories with PCA overlaps of *ca.* 0.7, so the Lammps simulation run using Shake may be regarded as dynamically indistinguishable from the Amber simulation, while the rRespa simulation is only slightly poorer.

Other Lammps simulations (results not shown) confirmed the trend for lower average energies compared with Amber6. To identify the source of this we performed an

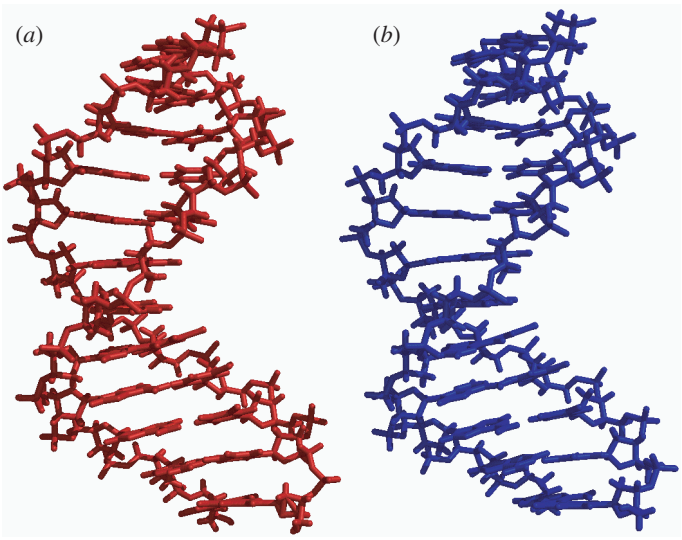

Figure 1. Comparison of the time averaged structures from 2 ns of MD on the dodecamer d(CTTTTGCAAAAG)$_2$ performed using (*a*) AMBER6 and (*b*) LAMMPS with rRESPA. The RMSD between the two structures is 0.58 Å.

Table 3. *Energy breakdown of* LAMMPS *simulations* (rRESPA *and* SHAKE) *compared with previous* AMBER *results*

(All values are in kcal mol^{-1} and are averages over 1000 snapshots taken from the MD simulations.)

system	bond	angle	dihed	1-4 VDW	1-4 EEL	VDW	EEL	EGB
AMBER	0.28	393.29	439.73	212.79	−1748.53	−407.63	−705.20	−2582.68
rRESPA	277.11	405.49	436.70	201.41	−1810.25	−406.83	−646.28	−2621.25
SHAKE	190.46	409.45	437.70	202.32	−1806.33	−408.24	−647.44	−2622.77

energy decomposition analysis (table 3). The bond energies differ markedly, but this is just due to the use or otherwise of SHAKE in the respective simulations, and can be discounted. The table shows that LAMMPS-generated structures have more negative 1-4 electrostatic energy (1-4 EEL) and solvation (EGB) terms, though this is partly countered by a less favourable general electrostatic energy (EEL) term. However, a similar analysis of the drug–DNA complex simulations (discussed below) revealed that though the overall energy difference of 40–50 kcal mol^{-1} between AMBER6 and LAMMPS simulations was still evident, this resulted from different balances between the same three key terms (1-4 EEL, EEL and EGB, results not shown). The one constant feature is a *ca.* 60 kcal mol^{-1} more negative value for the 1-4 electrostatic energy term in LAMMPS simulations. Since we have already shown that, given identical conformations of the DNA, both AMBER and LAMMPS calculate identical values for dihedral energy terms, the source of this discrepancy still remains to be determined. One possibility is that in the original AMBER6 simulations SHAKE was used to constrain all bond lengths, while in LAMMPS, to permit efficient parallelization, only bonds to hydrogen atoms are constrained. It could be that these constraints

Table 4. *Thermodynamic analysis of* LAMMPS *simulations compared with previous* AMBER*6 results*

(All values are averages over 1000 snapshots from each simulation and are in kcal mol^{-1}. E, energies recalculated using AMBER with the Generalized Born solvation model; TS_∞, Schlitter entropies ($T = 300$ K) extrapolated to infinite simulation time (see text). (*a*) AMBER6 results; (*b*) LAMMPS/rRESPA results; (*c*) LAMMPS/SHAKE results.)

	system	E	ΔE	$\Delta\Delta E$	TS_∞	ΔTS_∞	$\Delta\Delta TS_\infty$
(*a*)	free DNA	−4397.95			827.06		
			−28.70			28.43	
	1:1 complex	−4426.65		4.39	855.49		9.56
			−24.31			37.99	
	2:1 complex	−4450.96			893.48		
(*b*)	free DNA	−4443.38			873.39		
			−30.02			23.94	
	1:1 complex	−4473.40		−0.12	897.33		10.58
			−30.14			34.52	
	2:1 complex	−4503.54			931.85		
(*c*)	free DNA	−4435.32			875.80		
			−28.99			32.64	
	1:1 complex	−4464.31		2.57	908.44		7.02
			−26.42			39.66	
	2:1 complex	−4490.73			948.10		

disallow conformational adjustments that could relax unfavourable 1-4 interactions. However, since we find that LAMMPS gives very similar results using SHAKE and rRESPA (where no bonds are constrained), this seems unlikely.

(*c*) *Analysis of simulations on drug–DNA complexes*

Though we have clearly identified protocols for LAMMPS simulations that bring derived thermodynamic parameters into reasonable agreement with those calculated using AMBER, it is more important for most purposes that differences in enthalpies and entropies between systems or states are the same, irrespective of the MD 'engine' used to generate the configurations. To test this we performed additional simulations using LAMMPS on the 1:1 and 2:1 complexes between this DNA duplex and the minor groove binder Hoechst 33258. These simulations were performed analogously to the rRESPA and SHAKE simulations on the free DNA. As before, the trajectories were post-processed to calculate enthalpic plus solvation terms using the GB/SA method and configurational entropies calculated by the Schlitter approach. From these we were able to estimate, as before (Harris *et al.* 2001), the difference in free energy change between the first and second binding event ($\Delta\Delta G$), which has been determined by NMR (Gavathiotis *et al.* 2000) to be at least -4.0 kcal mol^{-1}. The agreement with the previous AMBER6 results is shown in table 4. AMBER6 simulations predicted that the binding is enthalpically anti-cooperative by *ca.* 4.4 kcal mol^{-1}. Using LAMMPS with SHAKE, binding is also predicted to be enthalpically anti-cooperative, but only by *ca.* 2.6 kcal mol^{-1}. But

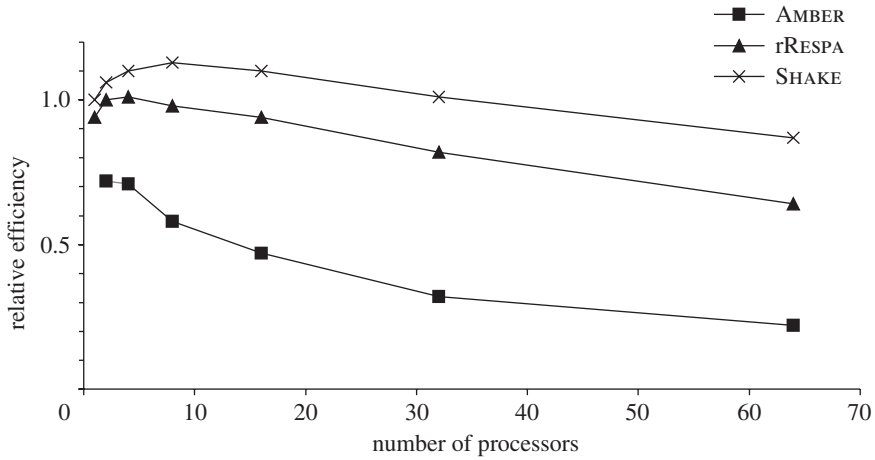

Figure 2. Relative efficiencies of simulations performed with AMBER and LAMMPS (the latter using rRESPA and SHAKE for time-stepping) run using different numbers of processors. All are scaled to the performance of LAMMPS with SHAKE on a single processor.

using LAMMPS with rRESPA, the simulations lead to the prediction that binding is essentially uncooperative ($\Delta\Delta E$ -0.1 kcal mol^{-1}). AMBER6 simulations predicted a strongly cooperative entropic term ($T\Delta\Delta S$ 9.6 kcal mol^{-1}) and the LAMMPS results show this too, bracketing the benchmark figure ($T\Delta\Delta S$ 1 kcal mol^{-1} more positive in the rRESPA simulation, 2.5 kcal mol^{-1} less positive in the SHAKE simulation). Note that, as before, configurational entropies from LAMMPS simulations tend to be greater than the AMBER counterparts. The fitting procedure used to estimate S_∞ also produces values for A (see above) that are *ca.* 10% higher, but the physical explanation of this difference is unclear. Taken together then, AMBER simulations predict an overall $\Delta\Delta G$ of -5.2 kcal mol^{-1}, while LAMMPS predicts $\Delta\Delta G$ to be -10.7 kcal mol^{-1} (using rRESPA) or -4.5 kcal mol^{-1} (using SHAKE). Overall we see that, as expected, while absolute energies and entropies from the different simulations can differ somewhat, differences, and double differences, are well reproduced. While it is clear that quantitatively the results are slightly different from the benchmark (more so for the rRESPA simulations than the SHAKE ones), the important qualitative conclusions from this study are the same: that the observed highly cooperative binding of Hoechst 33258 to this DNA dodecamer is entropy, not enthalpy, driven.

(*d*) *Analysis of computational efficiency*

The above analysis confirms that LAMMPS is reliable for such simulations, but is it useful? For this it must show a significantly improved performance on massively parallel computers, so that simulation problems that are difficult or impossible to handle using conventional codes can now be tackled routinely. To examine this we performed a series of simulations using different numbers of processors. In an ideal case, doubling the number of processors used for a calculation would halve the time required for solution. The results we obtained are shown in figure 2. The data are presented in terms of relative efficiencies, i.e. the time taken to complete the calculation compared with the theoretical (ideal) time. Perfect behaviour would therefore be represented by a line parallel to the x-axis. The data are also scaled, relative to the

time taken to complete the calculation on a single processor of the Cray T3E, using the LAMMPS code with SHAKE. Firstly, we observe that, independent of the number of processors used for the calculation, LAMMPS performs better than AMBER, and that LAMMPS using SHAKE performs better than LAMMPS using rRESPA. Secondly, we observe that, while AMBER performance falls off rapidly as the number of processors used increases, for LAMMPS performance actually improves for small increases (up to 4 or 8) in the number of processors used, i.e. using two processors instead of one, the time taken to complete the simulation more than halves. Above eight processors, LAMMPS performance begins to fall off somewhat, but always remains superior to that of AMBER. To put these numbers in perspective, a job which took 24 h to run on one processor using AMBER6 would be expected to take just 0.5 h on 32 processors using LAMMPS with SHAKE.

4. Summary

The AMBER98 force field, 'native' to the MD code AMBER, which is widely used and respected for MD studies of both DNA and proteins, has been successfully implemented in LAMMPS. The implementation has been validated, at least as far as the simulation of nucleic acid systems is concerned, by a careful analysis of MD data generated from a DNA system that has been the subject of detailed previous study using AMBER itself. To perform the validation we have had to consider carefully what constitutes similarity between two MD simulations and have settled on the following benchmarks which we consider testing, though not comprehensive. Firstly, and most trivially, calculations of static energies for snapshots of the system must be in agreement between the two MD codes. Secondly, over well-equilibrated portions of trajectories, averages of energies and energy components must be in agreement. Thirdly, the dynamical behaviour of the systems must be in agreement, and the measurement of PCA overlaps provides a useful method of checking this. Fourthly, non-enthalpic terms calculable from the MD ensembles must be in agreement, and the determination of configurational entropies provides this test.

With optimized simulation parameters, LAMMPS passes all these tests when compared with AMBER, and offers much improved performance in massively parallel environments. While AMBER and LAMMPS simulations can produce slightly different absolute values for enthalpic and entropic quantities, it is usually the case that enthalpy and entropy *differences* are the observables, and these are well reproduced. The validation of LAMMPS opens up significant new horizons for high-performance biomolecular computing. Order-of-magnitude increases in the sizes of problems that may be addressed—in terms of numbers of atoms per simulation, where scalability will be even more dramatic—are also now within reach. Order of magnitude increases in simulation time-scales will make new problems amenable to analysis through atomistic MD simulations. Conversely, for problems where current sizes and MD time-scales suffice, the time-to-solution may be reduced tenfold or more.

A final comment may be made concerning the more ambitious aims of a *computational systems biology* in which one tries to link molecular biological to cell biological and higher descriptions of living matter. To realize such an important dream, it will be necessary to find computationally efficient methods for bridging length- and time-scales from the smallest, electronic and molecular scales to the longer mesoscopic and

macrosopic length- and time-scales representative of cellular and intercellular dynamics. There is little doubt that any such approach will require the use of *multi-scale* modelling methods, in which descriptions at one level are fed in to those on higher (or indeed lower) levels. In principle, this may be done in either a hierarchical or a hybrid manner, the former by transferring information in an atemporal fashion between the levels concerned, the latter by the dynamical exchange of information when the physics of processes taking place on the different length and time-scales actually interacts. Such multi-scale modelling and simulation schemes have barely begun to be considered today in the life sciences, but they will surely require the use of high-performance computing methods, the exploitation of scalable codes of the kind considered here, and the probable deployment of computational grids (see, for example, http://www.realitygrid.org) to guarantee their ultimate reliability and efficiency.

We are grateful to Stephen Plimpton (Sandia National Laboratories, Albuquerque, NM, USA) for help with LAMMPS and useful discussions, and to Shantenu Jha for his comments on the manuscript. We thank the BBSRC initially, for a Class 3 grant, and, subsequently, the EPSRC, for a Class I grant (RealityGrid GR/R67699), both of which provided us with access to the Cray T3E-1200 at the UK National Supercomputing Service CSAR (Manchester, UK) on which many of our simulations were performed. We are also indebted to HEFCE for funding our SGI Onyx2 machine at University College London.

References

Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. 1993 *Proteins* **17**, 412–425.

Beveridge, D. L. & DiCapua, F. M. 1989 *A. Rev. Biophys. Biophys. Chem.* **18**, 431–492.

Boek, E. S., Coveney, P. V., Williams, S. & Bains, A. 1996 *Mol. Simulat.* **18**, 145–154.

Bostock-Smith, C. E., Harris, S. A., Laughton, C. A. & Searle, M. S. 2001 *Nucleic Acids Res.* **29**, 693–702.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. 1983 *J. Comput. Chem.* **4**, 187–217.

Case, D. A. (and 20 others) 1999 AMBER6. University of California, San Francisco, CA, USA.

Cheatham III, T. E., Cieplak, P. & Kollman, P. A. 1999 *J. Biomol. Struct. Dynam.* **16**, 845–862.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. 1995 *J. Am. Chem. Soc.* **117**, 5179–5197.

Cubero, E., Sherer, E. C., Luque, F. J., Orozco, M. & Laughton, C. A. 1999 *J. Am. Chem. Soc.* **121**, 8653–8654.

Delgado-Buscalioni, R. & Coveney, P. V. 2003 *Phys. Rev.* E **67**, 046704.

Duan, Y. & Kollman, P. A. 1998 *Science* **282**, 740–744.

Duan, Y. & Kollman, P. A. 2001 *IBM Syst. J.* **40**, 297–309.

Fersht, A. R. & Daggett, V. 2002 *Cell* **108**, 573–582.

Gavathiotis, E., Sharman, G. J. & Searle, M. S. 2000 *Nucleic Acids Res.* **28**, 728–735.

Haq, I., Ladbury, J. E., Chowdhry, B. Z., Jenkins, T. C. & Chaires, J. B. 1999 *J. Mol. Biol.* **271**, 244–257.

Harris, S. A., Gavathiotis, E., Searle, M. S., Orozco, M. & Laughton, C. A. 2001 *J. Am. Chem. Soc.* **123**, 12658–12663.

Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. 1996 *J. Phys. Chem.* **100**, 19 824–19 839.

Hess, B. 2000 *Phys. Rev.* E **62**, 8438–8448.

Hockney, R. W. & Eastwood, J. W. 1988 *Computer simulations using particles*. New York: Adam Hilger.

Humphrey, W., Dalke, A. & Schulten, K. 1996 *J. Mol. Graphics* **4.1**, 33–38.

Karplus, M. & McCammon, J. A. 2002 *Nat. Struct. Biol.* **9**, 646–652.

Kollman, P. A. 1993 *Chem. Rev.* **93**, 2395–2417.

Kollman, P. A. (and 14 others) 2000 *Acc. Chem. Res.* **33**, 889–897.

Plimpton, S. J. & Hendrickson, B. 1996 *J. Comput. Chem.* **17**, 326–337.

Plimpton, S. J., Pollock, R. & Stevens, M. 1997 Particle-mesh Ewald and rRESPA for parallel molecular dynamics simulations. In *Proc. 8th SIAM Conf. on Parallel Processing for Scientific Computing, Minneapolis, MN, 14–17 March 1997*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. 1977 *J. Comput. Phys.* **23**, 327–341.

Schlitter, J. 1993 *Chem. Phys. Lett.* **215**, 617–621.

Sherer, E. C., Harris, S. A., Soliva, R., Orozco, M. & Laughton, C. A. 1999 *J. Am. Chem. Soc.* **121**, 5981–5991.

Tuckerman, M. E., Jerne, B. J. & Martyna, G. J. 1991 *J. Chem. Phys.* **94**, 6811–6815.

Varnai, P. & Lavery, R. 2002 *J. Am. Chem. Soc.* **124**, 7262–7263.

Watson, J. & Crick, F. 1953 *Nature* **171**, 737–738.

Wattis, J. A. D., Harris, S. A., Grindon, C. R. & Laughton, C. A. 2001 *Phys. Rev.* E **63**, 061903.

Weiser, J., Shenkin, P. S. & Still, W. C. 1999 *J. Comput. Chem.* **20**, 217–230.

Wlodek, S. T., Clark, T. W., Scott, L. R. & McCammon, J. A. 1997 *J. Am. Chem. Soc.* **119**, 9513–9522.

Wu, X., Wang, S. & Brooks, B. R. 2002 *J. Am. Chem. Soc.* **124**, 5282–5283.

Yakushevich, L. V. 1998 *Non-linear physics of DNA*. Wiley.