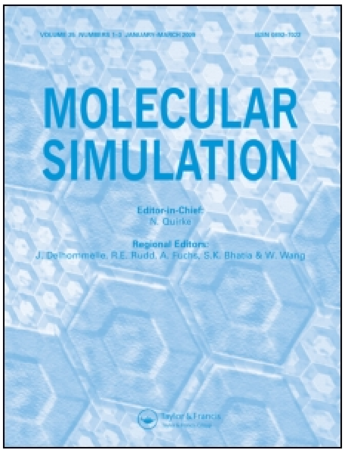


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Computer Simulation of Macromolecules

Julia M. Goodfellow[a]
[a] Department of Crystallography, Birkbeck College, London, UK

To cite this Article Goodfellow, Julia M.(1990) 'Computer Simulation of Macromolecules', Molecular Simulation, 5: 5, 277 — 291

To link to this Article: DOI: 10.1080/08927029008022414

URL: http://dx.doi.org/10.1080/08927029008022414

Molecular Simulation, 1990, Vol. 5, pp. 277–291
Reprints available directly from the publisher
Photocopying permitted by license only

COMPUTER SIMULATION OF MACROMOLECULES

JULIA M. GOODFELLOW

*Department of Crystallography, Birkbeck College, Malet Street,
London WC1E 7HX, UK*

(Received March 1990, accepted March 1990)

Computer simulation techniques are now an essential part of modern structural molecular biology. They are used in many different ways in order to study the conformation, dynamics and interactions of proteins and nucleic acids. In this paper, I shall review several of these applications and then focus on three specific areas, namely the conformation and dynamics of proteins including the use of free energy perturbation methods to study mutant proteins, the conformation and dynamics of DNA and DNA-drug complexes, and the use of computers with parallel architectures. Although simulation of molecules as large and complex as proteins and nucleic acids may be considered a grand challenge in itself, there are even greater challenges for the future.

KEY WORDS: Macromolecules, protein dynamics, parallel architectures.

INTRODUCTION

Computer simulation has been used for many years to study the dynamics of proteins, peptides and nucleotides [1]. Although initially these simulations were carried out *in vacuo*, the advent of supercomputers has led to an increase in the sophistication of the models and solvent is routinely included together with any counter-ions. These systems may be very large. For example, for even a small enzyme such as Ribonuclease A, there are over 1000 proteins atoms and 3000 water molecules.

More recently molecular dynamics algorithms are being used to refine X-ray crystallographic data and to obtain three dimensional structural information from 2D NOE NMR data. In these applications, molecular dynamics is being used at high temperatures to search conformational space given an experimental constraint. In the case of X-ray refinement an effective force is defined as the sum of the intermolecular forces and a suitably weighted X-ray force. The latter is a function of the observed and calculated structure factors. In a typical run, the protein is heated up to 3000° over 1 psec and then slowly cooled back to room temperature during another 1 psec. This simulated annealing procedure has been pioneered by Brunger [2, 3] and van Gunsteren [4, 5].

In structure determination in solution by 2D NMR techniques, the experimental restraints are the interproton NOEs. This technique has been used to define structures for small proteins [6] and nucleotides [7] in solution.

The modelling of proteins, whose sequence is known but whose structure is not, by homology with a known structure is becoming a very important technique. This is because the number of protein sequences is around 10,000 whereas there are only around 400 protein structures in the Brookhaven databank. In many cases of protein modelling, the initial protein structure is refined using energy minimization procedures [8] using the same potential energy functions as used in the molecular dynamics packages.

Another area in which computer simulation is important is that of DNA-drug complexes [9]. Although there are many oligonucleotide crystal structures, there are only a limited number of structures (whether from X-ray diffraction or 2D NMR) of drug-DNA complexes. For example, for the simple intercalator proflavine there is no structure for the complex with more than two DNA base-pairs. Many of these drugs show complex patterns of sequence specificity whose mechanism is not fully understood at the molecular level because of the absence of experimental structural data. In this case, computer simulation techniques can provide possible models which can always be validated by comparison with the available experimental data.

Perhaps one of the most exciting developments has been in the development of methods to estimate free energy differences [10]. This has been used to study the relative stability of mutant proteins [11] and the relative binding energy of ligands [12]. Although at present only relatively small changes can be studied, it has the potential to become an important technique in both protein engineering and drug design.

METHODS

There are several software packages which are routinely used in the simulation of macromolecules. These include AMBER (Kollman, UCSF), CHARMM (Karplus, Harvard) and GROMOS (van Gunsteren and Berendsen, Groningen). These contain inter- and intra-molecular force fields (including bond distance, bond angle, torsion angle, planarity and van der Waals', electrostatic and hydrogen bond terms). A distance cut-off of between 7.5 and 10.0 Å is usually applied in order to reduce the size of the calculation. Simulations can be carried out in NVT or NPT ensembles with or without periodic boundary conditions. Standard algorithms (e.g. Gear or Verlet) are used to solve Newton's laws of motion. Normally the SHAKE algorithm is applied which restrains certain bond distances in order that larger time steps can be employed. Typical time steps are around 1 to 2 fsec and simulations may run for up to 100 psec.

As well as using AMBER and GROMOS, we have also used Monte Carlo simulation algorithms including BOSS (Jorgensen, Purdue) and our own code BERNAL. These include algorithms for the calculation of free energy differences using the perturbation technique. We have used OPLS parameters and the TIP4P model of solvent in these calculations [13, 14].

Most of our calculations have been carried out on the CRAY XMP at ULCC. We have also used our own small in-house MEIKO computing surface with a total of 6 transputers for developing parallel algorithms as well as the Edinburgh concurrent computer project for testing and running on larger domains of up to 130 transputers.

RESULTS

(A) Simulation of Ribonuclease A in Solution

We have undertaken a molecular dynamics study of a small enzyme Ribonuclease A using the GROMOS package in order to study solvent both in the active site and on the surface of this macromolecule. This protein has been chosen because well-refined x-ray and neutron crystal structures are available [15] and it is relatively small for an enzyme, being only 13,500 Daltons. We have surrounded the protein molecule with over 3000 solvent molecules using truncated octahedra periodic boundary conditions. The crystal structure contains simple ions (sulphate and phosphate) in the active site but we have used a model complex with a more realistic cyclic cytosine nucleotide (referred to as CPP) generated by P. Thomas (Birkbeck College).

The initial optimization protocol led to only small changes (0.1 Å RMSD) in conformation compared to the crystal structure. A plot of energy against length of the simulation is shown in Figure 1. It can be seen that the energy is very flat from around 5 psec. We have also monitored the RMSD between the molecular dynamics structure and the optimized conformation (see Figure 2). It is apparent that changes in conformation are occurring between 40 and 50 psec. We chose to analyse the simulation between 20 and 40 psec while continuing with the dynamics run for a further 30 psec.

Although crystallography leads to an average static structure for a molecule, it can also provide information on the root mean square movement of atoms usually defined as a temperature factor or B value for macromolecules. We have compared the atomic motion for our simulation with the experimental values from the x-ray refinement (see Figure 3). The largest differences in conformation can be seen at the N and C terminal of the protein and in the region of residue 88. These three regions correspond to the parts of the protein structure with the largest deviation between the two independently refined structures [15]. Given that the simulation is in solution with no surrounding macromolecules, the overall agreement is very promising.

We have analysed the solvation of most charged and polar residues in this protein and find considerable agreement with experimental solvent distributions found by Thanki *et al.* [16]. In particular, we have looked at the solvation of two histidine residues in the active site namely His 119 and His 12. Histidine 12 forms a hydrogen

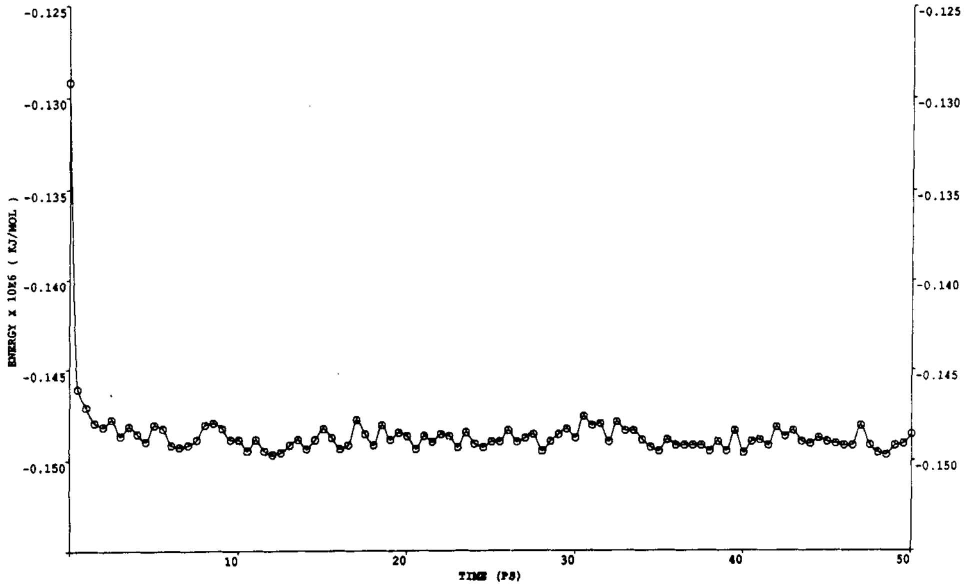


Figure 1 A plot of total energy versus length of simulation for Ribonuclease A in solution.

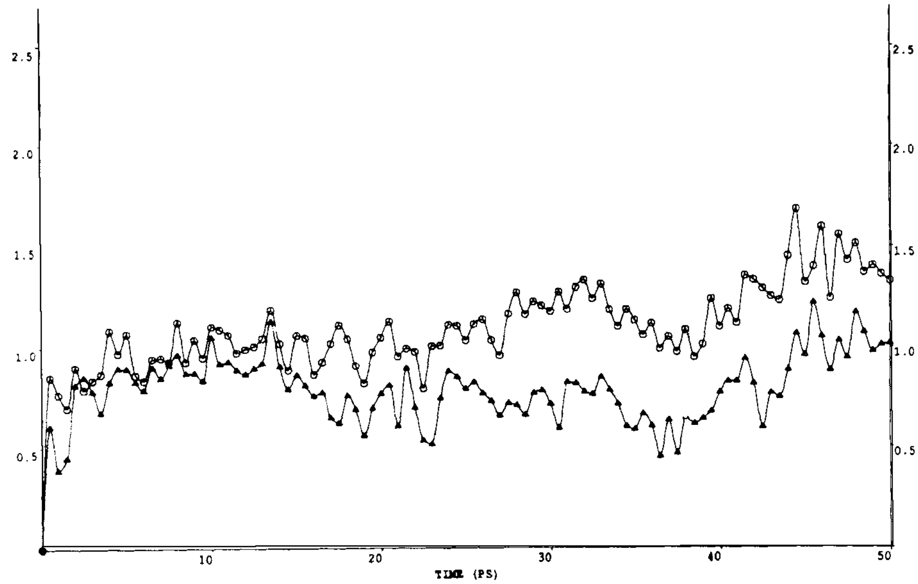


Figure 2 A plot of root mean square deviation (RMSD) between the optimized and simulated conformations versus the length of the simulation for Ribonuclease A in solution. The RMSD is calculated with active site residues superimposed. O-O-O represents the plot for superimposition of the inhibitor CCP and — — — represents the plot for the superimposition of CCP5 and two neighbouring side-chains.

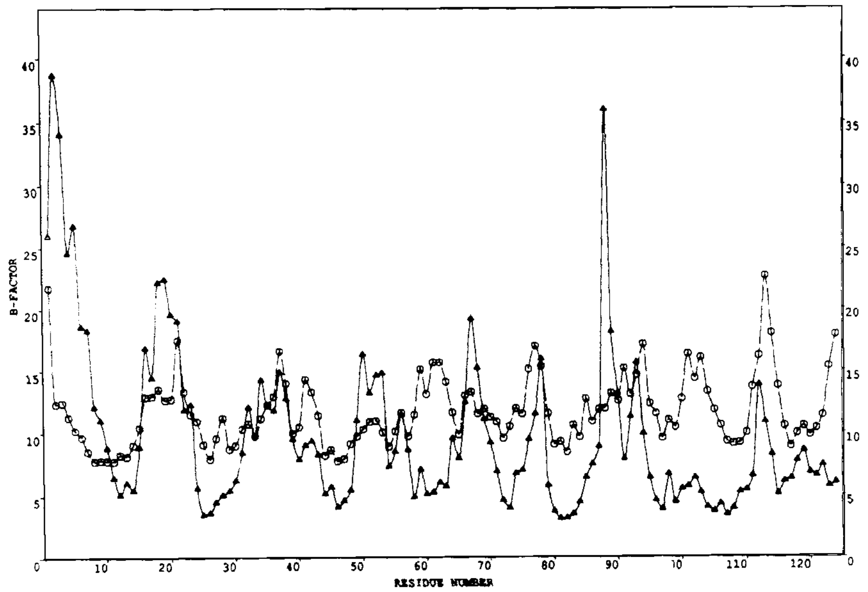


Figure 3 A plot of the crystallographic temperature B (which represents the root mean atomic fluctuations) for each residue. O-O-O is from the crystallographic refinement whereas the ▼—▼ is calculated from the simulation.

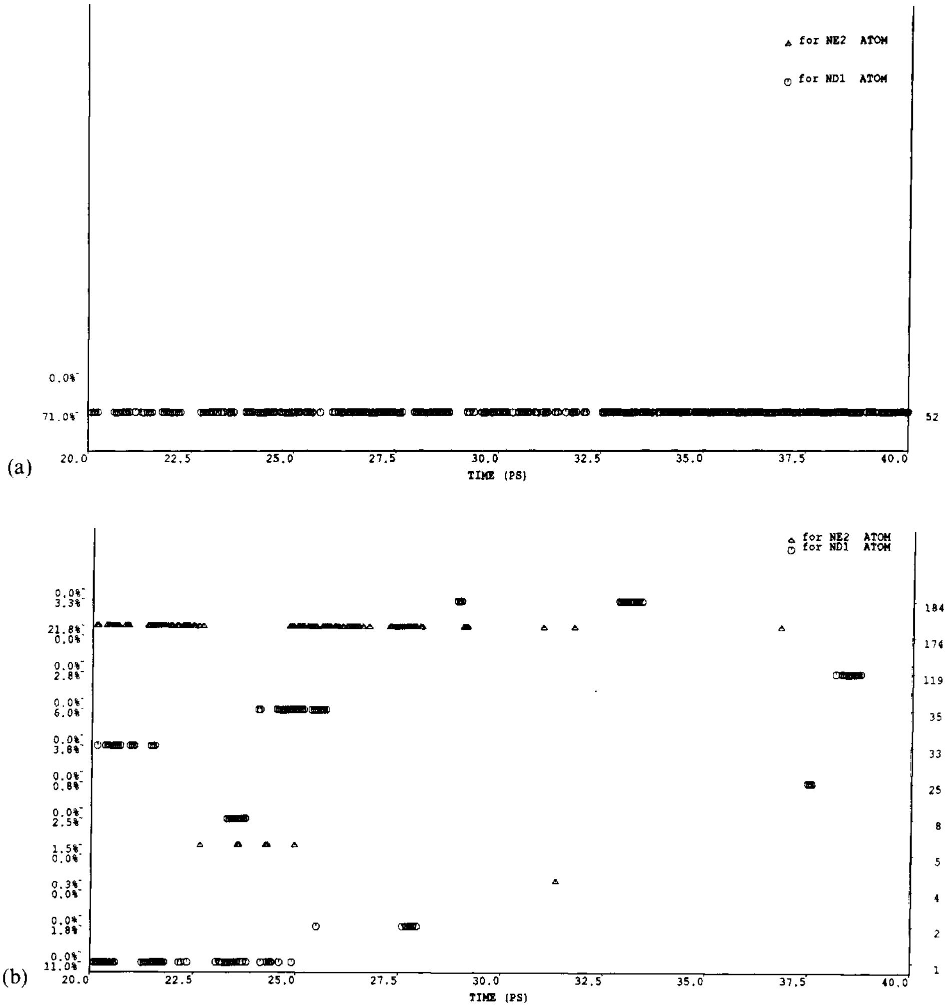


Figure 4 Plots of occupancy versus duration of the simulation for water molecules hydrogen bonding to (a) His 12 and (b) His 119.

bond between atom NE1 and O2′ of the sugar moeity of CCP. The other polar atom ND1 is hydrated by one water molecule (number 52) throughout the 20–40 psec of this analysis (Figure 4a). In contrast, His 119 which forms no direct hydrogen bonds to the substrate CCP shows a more complex pattern of hydration with up to 11 water molecules coming within hydrogen bonding distance of ND1 or NE2 atoms (Figure 4b). We are continuing with a more detailed analysis of both the conformation and dynamics of ribonuclease as well as its surrounding solvent.

(B) Drug-DNA Complexes

We have used the AMBER molecular dynamics package [17] to study the conforma-

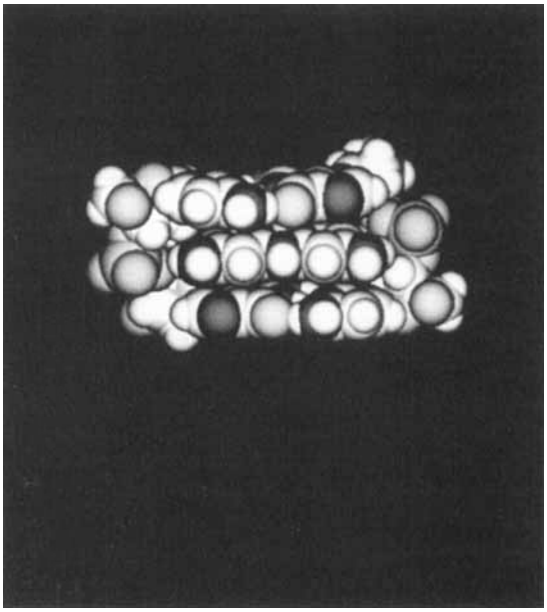


Figure 5 (*See* colour plate VIII). A diagram of dCpG-proflavine complex.

tion and dynamics of nucleotides and drug-nucleotide complexes. We started with simulations of crystals, using several unit cells, in order to compare the results from our simulations with experimental data. Two highly refined structures were chosen namely GpC nonahydrate and dCpG proflavine. The latter structure is illustrated in Figure 5.

The plots of energy against the length of the simulation and the root mean square deviation of the simulated to experimental structure are shown in Figure 6a and b. Our analysis of these simulations showed that the overall agreement between the simulated and experimental conformations is good. However, in both cases, small regions deviate from the crystal structure geometry. Restraining the solvent molecules to their experimental positions (but allowing the solute to move) leads to improved agreement between the simulated and experimental nucleotide conformation and relatively little change in the motion of the nucleotide atoms. For the dCpG-proflavine structure, the backbone torsion angles are very stable throughout the simulation with movement of between ± 6° and ± 11°. The least mobile torsion angles are related to sugar conformation whilst the most mobile is the torsion angle around the O5′–C5′ bond which is known to change on intercalation.

We have continued with these studies by investigating the larger nucleotide with the sequence d(CGATACGATACG) in solution both on its own and with the drug proflavine intercalated between the central CG base pairs (Figure 7 and Figure 8). In general, the dodecamer in solution shows greater flexibility than the dinucleotides with the crystal unit cells. Backbone torsion angles move between ± 10° and ± 15°. The phosphate atoms have larger root mean square movement than either the sugar or base atoms. This trend is consistent with results from X-ray crystallography on oligonucleotide structures. The sugar pucker is traditionally considered to be C3′-

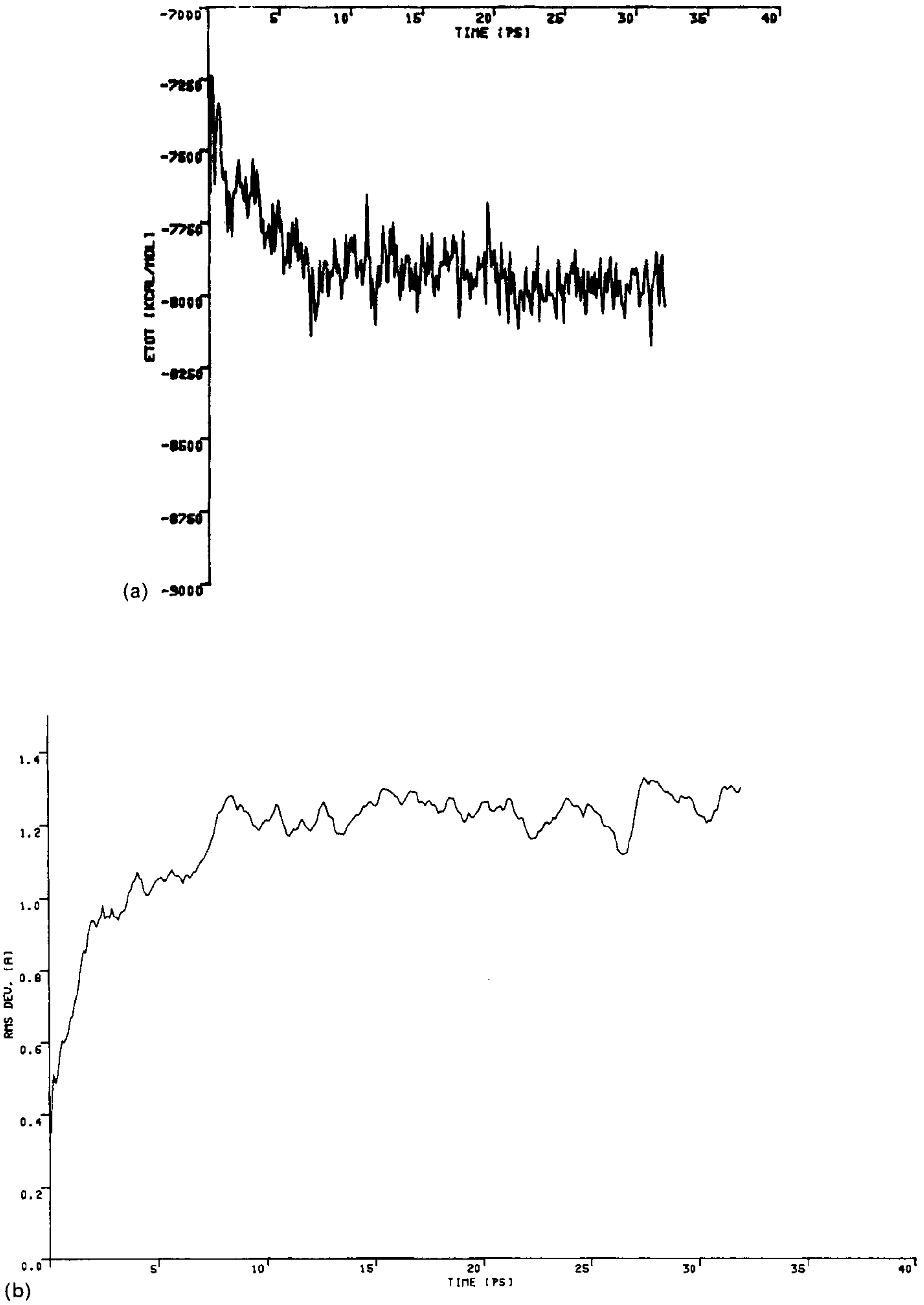


Figure 6 Plots of (a) total energy and (b) root mean square deviation in structure during the simulation of 4 unit cells of dCpG-proflavine.

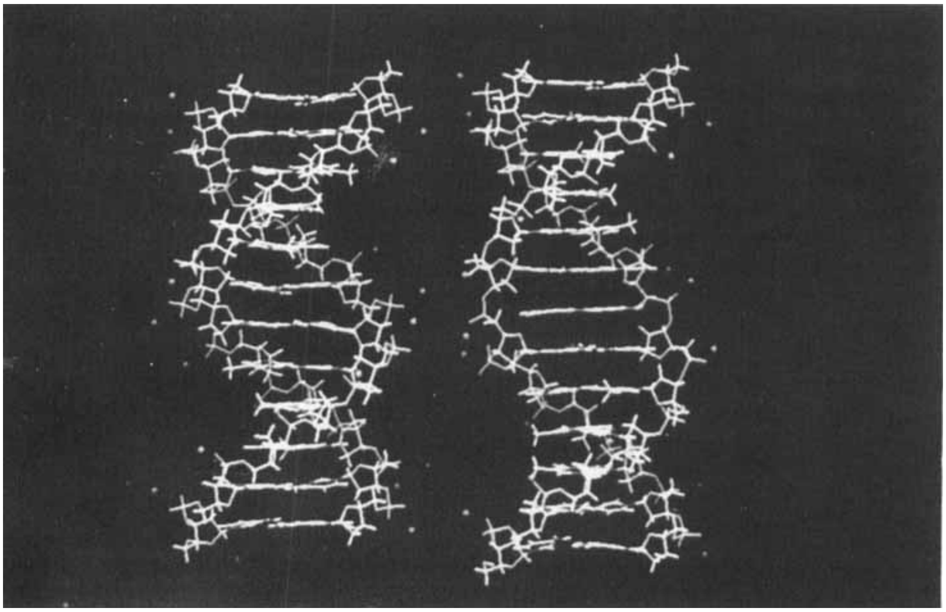


Figure 7 (*See* colour plate IX). A diagram of the dodecanucleotide dC etc.

endo and C2′-endo for 'A' and 'B' conformations respectively. However, the simulation shows that the sugar pucker is quite flexible with C2′-endo, C3′-endo and O1′ endo puckers occurring within an overall 'B'-like conformation. The other backbone torsion angles show considerable flexibility. Such trends are again consistent with local structural diversity which is being found more and more in X-ray crystal structures of oligonucleotides [18].

(C) Free Energy Estimates of Amino Acid Mutations

Modern genetic engineering has enabled us to make mutant proteins in which one or more amino acids are changed for those of another type. Often these changes are made to residues which are on the surface of the protein and thus one of the largest components to the change in free energy is that associated with the interaction between the solvent molecules and the amino acid side-chain. We have initiated a series of calculations which use the free energy perturbation method [10] in order to estimate the hydration related free energy changes as one amino acid is mutated to another within a polypeptide chain. The aims of these studies were two fold. First, we wished to study the reliability of such calculations and secondly we wished to look at the effect of different environments on the magnitude of the change in free energy.

We have considered the CH_3 to OH mutation both in the model system ethane to methanol and in a peptide chain where the side-chain valine is changed to that of threonine during the course of the simulation [11]. The results for these mutations are shown in Table 1. It can be seen that the total hydration related free energy change decreases from 8.3 kcal mol^{-1} for the ethane to methanol mutation to 7.05 kcal mol^{-1} for the valine to threonine mutation within the tripeptide (ala-X-ala) and decreases

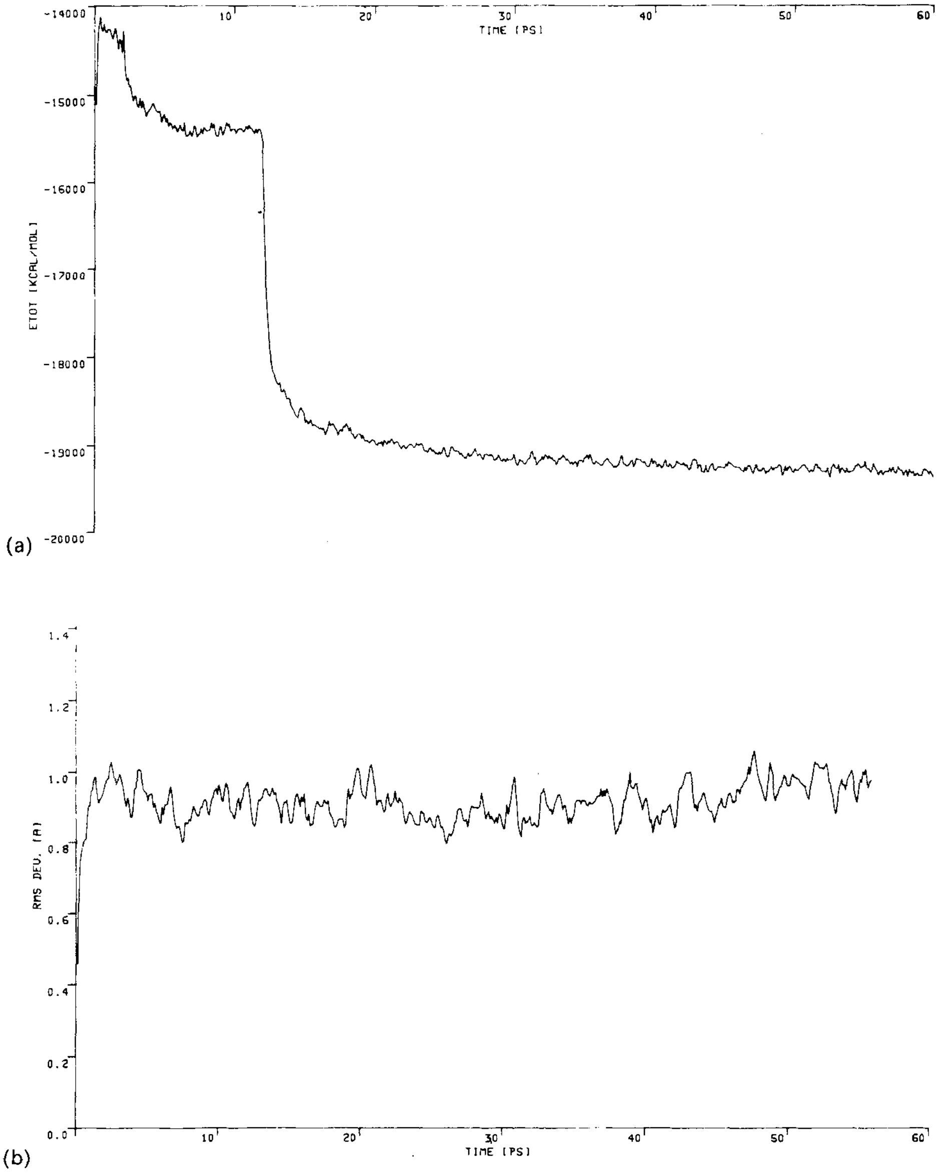


Figure 8 Plots of the total energy and root mean square deviation in structure during the simulation of the dodecanucleotide d(CGATACGATACG).

still further to 6.61 kcal mol^{-1} when we consider the valine to threonine mutation within a pentapeptide (ala-lys-X-lys-ala).

We have also considered the errors in such calculations. In general, only the difference in the forward and backward free energy change is considered. These were ± 0.3 kcal mol^{-1} and ± 0.18 kcal mol^{-1} for the ethane to methanol and valine to threonine changes respectively. However, we have employed a windowing technique

Table 1

System	Total free energy change (kcal mol^{-1}) Forward	Backward	Average
Ethane to methanol	8.68	−7.98	8.3 .32
Val to Thr I[a]	6.85	−7.20	7.05 .18
Val to Thr II[b]	6.79	−6.43	6.61 .18

(a) Valine to Threonine Mutation within the sequence Ala-X-Ala.
(b) Valine to Threonine Mutation within the sequence Ala-Lys-X-Lys-Ala.

Table 2

Interval	Free energy changes (kcal mol^{-1}) Forward	Backward
0. -0.125	–	−2.56 ± .14
0.125–0.25	1.85 ± .12	−1.37 ± .09
0.25 -0.5	1.76 ± .16	−1.81 ± .24
0.5 -0.75	0.36 ± .12	−1.12 ± .13
0.75 -1.0	0.34 ± .13	–

Table 3

Type	*Free energy change (kcal mol ^{1})* *Interval I*[a]	*Interval II*[b]
NVT1[c]	1.85 ± .12	−2.56 ± .14
NPT[d]	2.31 ± .07	−2.86 ± .08
NVT2[e]	1.56 ± .11	−2.27 ± .13

(a) Valine to threonine mutation within tripeptide for window 0 < → < 0.125.
(b) Valine to threonine mutation within tripeptide for window 0.125 < → < 0.25.
(c) NVT ensemble
(d) NPT ensemble
(e) NVT ensemble with different starting seed.

in which the average and standard deviation in the estimate of the free energy can be calculated for each window (Table 2). Typically, we have used five windows over 1,000,000 Monte Carlo steps. Analysis of the standard deviation per window, shows that the total errors for the forward and backward calculations (calculated as the square root of the sum of the variances) are 0.3 and 0.35 kcal mol^{-1} respectively. These estimates of the error are larger than that implied by comparison of the difference in total forward and backward free energies of 0.18 kcal mol^{-1}.

We have also found that the changes are dependent on the type of ensemble (NVT or NPT) and the starting seed (i.e. region of phase space) (Table 3) and also the size of the distance cut-off applied during the energy calculation. This leads us to believe that free energy protocols especially for complicated biological systems need to be optimized in order to give both precise and accurate results.

(D) Use of Computers with Parallel Architectures

We have been investigating the power of computers with parallel architectures to run simulation software. We started with a minimization algorithm – CARTE – which is

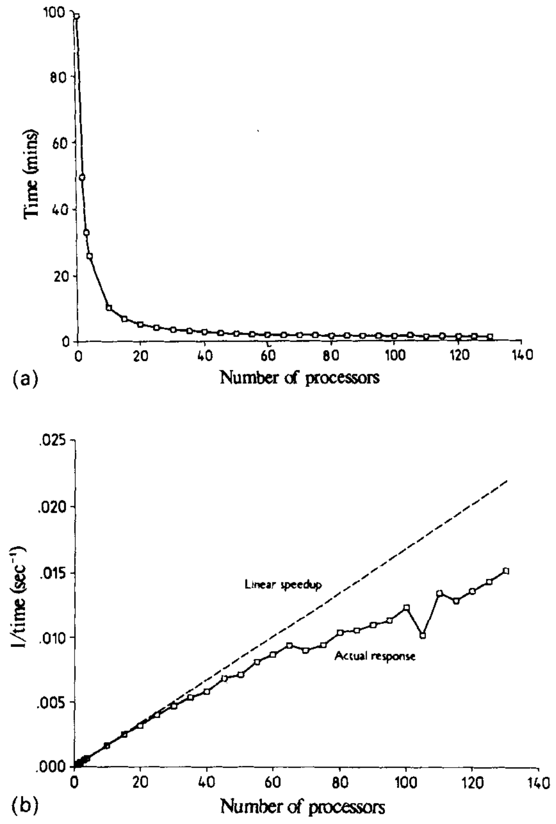


Figure 9 Plot of (a) time and (b) speed against number of transputers using a parallel version of the program CARTE.

used to study the solvation of polypeptides and nucleotides [19]. We have used a Fortran farm approach in which we have surrounded our Fortran code with an Occam harness in order that the code can run on any number of transputers [20–21]. We have developed this parallel code on our in-house Meiko Computing Surface and carried out further tests with up to 130 transputers at the Edinburgh Concurrent Supercomputer Project. Our results with CARTE are shown in Figure 9a and b.

We have also attempted to write parallel code for our in-house Monte Carlo program BERNAL. This has proved far more of a challenge. The main subroutine, E2 in Figure 10, involves the calculation of the potential energy of interaction between the randomly moved water molecule with all other atoms in the system under consideration. We have made several attempts to reduce the communications overheads which tend to reduce the efficiency of the program as can be seen in Figure 11. These strategies involve the change in data control between the master and slave processors during the energy evaluation subroutine E2 [22].

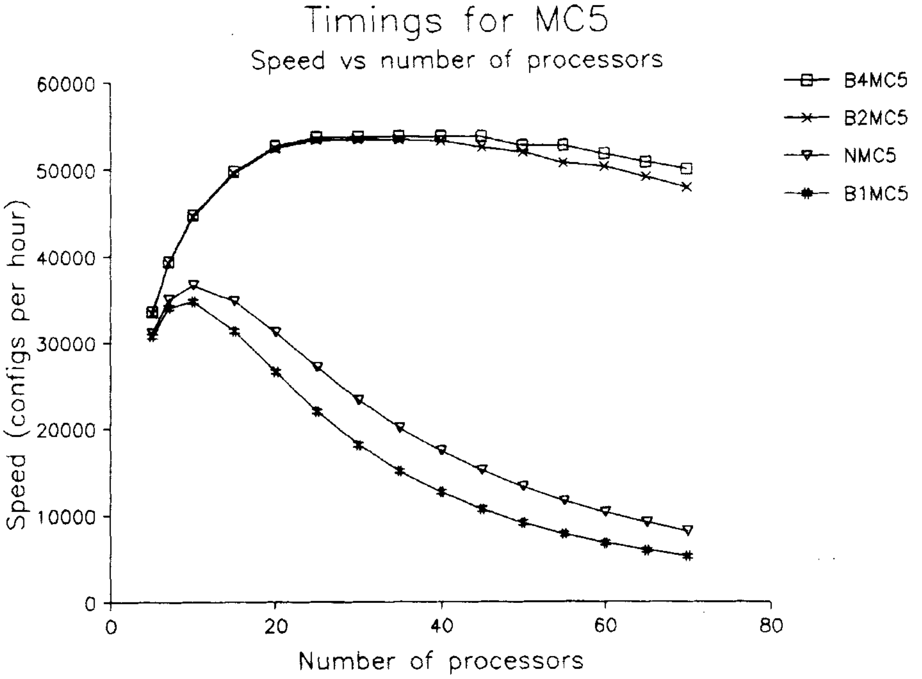
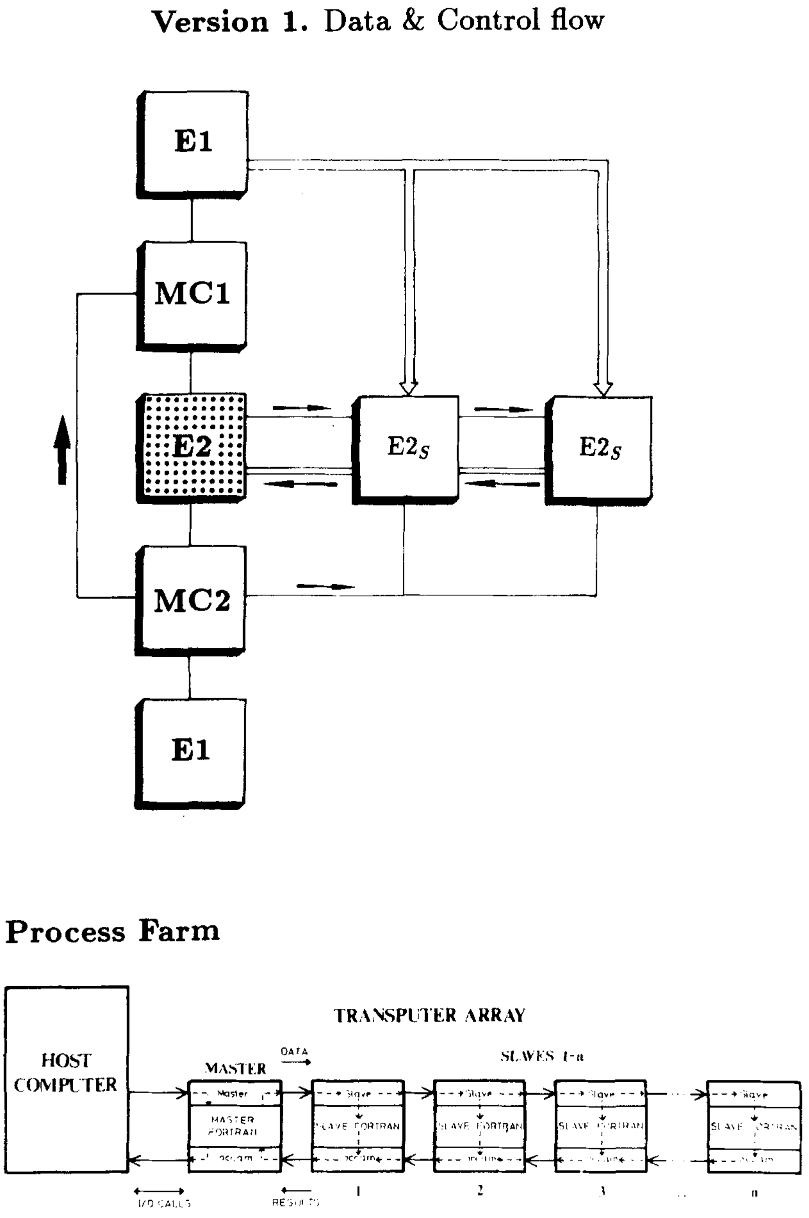


Figure 10 Plot of Speed versus number of transputers using parallel versions of the program BERNAL. NMC5 and B4MC5 are the results for the original code with minimum changes and the code with major changes to data control in routine E2. B1MC5 and B2MC5 are similar coding but with changes to the connections of the transputers.

DISCUSSION

It is clear that computer simulation of macromolecules can provide useful information on their conformation, dynamics and interactions. The use of simulation techniques is only likely to increase in the future as they are used more and more to study conformation and dynamics as well as being used in the determination of three dimensional X-ray and NMR structures. Although some of the systems described herein are relatively large and require considerable computer resources, they are actually rather small in comparison with average size enzymes or DNA molecules themselves. Moreover, one of the major features of interest for most macromolecules is not just the structure or dynamics of the molecule in isolation but its ability to interact with other molecules from small drugs to other macromolecules. In the future, we will want to study larger systems consisting of molecular complexes.

One of the main problems with simulation studies concerns the precision and accuracy of the calculations [23]. Accuracy can be assessed by comparison with experimental data such as the wealth of structural information from x-ray crystallography and 2D NMR. The two major parameters which affect the reliability of simulations are the interatomic potential functions and the sampling of phase space. In the future, we must develop and evaluate more sophisticated potentials than those currently employed in macromolecular simulations. We also need to continue to



Version 1. Data & Control flow

Process Farm

Figure 11 Diagram outlining the subroutines in the Monte Carlo program BERNAL and the Fortran Farm approach to parallelization.

assess whether the sampling is adequate presumably by comparing results from longer time periods and repeating calculations from different starting points.

The main factors contributing to the increased use of macromolecular simulation has arisen from both the availability of software packages and the increasing power of computers whether Cray XMPs in computer centres or the in-house workstations. For the future, we must also consider the power of machines with massively parallel architectures.

Acknowledgements

I would like to thank the Wellcome Trust for a University Research Leave Fellowship and the SERC for support under grants GR/E 72737, GR/F 81071, GR/E 51381, GR/D 86683, GR/E 47025 and GR/F 69581. I would also like to thank the people with whom I have worked on these projects including Drs S. Yoshioki, D.S. Moss, P. Herzyk, S. Neidle and M.A.S. Saqi, and Mr Doug Jones, Mr Mark Williams and Ms N. Thanki. I gratefully acknowledge support by the Edinburgh Concurrent Supercomputer Project.

References

[1] J.A. McCammon and S. Harvey in 'Dynamics of Proteins and Nucleic Acids', Cambridge University Press (1987).

[2] A.T. Brunger. 'Crystallographic Refinement by Simulated Annealing'. *J. Mol. Biol.*, **203,** 803–816 (1988).

[3] W.I. Weiss and A.T. Brunger. 'Crystallographic refinement by simulated annealing' in "Molecular Simulation and Protein Crystallography" (ed J.M. Goodfellow, K. Henrick and R. Hubbard) CCP4/CCP5 study weekend, SERC, Daresbury Laboratory, DL/SC1/R27 (1989).

[4] P. Gros, M. Fujinaga, A. Mattevi, F.M.D. Vellieux, W.F. van Gunsteren and W.G.J. Hol, 'Protein structure refinement by molecular dynamics techniques' in "Molecular Simulation and Protein Crystallography" (ed J.M. Goodfellow, K. Henrick and R. Hubbard) CCP4/CCP5 study weekend, SERC, Daresbury Laboratory, DL/Sc1/R27 (1989).

[5] W.F. van Gunsteren and H.J.C. Berendsen, BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, Groningen, The Netherlands (1987).

[6] G.M. Clore, A.M. Gronenborn, A.T. Brunger and M.J. Karplus, 'Solution Conformation of a Heptadecapeptide Comprising the DNA binding Helix F of the cyclic AMP Receptor Protein of E. Coli.' *J. Mol. Biol.*, **186,** 435–455 (1985).

[7] L. Nilsson, G.M. Clore, A.M. Gronenborn, A.T. Brunger and M.J. Karplus, 'Structure Refinement of Oligonucleotide by Molecular Dynamics with NOE Interproton Distance Restraints: Application to d(CGTACG)$_2$' *J. Mol. Biol.*, **188,** 455– (1986).

[8] A.M. Hemmings, S.I. Foundling, B.L. Sibanda, S.P. Wood, L.H. Pearl and T.L. Blundell, 'Energy Calculations on Aspartic Proteinases: Human Renin, Endothiapepsin and its complex with an Angiotensinogen Fragment Analogue, H142.', *Biochem. Soc. Trans.*, **13,** 1036–1041 (1985).

[9] S. Neidle, L.H. Pearl and J.V. Skelly, 'DNA structure and perturbation by drug binding', *Biochem. J.*, **243,** 1–13 (1987).

[10] T. Lybrand, J.A. McCammon and G. Wipff, 'Theoretical Calculations of Relative Binding Affinity in host-guest Systems', *Proc. Natl. Acad. Sci. USA*, **83,** 833–835 (1986).

[11] M.A.S. Saqi and J.M. Goodfellow, 'Free energy changes associated with amino acid substitution in Proteins', *Protein Engineering*, In Press (1990).

[12] D.A. Case, 'Dynamical Simulation of Rate Constants in Protein-Ligand Interactions', *Prog. Biophys. Molec. Biol.*, **52,** 39–70 (1988).

[13] W.L. Jorgensen and J. Tirado-Rives, 'The OPLS Potential Functions for Proteins', *J. Am. Chem. Soc.*, **110,** 1657–1670

[14] W.L. Jorgensen, J. Chandrasekhar, J. Madura, R.W. Impey and M.L. Klein, 'Comparison of Simple Potential Functions for Simulating Liquid Water', *J. Chem. Phys.*, **79,** 926–935 (1983).

[15] A. Wlodawer, N. Borkakoti, D.S. Moss and B. Howlin, 'Comparison of two Independently refined Models of Ribonuclease A', *Acta Cryst.*, **B42,** 379–387 (1986).

[16] N. Thanki, J.M. Thornton and J.M. Goodfellow, 'Distributions of Water Around Amino Acid Residues in Proteins', *J. Mol. Biol.*, **202,** 637–657 (1988).

[17] U.C. Singh, P.K. Weiner, J.W. Caldwell and P.A. Kollman, AMBER (UCSF version 3.0, Department of Pharmaceutical Chemistry, University of California, San Francisco (1986).

[18] F. Vovelle, R. Elliott and J.M. Goodfellow, 'Solvent bridging Sites in A- and B- DNA helices', *Intl. J. Biol. Macromol.*, **11**, 39–42 (1989).

[19] J.M. Goodfellow and F. Vovelle, 'Biomolecular Energy Calculations using Transputer Technology', *Eur. Biophys. J.*, **17**, 167–172 (1989).

[20] J.M. Goodfellow, D.M. Jones, R.A. Laskowski, D.S. Moss, M.A.S. Saqi, N. Thanki and R. Westlake, 'Use of Parallel Processing in the study of protein ... ligand binding', *J. Comp. Chem.*, **11**, 314–325 (1990).

[21] D.M. Jones and J.M. Goodfellow, 'Use of Transputers in the study of Macromolecular Interactions', in 'Applications of Transputers' (Eds. T.L. Freeman and C. Phillips) IOS, Amsterdam (1989).

[22] D.M. Jones and J.M. Goodfellow, 'Monte Carlo Simulations of Biomolecular Systems using transputer aways', in 'Transputer Applications 90' (Eds. D. Pritchard and C. Scott) IOS, Amsterdam (1990) In Press.

[23] M.A.S. Saqi and J.M. Goodfellow, 'Convergence Behaviour in Free Energy Simulations', Mol. Sim. In Press.