Computational Studies of DNA Structure and Recognition

Christina Rebecca Grindon

Thesis submitted to The University of Nottingham for the degree of Doctor of Philosophy, September 2003.

'Educated beyond common sense!'

Edwin Wheeldon

1961-2003

ACKNOWLEDGEMENTS

Firstly I would like to thank my supervisors Charlie Laughton and Malcolm Stevens. I would like to thank Charlie for always being so enthusiastic and optimistic throughout this research, especially through the times when I was sure it was never going to work (or get finished!). I would like to thank Malcolm for being such an inspiration and for enticing me to the pub every Friday (it didn't take much doing!).

A HUGE thank-you has to go to my predecessor, Sarah Harris, without whom this PhD would have been almost impossible. Without Sarah we would not have some of the analysis methods that have been vital to this research and I would not have had the many invaluable discussions needed to get my head around some of the physical and mathematical concepts they involve.

Special thanks go to the rest of the group, old and new, for keeping me sane (or should that be insane) and especially to Sean for the discussions that, in times of self doubt, made me realise I did actually know what I was doing.

There are a few collaborators I need to thank. On the LAMMPS project I would like to thank Peter C for giving us the opportunity to do the validation, Keir and Tom for help with the implementation of AMBER and getting the code running and Steve Plimpton for letting us have his code to play with in the first place. On the telomerase project I would like to everyone involved within Pharmacy and Mark, Evris and Huw from Chemistry. I would also like to thank Modesto and his group for their help, especially Blas and Manu, and for looking after me during my visits to Barcelona.

Finally I would like to thank my family and friends for their constant support and encouragement and especially to Nathan who, despite my preoccupations and mood swings these last few months, still seems to be happy to live with me!

ABSTRACT

This thesis involves the use of large scale molecular dynamics simulations and associated analysis techniques to study DNA structure and recognition.

LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) is a scalable molecular dynamics code including long-range Coulomb interactions that has been specifically designed to function efficiently on parallel platforms. Here we describe the implementation of the AMBER98 forcefield in LAMMPS and its validation for molecular dynamics investigations of DNA structure and flexibility against benchmark AMBER6 code results. Extended MD simulations on the hydrated DNA dodecamer $d(CTTTTGCAAAAG)_2$ and 1:1 and 2:1 drug complexes, which have 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. Also, compared to AMBER6, LAMMPS shows greatly improved scalability in massively parallel environments (Cray T3E).

The telomerase enzyme is active in 85-90% of human tumours and is therefore an important target in anti-cancer drug design. Telomerase acts at the telomeric regions of chromosomes adding successive (TTAGGG)n repeats causing immortalisation of the cell. Telomerase can be inhibited by the stabilisation of G-quadruplexes which, *in vitro* studies show, are formed in these telomeric regions. In order to minimise non-specific toxicity associated with this approach it is important that the drugs preferentially bind to quadruplex over duplex DNA. A series of novel polycyclic acridine salts have been synthesised within our laboratories that show this property. MD studies have been used to study alternative binding relationships of RHPS4 (our lead compound) to quadruplex and duplex DNA and to explore the differences in binding profiles of RHPS4 and its methyl derivative RHPS3. Analysis of extended simulations (≥ 3ns) has been carried out including evaluation of ΔG from enthalpic and entropic contributions, linear interaction energy, stacking interactions and molecular interaction potentials. "Correct" binding positions for RHPS4 in quadruplex and duplex DNA have been found and simulations and analysis of RHPS3 also carried out. Although the results are not conclusive and do not all agree with the experimental data we can conclude that quadruplex verses duplex selectivity is governed by a subtle balance between many factors, including electrostatic and vdW interactions, DNA flexibility and most probably the models used.

CONTENTS

CHAPTER 1 – THE IMPORTANT STRUCTURAL AND FUNCTIONAL ASPECTS OF DNA.

1.1 DNA structure and its biological significance

1.1.1 *The discovery of DNA*

One of the most significant scientific accomplishments of the last few centuries has been the discovery of DNA. In 1868, whilst researching the physiology of human lymph cells Friedrich Miescher discovered a phosphorous-containing substance which he named 'nuclein'. This was actually a nucleoprotein and it wasn't until 1889 that the protein free 'nucleic acid' was obtained by Richard Altmann (Portugal & Cohen, 1977). Over the next few decades the components of nucleic acids were discovered, the phosphate group, the sugar moiety and the bases – guanine (G), adenine (A), thymine (T) and cytosine (C) (Figure 1.1).

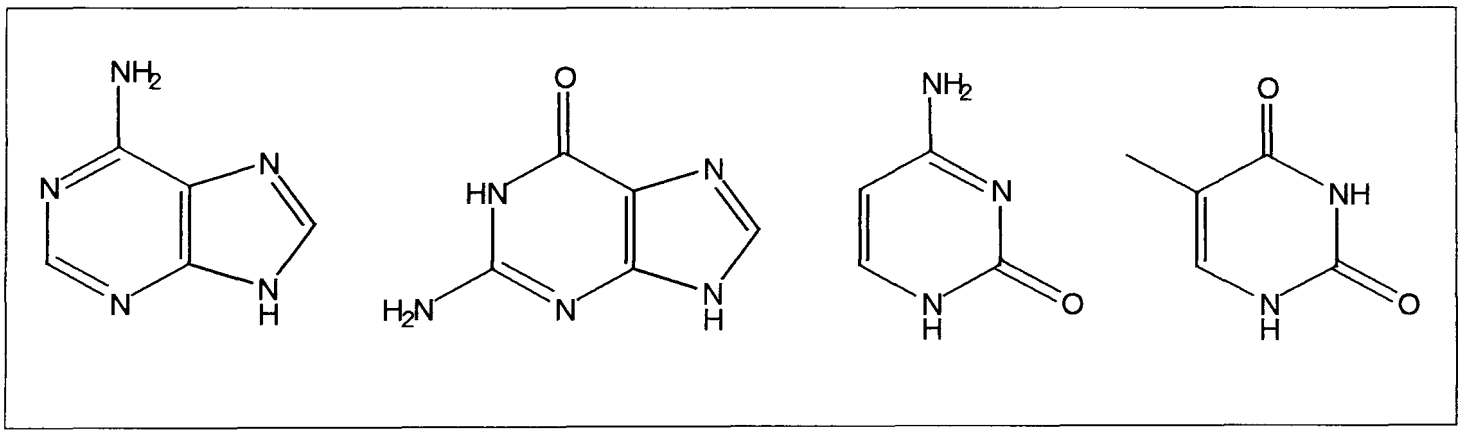


Figure 1.1 – The four bases of DNA, from left to right - adenine, guanine, cytosine and thymine.

Although at this point it was recognised that deoxyribonucleic acid (DNA) formed a major part of the cell's nucleus, the idea that its role was to carry genetic information was not known. This information did not emerge until there was a more in-depth knowledge of the structure of DNA.

It was 1953 when the structure of DNA was finally elucidated. Many scientists had been trying to discover the structure but it was Watson and Crick who deduced the double helical structure we all recognise today (Figure 1.2; Watson & Crick, 1953).

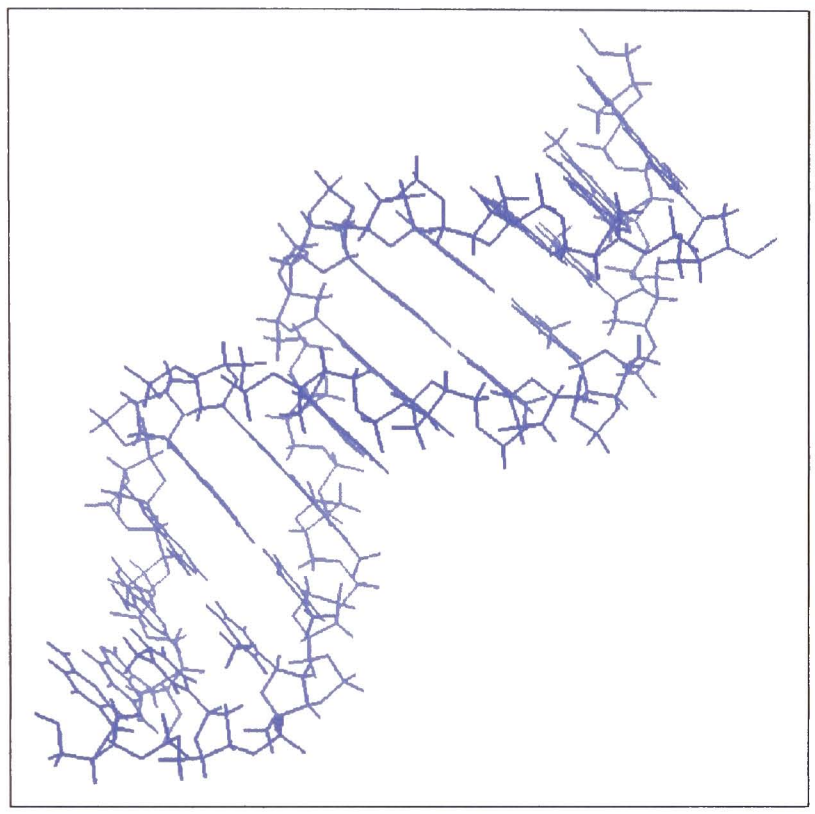


Figure 1.2 – The double helical structure of DNA.

The structure brought together ideas of several of the people studying DNA structure at that time. Astbury had deduced from his X-ray diffraction that flat nucleotides were spaced 3.34Å apart, standing out perpendicularly to the long axis (Astbury, 1947). Gulland postulated the presence of hydrogen bonds between purine and pyrimidine bases (although he believed the bases were in the enolic form not the keto form we now know they are in). Pauling attempted to create a helical model for DNA but he put the sugar phosphate backbone at its core with the bases pointing outwards (Pauling & Corey, 1953). Chargaff studied base composition and found that there were always equal numbers of purines (A or G) to pyrimidines (T or C) within a piece of DNA (Chargaff, 1950). Franklin studied X-ray diffraction patterns obtained from DNA fibres and found two forms of DNA depending on humidity levels, A-form at low humidity and B-form at high humidity (Franklin & Gosling, 1953). From this information she elucidated that both forms of DNA were highly crystalline and clearly helical in structure with phosphate groups on the

outside of the helix exposed to water and the bases on the inside of the helix (Blackburn and Gait, 1996).

1.1.2 The double helix

Watson and Crick took all this information on board and came up with the double helix. Two anti-parallel strands are wound around each other with the phosphate groups on the outside and the bases on the inside, as Franklin suggested. These strands are made up of nucleotides (phosphate group, deoxyribose sugar and hetrocyclic base) stacked upon each other in a 5', 3' arrangement (Figure 1.3).

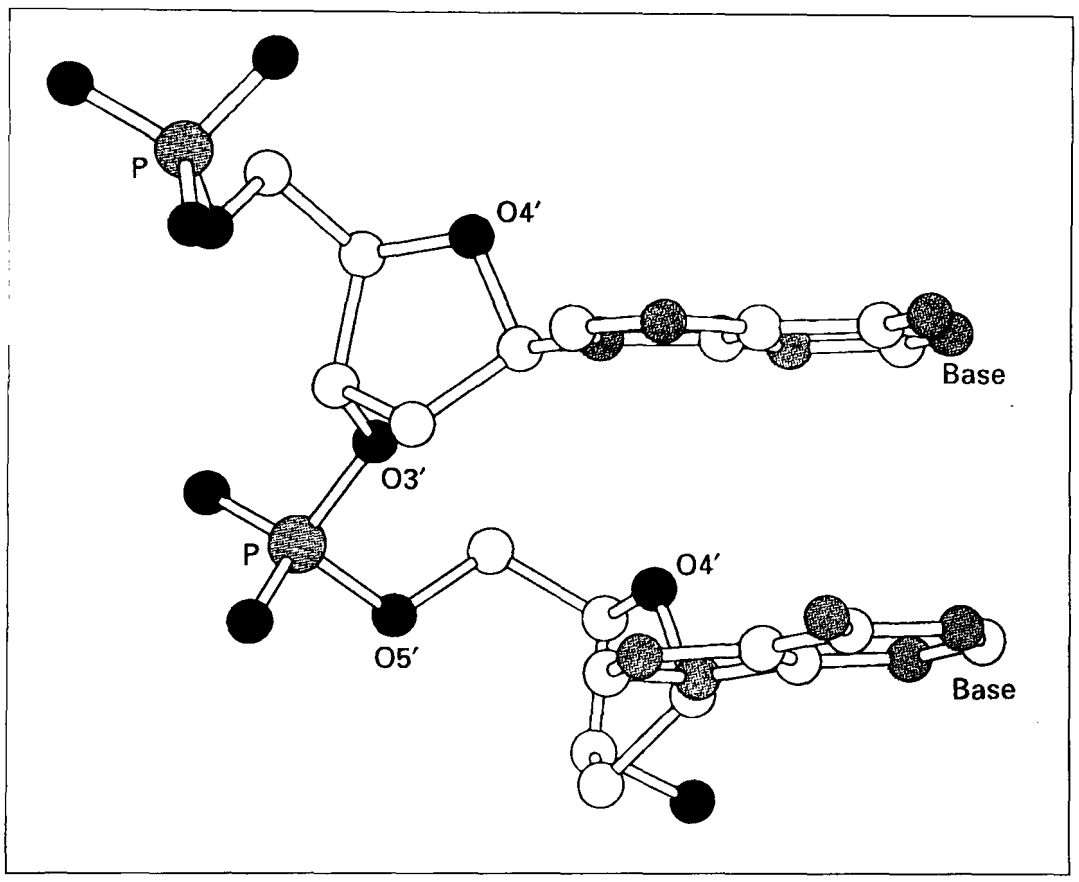


Figure 1.3 – The 5', 3' arrangement of nucleotide stacking (taken from Neidle, 1994).

Chargaff's rules were explained by complementary hydrogen bonding of bases - adenine with thymine and guanine with cytosine, now known as Watson-Crick base pairs (Figure 1.4). Combining the facts that there are two anti-parallel strands and the four bases always pair in a certain way Watson and Crick began to see how DNA could carry genetic information. At the end of their paper describing the double helix they wrote, "It has not escaped our

notice that the specific pairing we have postulated suggests a possible copying mechanism for the genetic material" (Watson & Crick, 1953).

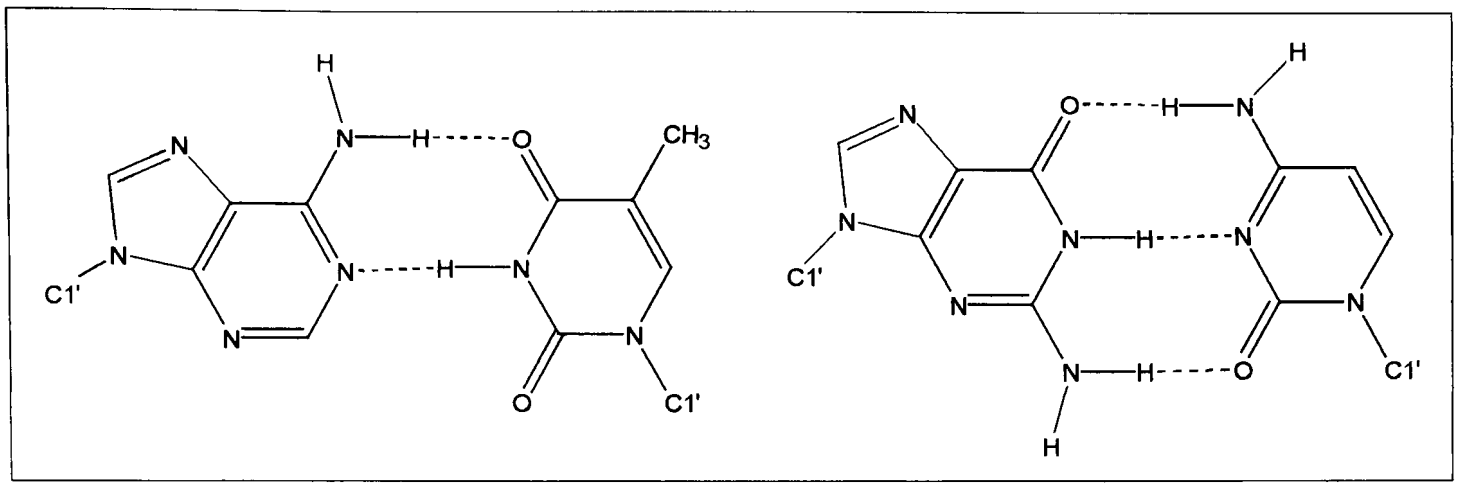


Figure 1.4 – Watson-Crick base pairing, from left to right – A-T and G-C, dashed lines represent hydrogen bonding.

A number of forces, mainly the hydrogen bonds between base pairs and the hydrophobic nature of these bases, hold the helical shape together. The hydrogen bonding between bases holds them in a planar arrangement so that they stack neatly upon each other with an inter stack distance of 3.4Å, van der Waals interactions also play a part in this interaction. The hydrophobicity of the bases becomes important when we think of the hydrated solvent environment DNA is normally found in. The bases are hydrophobic and therefore are found in the centre of the helix away from the solvent and the hydrophilic sugar phosphate backbone is open to the solvent.

The helix structure leads to the formation of grooves within the DNA, the major and minor grooves (Figure 1.5). These grooves are important for DNA functionality and bind various proteins and drugs as discussed in section 1.2.2. These grooves can be of different depths depending on the overall form of the DNA i.e. the grooves in B-form DNA are quite deep and well pronounced compared to other forms.

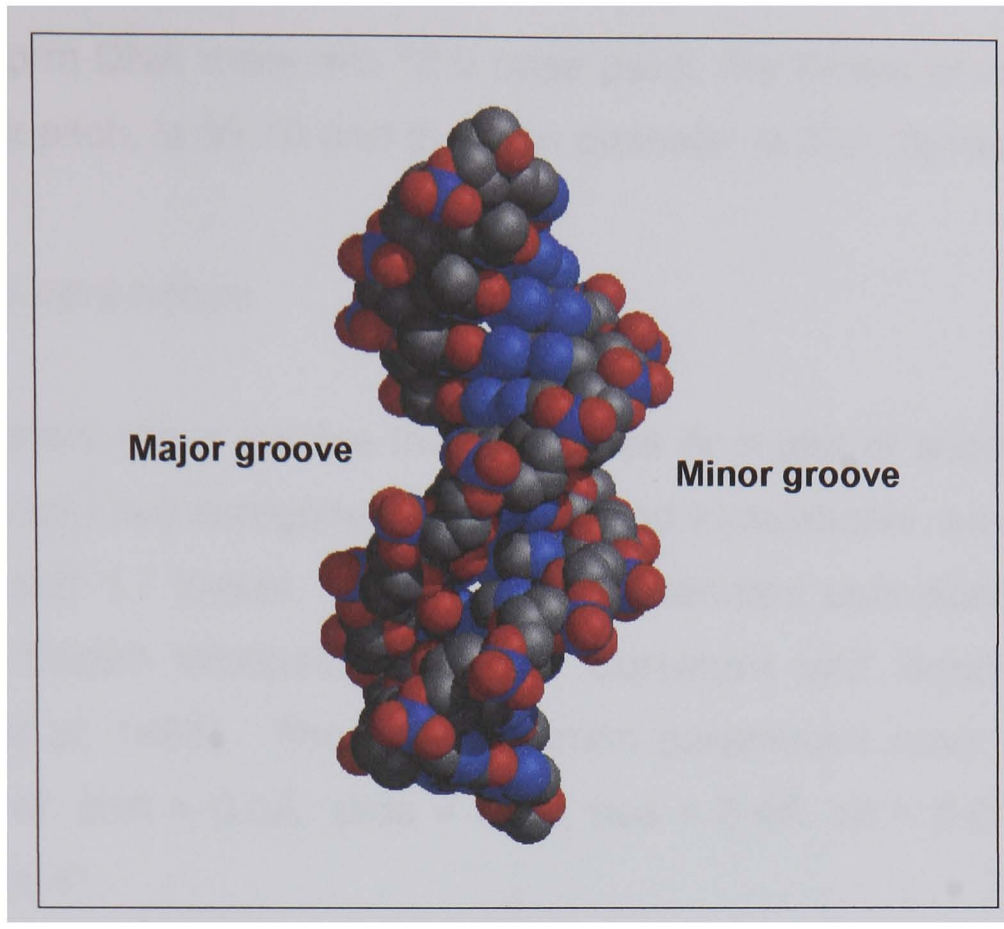


Figure 1.5 – Space filled representation of B-form DNA showing the major and minor grooves.

The B-form structure of the double helix, as described by Watson and Crick, is the most common structure formed by DNA (in solution), there are other forms of double stranded DNA including A-form and Z-form but these are not of importance to this study and will therefore not be discussed further.

1.1.3 Parameters describing the deformation of B-form DNA

The parameters that describe the deformation or flexibility of DNA can be divided into two groups; helical and base parameters. There are standard figures for these parameters within B-form DNA. Deviations from these standards are common and are what allow us to analyse and quantify the flexibility within DNA structure that can be needed for recognition (see later).

1.1.3.1 Helical parameters

B-form DNA forms mainly at high levels of humidity, for example, the cellular environment. It has right-handed (clockwise) helical rotation where the bases stack horizontally with respect to the helix axis. Throughout one full turn of

standard B-form DNA there are 10.5 base pairs, the height of one complete turn, the helix pitch, is 35.7Å and the helix diameter is 20Å (Sinden, 1994).

1.1.3.2 Base parameters

Base parameters either involve the two bases or a pair of successive base pairs and fall into two categories, rotational and translational, as described in Figures 1.6 and 1.7 below. These are the standard definitions as agreed upon at an EMBO Workshop on DNA Curvature and Bending in 1988 (Dickerson *et al*, 1989). The most common parameters have standard B-form values of: shift = 0.0Å, slide = 0.0Å, rise = 3.4Å, tilt = 6.0°, roll = 0.0° and twist = 34.4°.

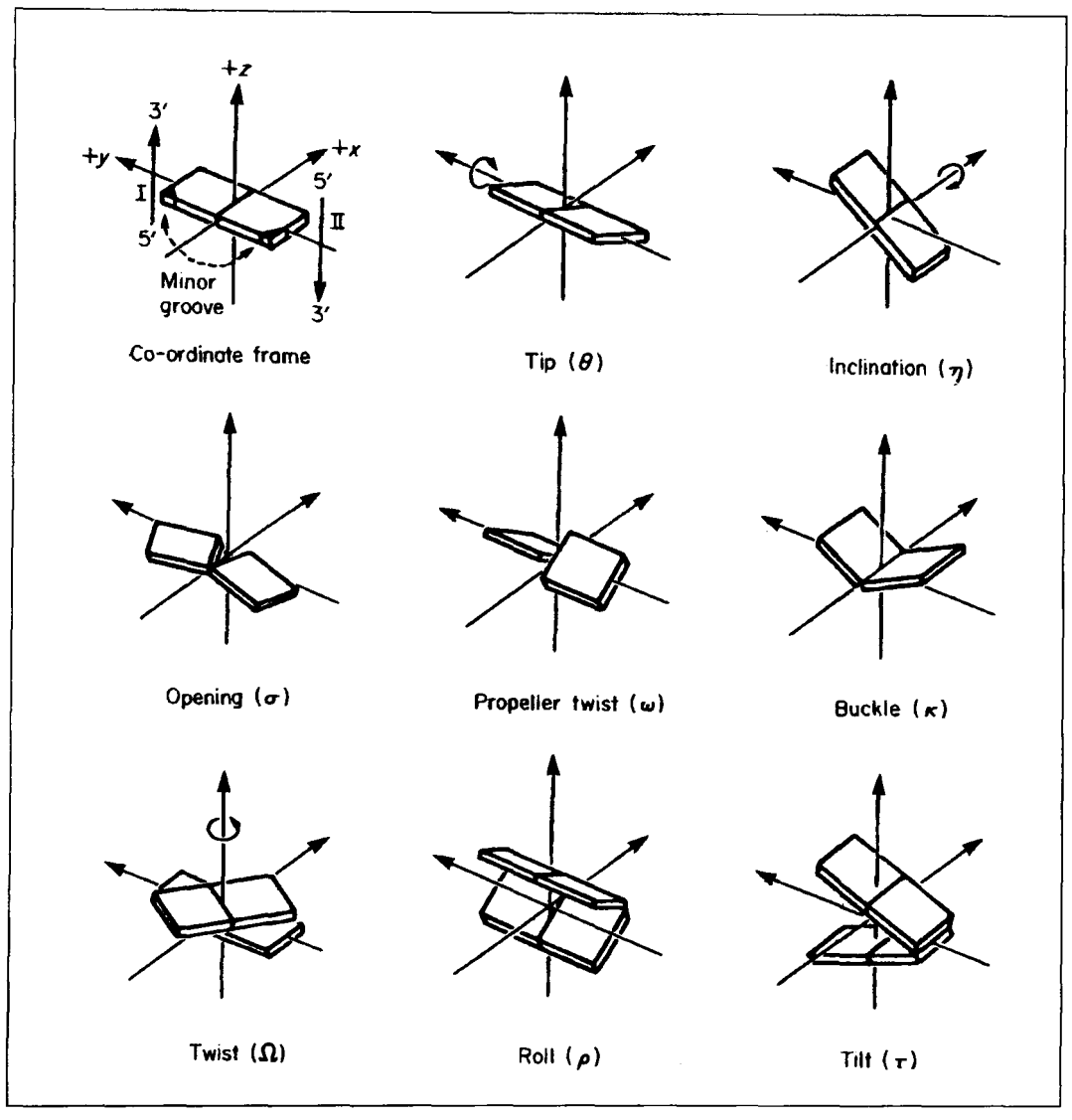


Figure 1.6 – Definitions of rotations involving 2 bases of a pair (top 2 rows) and 2 successive base pairs (bottom row). In the top row, the motions of the bases are co-ordinated, and in the middle row their motions are opposed. Columns at the left, centre and right describe rotations about the *z*, *y* and *x* axes respectively. The standard co-ordinate frame is defined at the upper left (taken from Dickerson *et al*, 1989).

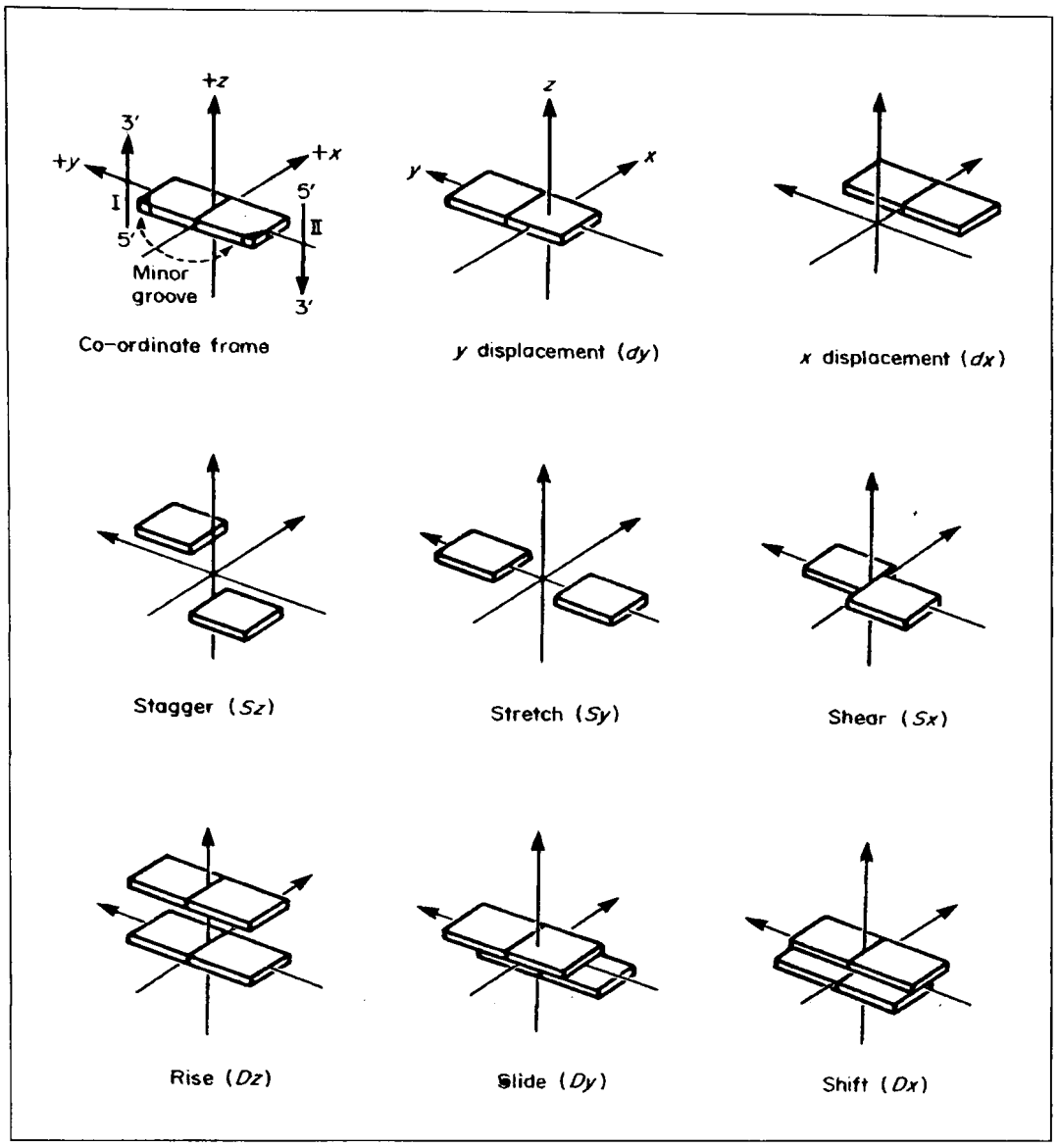


Figure 1.7 – Definitions of translations involving 2 bases of a pair (top 2 rows) and 2 successive base pairs (bottom row). In the top row, the motions of the bases are co-ordinated, and in the middle row their motions are opposed. Columns at the left, centre and right describe translations along the *z*, *y* and *x* axes respectively. The standard co-ordinate frame is defined at the upper left (taken from Dickerson *et al*, 1989).

There are other forms of DNA, discovered more recently, which exist by utilising more than two strands, three stranded (triplex) and four stranded (quadruplex) structures are also known. The existence of quadruplex DNA, its structure and its possible role in the fight against cancer will be discussed further in Chapter 4. The parameters describing the deformation and flexibility of these higher ordered DNA structures are not yet as common as those for the B-form duplex.

1.1.4 The biological importance of DNA

Once the structure of DNA was known, it became increasingly apparent that locked into the sequence of bases within the DNA of a cell was enough information to specify every cellular process within that cell i.e. a genetic code. DNA is able to pass on the information stored within its sequence via ribonucleic acid (RNA). Two of the most important roles of DNA and the two with which we may wish to interfere are replication and transcription, both of which read this genetic code.

1.1.4.1 Replication

When a cell divides, its DNA must create an exact copy of itself to give to the daughter cell, this process is known as replication. DNA replication is semi-conservative; the two daughter duplexes each contain one original strand from the parent DNA and one newly synthesised strand (Figure 1.8).

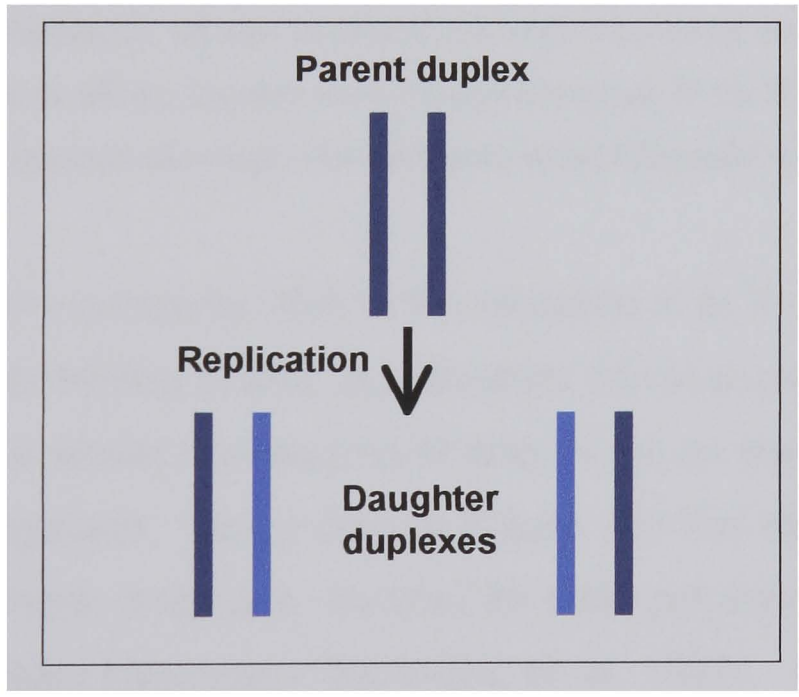


Figure 1.8 – Semi-conservative replication of DNA, daughter duplexes contain one original strand (from parent) and one new strand.

Replication occurs over one short stretch of DNA at a time, the parental strands separate at the point of replication and by doing so form a Y-shaped replication fork that travels down the DNA. At the replication fork the parental DNA begins to unwind (due to the actions of DNA gyrase and helicase

enzymes) and new strands begin to be synthesised bi-directionally (Figure1.9).

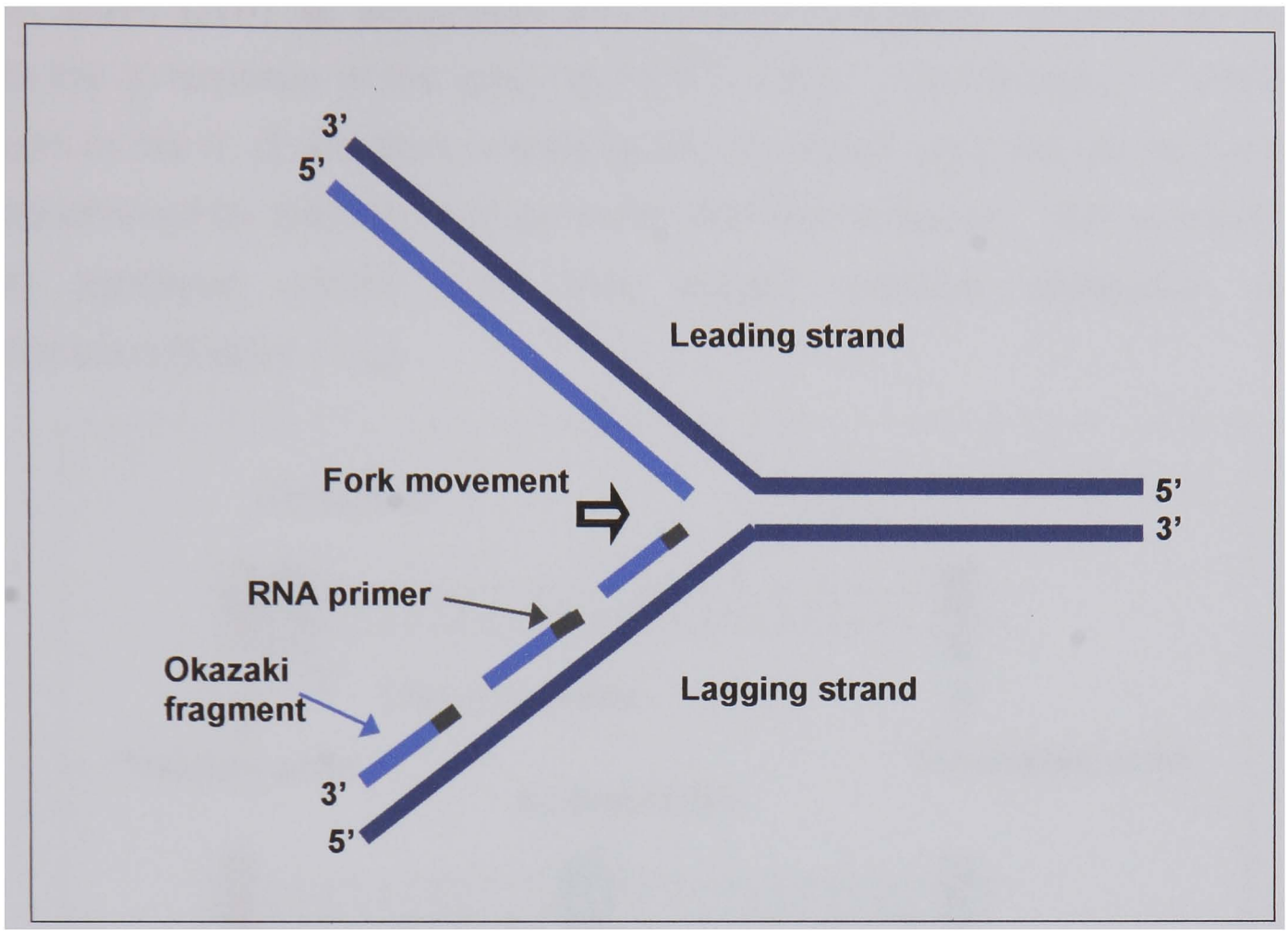


Figure 1.9 – Schematic of the replication fork showing continuous 5' to 3' synthesis on the leading strand and discontinuous 5' to 3' synthesis on the lagging strand through the formation of Okazaki fragments.

DNA polymerases synthesise DNA in the direction 5' to 3', therefore the 3', 5' parent strand, the leading strand, is replicated continuously. The replication of the 5', 3' parent strand, the lagging strand, is not as simple and requires a discontinuous approach. As shown in Figure 1.9 the lagging strand uses multiple RNA primers, these are required for DNA polymerases to initiate the synthesis of Okazaki fragments (Matthews *et al*, 1997). In the final step of the process these fragments are joined together by DNA ligase and a continuous strand is formed.

1.1.4.2 Transcription

Transcription is the primary stage of the biosynthesis of proteins. The genetic information encoded into a particular DNA sequence, is transcribed

onto messenger RNA (mRNA), through the action of DNA-dependent RNA polymerases. These enzymes require ribonucleoside triphosphates (ATP, GTP, UTP, CTP) as substrates and transfer nucleoside monophosphates onto the 3' terminus of the growing mRNA chain. Polymerisation therefore occurs in the 5', 3' direction and produces an mRNA chain whose sequence is determined by the base pairing within the DNA template. This process of RNA synthesis occurs over three stages: initiation, elongation and termination (Figure 1.10).

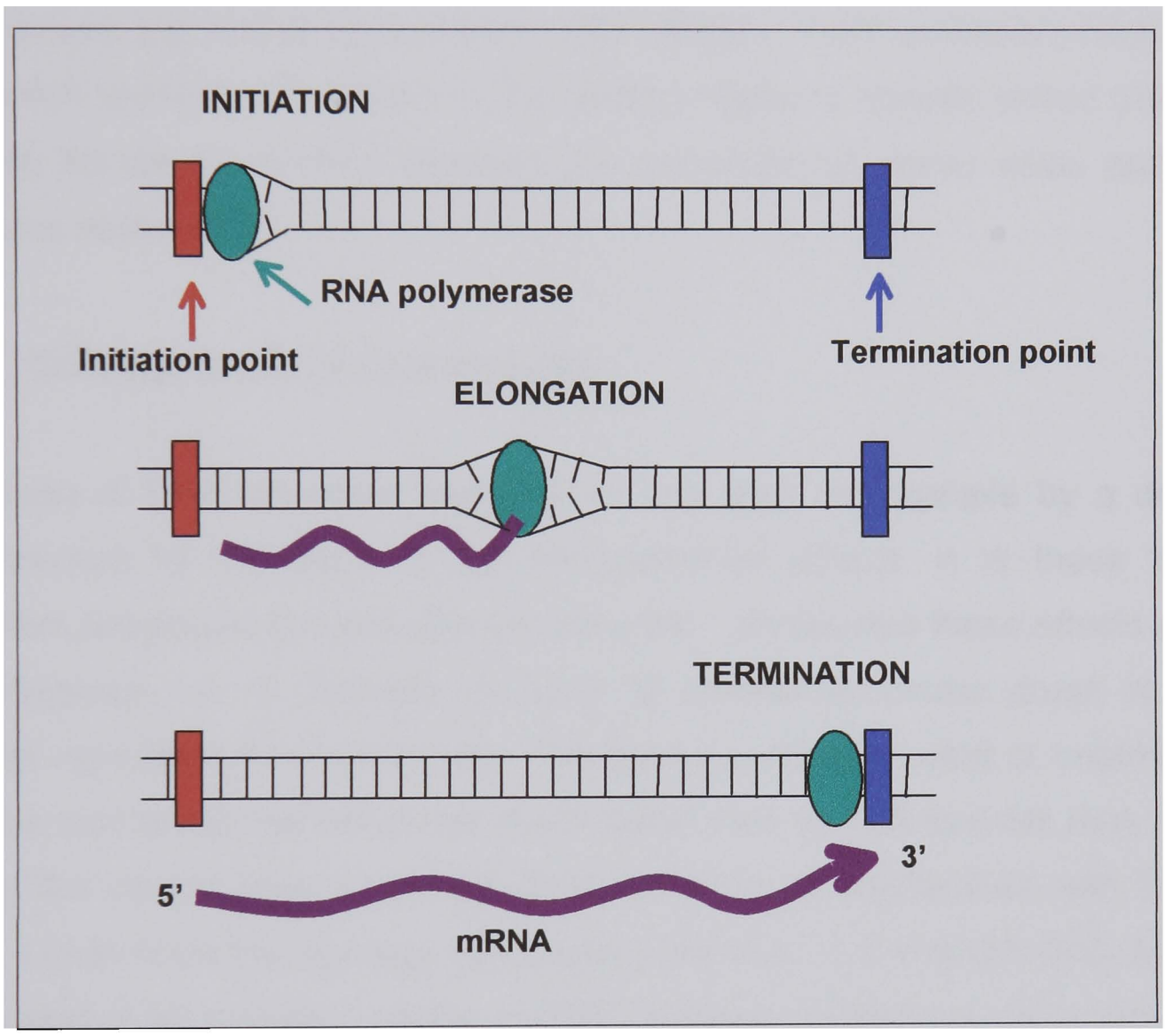


Figure 1.10 – Schematic showing the three phases of transcription, initiation, elongation and termination, leading to the synthesis of mRNA.

Initiation starts with RNA polymerase reversibly binding to the start of an encoding region of DNA. The enzyme is directed to these regions via specialised sequences known as promoters. This binding leads to unwinding of a portion of the double helix. The RNA polymerase then reads the template strand of the unwound DNA and mRNA synthesis begins. The second stage, elongation, involves the polymerisation of mRNA in the 5' to 3'

direction. During this process the newly synthesised mRNA strand trails as the double helix reforms behind the elongation complex. Termination occurs when the full gene has been transcribed. The RNA polymerase then stops binding to the DNA template and the mRNA strand that has been synthesised is released from the polymerase (Matthews *et al*, 1997).

Following on from transcription in the biosynthesis of proteins is the process of translation. The mRNA is translated via 3 lettered codons (the letters refer to the bases e.g. A,C,G,U) to transfer RNA (tRNA). Each amino acid has its own tRNA therefore the letters in the codons relate to specific amino acids. Peptide bonds are formed between the sequence of amino acids and a protein is assembled.

1.1.5 Consequences of DNA damage

If a piece of DNA becomes damaged in any way, for example by a drug administered to the body or by environmental effects, it is these two important processes that are affected the most. Sometimes these effects are advantageous i.e. a common property of chemotherapeutic drugs is to prevent replication from occurring. The theory behind this kind of treatment is that a cancerous cell will divide much faster than an ordinary cell thus it is mainly the cancer cells which are more sensitive to interference with their DNA. Other times the damage can cause problems i.e. if a single DNA base that is part of an encoding region of DNA becomes mutated, its transcription can lead to the wrong protein or a mutant protein being synthesised.

Cancer is one of the most common diseases caused by the transcription process going wrong. There are three types of genetic alteration or mutation which are known to lead to cancer, these mutations occur in oncogenes, tumour suppressor genes and genes that normally govern true replication of DNA e.g. DNA repair enzymes and cellular checkpoint genes (Oliff, 1999). Mutations arising in oncogenes generally result in what is known as "gain of function" changes to their encoded proteins (see example below), in contrast

to this mutations arising in tumour suppressor genes and DNA repair enzymes tend to result more in "loss of function" changes to their encoded proteins.

1.1.5.1 Example: The ras oncogene

An oncogene is described, in simple terms, as a gene that has the capability of causing cancer (Franks & Teich, 1995), one such oncogene is the *ras* oncogene. The normal *ras* gene encodes a GTP binding protein involved in signal transduction. This *ras* protein stimulates a cascade of events that culminate in the activation of nuclear transcription factors. The intrinsic activity of the *ras* protein is to convert active *ras*-GTP to inactive *ras*-GDP. A single nucleotide mutation to the ras gene (Franks & Teich, 1995) is enough to cause a defect in the activity of the *ras* protein such that it is always in the active *ras*-GTP form. This leads to the cell receiving a false signal to activate growth-promoting genes (Matthews *et al*, 1997). This leads to a gain in function where cells are able to grow indefinitely, i.e. they become immortal, a possible sign of cancer.

Mutated *ras* genes, of which there are three types – H-*ras*, K-*ras* and N-*ras*, are found in 20-30% of all human tumours but most commonly found in pancreatic cancer, colon cancer and adenocarcinoma of the lung (Oliff, 1999). *Ras* is a target for potential anti-cancer drugs and a number are now in development, for example, antisense oligonucleotides specific to the *ras* mutation of a gene which can be used to block mRNA translation (Leonard, 1997).

1.2 DNA Recognition

DNA recognition is a very important process within the human body. DNA sequences have to be recognised by the corresponding proteins and enzymes to enable the function to be carried out. For example, the initiation

stage of transcription could not take place if the RNA polymerase enzyme did not recognise the specific promoter sequence which signals the start of an encoding region of DNA.

The recognition of a piece of DNA can occur in two ways, firstly there is general (non-specific) recognition of any nucleotide or sequence, for example the packaging of DNA by histones, and secondly there is specific recognition whereby the ligand in question will recognise only a certain piece of DNA (nucleotide or sequence), for example restriction endonucleases which protect DNA by "cutting out" alien DNA (where the wrong base has been transcribed or a base has become mutated).

The mechanism of indirect readout (non-specific recognition) tends to occur through contacts made via the sugar-phosphate backbone, whereas direct readout (specific recognition) is more likely to occur through the same mechanism a single strand of DNA uses to recognise another - hydrogen bonding via the bases. Other mechanisms for readout of bases include electrostatic potential, steric effects and hydration.

1.2.1 Protein-DNA recognition

Protein-DNA interactions are occurring all the time within the body as almost all functions of DNA are carried out in conjunction with proteins. Most proteins tend to bind to the major groove as it is generally the site of direct readout of base sequence, also their structural motifs such as the α-helix are generally too large to fit into the minor groove, although some smaller proteins are known to bind in the minor groove, for example the TATA-box binding protein, discussed later.

There are three main types of protein involved in DNA interactions: regulatory, enzymatic and structural. The interaction of regulatory proteins involves the recognition of specific sequences of DNA to set off a cascade of events that can lead to gene expression (protein biosynthesis) or DNA

replication as described previously. Many of the proteins involved in these processes have an enzymatic component too, for example, the polymerases required for transcription and replication. Other enzymatic proteins that interact with DNA are the endo- and exo- nucleases whose function is to degrade DNA and RNA. Structural proteins bind to DNA and cause architectural effects, for example, the role of histones is to package DNA into chromosomes. This packaging involves the wrapping of around 150 base pairs of duplex DNA into almost two complete turns around the histone core (Bloomfield *et al*, 2000), highlighting the flexibility of DNA that can be required for protein interaction.

1.2.2 The role of the major and minor grooves in recognition

In order for direct readout of bases to occur the ligands need to have access to the bases via the grooves of DNA. Some of the hydrogen bond donors and acceptors of the bases are used in Watson Crick base pairing, but not all. As shown in Figure 1.11, the major groove has one hydrogen bond donor and two acceptor groups per base pair irrespective of the pairing.

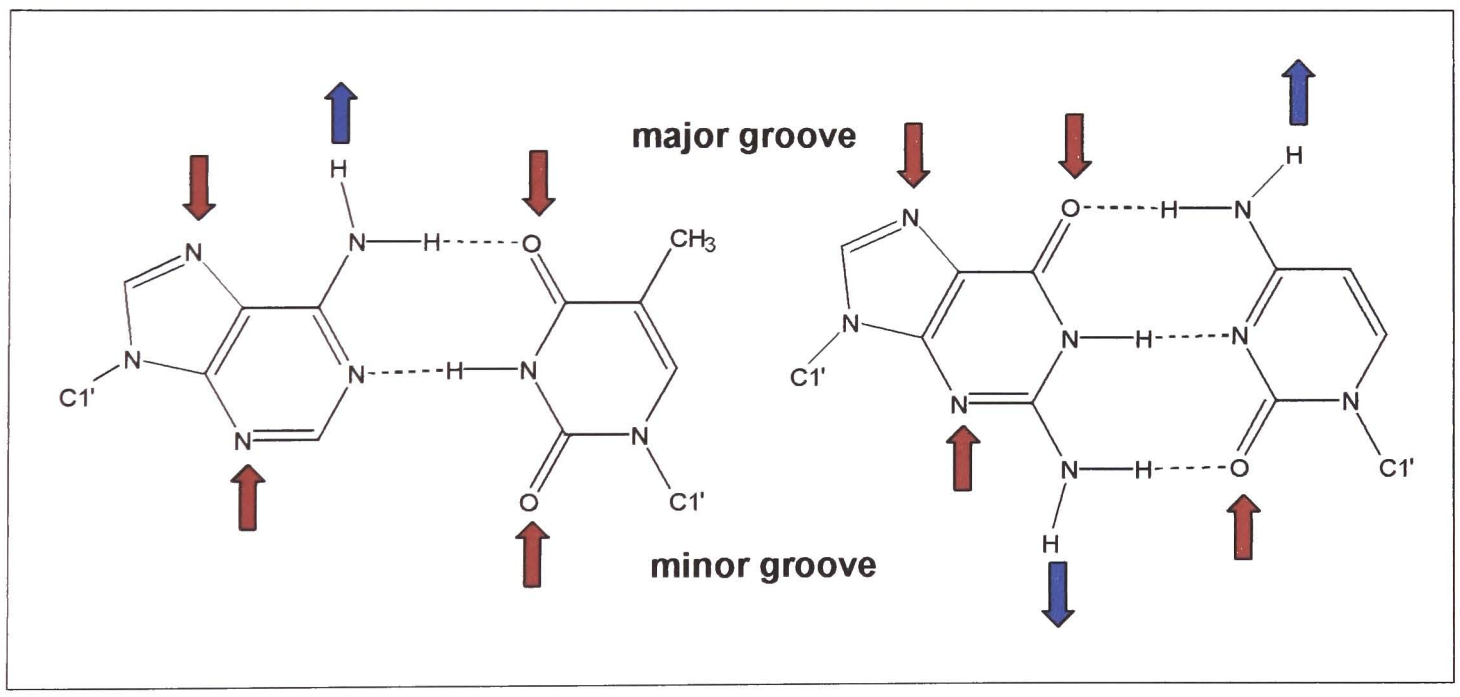


Figure 1.11 – The pattern of hydrogen bond donors (blue arrows) and acceptors (red arrows) in AT and GC base pairs.

CG and GC base pairs have different sequences of donors and acceptors, which means that they can be distinguished from the others by their

hydrogen bond donating and accepting patterns within the major groove. AT and TA base pairs have the same hydrogen bonding patterns in the major groove but can be distinguished from each other by the steric effect of the methyl group of thymine which introduces asymmetry into the groove (Neidle, 1994), it can also be involved in van der Waals interactions (Sinden, 1994).

The hydrogen bonding patterns in the minor groove are symmetric so discrimination between bases cannot be seen via just hydrogen bonding patterns; groove width and depth and electrostatic effects are also important for discrimination between bases in the minor groove, for example, AT containing sequences have narrower minor groove widths and GC containing sequences are more electron rich. Recently, White *et al* (1998) have used hairpin polyamides containing pyrrole (Py), imidazole (Im) and 3-hydroxypyrrole (Hp) to distinguish all four Watson-Crick base pairings. By forming four different ring pairings available with these molecules, a code to control sequence specificity can be achieved: Py/Im → CG; Im/Py → GC; Hp/Py → TA and Py/Hp → AT.

1.2.3 The role of flexibility in recognition

Recognition can require DNA to be very flexible to enable a ligand to interact, especially when the ligand is a protein (drugs tend to be much smaller and therefore do not require as much flexibility to interact), although this is not always the case as will be shown later. This flexibility comes from changes in the helical, and base parameters described earlier in section 1.1.3.

Dinucleotide steps display different degrees of flexibility depending on their sequence. The steps can be classified as either purine-pyrimidine (RY), purine-purine (RR) or pyrimidine-purine (YR) and these classifications can account for many conformational changes (Suzuki *et al*, 1996). YR steps (CA/TG, CG, TA) tend to have positive values of roll and high values of twist, RY steps (AC/GT, AT, GC) tend to have negative roll and low values of twist, and RR steps (AA/TT, AG/CT, GA/TC, GG/CC) are intermediate between

these two groups (Packer *et al*, 2000). El Hassan & Calladine (1997) have categorised these steps with respect to their flexibility: there are three rigid steps (AA/TT, AT and GA/TC); bistable steps (all G/C steps); and flexible steps (CA/TG and TA). These definitions were based on observations of deviation of slide, roll and twist from X-ray crystals structures.

1.2.4 Examples of protein-DNA recognition

1.2.4.1 The TATA-box binding protein

The TATA-box binding protein is known to bind to a specific sequence, the TATA box (TATAAAAG), of RNA polymerase II promoters forming a pre-initiation complex that denotes the starting point of transcription. Upon binding, the TATA box is severely bent towards the major groove (Figure 1.12) exposing on its opposite surface a very wide shallow minor groove which interacts with the ß-sheet of the protein (Kim *et al*, 1993 (a)). Interactions take place via the backbone and the minor groove. Side-chain/backbone interactions include salt bridges, water-mediated hydrogen bonds and van der Waals contacts with ribose groups. These interactions stabilise the unusually distorted DNA conformation within the complex. Interactions within the minor groove include phenylalanine base stacking which produces two dramatic kinks in the DNA, side-chain/base hydrogen bonding and hydrophobic and van der Waals contacts (Kim *et al*, 1993 (b)).

Figure 1.12 – The TATA-box binding protein bound to its corresponding TATA box sequence, TATAAAAG (PDB ID 1CDW).

1.2.4.2 The trp repressor

The *trp* repressor is involved in the transcriptional control of L-tryptophan levels in enteric bacteria. The binding of the *trp* repressor to the DNA operator sequence occurs almost entirely through backbone interactions including 24 direct and 6 solvent mediated hydrogen bonds. Unusually there are no direct hydrogen bonds to the bases that can explain the repressor's specificity to the particular sequence (Otwinowski *et al*, 1988). The binding of the *trp* repressor is also slightly unusual in that it does not result in much distortion of the DNA (Figure 1.13).

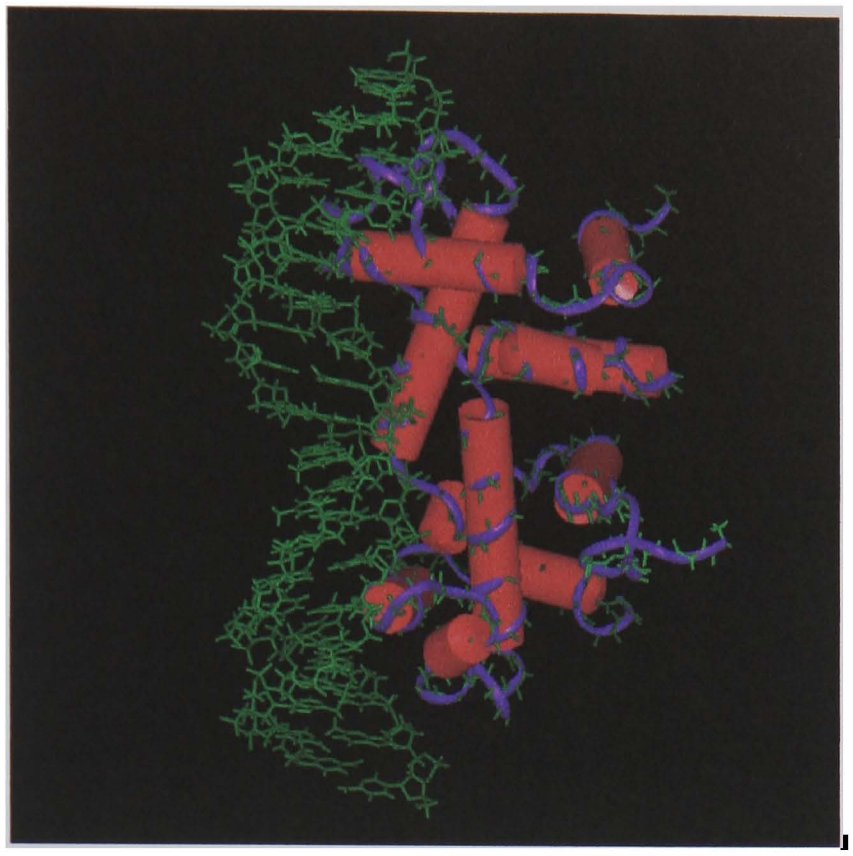


Figure 1.13 – The *trp* repressor/operator complex (PDB ID 1TRO).

1.2.5 Drug-DNA recognition

There are many naturally occurring and synthetically made molecules that bind strongly to DNA, some of these compounds are of medical importance (anti-tumour and antibiotic agents) and some are used for experimental reasons, e.g. the fluorescent stain ethidium bromide. These drug molecules are generally much smaller than proteins and have several ways of binding to DNA. The type of binding exerted by these molecules falls into one of two categories – irreversible covalent binding (alkylating agents) or reversible non-covalent binding (groove binders and intercalators) which are described below.

1.2.6 Alkylating agents

Alkylating agents covalently bind to DNA causing drug-DNA adducts which disrupt the natural DNA sequence. The purine bases are most susceptible to covalent attack, particularly guanines. The sites most at risk are O6, N6 and N7 in the major groove and N1 and N2 in the minor groove (Neidle, 1994), this is because these are the most nucleophilic sites and

alkylating/metallating agents tend to be electrophilic. These types of drugs are able to bind to a single site or multiple sites depending on their functionality, a classic example of this being the anti-cancer drug cisplatin (Figure 1.14). Cisplatin binds at the N7 atom of guanine and because of the two Cl groups it has available for electrophilic attack, can either bind to one N7 or cross-link two N7 atoms either on the same strand (intra-strand) or on opposing strands (inter-stand), similar patterns are seen with the nitrogen mustards (Figure 1.14).

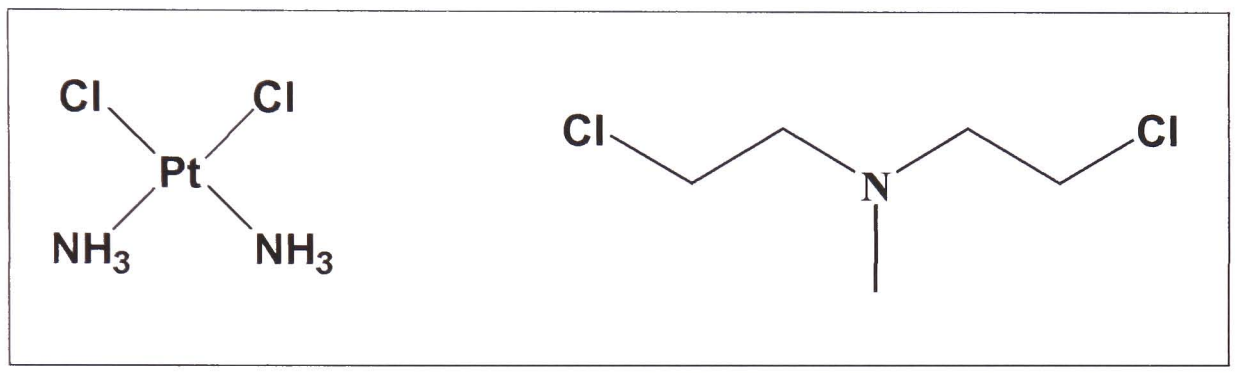


Figure 1.14 – Metallating/alkylating agents Cisplatin (left) and Nitrogen mustard (right).

This disruption of the DNA sequence and the formation of adducts can lead to strand breaks and other forms of DNA damage which in turn lead to disruption of replication and transcription.

1.2.7 Groove binders

Of the two grooves found in the structure of DNA it is the minor groove that is most often used for drug binding. The major groove is more commonly taken up by the binding of proteins and therefore drugs often have only the minor groove to bind to by default, also most groove binding drugs are too small to bind tightly to the major groove. Binding to the minor groove is non-covalent and governed by features such as isohelicity, charge and hydrogen bonding affinity.

The isohelicity is important as the grooves of DNA are curved and therefore molecules that follow this curvature are likely to bind more easily without

disrupting the DNA structure. Figure 1.15 shows the minor groove binding drugs Hoechst 33258 (discussed further in chapter 3) and netropsin, note their curved structure which is isohelical with DNA, therefore the binding of these drugs will lead to little distortion of the DNA.

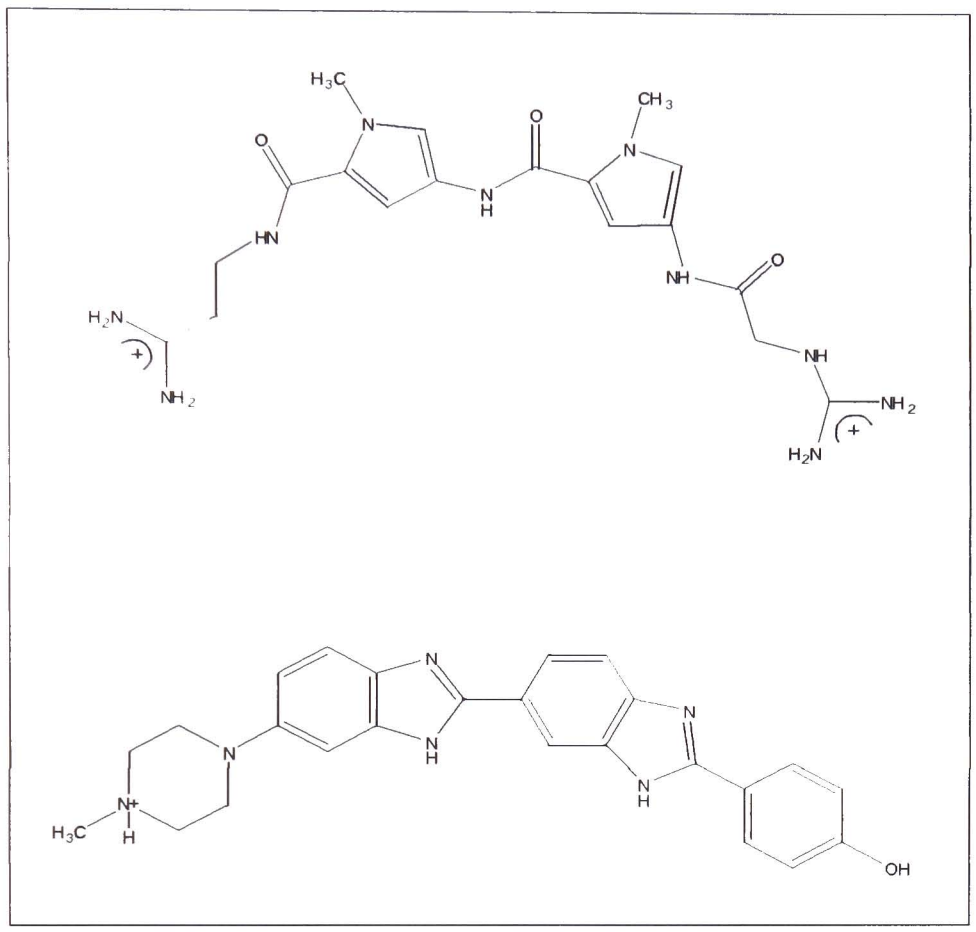


Figure 1.15 – Structures of netropsin (top) and Hoechst 33258 (below).

Groove binders often have a positive charge that forms interactions with the phosphate backbone and hydrogen bonding capabilities (amine NH groups) that interact with the N3 atoms of adenine and/or the O2 atoms of thymine, thus groove binders tend to bind more strongly to regions with AT sequences of DNA. AT regions are also narrower than their corresponding CG regions which leads to greater van der Waals interactions that help hold the drug molecules in place. It is through this specificity that groove binding drugs can be directed to certain sequences of DNA, for example, netropsin can bind to a region of four successive A or T bases.

1.2.7.1 Advanced groove binders

Although it became possible to target small sequences of DNA, what was really needed to help the fight against diseases such as cancer, were drugs

that could bind to extended sequences of DNA, therefore enabling the targeting of specific genes. In is known that to selectively recognise a single gene in human cells, a sequence of 17 base pairs must be recognised (Thuong & Helene, 1993).

Studies in the mid 1980's, suggested that by replacing the pyrrole (Py) rings of AT specific minor groove binders, such as distamycin or netropsin (Figure 1.15 above), with imidazole rings (Im) it should be possible to design ligands that could also recognise GC base pairs (Helene, 1998). Introducing a hydrogen bond acceptor (the extra N of imidazole, figure 1.16) to the drug leads to improved binding to guanine. This is because of improved hydrogen bonding and reduced steric hindrance of the protruding NH_2 group. This resulted in the group of polyamide minor groove binders called lexitropsins (Lown *et al*, 1986) which were made up of Py and Im rings linked together. Although it was now possible to have minor groove binders that recognised AT regions and GC regions there was little success in recognising extended sequences, this came later.

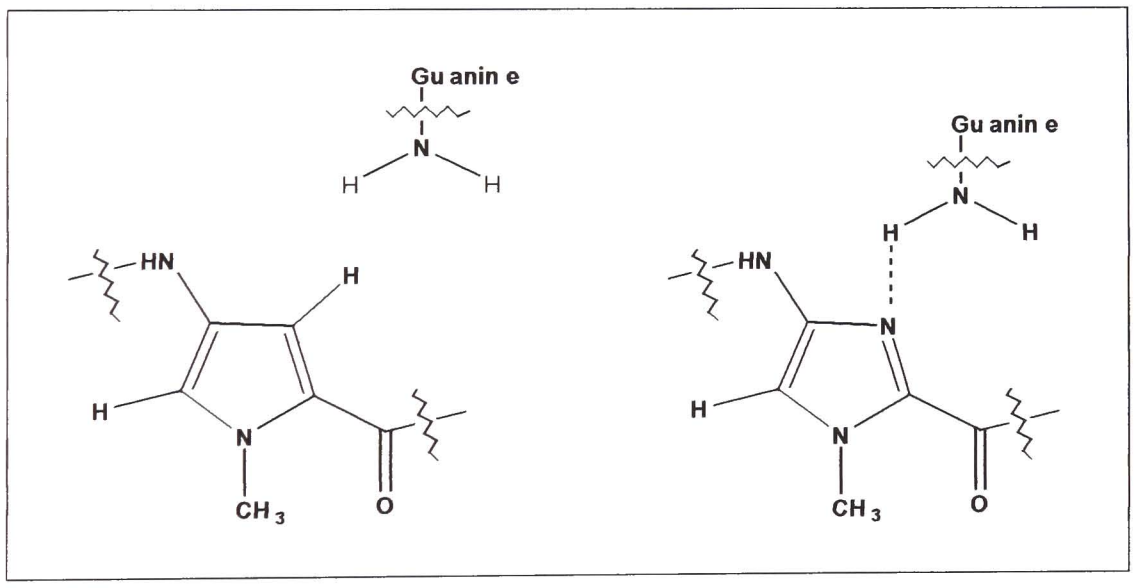


Figure 1.16 – Imidazole (right) forms a new hydrogen bond with the NH_2 of guanine whereas pyrrole (left) does not, the NH_2 of guanine is also sterically hindered by pyrrole.

The discovery that two molecules of distamycin could bind side-by-side in the minor groove of AT rich sequences of DNA lead to the linking of two of these monomers together to give exceptionally strong ligands for these sequences. Two types of ligands were discovered, hairpin polyamides (Mrksich &

Dervan, 1993) and cross-linked lexitropsins (Chen & Lown, 1994). Hairpin polyamides can distinguish between the four different Watson Crick base pairs by use of pairings of pyrrole (Py), 3-hydroxypyrrole (Hp) and imidazole (Im) ring systems (as discussed in section 1.2.2). More recently by combining these with benzimidazole (Bi), imidazo[4,5-*b*]pyridine (Ip) and hydroxybenzimidazole (Hz), double ring systems have been created with equal binding affinities for the specific base pairs (Renneberg & Dervan, 2003). With these recent developments it should now be possible to synthesise ligands that are able to recognise extended sequences of DNA and hopefully in the next few years it will be possible to target specific genes.

1.2.8 Intercalators

Intercalation is the insertion of a flat, usually heteroaromatic polycyclic ring system inbetween two base pairs of the double helix. Intercalators stack parallel to the base pairs above and below, their aromaticity taking advantage of the π stacking to help stabilise their insertion. The double helix is disrupted by this interaction, is extended by approximately 1 base pair's spacing (3.4Å) and unwinding of the helix is observed. A large number of antibiotics and anti-tumour agents come under the category of intercalators, they all have a chromophore, 2-3 six-membered rings in size, similar to a base pair.

1.2.8.1 A brief history of intercalators

The earliest and simplest intercalators are those based on the acridine skeleton, for example 9-aminoacridine and proflavine (Figure 1.17). These were used in early crystallography studies to determine how these types of drugs bound. It was discovered that these intercalators had anti-tumour properties, and so studies were carried out to find similar drugs that bound more strongly to DNA and were more potent against a wider variety of tumours.

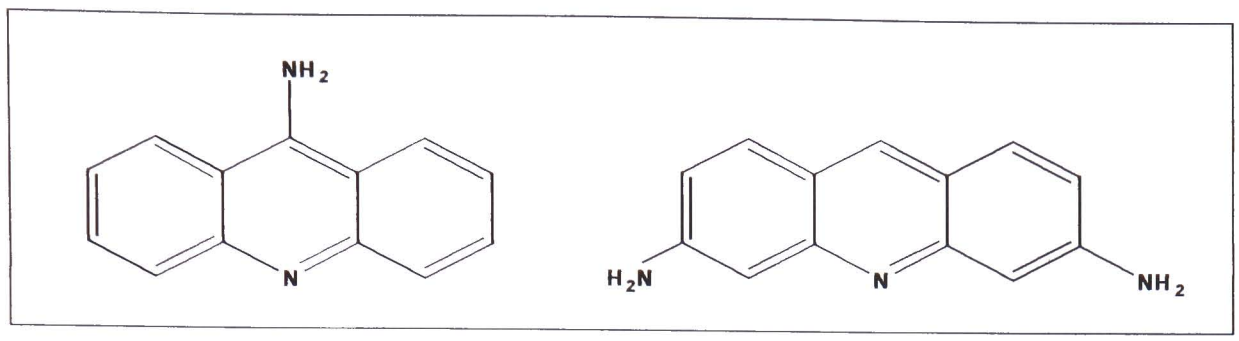


Figure 1.17 – Structure of the simple intercalators 9-Aminoacridine (left) and Proflavine (right).

Daunorubicin (Figure 1.18; DiMarco *et al*, 1964; Dubost *et al*, 1964) was one of the first intercalators commonly used in clinical treatment. Its derivative doxorubicin (Figure 1.18; Di Marco, 1975) also found use in the clinic as it has a much broader spectrum of anti-tumour activity despite only having one structural difference. Another drug which found its way into the clinic around this time was amsacrine (Figure 1.18; Cain & Atwell, 1974) which was used in the combination therapy of acute leukaemia.

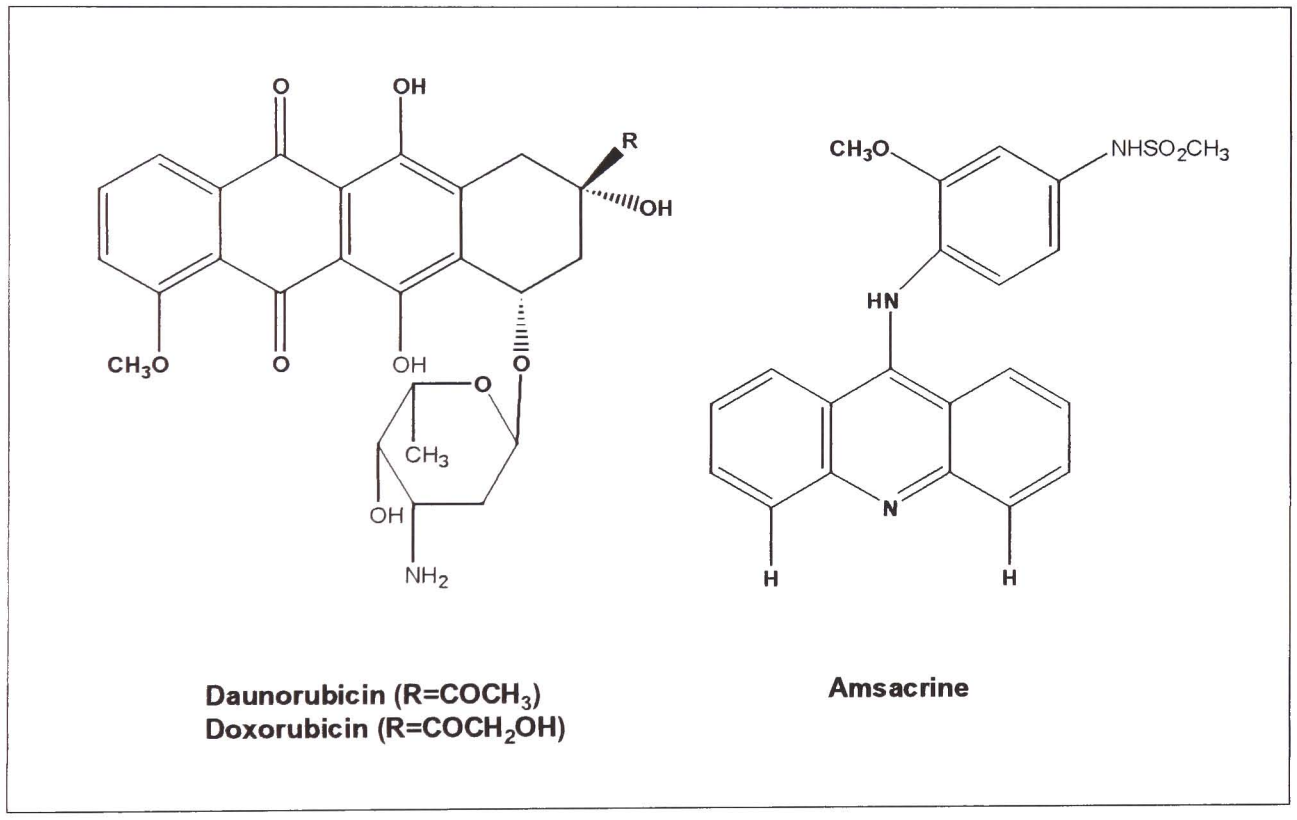


Figure 1.18 – Structures of Daunorubicin and Doxorubicin (left) and Amsacrine (right).

Intercalators with side chains that were able to protrude into the grooves of the DNA were synthesised, the rationale being that these would have better binding affinities for DNA and by attaching a basic side chain it would help

the solubility of these drugs under physiological conditions (Baguley, 1991). An example of this is the hybrid molecule between acridine carboxamide and amsacrine (Figure 1.19) which is thought to intercalate with the anilino group residing in the minor groove and the dimethylaminoethyl side chain in the major groove of DNA (Wakelin *et al*, 1990).

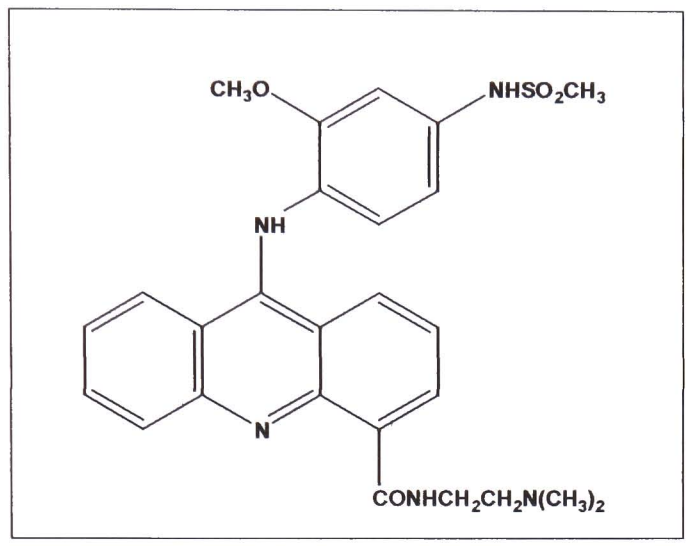


Figure 1.19 – Structure of the hybrid molecule between amsacrine and 9-aminoacridine carboxamide.

More recently, intercalators have been targeted towards four stranded quadruplex DNA with a view to stabilising its formation and inhibiting the telomerase enzyme (discussed further in Chapter 4).

1.2.8.2 Effect of intercalators on DNA

The binding of intercalators to DNA results in unwinding, lengthening and stiffening of the helix, thus introducing changes in topology that affect the interaction with associated enzymes and proteins. One such enzyme whose function is affected by intercalators is topoisomerase II (topo II). The role of topo II is to regulate DNA topological features, it cleaves a DNA duplex and then when the topology is correct reseals the strand breaks (Berger *et al*, 1996). One of its roles is to relieve torsional stress which builds up ahead of the replication fork during replication. It works alongside its isomer topoisomerase I whose job is to cleave one strand allowing the DNA to swivel before topo II cleaves both strands ready to reseal them when the torsional stress has been relieved (Matthews et al, 1997).

Intercalating drugs interfere with the breakage-reunion reaction of topo II by enhancing the formation of a DNA-enzyme complex in which the DNA strands are cleaved and covalently linked to the protein. The presence of elevated levels of these complexes provokes mutagenic and cell death pathways (Ferguson & Baguley, 1993).

1.3 References

Astbury W.T, (1947), *Symp. Soc. Exp. Biol,* **1**, 66.

Blackburn G. M. & Gait M. J, (1996), *Nucleic Acids in Chemistry and Biology, 2nd edition,* Oxford University Press.

Baguley B.C, (1991), *Anti-Cancer Drug Design,* **6**, 1-35.

Berger J.M, Gamblin S.J, Harrison S.C, Wang J.C, (1996), *Nature,* **379**, 225-232.

Bloomfield V. A, Crothers D. M, Tinoco Jr I, (2000), *NUCLEIC ACIDS – Structures, Properties and Functions*, University Science Books.

Cain B.F, Atwell, (1974), *Eur. J. Cancer,* **10**, 539-549.

Chargaff E, (1950), *Experientia*, **6**, 201-209.

Chen Y.H, Lown J.W, (1994), *J. Am. Chem. Soc,* **116**, 6995-7005.

Dickerson R.E, Bansal M, Calladine C.R, Diekmann S, Lavery R, Nelson H.C.M, Olsen W.K, Saenger W, Shakked Z, Sklenar H, Soumpasis D.M, Tung C-S, von Kitzing E, Wang A.H-J, Zhurkin V.B, (1998), *J. Mol. Biol,* **205**, 787-791.

Di Marco A, Gaetani M, Dorigotti L, Soldati M, Bellini O, (1964), *Cancer Chemoth. Rep,* **38**, 31-38.

Di Marco A, (1975), *Cancer Chemoth. Rep,* **6**, 91-106.

Dubost M, Ganter P, Maral R, Ninet L, Pinnert S, Preud'homme J, Werner G.H, (1964), *Cancer Chemoth. Rep,* **41**, 35-40.

El Hassan M.A, Calladine C.R, (1997), *Philos. Trans. Ser. A,* **355**, 43-100.

Ferguson L.R, Baguley B.C, (1993), *Environ. Mol. Mutagen,* **24**, 245-261.

Franklin R.E. & Gosling R, (1953), *Nature*, **171**, 740-741.

Franks L.M, Teich N.M, (1995), *Introduction to the Cellular and Molecular Biology of Cancer, 2nd edition,* Oxford University Press.

Helene C, (1998), *Nature,* **391**, 436-438.

(a) Kim Y, Geiger J.H, Hahn S, (1993), *Nature*, **365**, 512-520.

(b) Kim J.L, Nikolov D.B, Burley S.K, (1993), *Nature*, **365**, 520-527.

Leonard D.M, (1997), *J. Med. Chem,* **40** (19), 2971-2990.

Lown J.W, Krowicki K, Bhat U.G, Skorobogaty A, Ward B, Dabrowiak J.C, (1986), *Biochemistry,* **25**, 7408-7416.

Matthews H.R, Freedland R, Miesfeld R. L, (1997), *BIOCHEMISTRY – A Short Course,* Wiley-Liss, Inc.

Mrksich M, Dervan P.B, (1993), *J. Am. Chem. Soc,* **115**, 9892-9899.

Neidle S, (1994), *DNA Structure and Recognition,* Oxford University Press.

Oliff A, (1999), *BBA Reviews on Cancer*, **1423**, C19-C30.

Otwinowski Z, Schevitz R.W, Zhang R.G, Lawson C.L, Joachimiak A, Marmorstein R.Q, Luisi B.F, Sigler P.B, (1988), *Nature*, **335**, 321-329.

Packer M.J, Dauncey M.P, Hunter C.A, (2000), *J. Mol. Biol,* **295**, 71-83.

Pauling L. & Corey R. B, (1953), *Proc. Natl. Acad. Sci. USA*, **39**, 84-97.

Portugal F. H. & Cohen J. S, (1977), *A Century of DNA,* The Massachusetts Institute of Technology.

Renneberg D, Dervan P.B, (2003), *J. Am. Chem. Soc,* **125**, 5707-5716.

Sinden R.R, (1994), *DNA Structure and Function*, Academic Press, Inc.

Suzuki M, Yagi N, Finch J.T, (1996), *FEBS Letters,* **379**, 148-152.

Thuong N.T, Helene C, (1993), *Angew. Chem. Int. Edn Engl,* **32**, 666-690.

Wakelin L.P.G, Chetcuit P, Denny W.A, (1990), *J. Med. Chem,* **33** (7), 2039-2044.

Watson J. D. & Crick F.H.C, (1953), *Nature (London)*, **171**, 737-738.

White S, Szewczyk J.W, Turner J.M, Baird E.E, Dervan P.B, (1998), *Nature,* **391**, 468-471.

CHAPTER 2 – MOLECULAR MODELLING METHODOLOGY AND ANALYSIS TECHNIQUES.

2.1 Introduction to molecular mechanics

Molecular modelling of atoms and molecules can be done by either Quantum mechanical methods or molecular mechanical methods. Quantum mechanics is perhaps the "holy grail" of modelling techniques as it deals with the electrons in a system and can be used to obtain a full electronic structure. The calculations involved require a lot of computer power and therefore quantum mechanical models can only realistically be obtained for small molecules. To obtain representations of much larger molecules such as proteins and nucleic acids a simpler method needs to be used. Molecular mechanics methods ignore the electronic motions and calculate the energy of a system as a function of the molecular positions only (Leach, 2001). Molecular mechanics forcefields can, in some cases, give representations of molecules that are as accurate as quantum mechanical calculations but they cannot provide information about the electronic distribution of a system.

2.1.1 The forcefield

A molecular mechanics potential energy function, or forcefield, describes the structure and covalent connectivity of the molecules in a system. In atomistic molecular mechanics, each atom in the system is described as a point mass connected to another atom via a highly characterised spring. This is known as the "ball and stick" model. There are other types of molecular mechanics methods that use lower resolution models, therefore describing a molecule in less detail (Lafontaine & Lavery, 1999). For example, there are models that describe DNA as two helical strings attached to each other by bonds of varying magnitude (representing the fact that CG pairs bind more strongly than AT as they have an extra hydrogen bond). There are also models with even lower resolution, which describe DNA as just an elastic rod, these are

known as mesoscopic models. These lower resolution models are becoming more useful as we try to simulate longer stretches of DNA and other macromolecules for longer and longer timescales.

2.1.1.1 The forcefield equations

The forcefield equation is used to describe the nuclear motion of a system and can be broken down into the different bonded and non-bonded terms (Figure 2.1).

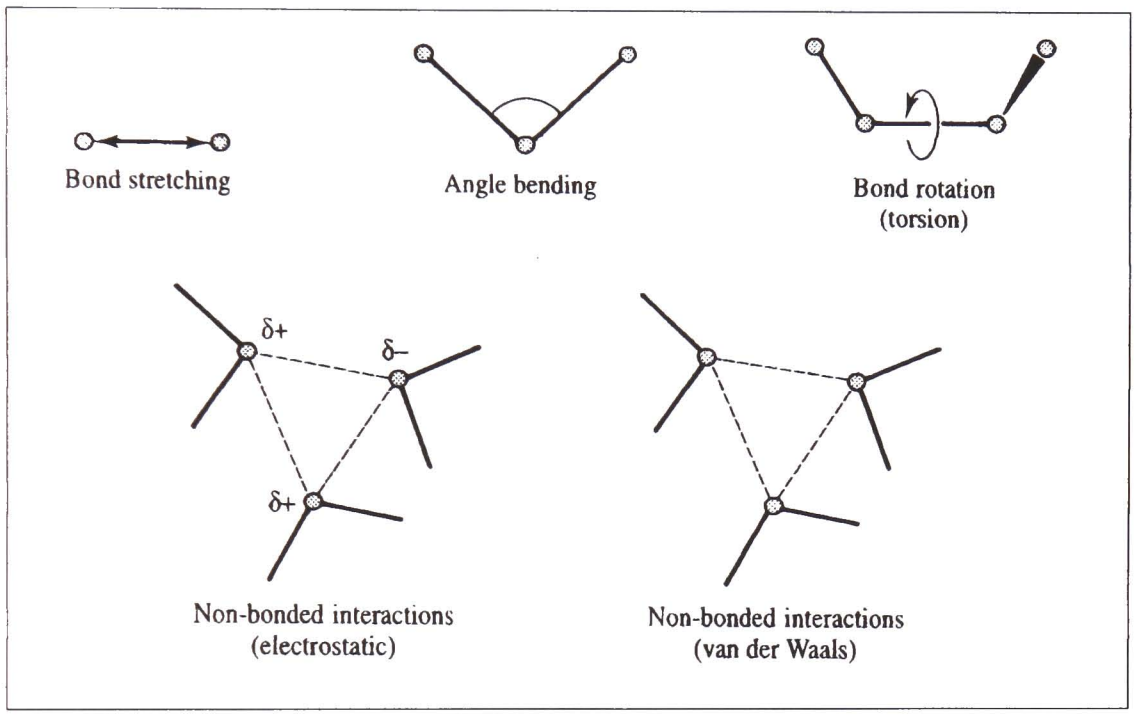


Figure 2.1 – Schematic representation of the key components of a molecular mechanics forcefield, bonded term (top row), non-bonded terms (bottom row) (taken from Leach, 2001).

2.1.1.2 Bonded terms

The bonded terms are made up of bond stretching, angle bending and dihedral angle terms. At the name suggests they are interactions between atoms that are connected either via a 1-2 (bond), 1-3 (angle) or 1-4 (dihedral) arrangement.

Bond stretching is modelled by a harmonic potential (Equation 1) that gives the variation in energy as the bond length (l) deviates from its equilibrium position (l_{eq}).

$$E_{bond} = \sum k_b \left(l - l_{eq}\right)^2$$

Equation 1 – Harmonic potential describing the bond energy. k_b is the force constant for the bond, l is the bond length and l_{eq} is the equilibrium bond length.

This equation gives a good approximation of energy around the equilibrium but when bond lengths increase dramatically from this, the equation is less accurate. This is because the harmonic potential used is similar to the Morse potential (model used to describe the potential energy curve for a typical bond) at equilibrium but deviates at higher bond lengths (Figure 2.2; Leach, 2001).

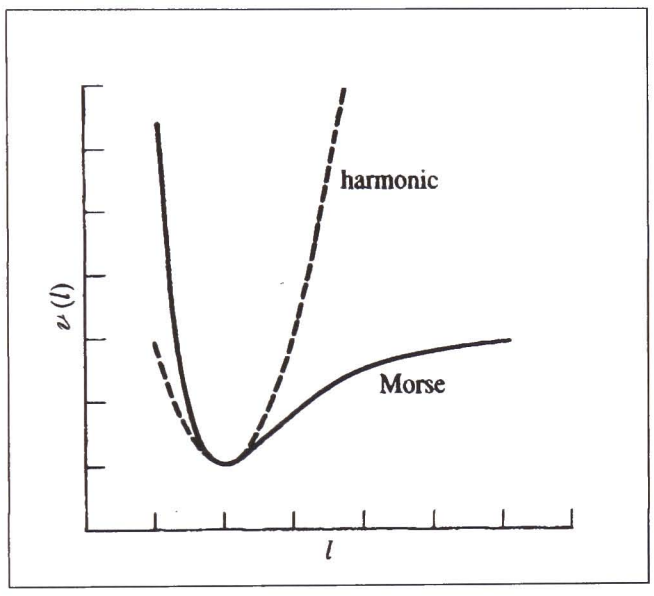


Figure 2.2 – Comparison of the harmonic potential with the Morse curve to show variation in bond energy (y axis) with interatomic separation (x axis) (taken from Leach, 2001).

The deviation of angles from their equilibrium values is also modelled by a harmonic potential (Equation 2). This equation is adequate for most applications.

$$E_{angle} = \sum k_\theta \left(\theta - \theta_{eq}\right)^2$$

Equation 2 – Harmonic potential describing the bond energy. k_θ is the force constant, θ is the bond angle and θ_{eq} is the equilibrium bond angle.

The final bonded term is the dihedral or torsional term. The dihedral term is more complicated than the other bonded terms and is calculated using a series of cosine functions that describe the energy as a function of rotation angle around a given bond (Equation 3). This energy function reproduces the preference for either staggered or eclipsed structures.

$$E_{dihed} = \sum V_n \left[1 + \cos\left(n\omega - \gamma\right)\right]$$

Equation 3 – Cosine function describing the dihedral energy. V_n is the barrier height, n is the multiplicity, ω is the torsion angle and γ is the phase factor.

One point to note about these bonded terms is that any electrostatic or van der Waals effects from 1-2 and 1-3 interactions are assumed to be included in the bond and angle terms and not in the non-bonded terms. However, with 1-4 interactions, the electrostatics and van der Waals effects must be calculated in their own right to give a full representation of the rotation about a dihedral angle (Cheatham & Kollman, 2000).

2.1.1.3 Non-bonded terms

Non-bonded terms, as their name suggests, do not depend on a bonding relationship between atoms, they are "through space" interactions. The non-bonded interactions in forcefields are generally represented by van der Waals and electrostatic interactions.

The van der Waals interaction is usually described by a Lennard-Jones potential that calculates the variation in energy upon interatomic separation. The Lennard-Jones potential is characterised by an attractive term and a repulsive term (Figure 2.3).

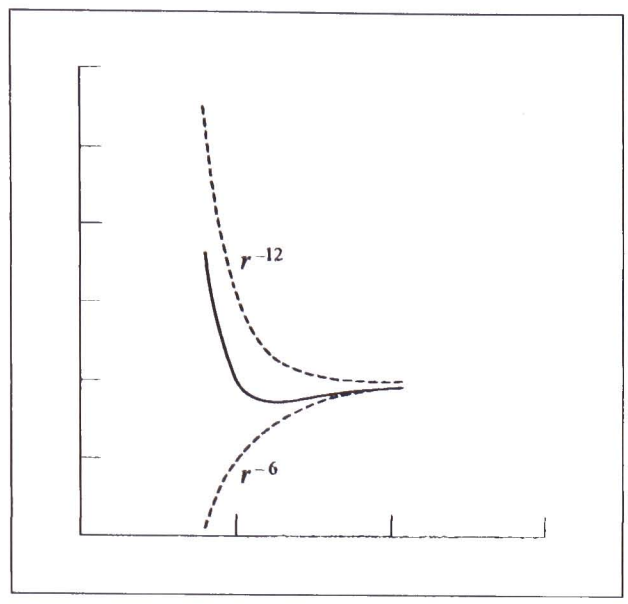


Figure 2.3 – The Lennard-Jones potential, constructed by an attractive component (r^6) and a repulsive component (r^{12}), shows variation in energy (y axis) with separation (x axis) (taken from Leach, 2001).

The attractive term is accounted for by an r^6 component and the repulsive term by a r^{12} component (Equation 4).

$$E_{vdw} = \sum_{ij} \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}$$

Equation 4 – Non-bonded van der Waals energy calculated using the Lennard – Jones 6-12 potential. *A* and C are constants based on the atoms involved and r_{ij} is the interatomic separation for the atom *i* and atom *j* pair.

The constants *A* and *B* are calculated for each atom pair via two parameters; the collision diameter at which the energy passes through a minimum (r_m) and the well depth (ε). The equations relating the constants to these parameters are $A = \varepsilon r_m^{12}$ and $B = 2\varepsilon r_m^{6}$. This potential can be rapidly calculated which is essential when dealing with large molecules with a large number of van der Waals interactions.

The remainder of the non-bonded interaction is accounted for by the electrostatic term. Areas of differing electronegativity within a molecule give rise to an uneven distribution of charge within that molecule. This charge distribution is represented by partial atomic charges designed to reproduce

the electrostatic properties of the molecule. The electrostatic interaction between two molecules is then the sum of the interaction between pairs of partial charges calculated using Coulomb's law. By knowing all the partial charges (q) and their positions (r) the electrostatic interaction can be computed (Equation 5).

$$E_{elec} = \sum_{ij} \frac{q_i q_j}{4\pi\varepsilon r_{ij}}$$

Equation 5 – Non-bonded electrostatic potential calculated via Coulomb's law. q_i and q_j are the partial charges of the atoms involved, r_{ij} is the interatomic separation (of positions r_i and r_j) and ε is the dielectric constant.

Partial charges are normally derived from *ab initio* calculations of molecular electrostatic potential (the RESP program (Cieplak *et al*, 1995) within the AMBER6 package (Case *et al*, 1999) has been developed to simplify this process). There are several ways of calculating electrostatics as will be discussed later.

2.1.1.4 The forcefield parameters

Molecular mechanics calculations require a set of parameters to be able to model a system and these are known as the forcefield parameters. Each atom in the system being modelled is assigned an atom type depending on the atom and what it is bonded to, for example, hydrogen bonded to an sp^3 carbon will have a different atom type to hydrogen bonded to an sp^2 carbon. These atom types are then used to describe the bonded and non-bonded interactions within the system. Taking the previous example of hydrogen bonded to sp^2 or sp^3 carbon, the bond length and force constant will be different for each of the different bonds therefore need to be parameters describing both types of bond. Parameters need to be set within the forcefield for every different bond type, angle type, dihedral type, van der Waals interaction and electrostatic interaction for the system being modelled.

The Amber98 forcefield (Cornell *et al*, 1995) has parameters defined for use in modelling biological molecules such as proteins and nucleic acids.
Once the forcefield parameters and equation have been established, these can then be used to generate molecular structures and obtain structural and energetic data for these structures. Two molecular mechanics methods commonly used for this are energy minimisation and molecular dynamics.

2.1.2 Energy Minimisation

Energy minimisation is used to find the lowest energy conformation for a molecule. There can be many different conformations of the same molecule and each will have a different energy. Conformations with steric clashes etc. tend to have high energies and therefore the conformation with the lowest energy is likely to give the most stable state. This variation in energy is known as the "potential energy surface", this surface can have many peaks and troughs but it is the energy at the bottom of the lowest trough that we are most interested in and this is known as the "global energy minimum" (Figure 2.4). The global energy minimum is very difficult to find as most minimisation algorithms can only find the nearest energy minimum to the starting structure. To get out of one trough and into another requires energy to be put into the system and will be discussed further in the section on Molecular Dynamics. There are two methods generally used in energy minimisation, these are the steepest descent method and the conjugate gradient method.

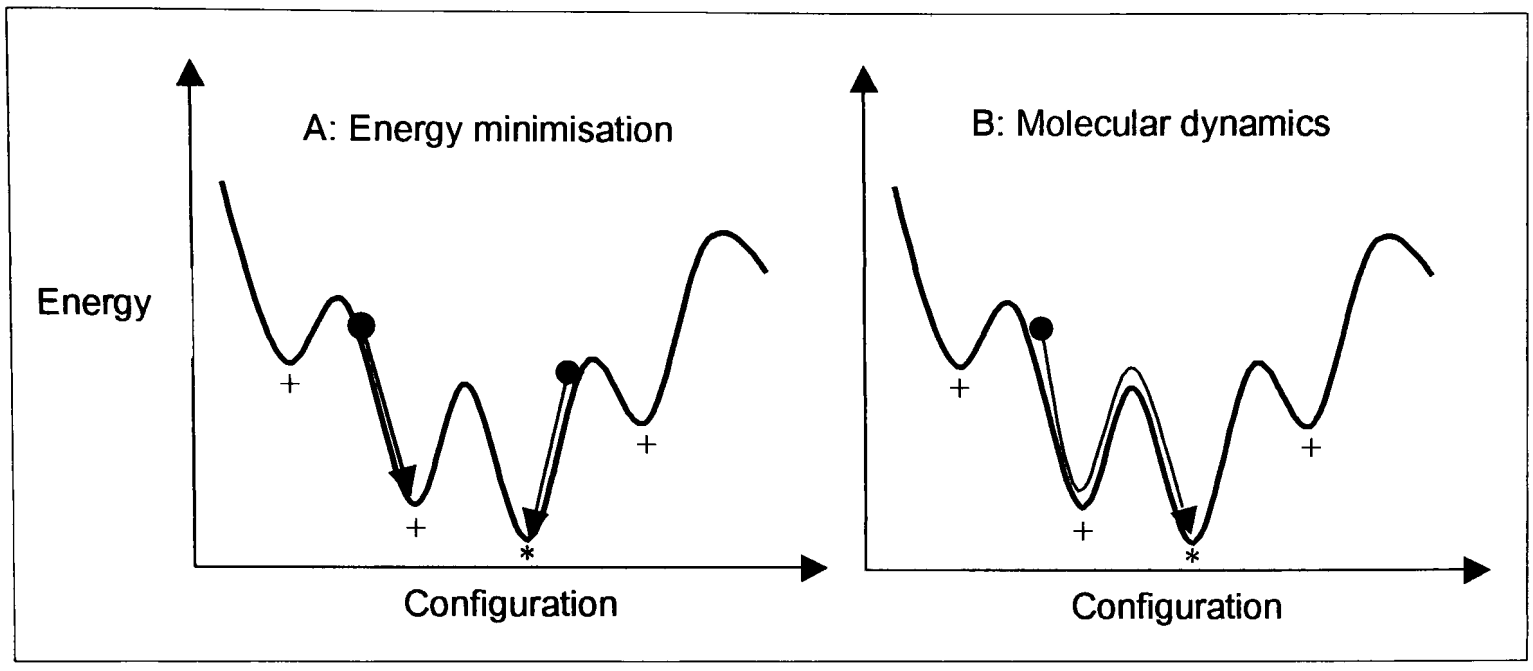


Figure 2.4 - In energy minimisation (A), different initial guesses (•) for the true structure may be optimised to local (+) rather than the global (*) energy minimum. Using molecular dynamics (B) the global energy minimum may be more reliably identified (Figure courtesy of Charles Laughton).

The steepest descent method, as its name suggests, finds the steepest direction on the energy surface to move down. A step down the slope is taken and if the energy decreases the step is increased by a factor until the energy starts to increase. A smaller value for the step length is then taken and the search continues until a minimum is found, although convergence around the minima is slow (Hirst, 1990). The conjugate gradient method is similar to the steepest decent in that the first step is equivalent. Subsequent steps are performed along a line that is a mixture of the current negative gradient and the previous search direction (Jensen, 1999) i.e. a conjugate gradient. This provides much better convergence when approaching a minimum although, as the previous gradient is stored to be used in the next calculation, it does require more computer power. It is therefore common to use both these methods when carrying out energy minimisation, steepest descent is used first as a quick route down the energy well and then conjugate gradient is used to home in more precisely on the minimum.

2.1.3 Molecular dynamics

Molecular dynamics (MD) is used to study the dynamical motion of a system over time. Successive conformations of a system are generated by integrating Newton's equations of motion, resulting in a trajectory that specifies how the positions and the velocities of all the atoms within the system vary with time. Each atom in the system is considered to be a point mass whose motion is determined by the forces exerted upon it by all the other atoms in the system, as described by Newton's first law, "A body continues to move in a straight line at constant velocity unless a force acts upon it" (Leach, 2001). The change in force between two atoms or molecules is calculated continuously with their separation. This continuous nature requires the equations of motion to be integrated by breaking the calculation down into a series of very short time steps (dt), typically 1 femtosecond. At each step the forces on the atoms are calculated (via solving the differential equations of Newton's second law, $F=ma$; Equation 6) and combined with the current positions and velocities to generate new positions and velocities as short time ahead. The force acting upon each atom is assumed to be constant during the time interval. The atoms are then moved to the new positions, an updated set of forces is computed and so on. This is how an MD simulation creates trajectories that describe how the particles in the system vary with time.

$$\frac{d^2x_i}{dt^2} = \frac{F(x_i)}{m_i}$$

Equation 6 – The motion of a particle of mass m_i, along a co-ordinate x_i, with $F(x_i)$ being the force on the particle in that direction.

2.1.3.1 Time steps

The size of the time step used to create the trajectory is very important as the smaller the time step used the better the approximation of the trajectory. A small time step however, means that more steps are required to propagate a

system of given time, i.e. the computational costs increase inversely with the size of the time step (Jensen, 1999). The size of the time step required depends upon the timescale of the motions under investigation. A typical time step is 1 fs because this is the smallest time scale of motions within a molecule and corresponds to high frequency bond stretching vibrations. In the simulation of biological macromolecules, the algorithm SHAKE (Ryckaert *et al*, 1997) is often used which constrains the bond lengths within a system so that a faster 2 fs time step can be used. In large systems such as proteins and nucleic acids, eliminating bond vibrations does not affect the integrity of the system but when angle-bending motions are constrained the integrity is no longer maintained (Schlick *et al*, 1997). The use of SHAKE therefore is restricted to bond lengths only but does reduce the computational costs required for simulations.

2.1.3.2 Periodic boundary conditions

It is commonplace in biological MD simulations to immerse a solute in a solvent box to represent the solvated environment of the molecule. In doing so one can also implement periodic boundary conditions where an infinite lattice is formed using this solvated box. To realistically model a solution, several hundreds of water molecules are needed, although this can lead to outer solvent molecules boiling off into space and to surface effects (Jensen, 1999). Periodic boundary conditions are used to overcome these effects. In the course of the simulation, as the molecule in the original box moves, its periodic image in all the other boxes moves in exactly the same way. If a molecule (or part of a molecule) leaves the original box, one of its periodic images will enter the box through the opposite face (Allen &Tildesley, 1987), therefore bulk properties ensue where there are never "outer molecules" and there is never a surface with which to interact.

2.1.3.3 Treatment of electrostatic interactions

The treatment of electrostatic interactions is of great importance when simulating highly charged systems such as nucleic acids. The polyionic nature of a DNA backbone, and its immersion in a solvent box, requires a particularly accurate treatment of the electrostatics especially long-range forces. Prior to 1995, simulations of DNA were plagued by instabilities caused by inaccurate truncation methods for calculating the Coulomb potential. These truncation methods used a finite cut-off distance of around 10-12Å, after which the interaction energy between non-bonded atoms was considered to be zero (Louise-May *et al*, 1996). Truncation was used, as it required less computational cost than the more accurate lattice sum methods.

Lattice sum methods can be used to calculate all non-bonded interactions by imposing a crystal-like periodicity to the MD system (similar to periodic boundary conditions). The Ewald summation method (Ewald, 1921) is one such method and works by splitting the interaction into short- and long-range contributions separated by a cut-off. The short-range term is evaluated directly while the long-range term is calculated in reciprocal space (Jensen, 1999). The related methods Particle Mesh Ewald (PME; Darden *et al*, 1993) and Particle-Particle Particle-Mesh (PPPM; Hockney & Eastwood, 1998) use the Ewald summation but are much faster due to the use of Fourier methods to calculate a smoothly varying long-range term (Sagui & Darden, 1999). It is these faster methods that lead to stable nucleic acid trajectories of nanosecond proportion in 1995 (Cheatham *et al*, 1995) and are still used in present day calculations.

The use of PME imposes a quasi-crystalline periodicity that could cause artificial stabilisation of the simulation system. This effect will depend upon the size of the periodic box, so that a box that is too small would produce crystal like behaviour and a box that is too large would expend unnecessary computational effort. Simulations have been performed on the DNA sequence $d(CGCGAAAAAACG)_2$ (Norberto de Souza & Ornstein, 1997) in

solution with differing sizes of solvent box (5, 10 and 15Å from the solute to the edge of the box). The root mean squared deviation (RMSD) of these three systems from the initial co-ordinates was monitored to detect if the flexibility of the system was altered by any crystal packing effects. No significant differences were found between the systems; therefore concluding that a minimum 5Å water layer is acceptable to allow normal dynamical behaviour of DNA in solution, although a cautionary 10Å layer is recommended as other systems may behave differently to the one used in the study.

2.1.3.4 Validation of MD methods for use on nucleic acids

To critically assess the feasibility of the MD program AMBER, and its related forcefield, for the study of DNA, Young *et al*, (1997) carried out a number of simulations, up to 5ns in length. The key issues they were concerned with were; treatment of boundary conditions, electrostatics, initial placement of solvent and run lengths. They chose the established Dickerson & Drew (1981) dodecamer d(CGCGAATTCGCG)$_2$ as a test model as it is often used for benchmarking systems in experimental and theoretical studies. The MD results showed a dynamically stable structure of B-form DNA, which compared favourably with crystallographic and NMR studies of d(CGCGAATTCGCG)$_2$. Molecules of solvent were mobile and able to gain access into the minor groove, reproducing the "spine of hydration". DNA bending was seen, with this bending supporting the "junction model" (for more details see Chapter 3), one of two mechanisms for bending proposed from experimental studies. Also sequence effects on groove widths were reproduced. Their findings show that accurate all-atom MD, stable on the ns timescale, using explicit solvation can give accurate descriptions of the DNA in solution and produce ideas about the nature of dynamic structure and solvation.

Many other simulations have been carried out to validate the use of MD, some of which are discussed in the next chapter. Others can be found in

reviews on MD studies of DNA by Cheatham & Kollman (2000) and by Cheatham & Young (2001).

Recently the forcefield parameters of AMBER have been updated (Cheatham *et al*, 1999) to more accurately represent sugar pucker phases and helical repeat. These modifications lead to improved agreement with experimental data.

2.1.3.5 Implicit solvation models

The solvent environment plays an important role in molecular structure and dynamics, for example, DNA duplex formation (and the formation of higher ordered DNA structures) and its subsequent structural type (e.g. B-form) is dependent on the ionic strength of the surrounding solvent. It is therefore important when simulating a biological molecule that the solvent is correctly described. Simulations normally use explicit solvation where every atom in the system (including all solvent atoms) is described. This tends to make computational costs high because of the need for accurate long-range electrostatic calculations. Implicit solvation models have been developed which treat the solvent as a continuum electrostatic model and is therefore a much faster method, as there are no long-range interactions to consider. The Generalised Born (GB; Still *et al*, 1990) model is one such model and has been implemented in the AMBER modelling package (Case *et al*, 1999) with an additional solvent accessible surface area term (SA) that can be turned on or off, as needs suffice (see below). The solvation free energy (G_{sol}) can then be described as a combination of a solute-solvent electrostatic polarization term (G_{pol}), a solvent cavity term (G_{cav}), and a solute-solvent vdW term (G_{vdW}) as in equation 7 (Still *et al*, 1990).

$$G_{sol} = G_{pol} + G_{cav} + G_{vdW}$$

Equation 7 – The GBSA equation describing solvation free energy.

The first G_{pol} term denotes the electrostatic contribution and can be calculated via equation 8, the GB equation.

$$G_{pol} = -\frac{1}{2}\left(1 - \frac{e^{-\kappa f_{GB}}}{\varepsilon}\right)\sum \frac{q_i q_j}{f_{GB}}$$

Equation 8 – The generalised Born equation, where κ is the Debye-Huckel screening parameter, f_{GB} depends upon effective Born radius (a_i) and the interaction distance between atoms (r_{ij}), ε is the dielectric constant and q_i and q_j are the partial charges.

The last two terms (G_{cav} and G_{vdW}) are associated with the formation of a cavity within the solvent and are related to SA by equation 9. If the free energy difference being computed is between two systems of very similar surface, the SA term can be switched off to reduce calculation time.

$$G_{cav} + G_{vdW} = \sum \sigma_k SA_k$$

Equation 9 – Equation describing the formation of a cavity within solvent, where k is the atom type and σ is a surface tension term.

The GBSA model has been used to simulate macromolecules such as DNA and proteins with reasonable results (Jayaram *et al*, 1988; Tsui & Case, 2001) and speed compared with explicitly solvated systems. Simulations of DNA can reproduce structural aspects quite well, in particular the transition from A-form to B-form DNA in low salt conditions (Tsui & Case, 2000), although a recent study within our group (Sands & Laughton, 2003) has shown that dynamical and especially thermodynamical properties are not always as well reproduced.

2.2 Analysis methods

2.2.1 Principal Component Analysis

Molecular dynamics simulations provide a vast amount of data which needs to be "mined" to extract relevant information. A method for extracting this information is Principal Component Analysis (PCA). PCA for the analysis of biomolecular simulations, in particular, was developed by the Berendsen group (Amadei *et al*, 1993) and has since proved to be a powerful tool in studying conformational behaviour in nucleic acids and proteins. It is a statistical method for analysing MD trajectories and can be used to find the components that make up the greatest overall contribution to the motion within a trajectory. It can also be used to obtain entropies via the Schlitter method and to gain insights into how similar trajectories of the same system are via the overlap method of Hess (see later).

A 3N x 3N covariance matrix of the Cartesian co-ordinates is generated from an MD simulation, then diagonalised to give 3N eigenvectors. The eigenvectors provide a vectorial representation of each component of structural deformation (Sherer *et al*, 1999), i.e. they indicate the direction of motions of the atoms. Each eigenvector has a corresponding eigenvalue that indicates the relative contributions made by the component to the motion as a whole. The eigenvectors associated with the highest eigenvalues can be selected for further analysis.

Projections of the trajectory along the major eigenvectors give useful information on equilibrium and conformational sampling (Wlodek *et al*, 1997). To ease the interpretation of the deformations associated with each principal component, short MD trajectories (animations) can be generated artificially by generating structures in which the projections vary linearly between the minimum and maximum values observed (Sherer *et al*, 1999). The resulting animations can be inspected visually using VMD (Humphrey *et al*, 1996).

2.2.1.1 PCA Overlap

PCA of a trajectory produces a description of the essential quasi-harmonic modes of deformation (eigenvectors) of the structure, as discussed above. Motion along each eigenvector relates to an orthogonal way in which the structure samples configurational space. If two separate simulations of a system are dynamically equivalent, they must sample this space the same way. This can be achieved if they have the same eigenvectors, or linear combinations of each other. For DNA the first ten eigenvectors from all simulations capture the majority of the important modes of flexibility of the system and so a comparison of these is sufficient (for proteins more may be required; Charles Laughton/Daniel Warner, personal communication). To quantify this comparison we used the eigenvector overlap measure of Hess (2000), calculated as the normalised sum of the dot products between a given number of eigenvectors from two separate simulations (Equation 10).

$$Overlap(a,b) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i.b_j)^2$$

Equation 10 – The Hess overlap equation, where a and b are the eigenvector matrices to be compared and n denotes the number of eigenvectors.

2.2.1.2 Calculation of entropy via the Schlitter method

Configurational entropies can be calculated from mass-weighted eigenvalues according to the method of Schlitter (Schlitter, 1993). The Schlitter method is an "upgrade" of the Karplus method developed much earlier (Karplus & Kushick, 1981), in which the entropy is described by the summation of the natural logarithm of the eigenvalues (Equation 11).

$$S = \frac{1}{2}k\sum \ln \lambda$$

Equation 11 – Simplified version of the Karplus equation, S is the entropy, k is the Boltzmann constant and λ is the eigenvalue.

The Karplus method leads to inaccuracies in entropy because of the small eigenvalues associated with the highest frequency modes. Although these inaccuracies do not contribute greatly to the overall dynamics of the system, they do contribute to the entropy as the logarithm becomes large and negative for these eigenvalues with values of much less than one. One possible solution to this problem was to use cut-offs, beyond which the eigenvalues were considered to be negligible. This method also leads to inaccuracies, because all the eigenvalues contribute to the entropy, the results which are obtained are biased heavily by where this cut-off is taken.

The Schlitter method differs from the Karplus method in two ways. Firstly the eigenvalues are mass weighted and secondly, to overcome the problem of large negative entropies for eigenvalues less that one, the logarithm of one plus the eigenvalue is taken (Equation 12) so that the logarithm of the eigenvalues is never negative.

$$S = \frac{1}{2} k \ln\left(1 + \frac{kTe^2}{\hbar^2}\lambda_m\right)$$

Equation 12 –The Schlitter equation, S is the entropy, k is the Boltzmann constant, T is the temperature (in K), e is the Euler number, $\hbar$ is Planck's constant/2Π and λ_m is the mass weighted eigenvalue.

Summation of the Schlitter equation for all eigenvalues therefore gives a more valid approximation of the configurational entropy of a system than the Karplus method. This has been shown in a comparison of the two methods by Schafer *et al* (2000).

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 to higher calculated configurational entropies. 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 (Equation 13).

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

Equation 13 – Function describing how to obtain the infinite entropy of a system, *S(t)* is the calculated entropy, S_{∞} is the infinite entropy, *t* is the simulation length and *n* and *A* are fitting parameters, *n* being related to the time correlation in the dynamics.

So from the fitting procedure, the entropy for a simulation of infinite length S_{∞} (in addition to the other parameters of the fit, *A* and *n*) can be estimated.

2.2.2 Linear Interaction Energy (LIE)

The Linear Interaction Energy (LIE) approach was developed by Aqvist and co-workers (1994), to provide an alternative to the time consuming Free Energy Perturbation (FEP) method for calculating relative or absolute binding affinities. In FEP the free energy is calculated for the perturbation of one molecule into another, thus the ΔG between two states is found. It is only possible to carry out FEP between molecules that differ slightly. The method requires the accumulation of ensemble averages along the perturbation path, which must be fine grained in order for the free energy to converge. In practise, drug design often deals with large molecules and derivatives that differ quite greatly and so a simplified method that could deal with these molecules was sought.

LIE is used to calculate free energy changes, using only electrostatic and van der Waals interactions, based on simulations of only two states. In the case of ΔG_{bind} the two states are the solvated ligand and the ligand bound to solvated protein/DNA. The method is based on the linear response approximation for electrostatic forces, which for polar solutions will yield harmonic free energy functions in response to changes in electric field

(Aqvist *et al*, 1994). The approximation to the electrostatic contribution of the binding free energy is shown in equation 14.

$$\Delta G^{el}_{bind} \approx \alpha \left\langle \Delta^{el}_{comp-drug} \right\rangle$$

Equation 14 – The electrostatic contribution to LIE, where α is a fitting parameter.

The value of α is determined from the knowledge that the electrostatic contribution to the solvation energy of a single ion is equal to half of the corresponding ion-solvent interaction energy. The value of $\Delta^{el}_{comp\text{-}drug}$, the difference in electrostatic interaction between the two drug environments, is obtained from simulations of the solvated drug and of the solvated drug-protein/DNA complex.

The vdW part of LIE represents the non-polar interactions and is calculated in a similar fashion to the electrostatic part. No analytical theory exists for the calculations of this part but there was indirect evidence that a similar linear treatment would work, using a different fitting parameter (Leach, 2001). Experimental free energies of solvation for various *n*-alkanes have an approximate linear dependency on the length of carbon chain. Also the mean vdW solute-solvent energies from simulations of n-alkanes showed a similar linear variation. From these observations Aqvist and co-workers hypothesised that a simple linear approximation similar to that of the electrostatics would be able to account for the non-polar binding contribution (Equation 15).

$$\Delta G^{vdW}_{bind} \approx \beta \left\langle \Delta^{vdW}_{comp-drug} \right\rangle$$

Equation 15 – The van der Waals contribution to LIE, where β is a fitting parameter.

The value of β was found by fitting the equation to experimental data, Aqvist and co-workers generated a β value of 0.161. The value of $\Delta^{vdW}_{comp\text{-}drug}$, the difference in vdW interaction between the two drug environments, is

obtained, as before, from simulations of the solvated drug and of the solvated drug-protein/DNA complex.

The two parts of the equation can then be put together (equation 16) to form the overall LIE equation.

$$\Delta G_{bind} = \alpha\left\langle \Delta^{el}_{comp-drug} \right\rangle + \beta\left\langle \Delta^{vdW}_{comp-drug} \right\rangle$$

Equation 16 – The LIE equation.

The LIE equation has been used in several studies to obtain free energies of binding, with reasonably good results (Hansson & Aqvist, 1995; Wall *et al*, 1999; to name a few). However, the transferability of the fitting parameters of Aqvist *et al* are not always successful, leading various groups to generate their own parameters (Paulsen & Ornstein, 1996; Jones-Hertzog & Jorgensen,1997). It was suggested at first that the vdW parameter β was forcefield dependent, but this was later discounted after simulations were repeated using different forcefields (CVFF, GROMOS, AMBER) and still gave the same value of β (Paulsen & Ornstein, 1996; Aqvist, 1996; Wang *et al*, 1999). Simulations on a number of differing systems lead to the conclusion that a fixed β parameter could not give results in agreement with experimental data in all cases. It was therefore suggested that the value of β is dependent on the system and its environment and if possible should be determined for each different system used. Less speculation has surrounded the other fitting parameter α as this is approximated analytically, although it has been noted by Hansson *et al* (1998) that the theoretical value of 0.5 decreases slightly as the number of hydroxyl groups increases.

2.2.3 Molecular Interaction Potential (MIP)

The calculation of the Molecular Interaction Potential (MIP) is based upon the quantum mechanical molecular electrostatic potential (MEP) with a further addition of a classical repulsion-dispersion term. This results in a method to represent accurate electrostatic interactions and also steric effects. MEP can

be described as the electrostatic component of the interaction energy between a molecule and a positive charge. Calculations of MEP give us valuable information about general topology, location and depth of MEP minima and the averaged information provided by charges derived from the fitting of the MEP and the Coulombic potential (Orozco & Luque, 1992). One of the biggest drawbacks of MEP is its inability to describe steric effects. This drawback restricts its use in molecular host-guest studies, which are popular for the screening of ligands in drug design. To overcome this, Orozco and Luque (1992) have added a Lennard-Jones type 6-12 repulsion–dispersion contribution to the total interaction energy to create the MIP (Equation 17).

$$V_{MIP}(r_1) = V_{REP}(r_1) + V_{DISP}(r_1) + V_{MEP}(r_1)$$

Equation 17 – The MIP equation, where r_1 is the position of the positive charge.

The MIP associated with the interaction of a probe molecule (of positive charge) with the time averaged structure of the subject molecule is calculated at points on a 3D grid surrounding the area of interest (binding site). Regions of negative potential describe favourable interactions between probe and subject molecule.

2.3 References

Allen M.P, Tildesley D.J, (1987), *Computer Simulation of Liquids,* Oxford University Press.

Amadei A, Linssen A.B.N, Berendsen H.J.C, (1993), *Proteins,* **17**, 412-425.

Aqvist J, Medina C, Samuelsson J-E, (1994), *Protein Eng,* **7**, 385-391.

Aqvist J, (1996), *J. Comp. Chem,* **17**, 1587-1597.

Case D.A, Pearlman D.A, Caldwell J.W, Cheatham T.E. III, Ross W.S, Simmerling C.L, Darden T.L, Merz K.M, Stanton R.V, Cheng A.L, Vincent J.J, Crowley M, Tsui V, Radmaer R.J, Duan Y, Pitera J, Massova I, Seibel G.L, Singh U.C, Weiner P.K, Kollman P.A, (1999), *AMBER 6,* University of California, San Francisco.

Cheatham T.E. III, Miller J.L, Fox T, Darden T.A, Kollman P.A, (1995), *J. Am. Chem. Soc,* **117**, 4193-4194.

Cheatham T.E. III, Cieplak P, Kollman P.A, (1999), *J. Biomol. Struct. Dyn,* **16**, 845-862.

Cheatham T.E. III, Kollman P.A, (2000), *Annu. Rev. Phys. Chem,* **51**, 435-471.

Cheatham T.E. III, Young M.A, (2001), *Biopolymers,* **56**, 232-256.

Cieplak P, Cornell W.D, Bayly C.I, Kollman P.A, (1995), *J. Comp. Chem,* **16** (11), 1357-1377.

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.

Darden T.A, York D, Petersen L.G, (1993), *J. Chem. Phys,* 98, 10089-10092.

De Souza O.N, Ornstein R.L, (1997), *Biophys. J,* **72**, 2395-2397.

Dickerson R.E, Drew H.R, (1981), *J. Mol. Biol,* **149**, 761-786.

Ewald P, (1921), *Ann. Phys,* **64**, 253-287.

Hansson T, Aqvist J, (1995), *Protein Eng,* **8**, 1137-1144

Hansson T, Marelius J, Aqvist J, (1998), *J. Comput.-Aided Mol. Des,* **12**, 27-35.

Hess B, (2000), *Phys. Rev. E*, **62**, (6) 8438-8448.

Hirst D.M, (1990), *A Computational approach to Chemistry,* Blackwell Scientific Publications.

Hockney W.G, Eastwood J.W, (1998), *Computer Simulations using Particles,* Adam Hilger, New York, NY.

Humphrey W, Dalke A, Schulten K, (1996), *J. Mol. Graphics,* **14** (1), 33-38.

Jayaram B, Sprous D, Beveridge D.L, (1998), *J. Phys. Chem,* **102**, 9571-9576.

Jensen F, (1999), *Introduction to Computational Chemistry,* John Wiley & Sons Ltd.

Jones-Hertzog D.K, Jorgensen W.L, (1997), *J. Med. Chem,* **40**, 1539-1549.

Karplus M, Kushick J.N, (1981), *Macromolecules,* **14**, 325-332.

Lafontaine I, Lavery R, (1999), *Curr. Opin. Struct. Biol,* **9**, 170-176.

Leach A.R, (2001), *Molecular Modelling, Principles and Applications,* 2nd edition, Pearson Publication Limited.

Louise-May S, Auffinger P, Westhof E, (1996), *Cur. Opin. Struct. Biol,* **6**, 289-298.

Orozco M, Luque F.J, (1992), *J. Comp. Chem,* **14**, 587-602.

Paulson M.D, Ornstein R.L, (1996), **9**, 567-571.

Ryckaert J.P, Ciccotti G, Berendsen H.J.C, (1997), *J. Comp. Phys,* **23**, 327-341.

Sagui C, Darden T.A, (1999), *Annu. Rev. Biophys. Biomol. Struct,* **28**, 155-179.

Sands Z.A, Laughton C.A, (2003) *Manuscript submitted to J. Am. Chem. Soc.*

Schafer H, Mark A.E, van Gunsteran W.F, (2000), *J. Chem. Phys,* **113**, 7809-7817.

Schlick T, Barth E, Mandzuik M, (1997), *Annu. Rev. Biomol. Struct,* **26**, 181-222.

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.

Still W.C, Tempczyk A, Hawley R.C, Hendrickson T, (1990), *J. Am. Chem. Soc,* **112**, 6127-6129.

Tsui V, Case D.A, (2000), *J. Am. Chem. Soc,* **122**, 2489-2498.

Tsui V, Case D.A, (2001), *Biopolymers,* **56**, 275-291.

Wall I.D, Leach A.R, Salt D.W, Ford M.G, Essex J.W, (1999), *J. Med. Chem,* **42**, 5142-5152.

Wang J, Dixon R, Kollman P.A, (1999), *Proteins Struct. Func. Gen,* **34**, 69-81.

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

Young M.A, Ravishanker G, Beveridge D.L, (1997), *Biophys. J,* **73**, 2313-2336.

CHAPTER 3 – THE VALIDATION OF LAMMPS FOR MOLECULAR DYNAMICS SIMULATIONS OF DNA

3.1 Review of DNA dynamics

DNA is not a static molecule, but highly flexible and dynamic. Subtle changes in the flexibility, and therefore shape, of DNA are important in the recognition process, especially for sequence specific recognition and it is because of this that the study of DNA dynamics, and not just structure, is so important. If we can begin to understand the dynamical behaviour of DNA we can begin to understand more about the complex nature of its functions.

While there are many biophysical experiments available to study structural properties of DNA, there is no experimental technique capable of generating a complete description of the dynamical behaviour of DNA. For this we need to turn to theoretical techniques such as Molecular Dynamics (MD) described in Chapter 2. MD can provide a theoretical description of both DNA structure and dynamics and, as such, is a useful tool not only for developing DNA models but also for helping to interpret experimental data.

3.1.1 The advent of MD as a tool for studying DNA dynamics

The first MD study of DNA was reported back in 1983 (Levitt, 1983) and since then the field has grown rapidly. Early studies of DNA did not include the solvent nor ionic environment, although the importance of water and counter-ions were recognised as an integral part of DNA structure (Westhof, 1988). The inclusion of the ionic environment did not arrive until better methods for the treatment of long-range electrostatics were available (see previous chapter). The treatment of electrostatics is crucial in the study of DNA because of its charged sugar-phosphate backbone. Also if a piece of DNA is to be solvated in a true representation of its environmental

surroundings (water and counter-ions), the long-range electrostatics must also be taken into account to obtain a stable structure.

Before ca. 1995, simulations of DNA were plagued by instabilities mainly due to the application of approximate methods that lacked the ability to correctly represent highly charged systems. Simulations of proteins with explicit solvent were already rather robust and reliable at this time (very few proteins have an overall charge of any magnitude) but simulations of DNA were characterised by distortion of duplex structures, broken base-pairing and misrepresented sequence specific fine structures (Cheatham & Kollman, 2000). Simulations of this era were also rather short (<500ps), which as we know today is not always long enough to reach equilibrium.

Since 1995, DNA simulations have become much more reliable. With more computing power than ever before, better forcefields (Cornell *et al*, 1995; MacKerell *et al*, 1995) and more accurate methods available especially for calculating long-range electrostatics (PME see previous chapter), it is now feasible to obtain stable DNA structures and dynamics to timescales of tens of nanoseconds (Auffinger & Westhof, 1998; Cheatham & Kollman, 2000; Beveridge & McConnell, 2000; Cheatham & Young, 2001). The reliability of these codes has been shown by the stability of simulations to longer time scales, reproduction of experimental data and the reproducibility of the simulation results.

3.1.2 Reproduction of experimental data via MD

The reproduction of experimental data is important not only to show the reliability of the codes and forcefields, but also to show the reliability of the method itself. The examples below show that MD, as a method, can reproduce experimental data including the topics of environmental effects (A-form – B-form transitions), structural flexibility (A-tract bending) and thermodynamic analysis (Co-operativity in DNA binding).

3.1.2.1 Example: A-form – B-form transitions

The major structural family of DNA involves the right-handed A- and B-form structures. There is a 5.7Å RMSD difference between canonical A- and B-forms with the major differences relating to sugar pucker, the angle between base pairs and the helix axis, the rise between base pairs (leading to an end-to-end length of ~30Å for B-form and ~23Å for A-form for a decamer) and the width of the minor groove (larger in A-form than B-form). The base parameters (see section 1.1.3.2) that have the largest deviations between forms are x-displacement (B-form = -1.0-0.0Å, A-form = -6.0-5.0Å) and inclination (B-form ~-6.0°, A-form >19°; Saenger, 1984).

It has been known for many years that DNA can adopt different forms depending of its environmental surroundings. The effect of the environment on DNA ranges from global changes based on the solvent and ionic concentrations leading to conversions between A- and B-form structures, to more local structural effects caused by changes in the helical parameters. It was this A-form to B-form transition which was used by Cheatham and Kollman to show that the Cornell *et al.* (1995) forcefield could reproduce the experimental data which showed that B-form structures are more stable at low salt and high humidity than A-form (Cheatham & Kollman, 1996).

Cheatham and Kollman carried out four unrestrained nanosecond length simulations in aqueous solution on the duplex $d(CCAACGTTGG)_2$, with the Cornell *et al.* (1995) forcefield. Two of the simulations were started from a canonical A-form structure and two from a canonical B-form. RMSD calculations showed that the average structures from the four trajectories converged to with 0.8-1.6Å of each other and to 3.1-3.6Å from the B-form X-ray structure reported for this sequence (Prive *et al*, 1991). The A-form to B-form transition takes place, in the simulations, on an approximately 500ps timescale (Experimentally, there are conflicting reports on the timescale of these transitions, studies on DNA films and fibres show that the transition can take hours or days (Szabo *et al*, 1996) whereas other experiments show the transition is rapid and highly reversible; Piskur & Rupprecht, 1995). At the

time this was the only demonstration of an A-B transition occurring during unrestrained MD, but since then a number of other studies have shown similar results (Cieplak *et al*, 1997; Miller & Kollman, 1997; Shields *et al*, 1997).

3.1.2.2 Example: A-tract bending

Curvature of DNA is an important topic in structural biology because it is this curvature, or bending, that allows DNA to bind to many of the proteins involved in its everyday functions, for example chromatin. Chromatin complexes are involved in the packaging of DNA into chromosomes. They are nucleoproteins involving ~145 base pairs of DNA wound around a histone protein core.

Some sequences are known to have an inherent tendency to bend, even when no protein is present, e.g. A-tract sequences. A-tracts involve runs of adenines on one strand and thymines on the other. The first observation of naturally occurring curved DNA was seen by Marini *et al* (1982) in kinetoplast DNA). This sequence had extensive repeats of A-tracts within it and further studies showed that bending occurred when A-tracts were repeated in phase with the helical repeat itself (Crothers *et al*, 1990) i.e. every 10-11 base pairs. This phasing is critical as it forces the A-tract to be on the same side of the helix all the time therefore allowing the small individual A-tract bends to magnify to a large overall bend. Gel studies demonstrate that A_6 A-tracts are bent by 17-21° in the direction of the minor groove (Koo *et al*, 1990).

The precise nature of this bending is not fully understood and has led to a number of models being put forward. The two main models are the junction model and the wedge model (Figure 3.1).

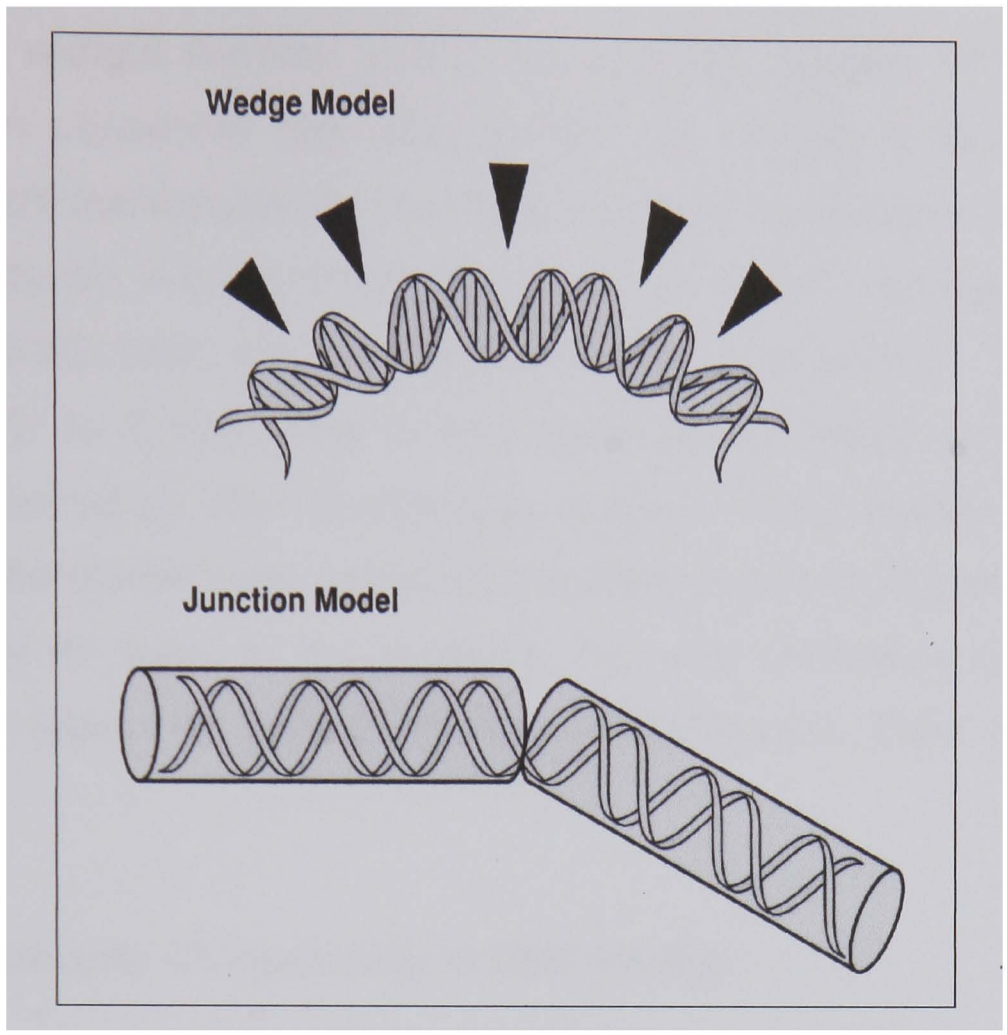


Figure 3.1 – Pictorial representations of the junction and wedge models (taken from Sinden, 1994).

The junction model postulates that bending is caused by two helical segments (A-tract and normal B-form) with differing base pair angular orientations meeting at a "junction". This abrupt change in structure leads to the helix axes being at an angle to each other, thus creating a bend. Therefore in this model the A-tract itself is straight and the bending occurs at the junctions. In contrast the wedge model proposes that a smooth global bend occurs due to small additive "wedges", composed of roll and tilt components of independent base pair steps. Therefore, in the wedge model the A-tracts themselves are bent. Several experiments have been carried out on these models and both have been predicted to be correct (Haran *et al*, 1994 and references therein).

Molecular modelling studies have been carried out to try to understand further the nature of DNA curvature and bending (for a review see Olsen & Zhurkin, 2000). To investigate A-tract bending, nanosecond simulations were carried out for a 25 base pair duplex with two successive A-tracts spaced by 11 base pairs (a full helix turn of B-form DNA). A control

sequence without A-tracts (composed of three repeats of the of *BamH1* recognition sequence) was also carried out (Young & Beveridge, 1998). Results from the simulations show axis bending to the extent of 16.5° per A-tract compared with experimental values of 17-21°, leading to an overall bend of 26-30° compared with 18° for the control sequence. The model also shows a 5' to 3' narrowing of the minor groove region of the A-tracts, a feature inferred by DNA footprinting studies. Other studies carried out in several laboratories have supported the idea of straight A-tracts with bending most likely to occur at the junctions, but also extending into the general sequence region as well (Beveridge & McConnell, 2000 and references therein).

3.1.2.3 Example: Co-operativity in DNA binding

In recent years it has become increasingly clear that some of 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*, 1997). And while it has been assumed that the major contribution to the entropy changes that accompany drug-DNA recognition comes from solvent reorganisation, it has been recently shown (Harris *et al*, 2001) 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.

It has been shown by combining NMR structure determination methods with extended molecular dynamics 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 is related to associated changes in minor groove width which 'clamp' the ligand in position.

More recently, taking advantage of the development of methods to quantify configurational entropy changes, a full thermodynamic analysis of how Hoechst 33258 binds to the DNA duplex d(CTTTTGCAAAAG)$_2$ (Harris *et al*, 2001) has been carried out. NMR titration experiments had shown that this DNA duplex binds two molecules of Hoechst (Figure 3.2), 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).

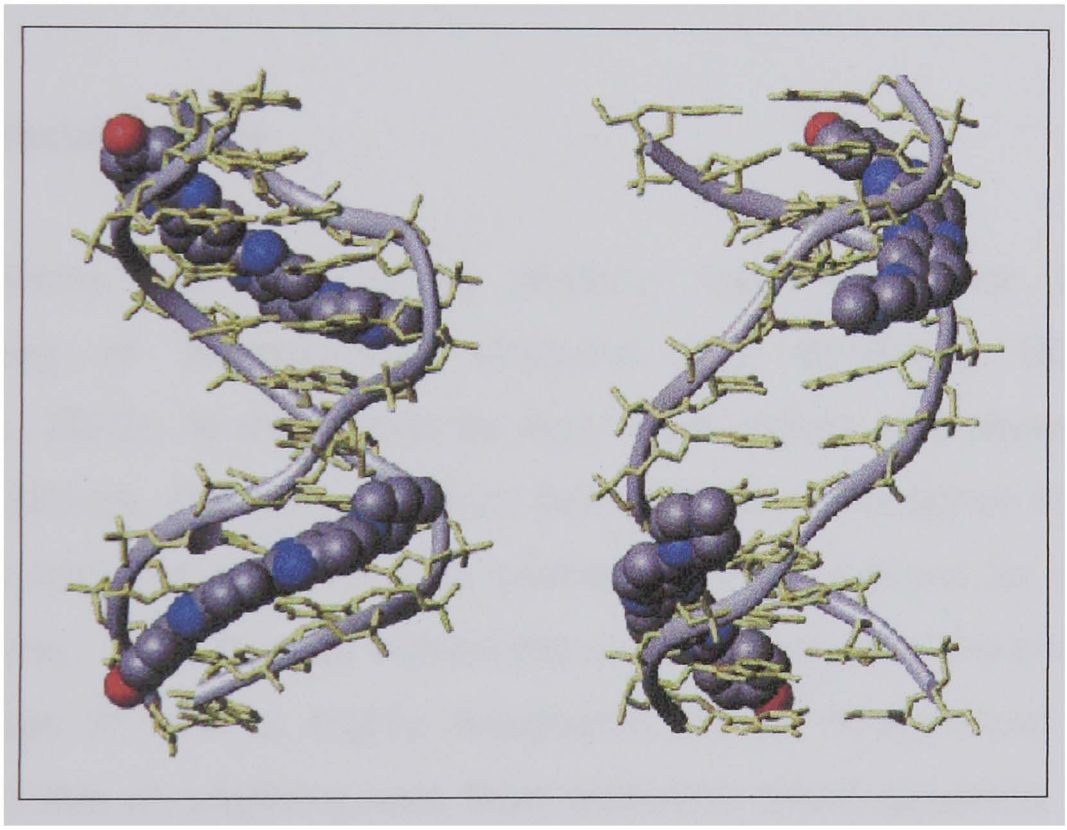


Figure 3.2 – Two views of the structure of the 2:1 complex of Hoechst 33258 bound to d(CTTTTGCAAAAG)$_2$ (taken from the NMR data).

Through a series of extended molecular dynamics simulations, it was shown that this co-operativity 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 calculations of this type is critically dependent on the possibility of obtaining extended molecular dynamics trajectories. It was found 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 a limit can be estimated quite reasonably by curve fitting, the process requires considerable extrapolation beyond currently accessible simulation timescales, 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 CPU time, this is not a realistic option for day-to-day studies.

3.1.3 Timescale issues

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) for increasing lengths of time. The rationale behind this 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 (Hawkins *et al*, 1996; Weiser *et al*, 1999) are now showing some promise.

The desire for extended simulations is driven by a number of factors. Firstly, MD simulations are often used as a form of structure optimisation, 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 minimisation 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). 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 the MD, and for large systems with important low-frequency modes of conformational flexibility, simulations of many nanoseconds at least may be required (*cf* the protein folding problem itself).

Many of the important conformational changes that DNA can undergo take place on the micro- to millisecond timescale (Table 3.1; Yakushevich, 1998) – for example the 'breathing' of the bases (see example below). A simulation of this length (1 microsecond) has been carried out of a protein folding intermediate in explicit solvent (Duan & Kollman, 1998), but this is the only one of its kind. Long simulation times (50-160ns; Daggett, 2000) have been observed for other peptides and small proteins but bigger and longer simulations of DNA are much more difficult to obtain with present codes. This is due to the need for more accurate solvation and electrostatic calculations, as already discussed, as these calculations use the most computer time and power. To obtain these types of simulation for DNA we will have to wait for technological advances to catch up with our requirements.

Time scale	Main types of internal motion
Picosecond	Short living motions and oscillations of atoms.
Nanosecond	Oscillations of small groups of atoms: sugars, phosphates, bases; bending and twisting of the double helix.
Microsecond	Winding and unwinding of the double helix; opening of base pairs.
Millisecond	Dissociation of the double helix; super helicity; overall rotation.
Second	Writhing; isomerisation; division of bacteria.

Table 3.1 - The time scales of the internal mobility of DNA (adapted from Yakushevich, 1998).

3.1.3.1 Example: Base pair breathing events

Base pair breathing occurs when 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 recognised by a DNA-binding protein. Currently, we cannot simulate this spontaneous process at the atomic level because of timescale issues. The rate-limiting step of base pair opening is partially related to the breaking of Watson-Crick hydrogen bonds. It is therefore expected that DNA containing base pairs without these stabilising hydrogen bonds will open more frequently. When thymine is replaced by its non-polar homologue difluorotoluene (F) the duplex is only minimally perturbed and NMR studies show that the modified base pair (Figure 3.3) forms, remains stacked and standard B-form DNA is maintained (Guckian *et al*, 1998). Studies in chloroform show no evidence that difluorotoluene forms hydrogen bonding interactions with adenine (Moran et al, 1997). Cubero *et al* (1999) support these findings and show that in a 10ns simulation of the duplex d(CTTTC**F**TTCTT)·d(AAGAAAGAAAG), numerous base pair opening and closing events occurred.

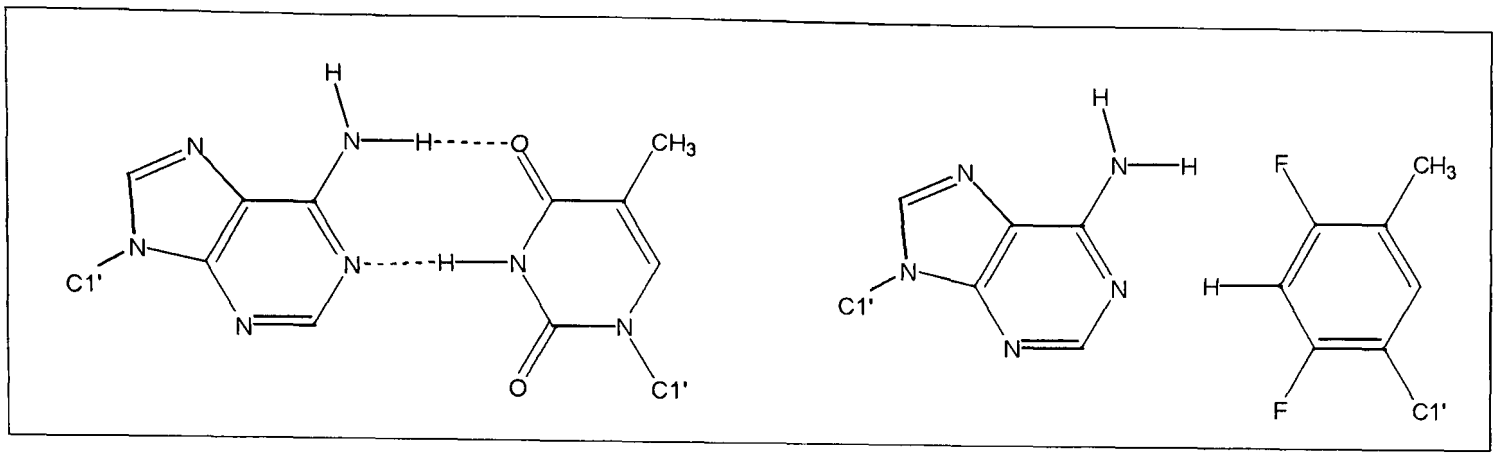


Figure 3.3 – Structures of standard AT (left) and its homologue AF (right).

Other methods of studying the phenomena of base pair opening include biasing the simulation to force the breathing event (Varnai & Lavery, 2002) or moving to non-atomistic representations where much longer timescales are computationally feasible (Wattis *et al*, 2001).

3.1.4 Parallel processing

Parallel processing provides the most obvious approach to reducing time-to-solution for calculations, but most common general-purpose molecular dynamics algorithms are not well suited to parallel computing, let alone massively parallel processing. Indeed, the more established MD life science packages (Case *et al*, 1999; Brooks *et al*, 1983) are long established codes that were not originally designed with parallel implementation in mind. The overwhelming majority of such 'legacy' codes have been parallelised 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 the size of both the model and the numbers 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 that highlights the importance of 'smart' algorithms in harnessing maximum benefit from modern parallel computers.

An alternative paradigm, which makes use of 'spatial domain decomposition' (Plimpton, 1995), 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, and not on less closely coupled clusters which have recently been gaining in popularity on grounds of cost.

3.2 Introduction to LAMMPS

LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator; Plimpton & Hendrickson, 1996) is a recently developed, highly scalable, MD code which implements spatial domain decomposition and is hence suitable for deployment on massively parallel supercomputers. In collaboration with Peter Coveney's group from Queen Mary University, London (now at UCL, London) we have tested the LAMMPS code for use in generating biological simulations using the AMBER forcefield parameters. Originally the LAMMPS99 code was used but subsequent editions of the code have been released, with useful applications added, therefore results are shown mainly for the LAMMPS2000 and LAMMPS2001 codes (website 1).

3.2.1 Validation using the Hoechst system

If LAMMPS and similar "parallel" codes are to gain widespread acceptance, it is vital that they are shown to produce results that agree with those obtained using the older, more established codes, such as AMBER. The aim of this section of the thesis is to use LAMMPS to reproduce as closely as possible

the AMBER results from the study of the Hoescht system (described earlier). Simulations were carried out, within LAMMPS, on the free DNA and the 1:1 and 2:1 drug DNA complexes (Figure 3.4) to ascertain whether the results could reproduce the phenomenon that the co-operativity within this system is due to entropic, rather than enthalpic, components of the recognition.

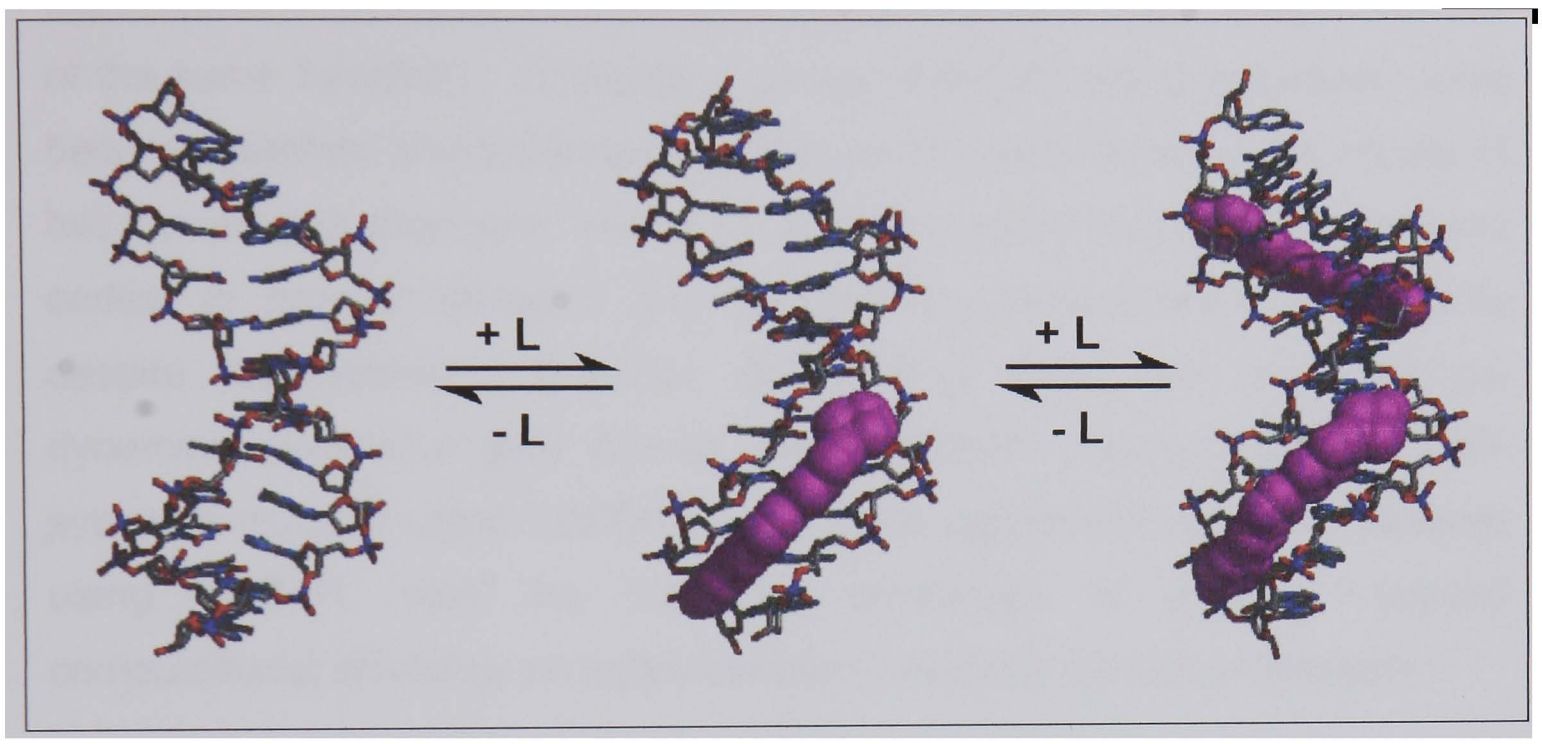


Figure 3.4 – Structures of the DNA dodecamer d(CTTTTCGAAAAG)$_2$ and the equilibria involved in the binding of Hoechst 33258 (L) to the sequence.

The AMBER simulations were performed on a single processor SGI Origin 200 machine using the AMBER6 (Case *et al*, 1999) suite of MD simulation programs and the associated Amber98 forcefield (Cornell *et al*, 1995), which is well validated for the simulation of both DNA and protein systems (Cheatham *et al*, 1999). The LAMMPS simulations were run on a massively parallel Cray T3E machine using the LAMMPS99, LAMMPS2000 and LAMMPS2001 codes as they became available to the group. We used an approach based on Principal Component Analysis (PCA; Amadei *et al*, 1993; Wlodek *et al*, 1997; Sherer *et al*, 1999; Cubero *et al*, 1999) 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 the various LAMMPS simulations.

Practical issues regarding differences in time-stepping, temperature coupling and treatment of long-range electrostatics, as well as the intrinsically chaotic nature of molecular dynamics (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 can never be expected to be identical to those run using the AMBER code, despite the use of the same forcefield. However, a series of benchmark comparisons have been established and implemented that can be made between the results of two simulations produced using two different architectures and/or software codes, to provide stringent and quantitative measurements of similarity despite this problem. However, as we shall show, with due care the dynamical behaviour (and derived thermodynamic parameters) of a DNA system simulated using LAMMPS is in good agreement with that obtained using AMBER, with the important advantage of greatly improved computational efficiency on tightly coupled massively parallel processors.

3.2.2 Implementation of the AMBER forcefield into LAMMPS

LAMMPS has the inherent flexibility required to calculate bonded and non-bonded components of molecular mechanics energy according to AMBER-style functions. The only change to the code required was to permit separate scaling factors to be applied to the van der Waals and electrostatic components of 1-4 interaction energies (this is now included in the current release of LAMMPS). Utility programs were produced to allow AMBER coordinate and topology files to be converted into their LAMMPS equivalents, and for LAMMPS trajectory files to be converted back into AMBER-style ones, for compatibility with MD analysis software.

3.3 Methods

3.3.1 Simulation protocol

All simulations in this study were run on up to sixty-four 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 counter-ions to establish electrical neutrality, and 1748 TIP3P 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 minimised using AMBER. Periodic boundary conditions were applied (initial box dimensions approximately 32Å x 36Å x 52Å) to a canonical (NVT) ensemble. Electrostatic interactions were calculated by the PPPM (Hockney & Eastwood, 1988) method using a grid order of 5 and a cut off of 9Å on the direct sum. A Lennard-Jones cut-off of 9Å was used for non-bonded interactions. The AMBER-specific treatment of 1-4 interactions (scaling non-bonded and electrostatic components differently) was handled through changes to the LAMMPS code (available now in the current LAMMPS release through use of the 'special bonds' directive). The simulation temperature was maintained through application of one of variety of temperature coupling options as detailed below. A dielectric constant of 1 was used, atomic positions were dumped every ps and the neighbour list was updated every 15 or 25 steps depending on the stability of the simulation.

AMBER simulations typically use SHAKE (Ryckaert *et al*, 1997) to constrain all bonds, permitting a 2fs integration timestep. SHAKE was not implemented in LAMMPS until the latest version (LAMMPS2001). Previous versions of the code did permit multiple time-stepping, using RESPA (Tuckerman *et al*, 1991, Plimpton *et al*, 1997) in which computationally expensive terms (in particular, the non-bonded interactions) are evaluated

less often than those requiring a small timestep (the bonded interactions). The earliest simulations, using LAMMPS99 however, did not use RESPA because at the time it was not compatible with the method of temperature coupling being used (Langevin) and therefore single time-stepping (1fs integration timestep) was used.

The protocol for LAMMPS MD studies using single time-stepping, and also the later studies using SHAKE, consisted of a 10 ps run in which the temperature of the system was raised from zero to 300K, then 1ns of dynamics at T=300K to ensure equilibration. All 2 ns production runs started with the configuration and velocities from this 1ns checkpoint. RESPA simulations required a slower warming run to avoid problems with neighbour list updates – 200ps warming from 0-100K, 200ps warming from 100-200K, 200ps warming from 200-300K, then 400ps equilibration at T=300K. Again, all 2ns production runs started with the coordinates and velocities from this 1ns time point.

3.3.2 Analysis methods

Separate solute and solvent temperatures in LAMMPS simulations were calculated from atomic velocities. Atomic Cartesian co-ordinate fluctuations and time-averaged structures were calculated using the AMBER utility program *ptraj* (website 2). Visualisation of trajectories was performed using VMD (Humphrey *et al*, 1996). Principal component analyses were performed according to the methods previously described (In Chapter 2) using in-house programs (Sherer *et al*, 1999; Cubero *et al*, 1999). The comparison of trajectories via the calculation of eigenvector overlaps was performed according to Hess (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 Generalised Born/Surface Area implicit solvation model (Tsui & Case, 2000), as previously described for the AMBER simulations (Harris *et al*, 2001). Configurational entropies were also

calculated from the mass-weighted eigenvectors according to the method of Schlitter as previously described in Chapter 2.

3.4 Results and discussion

3.4.1 Porting of the forcefield

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 forcefield 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 3.2).

ENERGY TYPE	AMBER (kcal/mol)	LAMMPS (kcal/mol)
Bond	0.0239	0.0239
Angle	399.8833	399.8833
Dihedral	438.7989	438.7989
Total VDW	2532.0560	2532.2143
Total Electrostatic	-27167.3825	-27167.0726
Total Energy	-23796.6204	-23796.1522

Table 3.2 – Static energy analysis, using both AMBER and LAMMPS, of a representative structure of the DNA duplex d(CTTTTGCAAAAG)$_2$.

Bond, angle and dihedral energies are identical, however 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 the same energies given identical input structures, it does not show that the calculated

forces are identical. The analysis of molecular dynamics simulations is obviously the most sensitive and practically important way of testing this.

3.4.2 Temperature control

AMBER typically employs the Berendsen algorithm (Berendsen *et al*, 1984) to control temperature but this is not available in LAMMPS. Two alternative thermostating methods are available in LAMMPS, Langevin (Schnieder & Stoll, 1978) and Nosé-Hoover (Nosé, 1984; Hoover, 1985). Initially we investigated Nosé-Hoover temperature control, but visualisation of the resulting trajectories indicated excessive motion of the DNA, compared to equivalent AMBER simulations. Calculation of atomic coordinate fluctuations confirmed this (Figure 3.5), the terminal bases in particular showing high mobility.

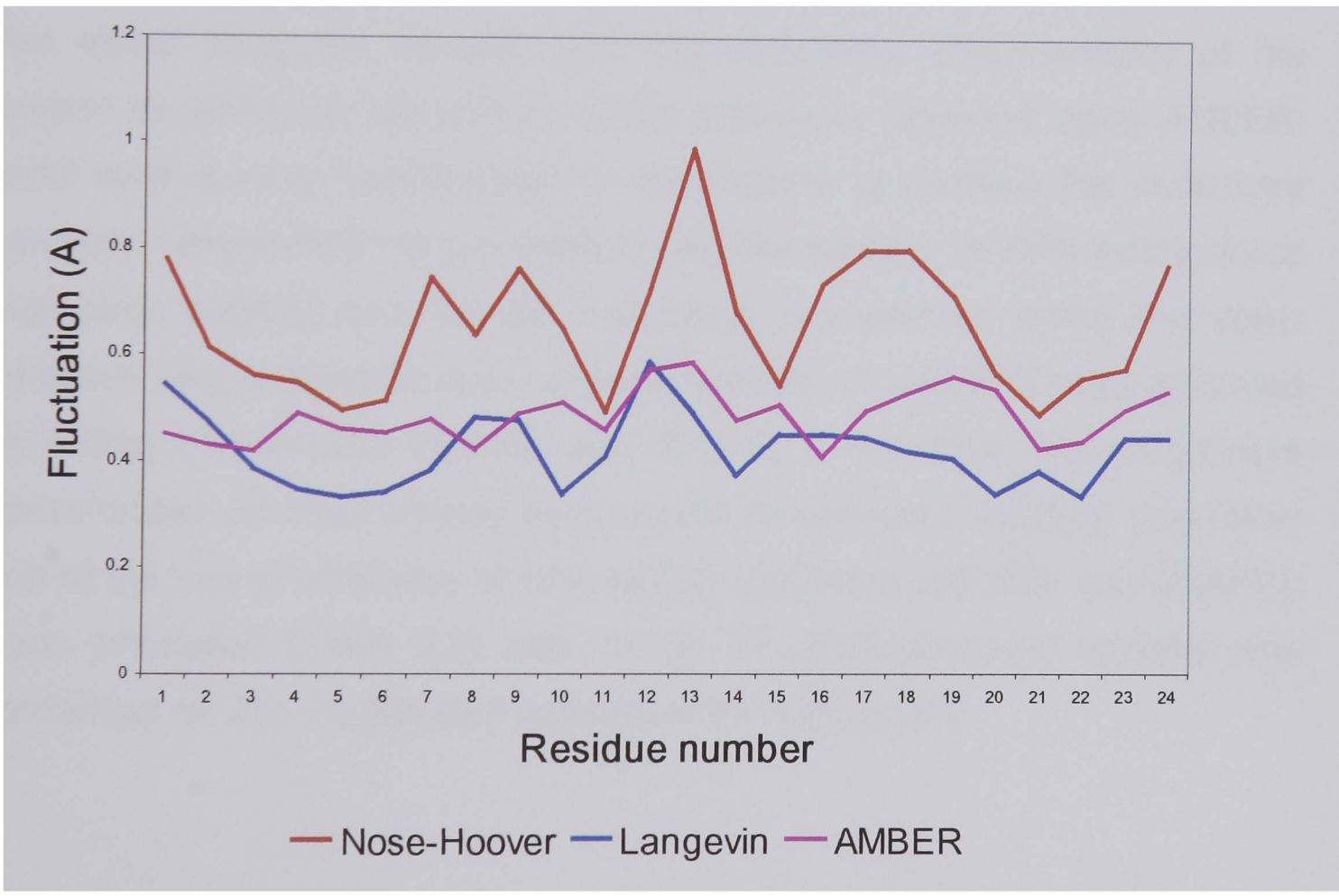


Figure 3.5 – Atomic Cartesian co-ordinate fluctuation data (averaged over all atoms for each base) for 10ps LAMMPS runs with Nosé-Hoover and Langevin temperature couplings compared to AMBER results (Berendsen coupling). Strand 1 runs from residue 1-12 and strand 2 (anti-parallel) from 13-24.

Calculation of individual solvent and solute temperatures revealed that the simulations suffered from the 'hot solute, cold solvent' problem. Switching to Langevin coupling, with separate scaling of solvent and solute temperatures, overcame this problem and led to simulations where the overall flexibility of the DNA was essentially indistinguishable from its behaviour in AMBER simulations. However, when this study began the use of Langevin coupling was incompatible with RESPA and SHAKE was not implemented, so stable simulations required the expense of the full evaluation of all energy terms every 1 fs time step. This obviously affected the practical comparison of attainable speeds with AMBER.

3.4.3 Analysis of simulations on the DNA alone – similarity analysis

With a suitable temperature coupling method established, extended simulations were carried out in order to determine values for key parameters that would bring the dynamic and thermodynamic characteristics of the system as close as possible to those previously obtained using AMBER. Initial studies were handicapped by the inability to combine the necessary Langevin temperature control method with the efficient RESPA energy/force evaluation method and SHAKE not being implemented within the code; however, later in these studies updated releases of LAMMPS were produced permitting use of both RESPA and SHAKE. Therefore, the parameters concentrated on most closely were a) the temperature coupling parameter and b) the use or otherwise of time saving algorithms (RESPA and SHAKE). Each simulation (Table 3.3) was run on 64 processors and analysis was carried out on 2ns equilibrated portions of the trajectories.

SIMULATION ID	TEMP COUPLING (fs^{-1})	RESPA/SHAKE?
LAMMPS 1	0.01	NONE
LAMMPS 2	0.001	NONE
LAMMPS 3	0.0001	NONE
RESPA 1	0.01	RESPA
RESPA 2	0.001	RESPA
RESPA 3	0.0001	RESPA
SHAKE	0.001	SHAKE

Table 3.3 – Details of the simulations carried out on the free DNA.

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 co-ordinate deviation 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 to that obtained using AMBER. For this, each snapshot from the 2 ns trajectories was stripped of its waters and counter-ions, and used as the input for a single-point energy calculation in AMBER, using the GB/SA method to provide a solvation term. This permitted a direct comparison with previous AMBER data for this DNA sequence. Thirdly, the configurational entropy of each LAMMPS trajectory was calculated, and compared to 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. Fourthly, the dynamics of each simulation was investigated in more detail through the comparison of principal components (see previous chapter).

3.4.3.1 Results from the similarity analysis

The results obtained are shown in Table 3.4. The coupling parameter for the Berendsen thermostat used in the reference AMBER studies was expected

to be equivalent to a 0.001 fs^{-1} coupling parameter for the Langevin thermostating used in LAMMPS. However, initial simulations performed without RESPA or SHAKE indicated that a weaker temperature coupling (0.0001 fs^{-1}) gave results slightly closer to those obtained using AMBER. This weakening in the temperature coupling had no discernable effect on the average energetics of the DNA dodecamer duplex, but, unsurprisingly, led to increasing configurational entropy. Looser temperature coupling also led to the major modes of deformation of the DNA resembling more closely those obtained using AMBER (increased PCA overlap).

SYSTEM	RMSD (Å)	AVERAGE ENERGY (bond energy removed) kcal/mol	T*ENTROPY (TS) kcal/mol	PCA OVERLAP
AMBER	n/a	-4397.95	693.67	n/a
LAMMPS 1	0.95	-4442.87	589.79	0.4214
LAMMPS 2	0.65	-4444.15	637.56	0.6309
LAMMPS 3	0.63	-4443.27	646.17	0.6678
RESPA 1	2.24	-4437.09	695.99	0.6211
RESPA 2	0.58	-4443.38	727.13	0.6315
RESPA 3	0.73	-4443.33	713.28	0.6867
SHAKE	0.59	-4435.32	732.07	0.6796

Table 3.4 – Benchmark comparisons for LAMMPS simulations of d(CTTTTGCAAAAG)2 against the previous AMBER results.

When the equivalent simulations were repeated using RESPA, there were some differences. Firstly, while the tightest temperature coupling (0.01 fs^{-1}) clearly led to a poor average structure and an average energy which differed from the other LAMMPS simulations, the two weaker coupling constants provided trajectories with varying merits. A coupling constant of 0.001 fs^{-1} provided the best (lowest RMSD) time-averaged structure (Figure 3.6) but a coupling constant of 0.0001 fs^{-1} provided a trajectory with principal components closest to those obtained using AMBER.

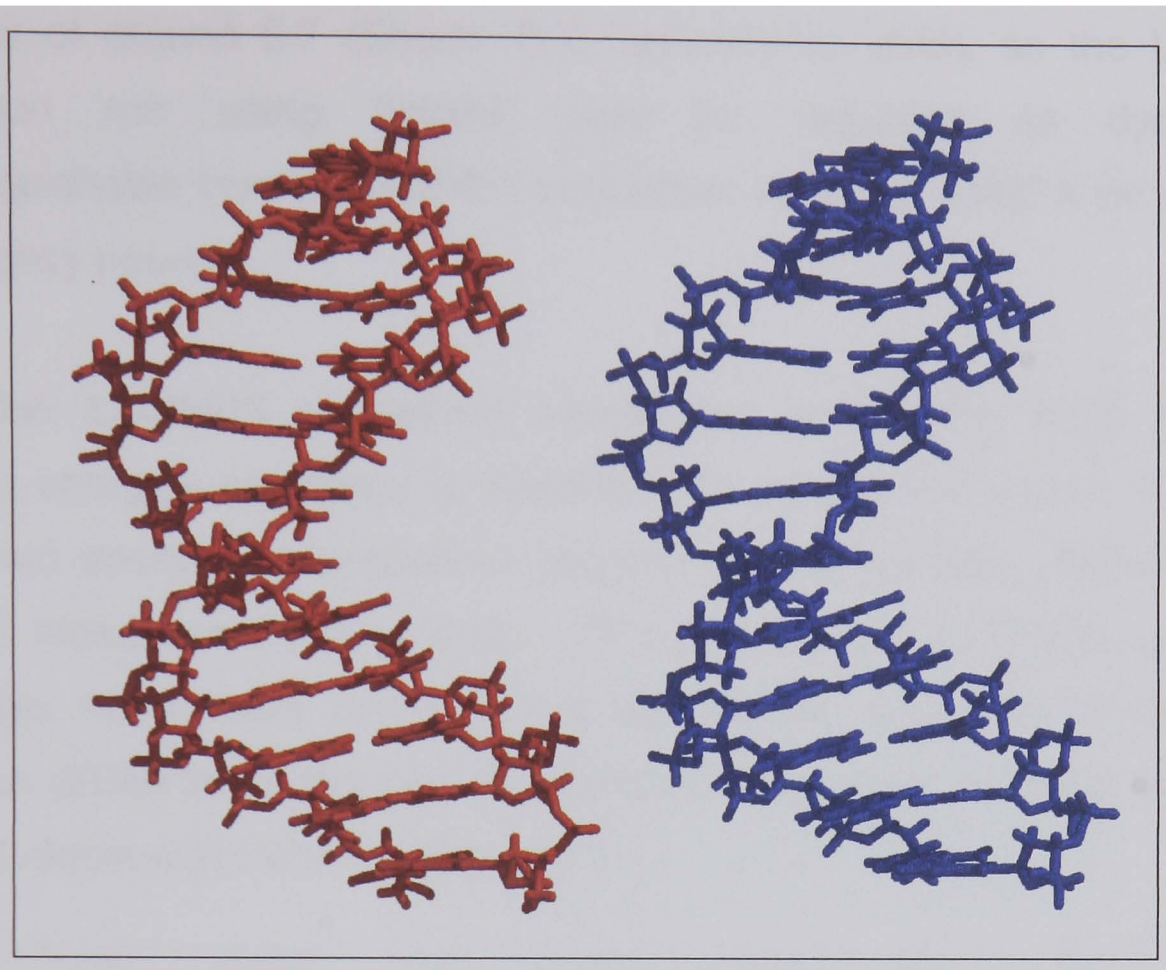


Figure 3.6 – Time averaged structures of the free DNA from RESPA 2 (red) and from AMBER (blue) the RMSD value between these is 0.58Å.

We decided at this point, that a temperature coupling of 0.001 fs^{-1} was to be used for all further simulations as the results from RESPA 2 were reasonably good and this was the same temperature coupling as used in the AMBER simulations, therefore a direct comparison could be made. The rest of this section will deal with similarities between AMBER and the SHAKE and RESPA simulations carried out with this temperature coupling.

Clearly, using LAMMPS with either SHAKE or RESPA, the time-averaged structure of the DNA is very similar to that obtained using AMBER. Experience within the group (unpublished data) shows that RMSD values of this magnitude (less than 0.6 angstroms) are frequently observed between independent simulations of the same system using the same MD code. However, the average energy of the two LAMMPS simulations are around 65 kcal/mol more negative than that of the AMBER simulation, and their entropy (TS at 300K) is around 36 kcal/mol more positive. The PCA overlap values are good – it is typically observed that separate independent simulations of

the same system using the same MD code lead to trajectories with PCA overlaps of around 0.7 (Sherer E.C, unpublished data), so the LAMMPS simulation run using SHAKE may be regarded as dynamically indistinguishable from the AMBER simulation, while the RESPA simulation is only slightly poorer.

The other LAMMPS simulations carried out confirm the trend for lower average energies compared to AMBER. To identify the source of this, we performed energy decomposition analysis on the AMBER, RESPA 2 and SHAKE simulations (Table 3.5). This shows that LAMMPS generated structures have more negative 1-4 electrostatic energy (1-4 EEL) and solvation (EGB) terms though this is partly countered for by a less favourable general electrostatic energy (EEL) term.

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
RESPA 2	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

Table 3.5 – Energy decomposition of LAMMPS simulations (RESPA and SHAKE) compared to previous AMBER results. All values are in kcal/mol and are averages over 1000 snapshots taken from the MD simulations)

However, a similar analysis of the drug-DNA complex simulations (discussed below) revealed that the overall energy difference of 40-50 kcal/mol between AMBER and LAMMPS simulations was still evident, this resulted from different balances between the same three key terms (1-4 EEL, EGB and EEL). The one constant feature is an approximately 60 kcal/mol 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 terms, the source of this discrepancy still remains to be determined. One possibility is that in the original AMBER simulations SHAKE was used to constrain all bond lengths, whereas in LAMMPS, to permit efficient parallelisation, only bonds to hydrogen atoms are constrained. However, since we find that

LAMMPS gives very similar results using SHAKE and RESPA (where no bonds are constrained) this seems unlikely.

3.4.4 *Analysis of drug-DNA complexes – similarity and thermodynamic analysis*

Though clearly we have 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 code used to generate the configurations. To test this additional simulations were performed using LAMMPS on the 1:1 and 2:1 complexes between this DNA duplex and the minor drug binder Hoechst 33258.

These simulations were performed analogously to the RESPA 2 and SHAKE simulations on the free DNA. Two sets of analysis were carried out on these simulations. Firstly, a similarity analysis was carried out on the 1:1 and 2:1 complex simulations. The drug atoms were stripped out, leaving just the DNA molecule and the resulting trajectories were put through the same analysis as the free DNA, as described earlier. Secondly, the complete trajectories were put through a thermodynamic analysis. As before, the trajectories were post processed to calculate enthalpic plus solvation terms using the GB/SA method and configurational entropies were calculated via the Schlitter approach. From these we were able to estimate, as before (Harris *et al*, 2001), the differences in the free energy change between the first and second binding event ($\Delta\Delta G$), which has been shown by NMR (Gavathiotis *et al*, 2000) to be at least –4.0 kcal/mol.

3.4.4.1 Results from similarity analysis

The agreement with the previous AMBER results is shown in Table 3.6. These results show that the RMSD values get smaller (showing an increase in similarity in the time-averaged structures) and the PCA overlaps get larger

(showing increased dynamical similarity) when the drugs bind to the DNA. This is as expected as the DNA gets more rigid as the drugs bind to its minor groove. The RESPA and SHAKE results show excellent RMSD and PCA overlap values.

(a) 1:1 drug-DNA complex results

SYSTEM	RMSD (Å)	AVERAGE ENERGY (bond energy removed) kcal/mol	T*ENTROPY (TS) kcal/mol	PCA OVERLAP
AMBER	n/a	-4397.56	678.67	n/a
RESPA	0.68	-4437.67	704.98	0.6972
SHAKE	0.50	-4433.84	715.47	0.7194

(b) 2:1 drug-DNA complex results

SYSTEM	RMSD (Å)	AVERAGE ENERGY (bond energy removed) kcal/mol	T*ENTROPY (TS) kcal/mol	PCA OVERLAP
AMBER	n/a	-4394.60	669.39	n/a
RESPA	0.51	-4443.38	693.00	0.7854
SHAKE	0.53	-4427.85	700.53	0.7743

Table 3.6 – Benchmark comparisons of the 1:1 and 2:1 drug-DNA complexes against the previous AMBER results.

As stated earlier there is still a 40-50 kcal/mol difference in the enthalpies calculated from LAMMPS from those calculated from AMBER. As before, energy decompositions were carried out (Table 3.7) and the major differences were again in the 1-4 EEL, EGB and EEL terms. The entropies generated through LAMMPS were again high in comparison to AMBER. However, as mentioned earlier, we are more interested in the differences in enthalpies and entropies between states, in this instance, between the free DNA and 1:1 drug-DNA complex and between the 1:1 and 2:1 drug-DNA complex.

(a) 1:1 drug-DNA complex decomposition

SYSTEM	BOND	ANGLE	DIHED	1-4 VDW	1-4 EEL	VDW	EEL	EGB
AMBER	0.12	432.54	461.97	230.59	-1709.86	-475.33	-971.93	-2394.76
RESPA	298.12	442.80	462.94	217.21	-1770.69	-478.05	-965.70	-2381.92
SHAKE	203.61	447.06	464.73	218.43	-1772.91	-475.97	-986.95	-2358.77

(b) 2:1 drug-DNA complex decomposition

SYSTEM	BOND	ANGLE	DIHED	1-4 VDW	1-4 EEL	VDW	EEL	EGB
AMBER	0.13	470.08	488.36	247.96	-1670.00	-541.97	-1353.49	-2092.03
RESPA	319.71	476.86	490.65	233.43	-1733.69	-543.70	-1316.07	-2111.02
SHAKE	218.31	486.31	492.37	234.98	-1731.35	-549.01	-1349.81	-2077.61

Table 3.7 - Energy decomposition of LAMMPS simulations of the 1:1 and 2:1 drug-DNA complexes, compared to previous AMBER results. All values are in kcal/mol and are averages over 1000 snapshots taken from the MD simulations)

3.4.4.2 Results from the thermodynamic analysis

The agreement with the previous AMBER results is shown in Table 3.8. AMBER simulations predicted that the binding is enthalpically anti co-operative, $\Delta\Delta E$ equal to 4.4 kcal/mol. Using LAMMPS with SHAKE, binding is also predicted to be enthapically anti co-operative, but only by 2.6 kcal/mol. Using LAMMPS with RESPA, the simulations lead to the prediction that binding is essentially unco-operative with a $\Delta\Delta E$ of –0.1 kcal/mol. AMBER simulations predicted a strongly co-operative entropic term, $\Delta\Delta TS$ of 9.6 kcal/mol, and the LAMMPS results show this too, with $\Delta\Delta TS$ 1 kcal/mol more positive in the RESPA simulation and 2.5 kcal/mol less positive in the SHAKE simulation.

(a) AMBER6 results

SYSTEM	E	ΔE	ΔΔE	TS_∞	ΔTS_∞	$\Delta\Delta TS_\infty$
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) LAMMPS/RESPA results

SYSTEM	E	ΔE	ΔΔE	TS_∞	ΔTS_∞	$\Delta\Delta TS_\infty$
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) LAMMPS/SHAKE results

SYSTEM	E	ΔE	ΔΔE	TS_∞	ΔTS_∞	$\Delta\Delta TS_\infty$
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		

Table 3.8 – Thermodynamical analysis of LAMMPS simulations compared to previous AMBER6 results (T = 300 K). All values are averages over 1000 snapshots from each simulation and are in kcal/mol.

Taken together, the AMBER simulations predict an overall ΔΔG of –5.2 kcal/mol, while our LAMMPS simulations predict ΔΔG to be –10.7 kcal/mol (using RESPA) and –4.5 kcal/mol (using SHAKE), ΔΔG terms are calculated from $\Delta\Delta E - \Delta\Delta TS_\infty$ values, where the E terms equate with system enthalpy. 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 RESPA simulations than the SHAKE ones), the important qualitative conclusions from the study are the same: that the observed highly co-operative binding of Hoechst 33258 to this DNA dodecamer is entropy, not enthalpy driven.

3.4.5 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 a series of 10ps simulations were performed using different numbers of processors using both AMBER and the various versions of LAMMPS. In an ideal case, doubling the number of processors used for a calculation would halve the time required for solution. However, as discussed in the introduction, this is seldom the case in practice.

The efficiency of each simulation was calculated according to the equation:

$$\text{Efficiency} = np*t/l,$$

where np is the number of processors, t the wallclock time taken for completion of the computation (in seconds), and l is the length of the simulation (in femtoseconds).

The results are shown in Figure 3.7, with efficiencies normalised so that they are relative to the performance of LAMMPS using SHAKE when run on a single node of the Cray T3E.

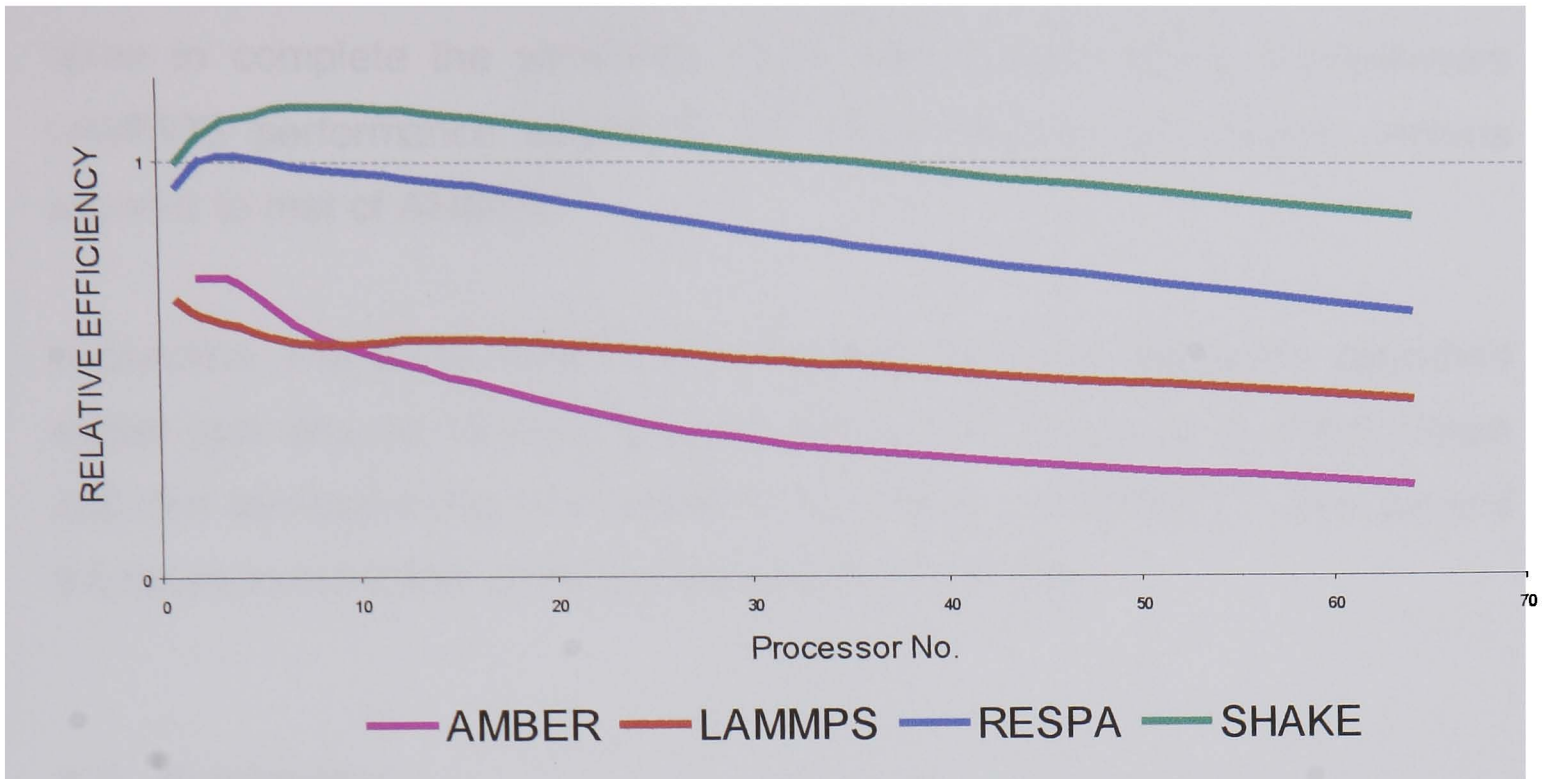


Figure 3.7 - Relative efficiencies of simulations run using different numbers of processors. All are scaled to the performance of LAMMPS with SHAKE on a single processor.

Amber clearly performs poorly in this environment; its efficiency falls off rapidly when the simulation is run on over four nodes (Note – due to memory limitations it was not possible to run AMBER on one processor of the T3E). Without RESPA or SHAKE, and on up to four processors, LAMMPS performs slightly worse than AMBER, in that the simulations take longer and the efficiency drops away even faster. However, above four processors the efficiency of LAMMPS stabilises to the extent that, above eight processors the simulations run faster than the corresponding AMBER ones. This is almost certainly a cache effect, whereby on more than four processors the whole system can be fitted into cache memory while on fewer processors it cannot.

The use of RESPA or SHAKE within LAMMPS is clearly advantageous; when LAMMPS uses either RESPA or SHAKE it outperforms AMBER. LAMMPS using SHAKE performs better than LAMMPS using RESPA and both these perform better than LAMMPS with neither, which only starts to out perform AMBER above 8 processors. The LAMMPS performance actually improves (RESPA and SHAKE) for small increases (up to 4 or 8) in the number of

processors used – i.e. when using two processors instead of one, the time taken to complete the simulation more than halves. Above 8 processors LAMMPS performance begins to fall off somewhat, but always remains superior to that of AMBER.

In practice, 1ns of the AMBER simulation of the 12mer sequence described earlier took around 10 days to complete on one processor of a SGI Origin 200; the corresponding 1ns LAMMPS simulation, using SHAKE, took around 6.6 hours to complete on 64 processors of a Cray T3E.

3.5 Summary

The Amber98 forcefield, 'native' to the MD code AMBER which is widely used and respected for molecular dynamics 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 that 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 optimised simulation parameters, LAMMPS passes all these tests when compared to 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 timescales will make new problems amenable to analysis through atomistic molecular dynamics simulations. Conversely, for problems where current sizes and MD timescales suffice, time-to-solution may be reduced ten-fold or more.

3.6 References

Amadei A, Linssen A.B.M, Berendsen H.J.C, (1993), *Proteins*, **17**, 412-425.

Auffinger P, Westhof E, (1998), *Curr. Opin. Struct. Biol*, **8**, 227-236.

Berendsen H.J.C, Postma J.P.M, van Gunsteren W.F, Dinola A, Haak V.R, (1984), *J. Chem. Phys*, **81**, 3684-3690.

Beveridge D.L, DiCapua F.M, (1989), *Annu. Rev. Biophys. Biophys. Chem,* **18**, 431-492.

Beveridge D.L, McConnell K.J, (2000), *Curr. Opin. Struct. Biol*, **10**, 182-196.

Brooks B.R, Bruccoleri R.E, Olafson B.D, States D.J, Swaminathan S, Karplus M, (1983), *J. Comp. Chem,* **4** (2), 187-217.

Bostock-Smith C.E, Harris S.A, Laughton C.A, Searle M.S, (2001), *Nucleic Acids Res,* **29**, 12658-12663.

Case D.A, Pearlman D.A, Caldwell J.W, Cheatham T.E. III, Ross W.S, Simmerling C.L, Darden T.L, Merz K.M, Stanton R.V, Cheng A.L, Vincent J.J, Crowley M, Tsui V, Radmaer R.J, Duan Y, Pitera J, Massova I, Seibel G.L, Singh U.C, Weiner P.K, Kollman P.A, (1999), *AMBER 6*, University of California, San Francisco.

Cheatham T.E. III, Kollman P.A, (1996), *J. Mol. Biol,* **259**, 434-444.

Cheatham T.E. III, Cieplak P, Kollman P.A, (1999), *J. Biomol. Struct. Dyn,* **16**, 845-862.

Cheatham T.E. III, Kollman P.A, (2000), *Annu. Rev. Phys. Chem,* **51**, 435-471.

Cheatham T.E. III, Young M.A, (2001), *Biopolymers,* **56**, 232-256.

Cieplak P, Cheatham T.E. III, Kollman P.A, (1997), *J. Am. Chem. Soc,* **119**, 6722-6730.

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.

Crothers D.M, Haran T.E, Nadeau J.G, (1990), *J. Biol. Chem,* **265**, 7093-7096.

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

Daggett V, (2000), *Curr. Opin. Struct. Biol,* **10**, 160-164.

Duan Y, Kollman P.A, (1998), *Science,* **282**, 740-744.

Duan Y, Kollman P.A, (2001), *IBM Systems Journal*, **40** (2), 297-309.

Franklin R.E, Gosling R, (1953), *Nature,* **171**, 740-741.

Fersht A.R, Daggett V, (2002), *Cell* , **108** (4), 573-582.

Gavathiotis E, Sharman G.J, Searle M.S, (2000), *Nuc. Acids. Res,* **28** (3), 728-735.

Guckian K.M, Krugh T.R, Kool E.T, (1998), *Nat. Struct. Biol,* **5**, 954-959.

Haran T.E, Kahn J.D, Crothers D.M, (1994), *J. Mol. Biol,* **244**, 135-143.

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**, 19824-19839.

Haq I, Ladbury J.E, Chowdhry B.Z, Jenkins T.C, Chaires J.B, (1997), *J. Mol. Biol,* **271**, 244-257.

Hess B, (2000), *Phys. Rev. E*, **62**, (6) 8438-8448.

Hockney R.W, Eastwood J.W, (1988), *Computer Simulations Using Particles*; Adam Hilger, New York, NY.

Hoover W.G, (1985), *Phys. Rev. A,* **31**, 1695-1697.

Humphrey W, Dalke A, Schulten K, (1996), *J, Molec. Graphics,* **4.1**, 33-38.

Karplus M, McCammon J.A, (2002), *Nature Struct. Biol,* **9**, 646-652.

Kollman P.A, (1993), *Chem. Rev,* **93**, 2395-2417.

Kollman P.A, Massova I, Reyes C, Kuhn B, Huo S.H, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case D.A, Cheatham T.E. III, (2000), *Accounts Chem. Res,* **33**, 889-897.

Koo H.S, Drak J, Rice J.A, Crothers D.M, (1990), *Biochemistry,* **29**, 4227-4234.

Levitt M, (1983), *Cold Spring. Harb. Symp. Quant. Biol,* **47**, 251-262.

MacKerell A.D, Wiorkiewicz-Kuczera J, Karplus M, (1995), *J. Am. Chem. Soc,* **117**, 11946-11975.

Marini J.C, Levene S.D, Crothers D.M, (1982), *Proc. Natl. Acad. Sci. USA,* **79**, 7664-7668.

Miller J.L, Kollman P.A, (1997), *Biophys. J,* **73**, 2702-2710.

Moran S, Ren R.X-F, Kool E.T, (1997), *Proc. Natl. Acad. Sci. USA,* **94**, 10506-10511.

Nosé S, (1984), *J. Chem. Phys,* **81**, 511-519.

Olsen W.K, Zhurkin V.D, (2000), *Curr. Opin. Struct. Biol*, **10**, 286-297.

Piskur J, Rupprecht A, (1995), *FEBS Lett,* **375**, 174-178.

Plimpton S.J, (1995), *J. Comp. Chem,* **117**, 1-19.

Plimpton S.J, Hendrickson B, (1996), *J. Comp. Chem,* **17**, 326-337.

Plimpton S.J, Pollock R, Stevens M, (March 1997) *Proc. of the 8th Conf on Parallel Processing for Scientific Computing*, Minneapolis, MN.

Prive G.G, Yanagi K, Dickerson R.E, (1991), *J. Mol. Biol,* **217**, 177-199.

Ryckaert J.P, Ciccotti G, Berendsen H.J.C, (1997), *J. Comp. Phys,* **23**, 327-341.

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

Schnieder T, Stoll E, (1978), *Phys. Rev. B,* **17**, 1302-1322.

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

Shields G.C, Laughton C.A, Orozco M, (1997), *J. Am. Chem. Soc,* **119**, 7463-7469.

Sinden R.R, (1994), *DNA Structure and Function,* Academic Press, Inc.

Szabo A, Shi B, Lee S.A, Rupprecht A, (1996), *J. Biomol. Struct. Dyn,* **13**, 1029-1033.

Tsui V, Case D.A, (2000), *J. Am. Chem. Soc,* **122**, 2489-2498.

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.

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

Website 1 – www.cs.sandia.gov/~sjplimp/main.html accessed July 2003.

Website 2 – www.amber.scripps.edu/doc6/ptraj.html accessed July 2003.

Weiser J, Shenkin P.S, Still W.C, (1999), *J. Comp. Chem,* **20**, 217-230.

Westhof E, (1988), *Annu. Rev. Biophys. Biophys. Chem,* **17**, 125-144.

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*, John Wiley & Sons Ltd, Chichester.

Young M.A, Beveridge D.L, (1998), *J. Mol. Biol,* **281**, 675-687.

CHAPTER 4 – MOLECULAR DYNAMICS APPROACHES TO CALCULATING BINDING AFFINITIES - APPLIED TO A NOVEL CLASS OF TELOMERASE INHIBITORS.

4.1 Telomeres and Telomerase

4.1.1 The role of telomeres

Telomeres are guanine rich sequences of DNA found at the end of chromosomes. In humans, the terminal approximately 10 kilobases (kb) of chromosomes are made up of tandem repeats of the telomeric sequence TTAGGG (Blackburn, 1991). These telomeres protect the chromosomes ends from end-to-end fusion, exonuclease degradation and aberrant recombination (Perry & Jenkins, 1999 and references therein), and by doing so maintain the integrity of the genetic information stored in the DNA. Telomeric DNA is generally double stranded except for the extreme 3' end of the telomere, which consists of a single stranded overhang. This single strand is formed during DNA replication because DNA polymerase cannot fully replicate the extreme 3' end during lagging strand synthesis. This is known as the "end replication problem" (Watson, 1972) and results in the loss of around 50-100 base pairs per cell division. This could lead to dangerously short telomeres and the possibility that genes would not be fully replicated causing major problems upon transcription. Cells do not allow this to happen, they are not able to divide indefinitely and have a pre-determined proliferative lifespan to stop this occurring, termed the Hayflick limit (Hayflick, 1961). Human somatic cells can undergo around 50-80 cell divisions before reaching their Hayflick limit, at which time they undergo cell cycle arrest (senescence) and stop dividing.

Although senescence is an irreversible process, there are cells that are able to escape the process. Once the Hayflick limit is reached within a cell, the telomere shortening signal is transduced to tumour suppressor pathways

controlled by p53 and retinoblastoma (Rb) which lead to senescence or apoptosis (Leong & Seow, 2001). If p53 and Rb are inactivated by viral oncogenes, the Hayflick limit can be overcome, cells continue to divide and their telomeres shorten until they reach a second barrier, crisis. At this point there is genomic instability and most cells die, although in rare cases cells can escape crisis and become immortal. The unrestricted proliferation shown by cells with have passed crisis is associated with the expression of the telomerase enzyme which is able to stabilise the length of their telomeres (Wright & Shay, 1992), implicating this enzyme in the development of tumours.

Telomerase is also active in around 85-90% of human tumours (Kim *et al*, 1994) and is responsible for the immortality of these cells. It is not normally detected in normal somatic cells (exceptions include the immune system, skin, intestinal lining and hair follicles; Blackburn, 2000) but is found in stem cells which also maintain the length of their telomeres through activation of the telomerase enzyme.

4.1.2 The telomerase enzyme

Telomerase is a ribonucleoprotein reverse transcriptase enzyme which is capable of synthesising telomeric repeats, thus maintaining the length of telomeres through an indefinite number of cell divisions. It is made up of two core subunits and a number of other associated protein subunits (Figure 4.1).

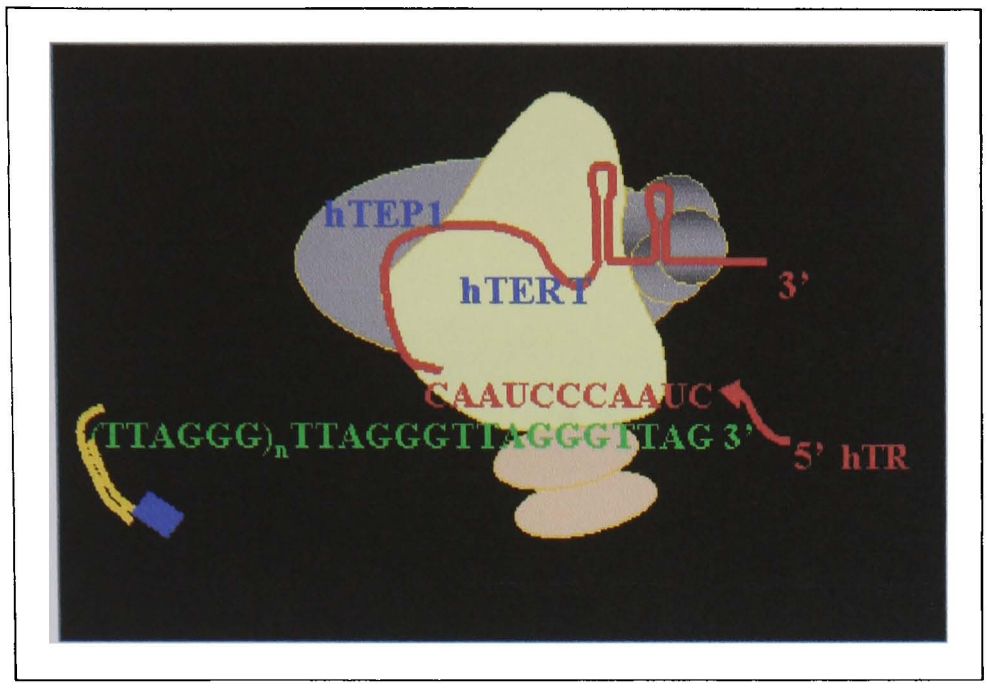


Figure 4.1 – The major subunits of the telomerase enzyme (adapted from Shay & Wright, 1999).

The first core subunit is a human telomerase RNA part (hTR) which contains an 11 bp sequence that is the template for the telomeric repeats added to the chromosome. The second core subunit is a human telomerase reverse transcriptase (hTERT) part that catalyses nucleotide polymerisation (Corey, 2000). Both hTR and hTERT are required for telomerase function but it is the presence of hTERT that is responsible for the expression of telomerase activity as all human somatic cells already contain hTR but not hTERT (Feng *et al*, 1995).

Telomerase adds the TTAGGG repeats to telomeric DNA by attaching to the ends of chromosomes and then utilising its internal hTR template to specify the sequence of added nucleotides, it then moves downstream and re-positions its template for further additions of the telomeric repeat (Figure 4.2).

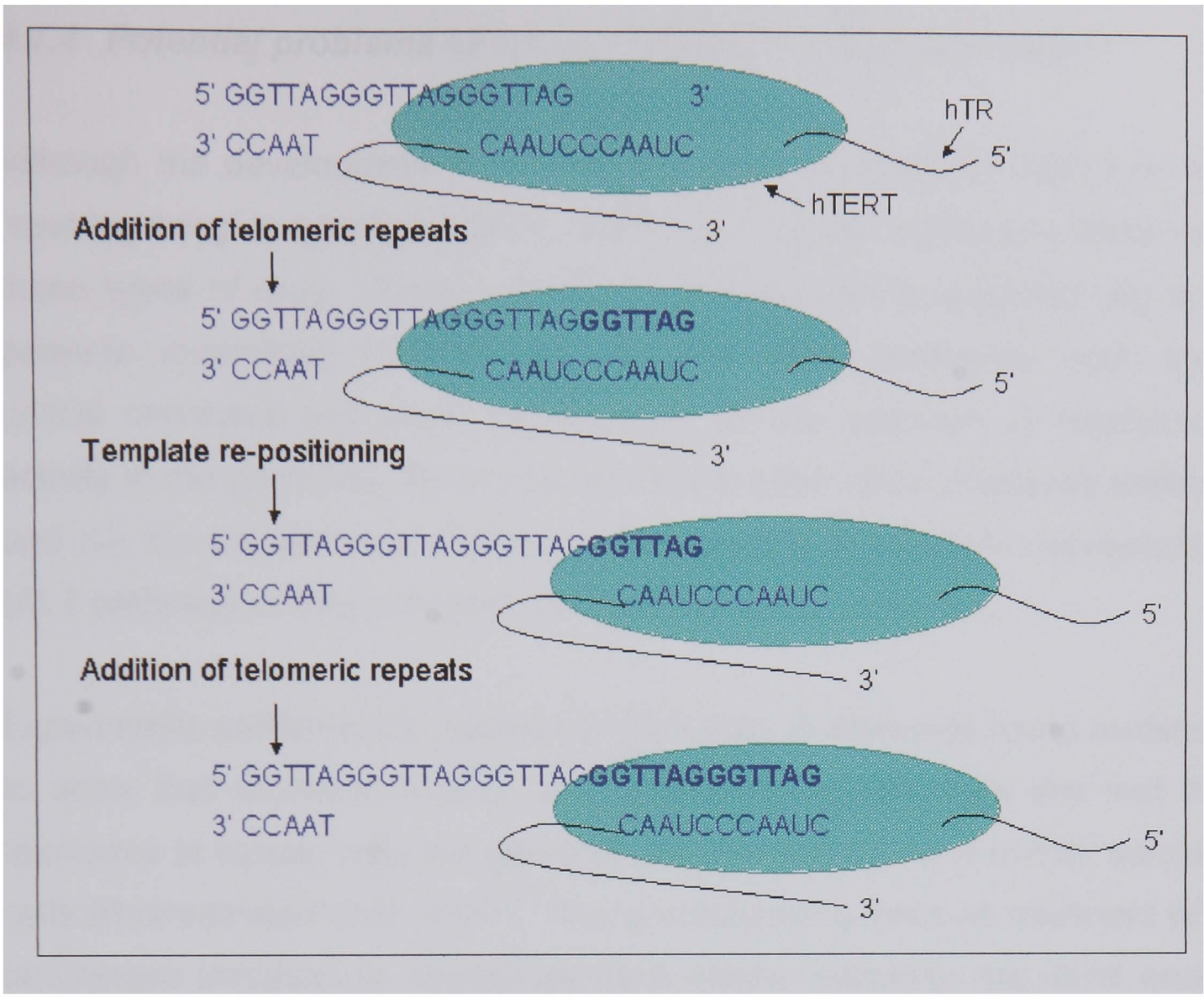


Figure 4.2 – The addition of telomeric repeats (adapted from White *et al*, 2001).

4.1.3 Telomerase as an anti-cancer target

The observations of telomerase activity in tumours along with no activity in the vast majority of normal cells have led many people to propose the enzyme as an attractive anti-cancer target (Neidle & Kelland, 1999; Perry & Jenkins, 1999 and references therein). The lack of telomerase activity in normal cells makes telomerase a potential highly specific target, as in theory, a telomerase inhibiting drug would not have any discernible effects on normal cells within the body.

Telomerase activity has also been proposed as a molecular marker for the early detection of cancer, and as a prognostic indicator of disease outcome and patient response to chemotherapy (Kerwin, 2000).

4.1.4 Potential problems of telomerase as an anti-cancer target

Although the development of telomerase inhibitors has been hailed as the "next big thing" in cancer research, there are a number of potential flaws with these types of drug. Issues of concern include: (i) the expected lag time between telomerase inhibition and the time when telomeres reach their critical shortness and enter senescence; (ii) the detection of telomerase activity in the germ line, stem cells and some other cells (discussed earlier); and (iii) the existence of alternative mechanisms of telomere maintenance (ALT pathway) raising concerns over possible drug resistance.

Experiments carried out to determine the length of telomeres found evidence to show that telomere lengths vary greatly across cell lines and that the telomeres in cancer cells are generally shorter than those of normal somatic cells (Brummendorf et al, 2000). This is encouraging news as treatment with telomerase inhibitors to specific cancers whose telomeres are short would result in rapid senescence and/or apoptosis in tumour cells well before any toxicities could occur to the germ line and stem cells which have longer telomeres. Also, to date there is no evidence of drug-resistance emerging (e.g. by the ALT pathway) in cancer cells exposed to telomerase inhibiting drugs (Kelland, 2000).

4.2 Telomerase inhibition

A telomerase inhibitor should possess the ability to reduce telomerase activity in cell extracts, preferably at the sub-micromolar level. The PCR based telomerase repeat amplification protocol (TRAP; Kim *et al*, 1994) assay has been developed to measure telomerase activity.

The classic model for telomere erosion stipulates that after ~20 cell divisions, cells will have critically short telomeres and will enter the senescent state. This has been confirmed by experiments on HeLa cells, where these cells lost telomeric DNA and began to die after 23-26 divisions (Feng et al, 1995).

The model predicts that long-term exposure of tumour cells to telomerase inhibitors, at levels which don't cause acute cytotoxicity, should induce telomere shortening and growth arrest (the rate depending upon initial telomere length). Chemically related molecules that do not have any effect on inhibition should not cause telomere shortening or senescence (Neidle & Kelland, 1999).

There are a number of ways to inhibit telomerase, including antisense approaches, reverse transcriptase inhibitors and quadruplex stabilising drugs. Antisense approaches and reverse transcriptase inhibitors will be discussed here and quadruplex stabilising drugs will be discussed later in section 4.3.2.

4.2.1 Antisense approaches

Antisense approaches typically target the RNA component hTR, especially the 11 base pair sequence that encodes the telomeric repeat. These types of drugs consist of short pieces of DNA which are complementary to the target RNA. Their mode of action is to bind to the RNA via Watson Crick base pairing and inhibit the translation of the RNA sequence, hence no new telomeric repeats are formed. The template RNA part of telomerase must be exposed for new telomeric repeats to be added, therefore the target is readily accessible.

As with any antisense strategy, there is the problem of delivery to the cell, without a transfecting agent these types of drug do not readily enter cells. Also if they do manage to get into the cells they are prone to degradation by exo and endo-nuceases as they look like segments of broken DNA.

A variety of studies have been carried out on antisense oligonucleotides which target the template and non-template regions of hTR. Many have reported a reduction in telomerase activity, but most have not reported reduced telomere length with continued treatment (White *et al*, 2001 and

references within). The use of protein nucleic acids (PNAs) hybridised with DNA has probably produced some of the best results to date. Improved methods of delivery involving cationic lipids (Hamilton *et al*, 1999) lead to the exposure of PNA in cells. This lead to telomerase inhibition (with IC_{50} values in the *n*M range), telomere shortening and onset of apoptosis, dependent on initial telomere length (Herbert *et al*, 1999).

4.2.2 Reverse transcriptase inhibitors

Reverse transcriptase inhibitors are able to incorporate into viral DNA and block chain elongation using the reverse transcriptase enzyme. Telomerase has become a target for these types of drug because of its reverse transcriptase (hTERT) activity. Inhibition or prevention of enzyme activity by chain terminating nucleoside triphosphates such as 3'azido-3'deoxythymidine (AZT, already in use for the treatment of AIDS) has been proposed. Studies have shown these types of drugs commonly slow the growth of cells in culture but they do not generally lead to telomere shortening or senescence. Dideoxyguanosine inhibitors have shown some shortening of telomeres and potent telomerase inhibition, with an IC_{50} value of 8.6 μM against A2780 cell extracts (Perry & Jenkins, 1999).

Problems with this type of inhibitor tend to centre on their lack of selectivity. Other polymerases can be inhibited, including *Taq* polymerase which is used in the amplification step of the TRAP assay. Therefore, inhibition of *Taq* polymerase could lead to false-positive results from the TRAP assay when analysing these drugs (Kelland, 2000).

4.3 Quadruplex DNA

The guanine rich telomeric overhang found at the ends of chromosomal DNA is thought to adopt a higher-ordered quadruplex structure. Stabilisation of this quadruplex structure affects the function of telomerase, as it needs a linear strand of DNA to be able to access the hTR RNA template enabling the

enzyme to add telomeric repeats (Zahler *et al*, 1991). Drugs that stabilise this quadruplex structure are also being developed as telomerase inhibitors and will be discussed below. Firstly we need to explore the quadruplex structure in more detail.

4.3.1 Quadruplex structure

Sequences of DNA rich in guanine are able to form quadruplex structures whose principal structural motif is the G-quartet. The G-quartet consists of four guanine bases in a cyclic planar arrangement, held together by Hoogstein hydrogen bonds (Figure 4.3). Two or more of these G-quartets can stack upon each other to form four stranded G-quadruplexes.

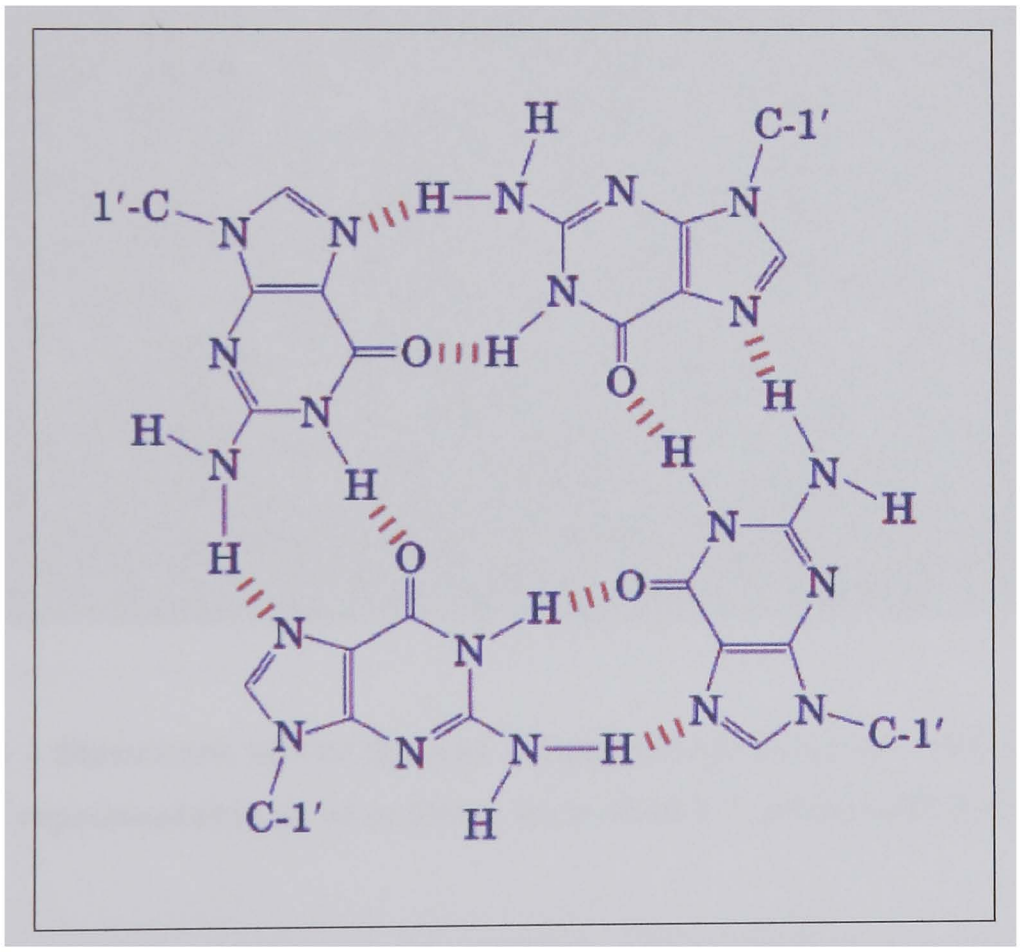


Figure 4.3 – The G-quartet structure with Hoogstein hydrogen bonds shown in red.

It is not yet clear whether quadruplex structures actually form *in vivo* but experimental studies (for example NMR, PAGE, CD) have shown that they readily form, *in vitro*, under physiologically compatible conditions of temperature, pH, ionic strength and in the presence of predominant intracellular cations (Hardin *et al*, 1992). The presence of cations is thought

to play a crucial role in the stability of quadruplexes. Structural and thermodynamic studies have shown that cations induce and stabilise quadruplex formation in the order $Cs^+ < Rb^+ < K^+ < Na^+ < Li^+ <$ for monovalent cations and $(K^+) > Ca^{2+} > Mg^{2+}$ for divalent cations. The cations are thought to bind in a central channel between the quartets (Figure 4.4) and are co-ordinated by electronegative guanine O6 atoms. This central channel is a unique feature of G-quartets and is thought to distinguish telomeric DNA structures from other DNA structures that are based primarily on hydrogen bonding interactions (Williamson, 1994).

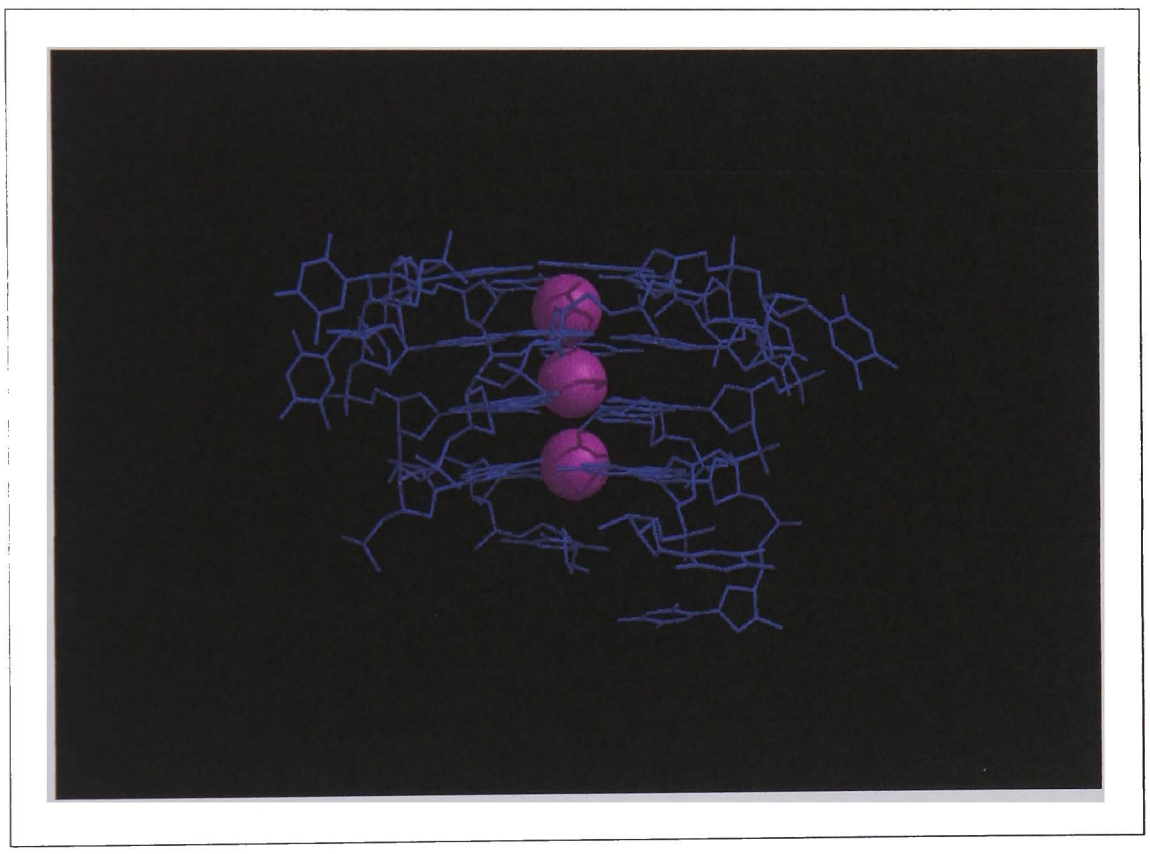


Figure 4.4 – Structure of the human telomeric quadruplex with space-filled representations of sodium ions in the central channel.

Molecular modelling studies by Spackova *et al* (1999) have shown that the number of ions present can have effects on stability. Molecular dynamics simulations were carried out on quadruplex with 3 Na^+ ions, 2 Na^+ ions and no Na^+ ions. The results showed a decline in stability as ions were reduced, leading to an unstable structure which began to fall apart when no ions were present. Analysis of these simulations by PCA (Chapter 2) confirmed that this disruption of the quadruplex structures was categorised by "slipped" and "spiral" structures (Grindon, unpublished results).

Although the G-quartets form the major structural feature of quadruplex DNA, it is also able to form a number of different morphological structures characterised by the number of strands used to form the quadruplex and variation of the sugar glycosidic torsion angle. One, two or four strands can be used to form varying types of quadruplex structure as shown in Figure 4.5. Single stranded structures are known as intramolecular structures whereas two and four stranded structures are known as intermolecular with two stranded structures often referred to as hairpin structures.

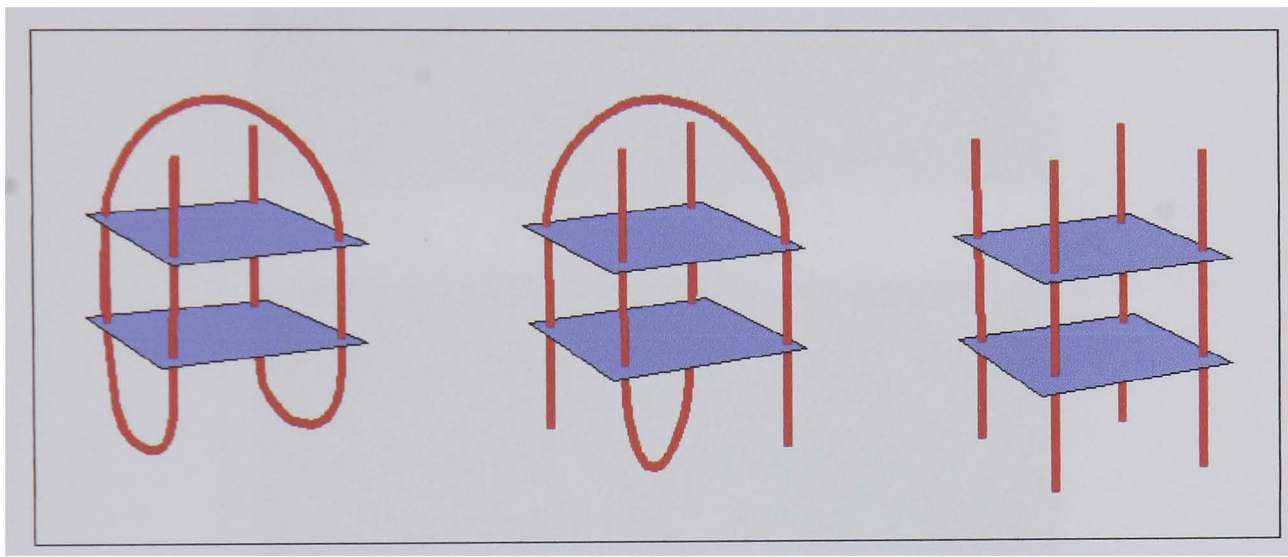


Figure 4.5 – Structures of 1,2 and 4 stranded quadruplexes.

Another source of structural variation is the relative arrangement of the guanine strands, for example, all parallel, adjacent parallel/anti-parallel and alternating parallel/anti-parallel, depending on their directionality (Neidle & Parkinson, 2003).

The loops connecting guanine tracts in intramolecular and two stranded intermolecular quadruplex formations can run in a number of different ways, normally connecting diagonally or edge-wise (Simonsson, 2001). The human telomeric sequence in solution with Na^+ ions, as shown by NMR (Wang & Patel, 1992), is thought to adopt an intramolecular alternating parallel/anti-parallel structure, with edge-wise TTA loops at the bottom and a diagonal TTA loop at the top (Figure 4.6).

More recently a crystal structure of the human telomeric sequence has been characterised with K^+ ions (Parkinson *et al*, 2002) and shows a very different

intramolecular parallel structure with the TTA loops around the sides of the G-quartet core (Figure 4.6).

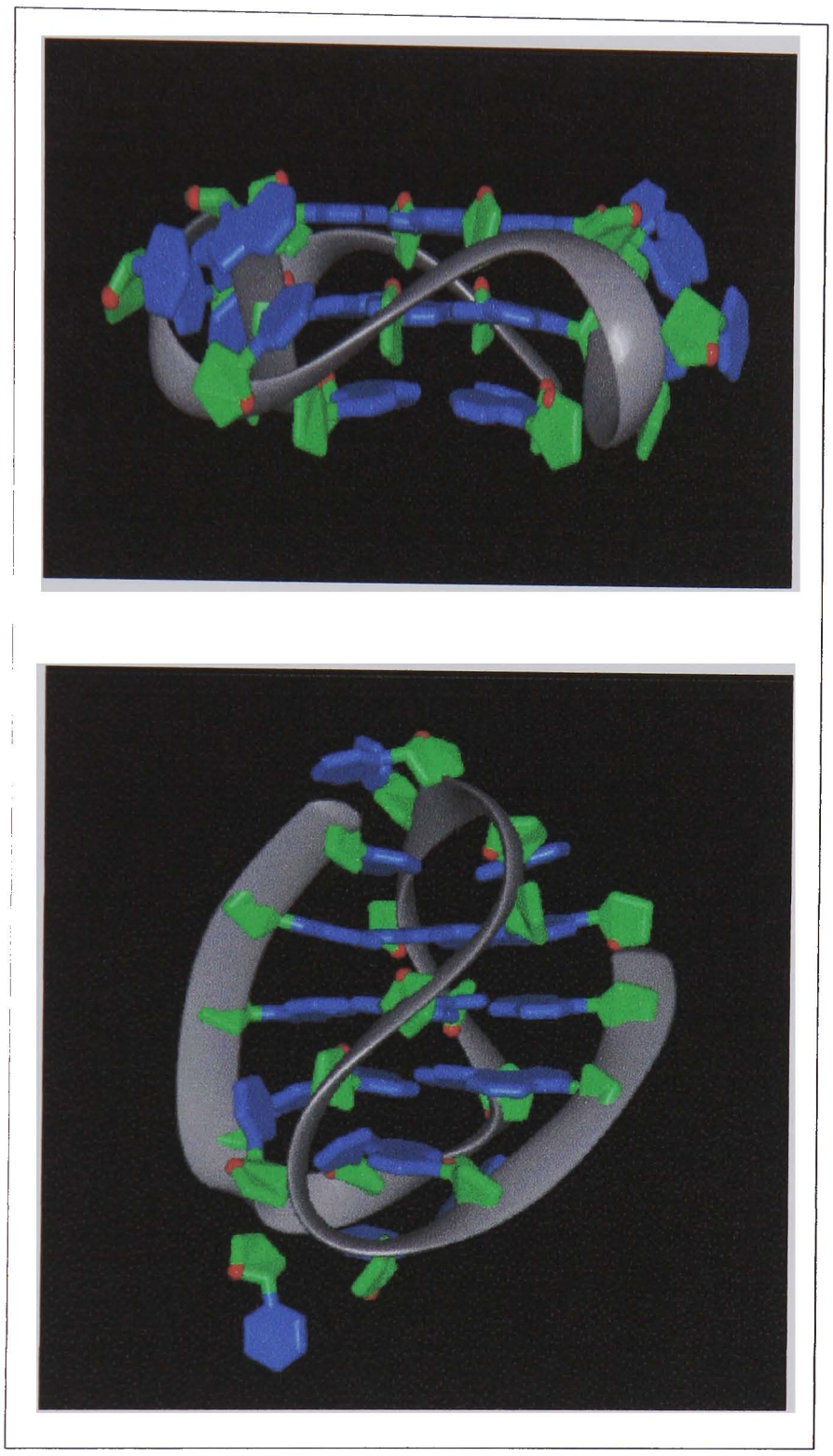


Figure 4.6 – The crystal structure (top) with parallel strands and TTA loops around the side and the NMR structure (bottom) with anti-parallel strands and TTA loops at the top and bottom.

The final structural feature of quadruplexes is due to the angle of the glycosidic torsion and results in differing groove widths. Bases in normal B-form DNA are found exclusively in the *anti* conformation whereas guanines involved in G-quartet formation can adopt either *syn* or *anti* conformations (Figure 4.7).

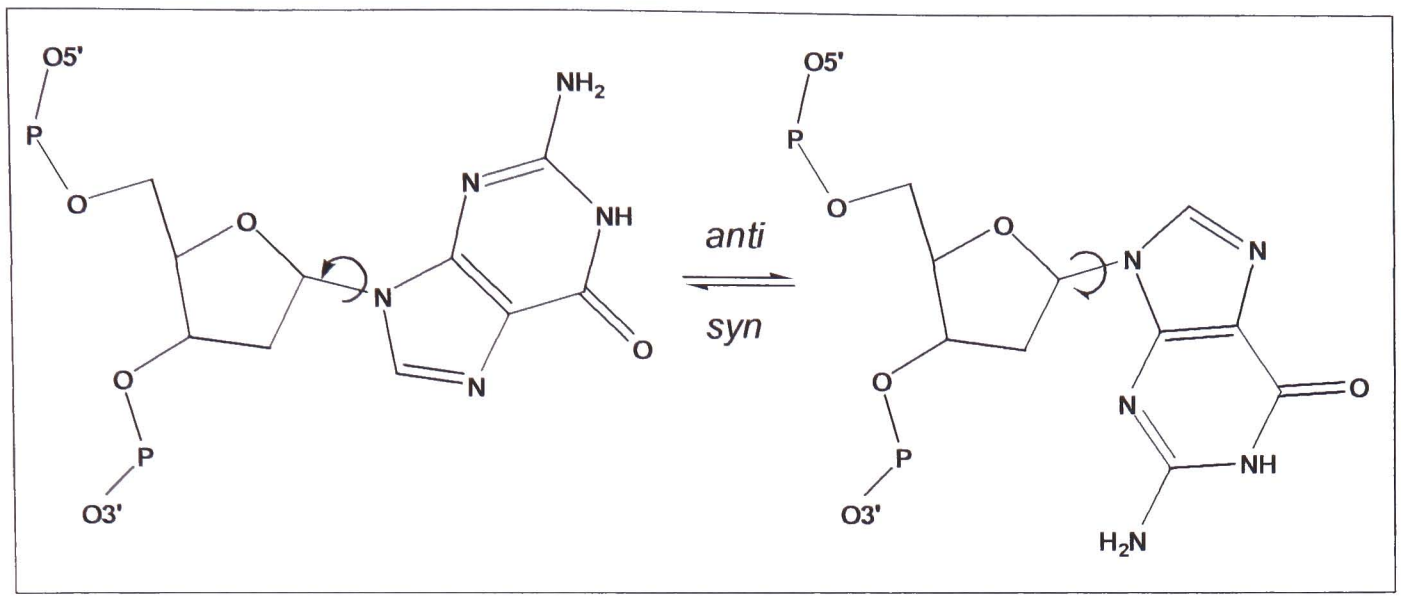


Figure 4.7 – Rotation around the glycosidic bond enabling a guanine base to interconvert between *syn* and *anti* conformations.

Different groove widths are found depending on the *syn-anti* conformations around a G-quartet and can result in wide, narrow and medium groove widths. If all four strands involved in a G-quartet are parallel (as in the telomeric crystal structure), the four grooves are all of medium size due to the fact that bases on parallel strands must always have the same glycosidic torsion angle. Bases on anti-parallel strands must then have opposite glycosidic torsion angles resulting in combinations of wide, medium and narrow grooves as shown in Figure 4.8. The NMR structure of the anti-parallel intramolecular telomeric quadruplex has two medium, one wide and one narrow groove.

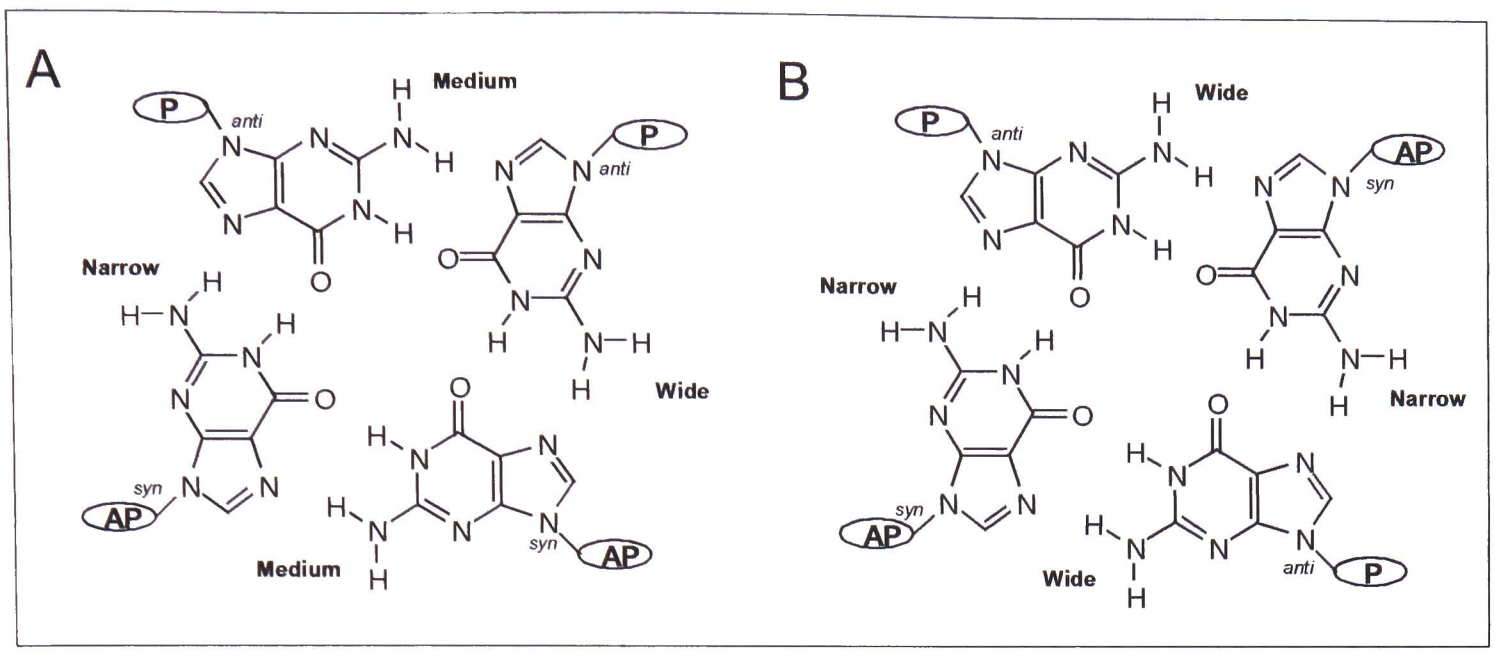


Figure 4.8 – A shows groove widths for two adjacent parallel (P) and anti-parallel (AP) strands and B shows groove widths for alternating parallel (P) and anti-parallel (AP) strands (adapted from Simonsson, 2001).

4.3.1.1 Regions of quadruplex DNA outside of telomeres

Quadruplex structures are not just found at telomeric regions of DNA. They are also thought to exist in the *c-myc* promoter, the fragile X syndrome triplet repeat and the thrombin binding aptamer, for example. These all contain guanine rich sequences, although not all have G_4 quartets as described above. For example the fragile X syndrome triplet repeat of $d(CGG)_n$ forms GCGC quartets (Kettani *et al*, 1995).

The thrombin binding aptamer d(GGTTGGTGTGGTTGG) has been shown by NMR to form a quadruplex structure. The quadruplex consists of two G4 quartets connected by two TT loops at one end and a TGT loop at the other, also a potential T.T base pairing is formed between the two TT loops (Macaya *et al*, 1993).

The c-myc promoter region of the human *c-myc* oncogene forms a stable Watson-Crick double helix under physiological conditions, although addition of potassium or a ligand can induce quadruplex formation. Stabilisation by potassium is a slow process but Rangan *et al* (2001) discovered that upon addition of a small molecule they had synthesised, PIPER, they could rapidly form quadruplex structures from the *c-myc* promoter region.

4.3.2 Inhibition of telomerase by quadruplex stabilising drugs

The requirement of a positive ion to stabilise quadruplexes has led to research into ligands that can mimic this positive ion and therefore enforce stability of the quadruplex. The development of quadruplex stabilising drugs has largely stemmed from experience with intercalating drugs for duplex and triplex DNA. Most drugs to date are based on a planar multi-ringed chromophore where binding is expected to occur due to π orbital overlap with the G-quartets. As with duplex intercalators, side chains can be added to help with binding. This approach was supported by the fact that ethidium bromide, an intercalating dye which binds to duplex DNA, was also found to bind to quadruplex DNA with slightly better binding affinities (ΔG = -6.8 kcal/mol for duplex and ΔG = -7.3 kcal/mol for quadruplex DNA; Guo *et al*, 1992).

Evidence to support the mechanism of quadruplex stabilising drugs as potential telomerase inhibitors came from the observation that inhibition only occurred after 3/4 telomere repeats had been synthesised. This was consistent with the requirement of four repeats of human telomere d(TTAGGG) for an intramolecular quadruplex structure to form (Sun *et al*, 1997).

Examples of drugs which can stabilise quadruplex DNA are porphyrins, perylene type compounds, telomestatin, anthraquinones and acridines. Below are a few examples with details of their merits and limitations.

4.3.2.1 Example: Porphyrins

The porphyrin TMPyP4 and its structural isomer TMPyP2 (Figure 4.9) have been evaluated for their telomerase inhibition properties. TMPyP4 is a potent telomerase inhibitor with an IC_{50} value of 6 μM whereas TMPyT2 does not inhibit telomerase (IC_{50} ~ 250 μM; Izbicka *et al*, 1999 (a)). The differences in these properties are thought to be due to the differences in interaction with G-quartets. TMPyP2 is sterically hindered from stacking interactions within

the quartet region whereas TMPyP4 is not. This is due to the N^+-methyl groups of the rings pointing out into the groove regions of G-quartets (as can be seen in Figure 4.9) where they can also make electrostatic interactions with the negatively charged backbone (Anantha *et al*, 1998).

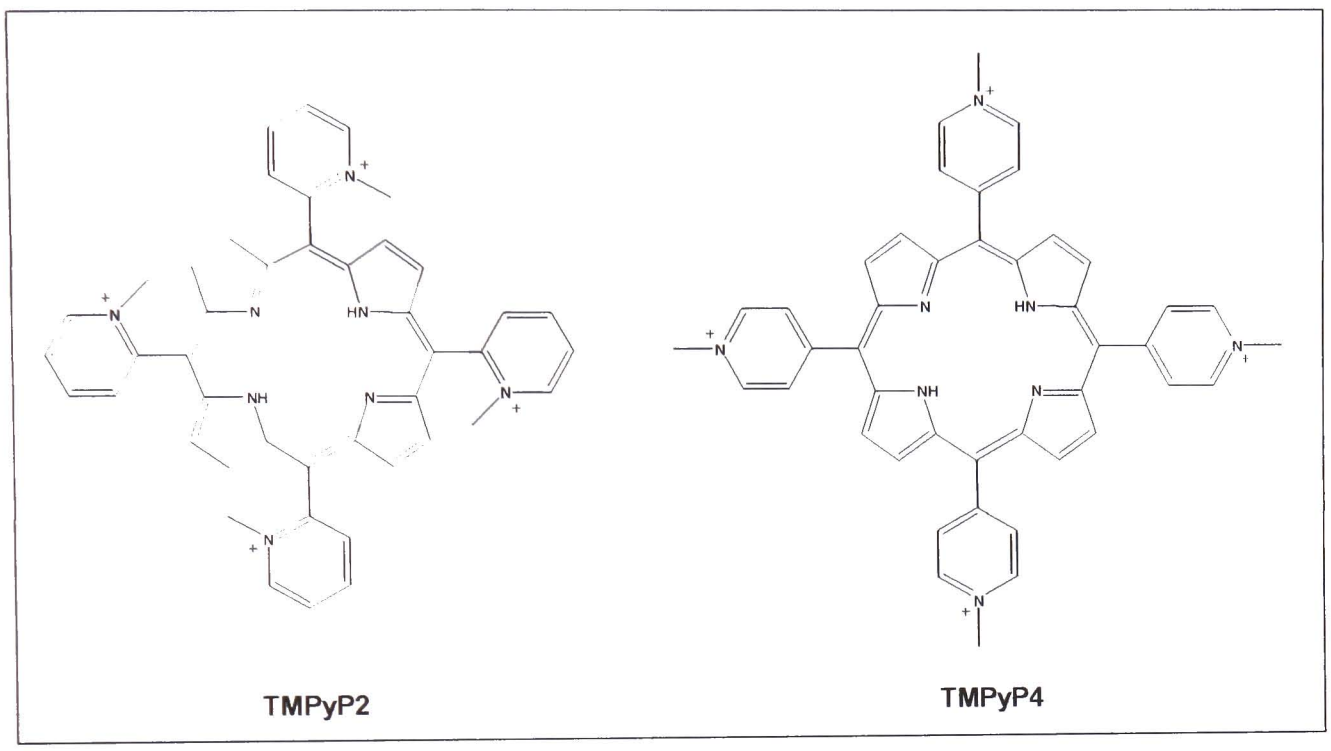


Figure 4.9 – Structures of the porphyrins TMPyP2 and TMPyP4.

Molecular modelling studies have shown that the most favourable binding site for these compounds is end-stacking intercalation in the loop region of an intramolecular quadruplex and not intercalation between G-quartets (Han *et al*, 2001). A contradictory study using Isothermal titration calorimetry and spectroscopic techniques has shown that intercalation between the G-quartets is the most favoured binding site for these types of drugs (Haq *et al*, 1999).

4.3.2.2 Example: Perylenes

The perylene type compound, PIPER (Figure 4.10), has been identified as a potent telomerase inhibitor and has also been shown to accelerate the formation of a variety of quadruplex DNA structures (Han *et al*, 1999). NMR studies of PIPER and the parallel quadruplex d(TTAGGG)4 have shown that a 2:1 DNA-drug complex is formed with the drug sandwiched between the two G-quartets at the ends of the quadruplexes with the positively charged

side chains located in the grooves (Fedoroff *et al*, 1998), there is no intercalation between G-quartets.

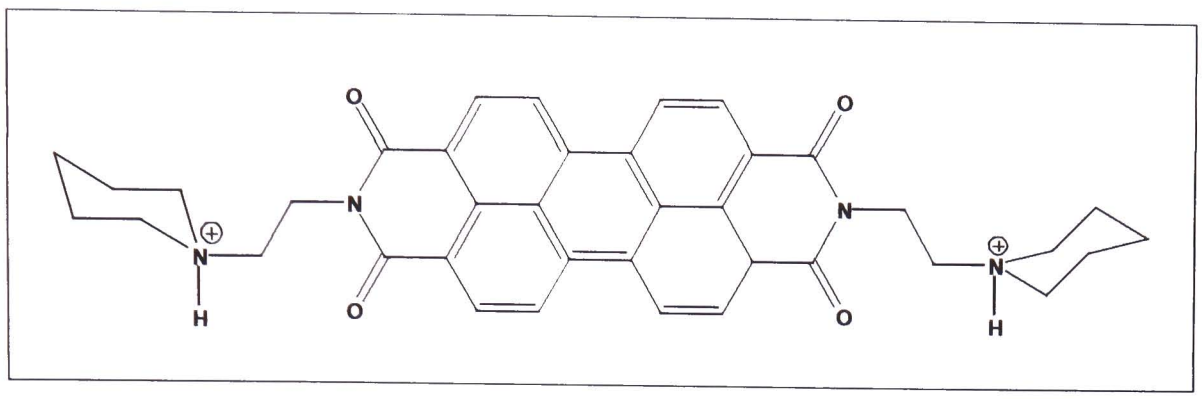


Figure 4.10 – Structure of PIPER.

4.3.2.3 Example: Telomestatin

The screening of a wide range of compounds for potent telomerase inhibition led to the isolation of telomestatin (Figure 4.11) from *Steptomyces anulatus*. Telomestatin specifically inhibited telomerase, without affecting DNA polymerases and reverse transcriptases such as *Taq*, with an IC_{50} value of 0.005 μM (Shin-ya *et al*, 2001). These results showed that telomestatin was the strongest and most specific telomerase inhibitor ever reported. Further studies carried out by Kim *et al* (2003) have shown that telomerstatin also has a preference for intramolecular quadruplexes (like the human telomeric quadruplex) over other types of quadruplex structure.

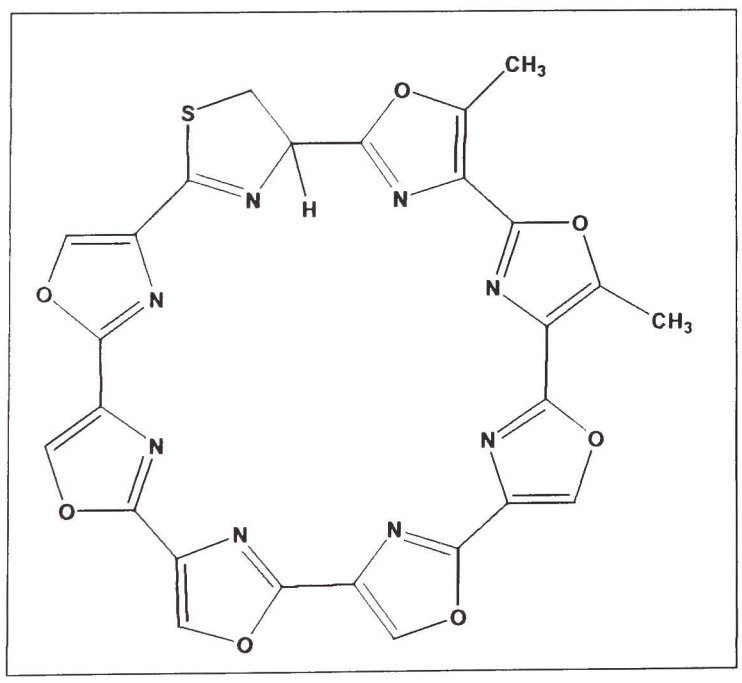


Figure 4.11 – Structure of Telomestatin.

4.3.2.4 *Example: Anthraquinones*

The anthraquinones were some of the first drugs to be used as quadruplex binders after they were originally studied for their duplex binding nature (Collier & Neidle, 1988). A wide range of substituted amido-anthraquinones have been synthesised and evaluated via the TRAP assay (Perry *et al*, 1998 (a)). Whilst the number and nature of the substituents seems to be important for activity, their positions around the anthraquinone chromophore (Figure 4.12) is much less so (Neidle *et al*, 2000). The best of these compounds have IC_{50} values in the range 1-10 μM (Perry *et al*, 1998 (b)). The down side of these compounds is that they inhibit *Taq* polymerase, an indication of reduced telomerase selectivity.

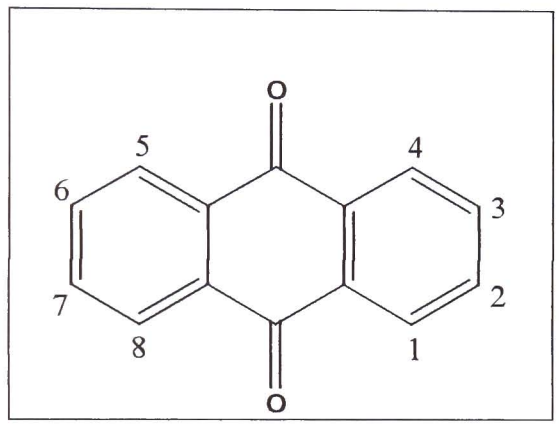


Figure 4.12 – The anthraquinone chromophore with numbered substitution sites.

4.3.2.5 *Example: Acridines*

Like the anthraquinones, a number of substituted acridines have been synthesised for use as quadruplex stabilising drugs (Harrison *et al*, 1999; Read *et al*, 1999). Di-substituted acridines were found to have similar properties to the anthraquinones but tri-substituted acridines (Figure 4.13) have some interesting results. Whilst the di-substituted compounds had IC_{50} values in the 10s of μMs adding an extra substituent reduced this greatly with the most potent inhibitor having an IC_{50} value of 0.006 μM (Read *et al*, 2001). These 3,6,9 tri-substituted acridines also had low potency in cytotoxicity assays, showing that they have less affinity for duplex DNA (confirmed by their binding constants – K(duplex) in the 10^5 region and K(quad) in the 10^7

region). These were the first quadruplex stabilising drugs to show selectivity of quadruplex over duplex. An added bonus was that these compounds showed no inhibition of *Taq* polymerase.

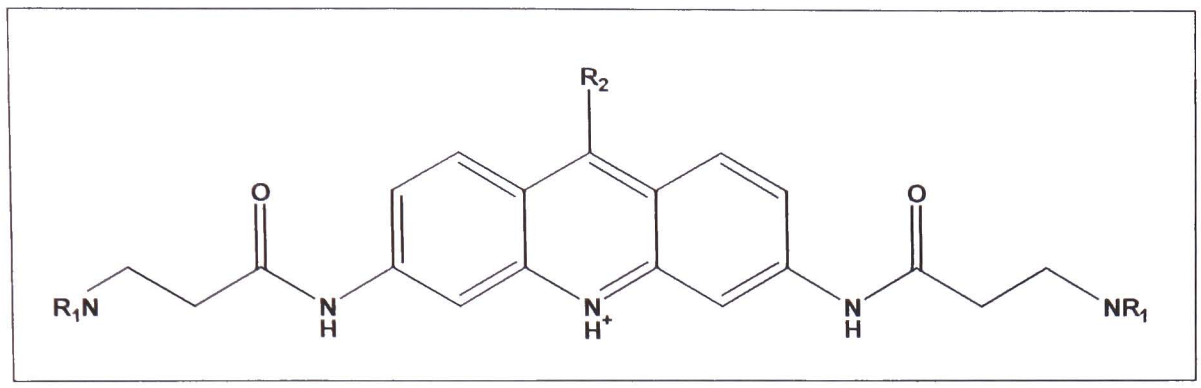


Figure 4.13 – The substitution pattern of the 3,6,9-tri-substituted acridines.

4.3.3 The issues surrounding selectivity

The issue of selectivity is very important for quadruplex stabilising drugs, as stabilisation of duplex DNA would result in unwanted side effects, as seen with conventional unselective chemotherapy. A drug needs to have a low molar IC_{50} from telomerase inhibition studies but not as low molar GI_{50} (50% growth inhibition) from cytotoxicity studies, a rough guide is that $IC_{50} < 0.1$ GI_{50} (Neidle *et al*, 2000). This is because of the suggested need for these types of inhibitors to be administered over a number of rounds of cell division before telomere erosion and senescence occurs. The concentration of inhibitor will have to be significantly below acute toxicity levels otherwise general cytotoxic cell kill will take place.

Most of the quadruplex stabilising drugs reported to date, do not have selectivity for quadruplex over duplex DNA and unsurprisingly their levels of cytotoxicity and telomerase inhibition have been found to be comparable (Read *et al*, 2001). Modest selectivity for quadruplex has been seen with some substituted porphyrins (Arthanari *et al*, 1998). More pronounced selectivity has been seen with some polycylic quino-acridines synthesised in our labs in Nottingham (Gowan *et al*, 2001; Heald *et al*, 2002 (a)) and by the 3,6,9 tri-substituted acridines discussed above.

Drugs are now being developed to bind specifically to quadruplexes by means of substituent side chains that are designed to interact with the four grooves of quadruplex DNA (Figure 4.14). Synthesis and subsequent biochemical testing of these compounds is a time consuming process. Molecular modelling approaches are being developed (Read & Neidle, 2000; Cairns *et al*, 2002) with the hope that they will be able to predict binding affinities of potential new selective compounds and so direct the synthesis of these compounds.

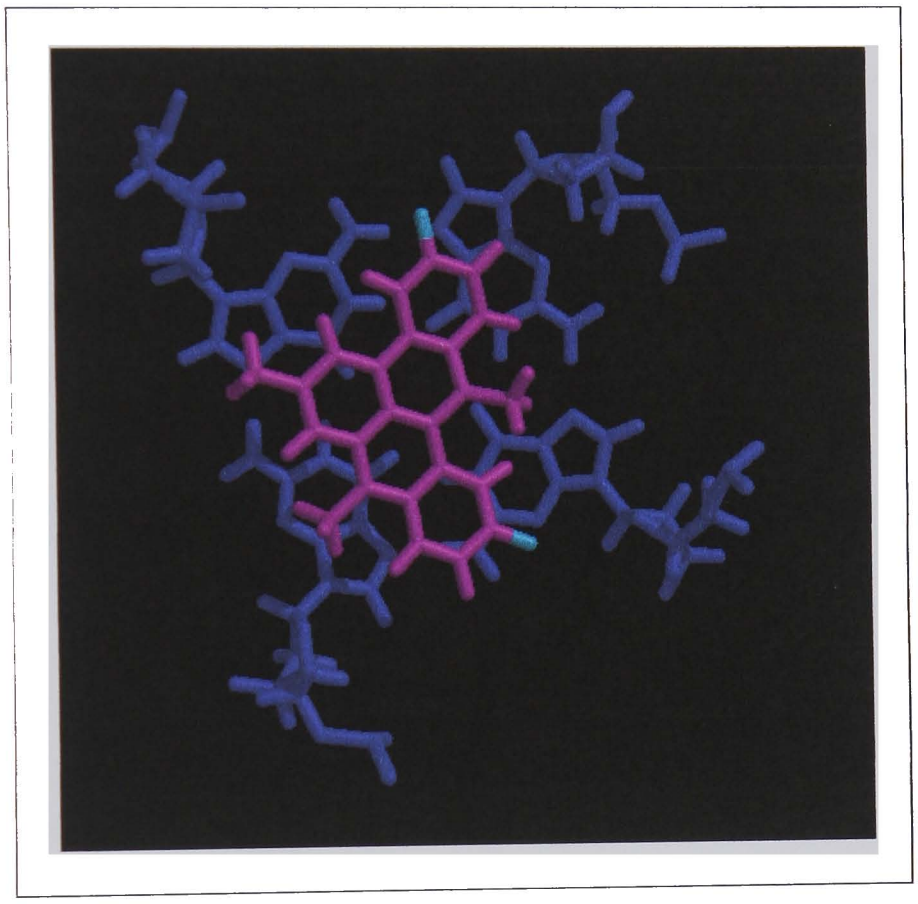


Figure 4.14 – Overlay of a quadruplex stabilising drug, (RHPS4; magenta) with a G-quartet (blue). The cyan regions of the drug indicate fluorines which could be replaced by substituents which would interact more strongly with the grooves.

4.4 Energetics of DNA-drug interactions

Understanding DNA-drug interactions, and the thermodynamics of these reactions, is of fundamental and practical importance to the study of drug design. On a practical level, thermodynamics can offer key insights into the molecular forces that drive DNA-drug complex formation that cannot be calculated from structural studies alone. Fundamentally, these interactions

can reveal predicted binding free energies and details about the components of free energy itself (Chaires, 1998).

The energetics of DNA-drug interactions have been studied and it has been found that, in general, the forces driving the interaction of both groove binders and intercalators are enthalpic. Drugs which edge bind or only partially insert into the helix are more likely to be entropically driven (Haq & Ladbury, 2000). The magnitude of the energetics is also of importance, for example, a large favourable enthalpy indicates formation of hydrogen bonds and/or vdW contacts, whereas a less favourable enthalpy is more likely to indicate removal of bound water molecules into the bulk solution.

A model for intercalative binding (based on a model for protein-ligand interactions; Ross & Subramanian, 1981) shows the energy breakdown upon the different steps involved in binding a drug. The formation of an intercalation complex requires three steps each with their own energetic profile (Figure 4.15).

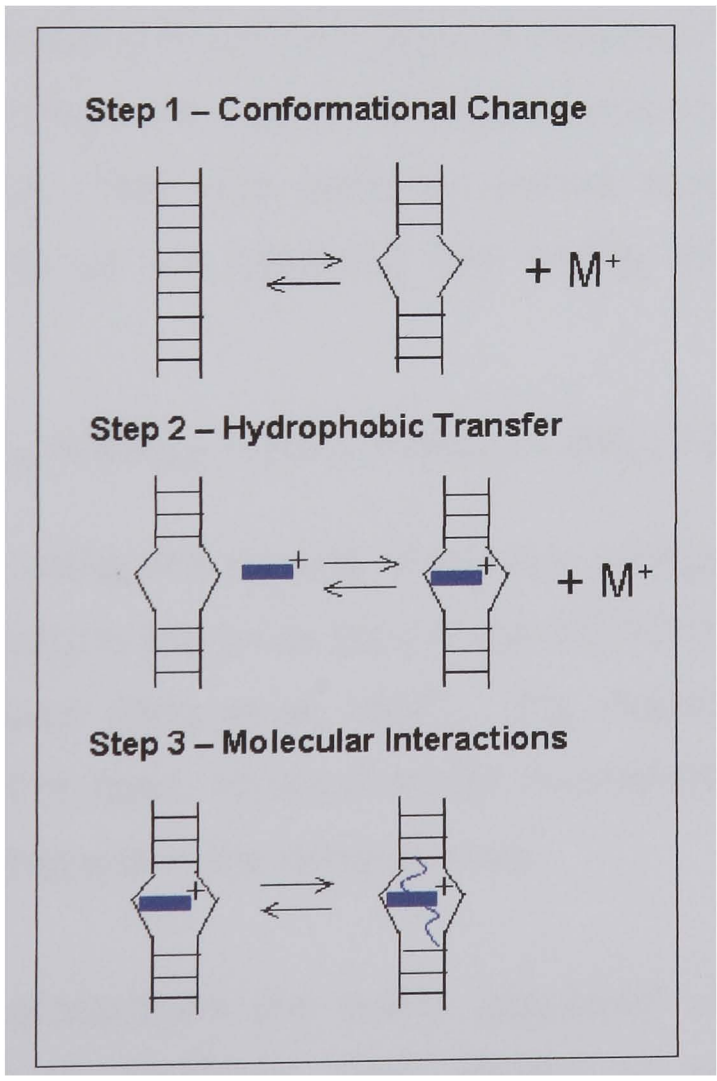


Figure 4.15 – Steps involved in the intercalation of a drug (adapted from Chaires, 1998).

Step 1 involves the formation of a binding site within the DNA (ΔG_{conf}). During this step the DNA must unwind and base pairs separate to form the cavity into which the drug will bind. Also at this step, there will be a release of water and counter-ions, due to the change in charge density as the phosphate groups of DNA are separated (ΔG_{pe}). This conformational change is associated with a positive free energy change. Step 2 involves the transfer of the drug from solution into the binding site (ΔG_{hyd}). This is a hydrophobic process as the non-polar aromatic ring system of the drug is removed from solution and placed into the cavity formed in step 1. This hydrophobic transfer is associated with a negative free energy change and will also result in release of more counter-ions if the drug is positively charged. The final step involves molecular interactions such as hydrogen bonding, vdW, electrostatic, and stacking interactions to tether the drug into position (ΔG_{mol}). This step is also associated with a negative free energy change. Each of these steps contributes to the overall binding free energy.

The method of parsing free energy into contributions (Chaires, 1998 and references within) involves these four types of contributions, but also another type – a contribution from the loss of translational and rotational motion upon drug binding (ΔG_{t+r}). This then gives an overall equation describing the contributions making up the observed free energy change upon binding ΔG_{obs}.

$$\Delta G_{obs} = \Delta G_{conf} + \Delta G_{t+r} + \Delta G_{hyd} + \Delta G_{pe} + \Delta G_{mol}$$

A detailed analysis using this parsing of the free energy was carried out on Hoechst 33258 binding to the minor groove the d(CGCAAATTTGCG)$_2$ duplex to form a 1:1 complex (Haq *et al*, 1997). The majority of the favourable binding energy, in this case, arises from the hydrophobic transfer of ligand from solvent to binding site in the minor groove.

Not all of these contributions are easily calculated via experimental and theoretical techniques therefore, free energy of binding is generally

calculated using binding constants (K_b) or enthapic (ΔH) and entropic ($T\Delta S$) values via the following equations.

$$\Delta G_{bind} = -RT \ln K_b \qquad \Delta G_{bind} = \Delta H - T\Delta S$$

4.4.1 Calculations of ΔG_{bind} by experimental methods

To measure binding free energies by means of binding constants (K_b) there are many experimental techniques used, for example, Spectroscopic titrations, surface plasmon resonance (SPR) and Isothermal titration calorimetry (ITC).

Spectroscopic titrations involve the titration of a drug into a known concentration of DNA (or vise versa) to construct equilibrium binding isotherms. The data can then be analysed using thermodynamic models to obtain the binding constant (Chaires, 2001).

SPR involves immobilizing one of the reacting species (typically the DNA) onto a gold sensor chip. The other component (the drug) is then passed over the chip and binding interactions are detected by measuring the refractive index of the chip. The data is analysed by construction of a binding curve (Davis & Wilson, 2001).

In Isothermal titration calorimetry (ITC) a binding curve is obtained, defined by the amount of heat released or absorbed as a function of the total concentration of ligand, and analysed using an appropriate model (Haq, 2002 and references therein). ITC can also be used to calculate changes in enthalpy and hence entropic contributions can also be inferred via the equation above.

4.4.2 Calculation of ΔG_{bind} by computational techniques

The calculation of ΔG_{bind} by computational techniques has been hampered by the lack of a universally accepted method for calculating entropic terms.

Many methods have been developed to get round this problem including LIE (discussed in chapter 2) and MMPBSA (part of the AMBER6 suite of programs; Case *et al*, 1999), which do not involve the calculation of configurational entropy terms but implicitly take into account solvent associated terms.

There have been few groups who have tried to calculate ΔG_{bind} for quadruplex stabilising drug via computational techniques. Cairns *et al*, (2002) studied the binding of anthraquinones to quadruplex DNA via computational binding enthalpies ΔE_{bind}. Their binding enthalpies were calculated using the equation $\Delta E_{bind} = E_{complex} - (E_{drug} + E_{DNA})$, a similar method to how we have calculated ΔG_{bind} (see section 4.7.1 later) but without the need for entropic data. Although the method neglects any entropic contributions, they have found good correlations between experimental and computational data.

Read *et al*, (1999; 2001) have used a method adapted from the LIE approach to calculate binding affinities for their acridine based drugs. They have found that although the actual energies are not identical to results from experiments, the rankings from this method are in accord with biochemical measurements of telomerase inhibition and binding enthalpies determined via ITC.

In the rest of this chapter we will evaluate a number of different computational methods for predicting binding affinities and interactions using quadruplex DNA and two new polycyclic quino-acridines synthesised within our labs in Nottingham.

4.5 Polycyclic quino-acridines – a novel class of telomerase inhibitors

At Nottingham, a series of biologically active polycyclic quino-acridines have been developed (Heald, 2002 (b)) based on acridine compounds. The lead compound of the series is RHPS4, a difluorinated derivative (Figure 4.16)

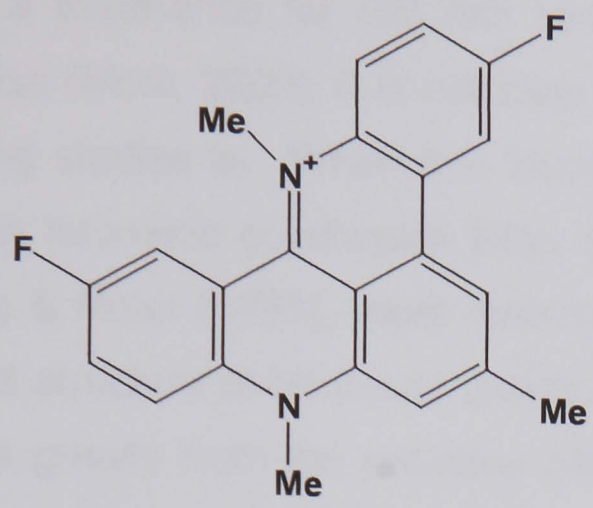
which shows high and selective affinity, *in vitro*, for quadruplex DNA and good telomerase inhibitory activity via the TRAP assay (Kim *et al*, 1994). A related molecule RHPS3 (Figure 4.16), shows much less selectivity for quadruplex over duplex DNA, but a similar TRAP assay result. Figure 4.16 shows the structures of RHPS4 and RHPS3 along with experimental results including TRAP assay, GI_{50}, Competition dialysis and SPR binding.

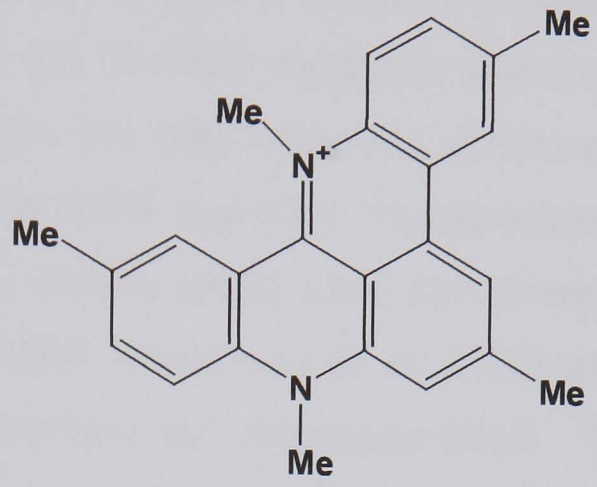
As already stated both drugs have similar TRAP assay results with RHPS3 being a slightly better telomerase inhibitor. The GI_{50} results are higher for RHPS4 than RHPS3 indicating that RHPS4 is more selective to quadruplex DNA and therefore less cytotoxic than RHPS3. The competition dialysis results (Modi, 2002) show that RHPS4 has a greater affinity for higher ordered structures than RHPS3 and the SPR binding constants (K_b; Wilson, unpublished results) show that the selectivity of RHPS4 for quadruplex over duplex DNA is due to RHPS4 having stronger binding to quadruplex and poorer binding to duplex DNA in comparison to RHPS3.

Lack of selectivity for quadruplex DNA is undesirable as it is correlated with general unselective cytotoxicity, an unacceptable phenomena for a selective drug. It is because although structurally very similar, there are major differences in binding and selectivity for RHPS4 and RHPS3 (described above), that these two molecules were chosen to be studied by molecular modelling. The aim of this modeling being to try to understand the origins of their differing affinities, so as to guide the future development of the series of drugs.

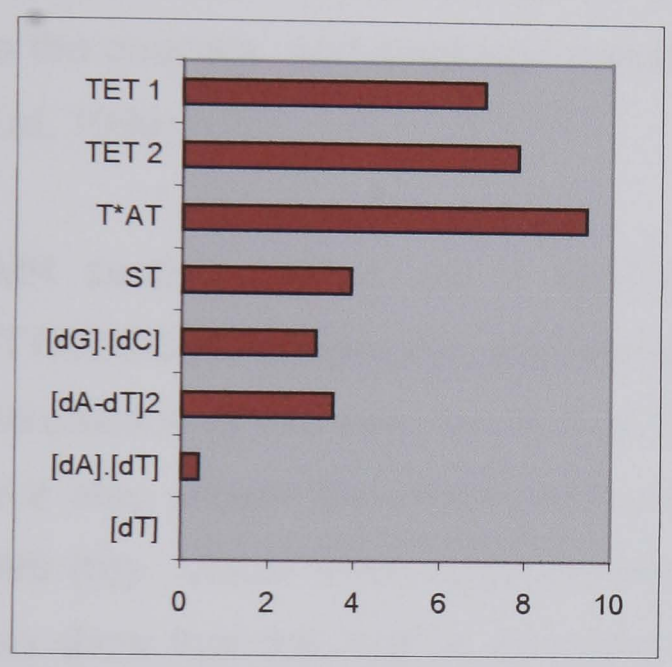
RHPS4 **RHPS3**

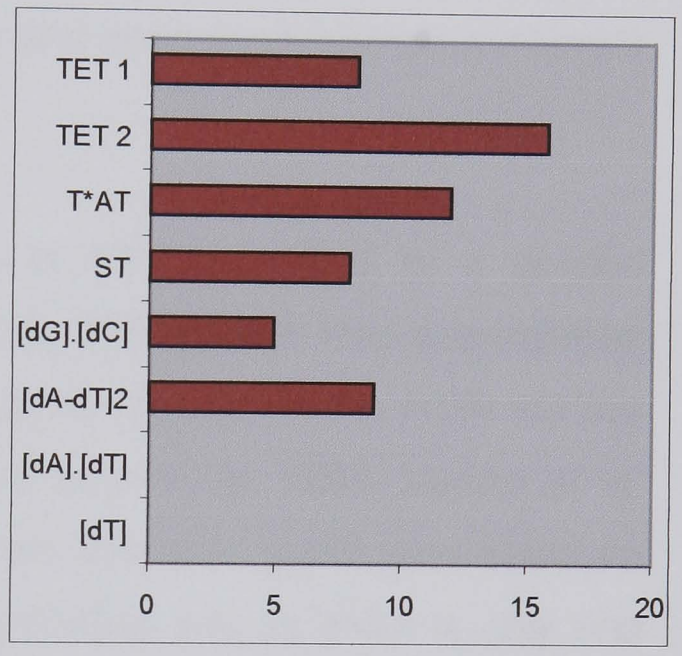
A





B





Key – Y axis: TET1 = $(T_2G_{20}T_2)_4$ quadruplex; TET2 = $AG_3(TTAGGG)_3$ quadruplex; T*AT = poly (dT).poly (dA).poly (dT) triplex; ST = Salmon testes duplex; [dG].[dC] = poly (dG).poly (dC) duplex; [dA-dT]2 = poly (dA.dT)2 duplex; [dA].[dT] = poly (dA).poly (dT) duplex; [dT] = poly (dT) single strand. X axis: µM bound drug.

C

Drug	TRAP IC_{50}	GI_{50}	K_b duplex	K_b quad
RHPS4	0.31 µM	13.18 µM	3.4×10^5	11.0×10^6
RHPS3	0.25 µM	0.40 µM	5.4×10^5	7.4×10^6

Figure 4.16 – A shows the structures, B shows the competition dialysis results and C shows the TRAP, GI_{50} and binding constants for RHPS4 and RHPS3.

4.5.1 Rationale for modeling studies

While there is good evidence that the quino-acridines bind to duplex DNA (with a preference for GC rich sequences) through a simple intercalation process (Modi, 2002), it is not clear what the "correct" model for quadruplex binding studies is. When this study began the only observed structure for human telomeric quadruplex DNA was the NMR structure characterised by Wang & Patel (1993), more recently the Neidle group have put forward a crystal structure of telomeric quadruplex DNA (Parkinson *et al*, 2002) which differs greatly from the previous NMR structure (as discussed later). This NMR structure has many sites for possible binding of ligands; intercalation via the quartets, end-stacking intercalation and even groove binding (Mergny *et al*, 1999).

NMR studies carried out in Nottingham of RHPS4 bound to a parallel $d(TTAGGGT)_4$ quadruplex (Gavathiotis *et al*, 2001) show that end-stacking intercalation is the most favored of these binding sites. Other NMR studies have also shown this phenomenon (Wheelhouse *et al*, 1998; Izbicka *et al*, 1999 (b)). Visual observations of the human telomeric quadruplex structure also show that the "top" is the most likely binding site as there is only one loop obstructing access to the binding region, whereas there are two at the "bottom". Therefore for all quadruplex studies the NMR structure was used and the "top" end-stacked binding region was explored further (Figure 4.17).

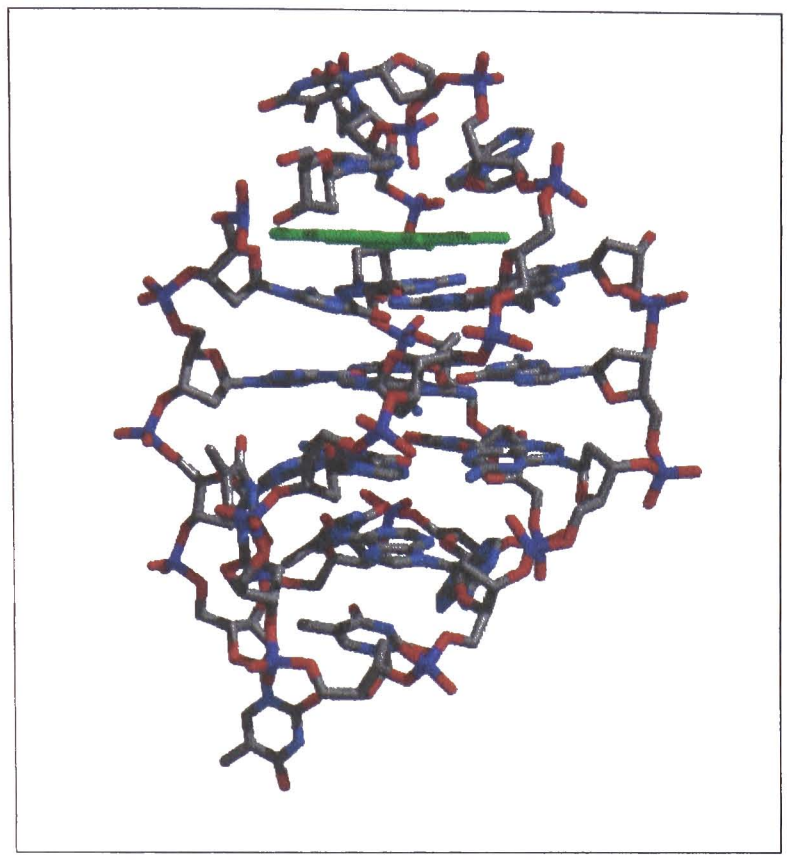


Figure 4.17 – Model of the NMR structure showing the binding of RHPS4 in the "top" end-stacked binding site.

For the duplex studies the hexamer $d(CGCGCG)_2$ was used with the middle CG binding region explored further. The aim of these initial modeling studies was to find the "correct" binding orientations for the RHPS4 within the chosen binding sites of both quadruplex and duplex DNA. Once the "correct" binding orientations were found it was possible to model RHPS3 to look for correlations between theoretically derived and experimentally derived binding affinities. Due to time limitations it was not possible to model any other drugs from the series, nor was it possible to model any of the drugs in the crystal structure.

4.6 Methods

4.6.1 Generation of models

The quadruplex model was taken from Patel's NMR structures found in the Protein Database (PDB ID no. 143D). Of the six structures, the one with the lowest average RSMD when compared to all other structures was chosen. A series of restrained energy minimisations were carried out to create an

intercalation site of approximately 6.8Å (equivalent to two base steps). The ligand was then manually docked into the binding site and two sodium ions also docked in-between the quartets to help stabilisation. The system was electrically neutralised by addition of sodium counter-ions and immersed in a periodic box of TIP3P water molecules (approximate size, 47Å x 58Å x 47Å) before a three step minimisation was carried out (step 1 – DNA restrained, step 2 – loop region of DNA restrained, step 3 – no restraints) to obtain a low energy structure. To obtain the different orientations of the ligand within this binding site the visualisation program Midas (Ferrin *et al*, 1988) was used in conjunction with Crystal Eyes (website 1) stereo specs for 3-D visualisation. Eight different orientations were found (Figure 4.18) and the resulting systems subjected to a further two step minimisation (step 1 – DNA and ligand restrained, step 2 – no restraints) to allow rearrangement of water molecules, ensuring low energy structures with no steric clashes.

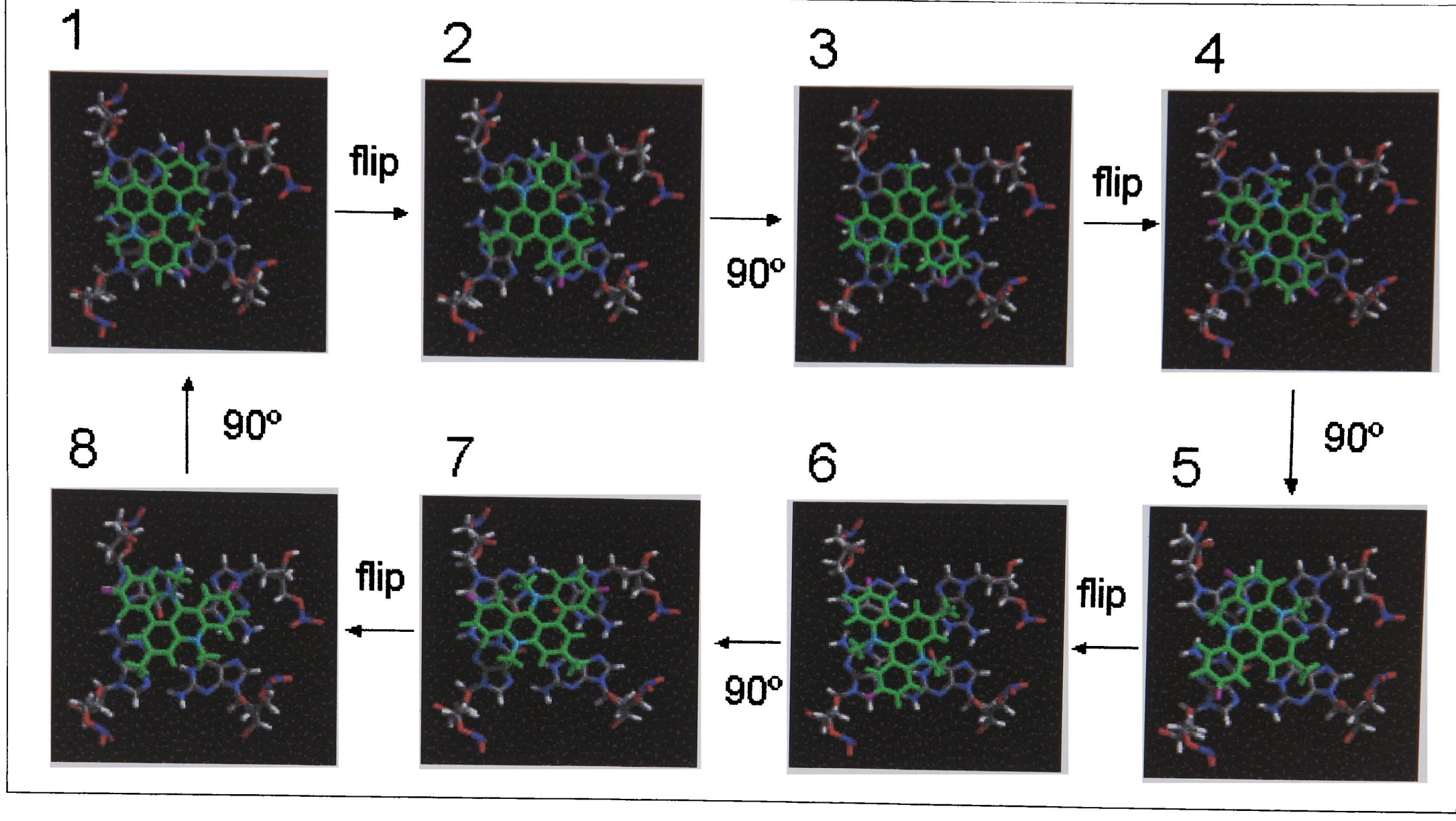


Figure 4.18 – The 8 different orientations of RHPS4 within quadruplex DNA and how they were obtained (drug coloured green with fluorines coloured magenta).

The duplex model was generated using the *nucgen* program, within the AMBER6 suite of programs (Case *et al*, 1999), in standard B-form. The intercalation site was created, ligand docked, and the system solvated (approximate periodic box size, 45 x 44 x 49Å) and minimised as described above. Three different orientations of ligand were found (Figure 4.19) via the use of Midas and Crystal Eyes, and final minimisation carried out as above.

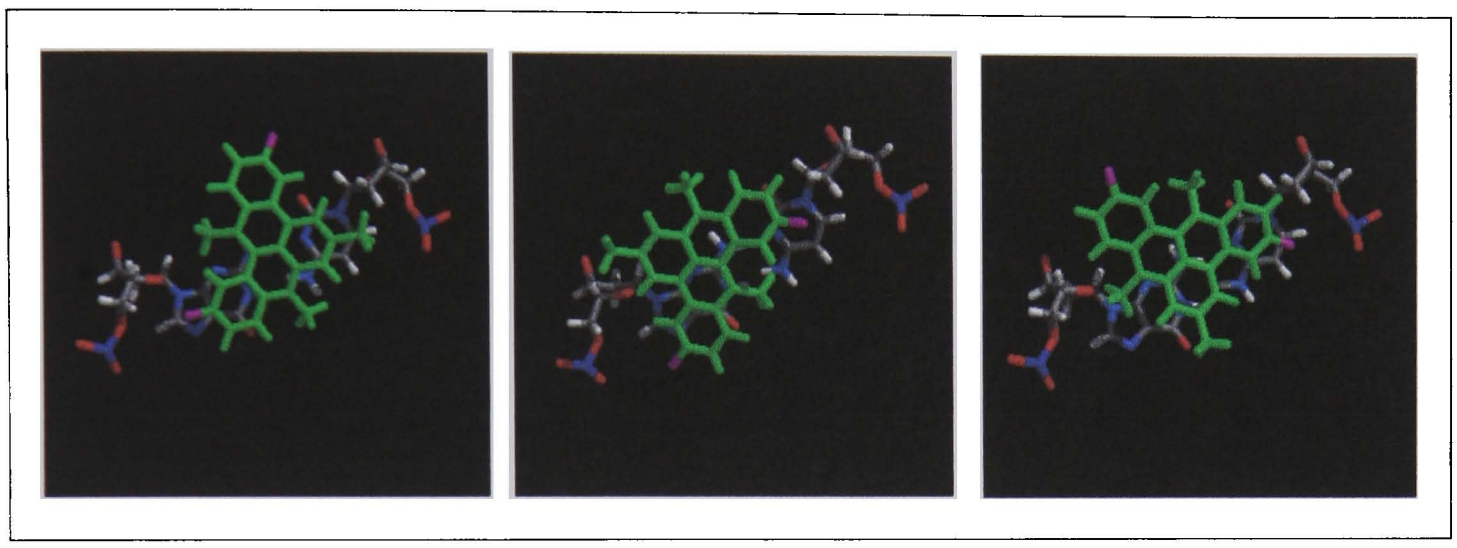


Figure 4.19 – The 3 different orientations of RHPS4 with duplex DNA (drug coloured green with fluorines coloured magenta).

All models were initially created with RHPS4 bound, for RHPS3 models the ligand was replaced prior to final minimisation.

4.6.2 Simulation protocol

All simulations were performed using the AMBER6 suite of programs and the associated Amber98 force field used to describe the DNA and solvent. The HF/6-31G*/RESP methodology was used to derive charges for RHPS4 and RHPS3 (Bayly *et al*, 1993). Starting structures were taken from the end of the final minimisations carried out upon building the models. 10ps warming runs were carried out followed by 500ps of equilibration before production runs of 3ns were performed with constant pressure (1atm) and temperature maintained at 300K by Berendsen coupling. SHAKE was used to constrain all bonds, allowing a 2fs time step for integration of Newton's equations. The electrostatics were calculated via the PME method and van de Waals interactions evaluated with a 9.0Å residue based cut-off.

Similar simulations were also carried out for the drug on its own and quadruplex and duplex DNA (without binding site) on their own for the purposes of calculating free energies. Overall 17 simulations were carried out in the study, on systems ranging from 1997 atoms (solvated RHPS4) up to 9057 atoms (solvated quadruplex with RHPS3 bound), representing a total equilibrated simulation time of over 50 ns.

4.6.3 Analysis methods

To find the "correct" binding orientations four different analysis methods were used. Energy analysis was done by using the implementation of the GB/SA method based on the MD trajectories obtained by using explicit solvation as described earlier (Section 3.3.2). Configurational entropies were calculated by the method of Schlitter as previously described (Chapter 2). Free energies were then derived via enthalpies and entropies from the three simulations (drug only, DNA only, drug-DNA) using the equation $\Delta G = \Delta H - T\Delta S$. The LIE method (described in Chapter 2) was also used to obtain values of free energy. The *anal* program (within the AMBER6 suite) was used to calculate electrostatic and van der Waals (vdW) interaction energies between the drug and the rest of the system for the drug only and drug-DNA simulations. These values were put into the LIE equation with the parameters used by Aqvist *et al* (1994). Stacking interaction energies were obtained, again by use of *anal.* Electrostatic and vdW interaction energies were calculated between the drug and the quartet/base pair above and below to give the energy associated with stacking. Finally the MIP method (described in Chapter 2) was used to give another form of interaction energy. A static structure of DNA was used with the drug acting as probe. Interaction energies were calculated at different points of a grid within the binding site.

4.7 Results and Discussion

4.7.1 Full evaluation of ΔG (RHPS4)

For the full evaluation of free energy for the systems, three simulations were required for each system. Simulations were carried out for the drug-DNA complexes, the DNA (both quadruplex and duplex) and the drug. From these simulations, solvated enthalpies and configurational entropies were obtained and changes in these values, and the free energy, upon drug binding calculated via the following equations (where $E = (H + G_{solv})$):

$$\Delta G_{bind} = G_{comp} - (G_{DNA} + G_{drug})$$
$$T\Delta S_{bind} = TS_{comp} - (TS_{DNA} + TS_{drug})$$
$$\Delta E_{bind} = E_{comp} - (E_{DNA} + E_{drug})$$

The results from these calculations for both quadruplex and duplex DNA are shown in Table 4.1. As can be clearly seen for both the quadruplex and duplex systems the changes in entropy upon binding are unreasonably spread, some being positive and some negative (the negative values of TΔSbind show that system gets more flexible upon drug binding, the positive values show stiffening of the DNA upon binding). This spread in entropy values leads to a wide spread in free energy values and so the solvated enthalpies are perhaps the most reliable data from this method, with position 2 being most favoured for quadruplex and position 2 also for duplex DNA.

(a) Quadruplex results (RHPS4)

Drug position	ΔE_{bind} (kcal/mol)	$T\Delta S_{bind}$ (kcal/mol)	ΔG_{bind} (kcal/mol)
1	-14.08	55.43	-69.51
2	-28.68	17.50	-46.18
3	-22.95	36.56	-58.83
4	-24.49	15.39	-39.88
5	-20.04	32.40	-52.44
6	-19.54	22.43	-41.97
7	-14.05	17.34	-31.39
8	-23.86	13.05	-36.91

(b) Duplex results (RHPS4)

Drug position	ΔE_{bind} (kcal/mol)	$T\Delta S_{bind}$ (kcal/mol)	ΔG_{bind} (kcal/mol)
1	-15.05	33.57	-49.07
2	-23.72	-13.02	-10.70
3	-23.02	-0.96	-22.06

Table 4.1 – Results from the full evaluation of ΔG_{bind}.

This spread in entropy results is completely unexpected as only the position of the drug is changed in each case and prior to each simulation an energy minimization has been carried out to ensure any unfavourable steric clashes are removed. To investigate further the unreasonable spread of the entropies we removed the drug from one of the quadruplex systems and one of the duplex systems and recalculated their entropies. If the differences in entropy are very large is likely to be due to a large amount of movement in the drug, causing the fluctuation in entropy. The calculated entropy differences in each case were approximately equal to the entropy of the drug alone; therefore the fluctuation in entropy must come from movement within the DNA itself.

4.7.2 *Evaluation of ΔG by Linear Interaction Energy (RHPS4)*

For the evaluation of free energy by LIE two simulations were required, the drug-DNA complex and the drug alone. Values for the electrostatic and vdW interactions of the drug with the rest of the system were calculated (for the drug-DNA complex, interaction between drug and DNA plus solvent and for the drug alone, the interaction between drug and solvent). The free energy was found by inputting these values into the following equation:

$$\Delta G_{bind} = \alpha \left(\langle E^{el} \rangle_{complex} - \langle E^{el} \rangle_{drug}\right) + \beta \left(\langle E^{vdW} \rangle_{complex} - \langle E^{vdW} \rangle_{drug}\right)$$

The parameters α and β were taken to be 0.5 and 0.161 respectively as used by Aqvist *et al* (1994), and the resulting values of ΔG_{bind} are shown in Table 4.2, the most favoured positions being 2 and 6 for quadruplex and position 2 for duplex DNA.

These parameters are quite system dependent (see section 2.2.2) but, as only the position of the drug was varied and not the drug itself, system specific parameters could not be obtained for this system.

(a) Quadruplex results (RHPS4)

Drug position	ΔG_{bind} (kcal/mol)
1	-13.17
2	-13.58
3	-12.55
4	-13.28
5	-10.44
6	-13.67
7	-12.12
8	-12.09

(b) Duplex results (RHPS4)

Drug position	ΔG_{bind} (kcal/mol)
1	-13.58
2	-15.34
3	-14.66

Table 4.2 – Results from the evaluation of ΔG_{bind} by LIE.

4.7.3 Stacking interactions (RHPS4)

Only one simulation, the drug-DNA complex system, was required for the calculation of stacking interaction energies. Electrostatic and vdW interaction energies were calculated (by Jose Ramon Blas, University of Barcelona) between the drug and the quartet below plus base pair above for quadruplex (Figure 4.20) and between the drug and base pairs above and below for duplex DNA. These interaction energies were then combined to give a stacking interaction energy, E_{stack}.

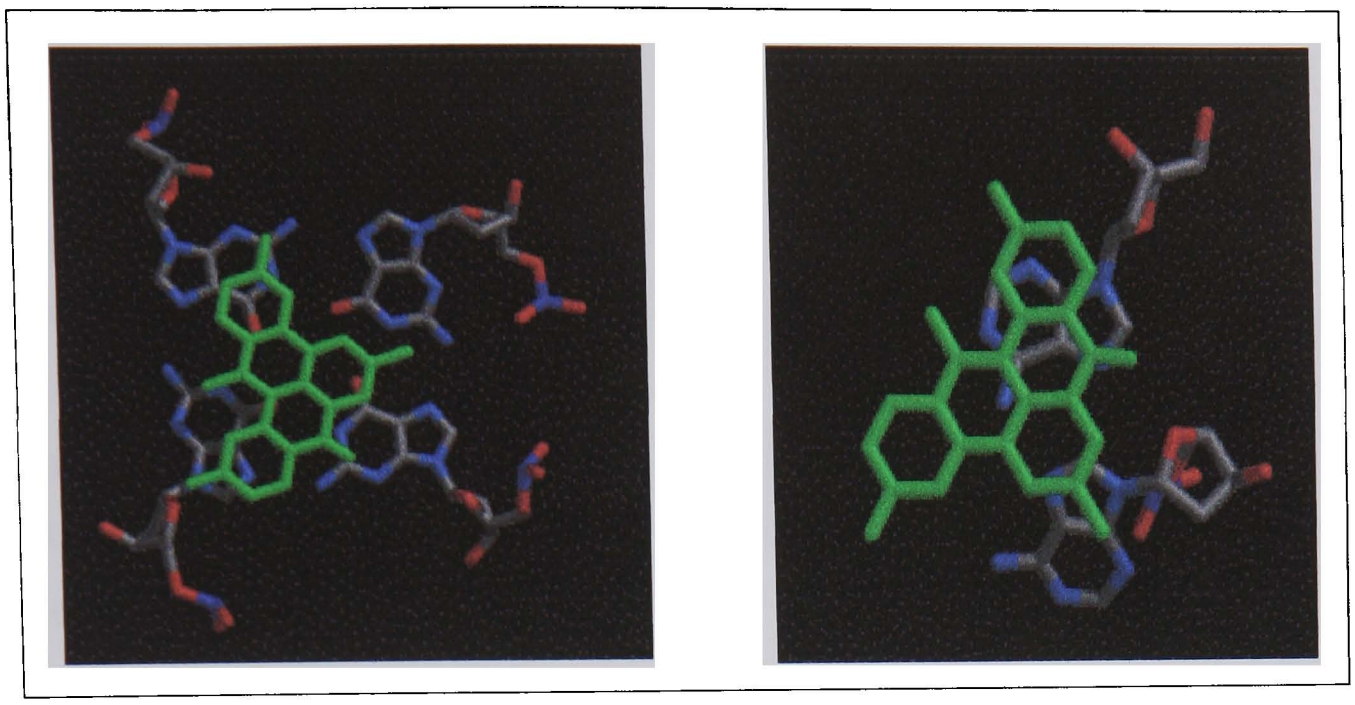


Figure 4.20 – Drug (coloured green) stacked with quartet below (left) and with base pair above (right) in quadruplex DNA.

The results from the stacking interaction can be seen in Table 4.3 which shows that for quadruplex DNA, position 6 is most favourable and for duplex DNA, position 2 is most favourable.

(a) Quadruplex resuls (RHPS4)

Position	E_{stack} (kcal/mol)
1	-216.46
2	-240.37
3	-235.11
4	-233.88
5	-211.26
6	-245.23
7	-203.54
8	-236.18

(b) Duplex results (RHPS4)

Position	E_{stack} (kcal/mol)
1	-189.30
2	-197.05
3	-192.08

Table 4.3 – Results of the stacking interaction.

4.7.4 Molecular Interaction Potential (RHPS4)

No simulations were required to carry out the MIP calculations, only a static structure. RHPS4 was used to probe the binding site, with interaction potentials calculated at the points on a 3D grid describing the binding site. We used various different grid spacings to try to get convergence in the potential energies. As can be seen in Figure 4.21, the limits of the technique (0.4Å grid spacing) are reached before the results converged. We therefore did not use this method further.

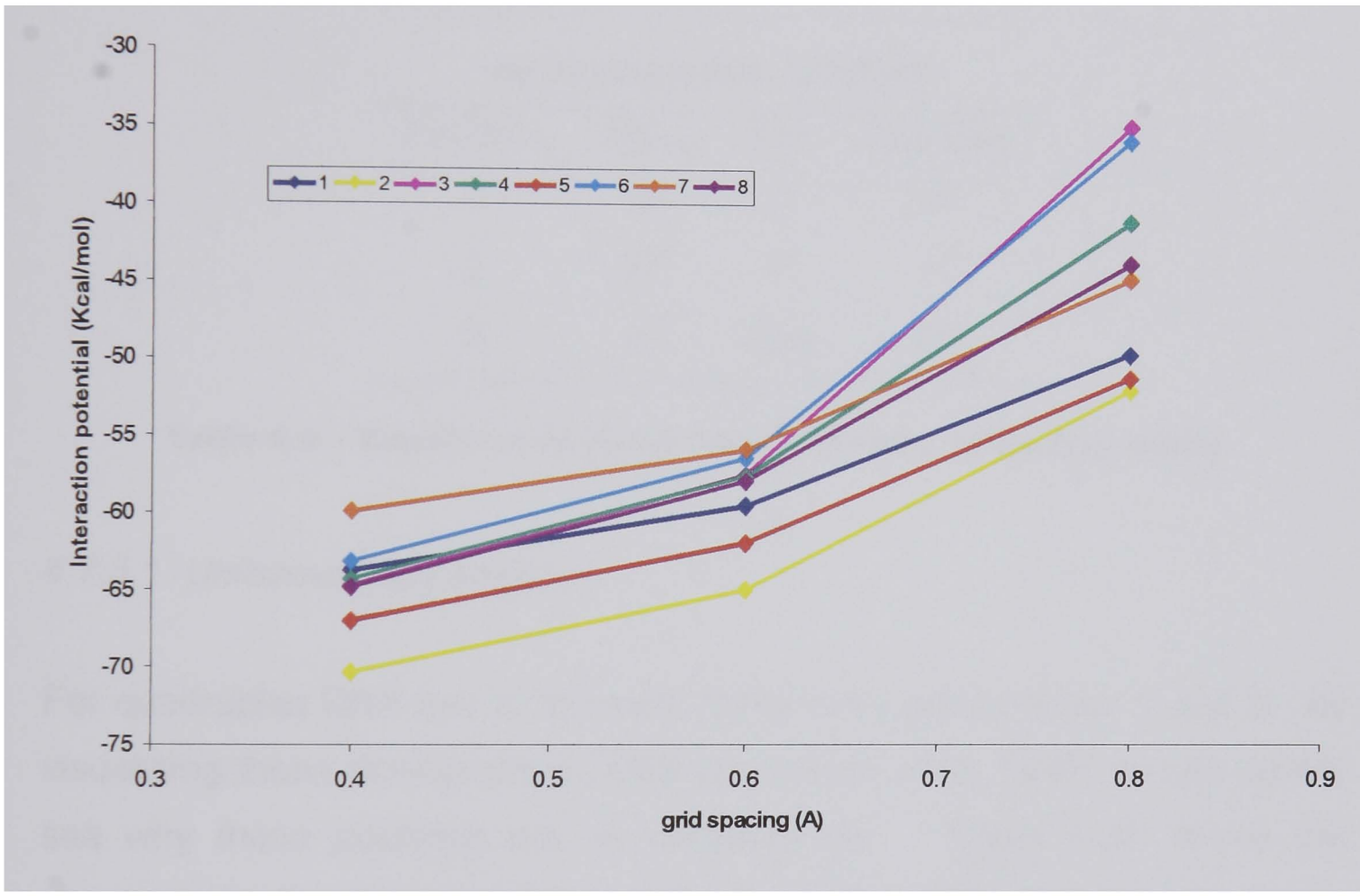


Figure 4.21 – Graph showing the results from MIP analysis.

4.7.5 Determination of "correct" orientations

To determine the "correct" positions we took the results from three of the four methods, all except the MIP method, and ordered them according to their binding abilities. Table 4.4 shows this ordering. We can see that for quadruplex DNA positions 6 and 2 are the most favourable and for duplex DNA position 2 is clearly most favoured (for views of positions see section 4.6.1, figures 4.18 and 4.19).

(a) Quadruplex results (RHPS4)

Position	ΔE_{bind}	LIE	Stacking
1	8th	4th	6th
2	**1st**	2nd	2nd
3	4th	5th	4th
4	2nd	3rd	5th
5	5th	8th	7th
6	6th	**1st**	**1st**
7	7th	6th	8th
8	3rd	7th	3rd

(b) Duplex results (RHPS4)

Position	ΔE_{bind}	LIE	Stacking
1	3rd	3rd	3rd
2	**1st**	**1st**	**1st**
3	2nd	2nd	2nd

Table 4.4 – Results of analysis ranked in order of binding ability.

4.7.5.1 Unfavourable positions

For quadruplex DNA two of the most unfavoured positions are 1 and 5. By visualising these simulations in VMD (Humphrey *et al*, 1996) we can clearly see why these positions are so unfavourable. Figure 4.22 shows the structures at the start and finish of the 3ns simulations and we can clearly see that for position 1 there is disruption of the terminal adenine and for position 5 the drug is being expelled from the binding site.

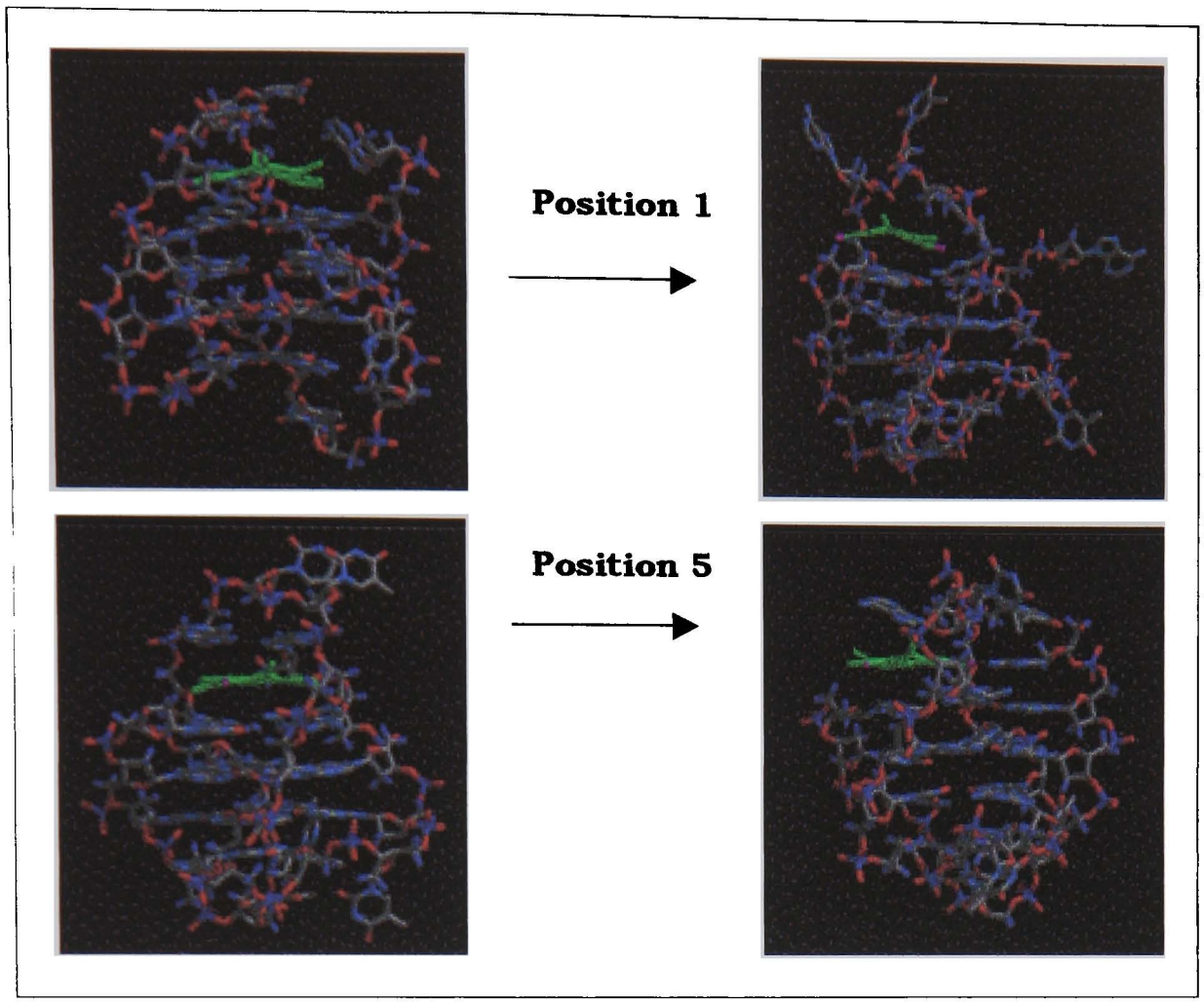


Figure 4.22 - Structures from the start (left) and finish (right) of simulations for the unfavoured positions 1 and 5.

4.7.5.2 Comparison to NMR

We have shown that binding positions 2 and 6 are energetically favoured most in quadruplex DNA. Comparison to the NMR structure of Gavathiotis *et al* (2001), shows that position 6 most closely resembles the binding mode seen in the NMR structure (similar position, opposite face of the quartet, as can be seen in Figure 4.23). We therefore used position 6 for all further analysis of the quadruplex binding mode.

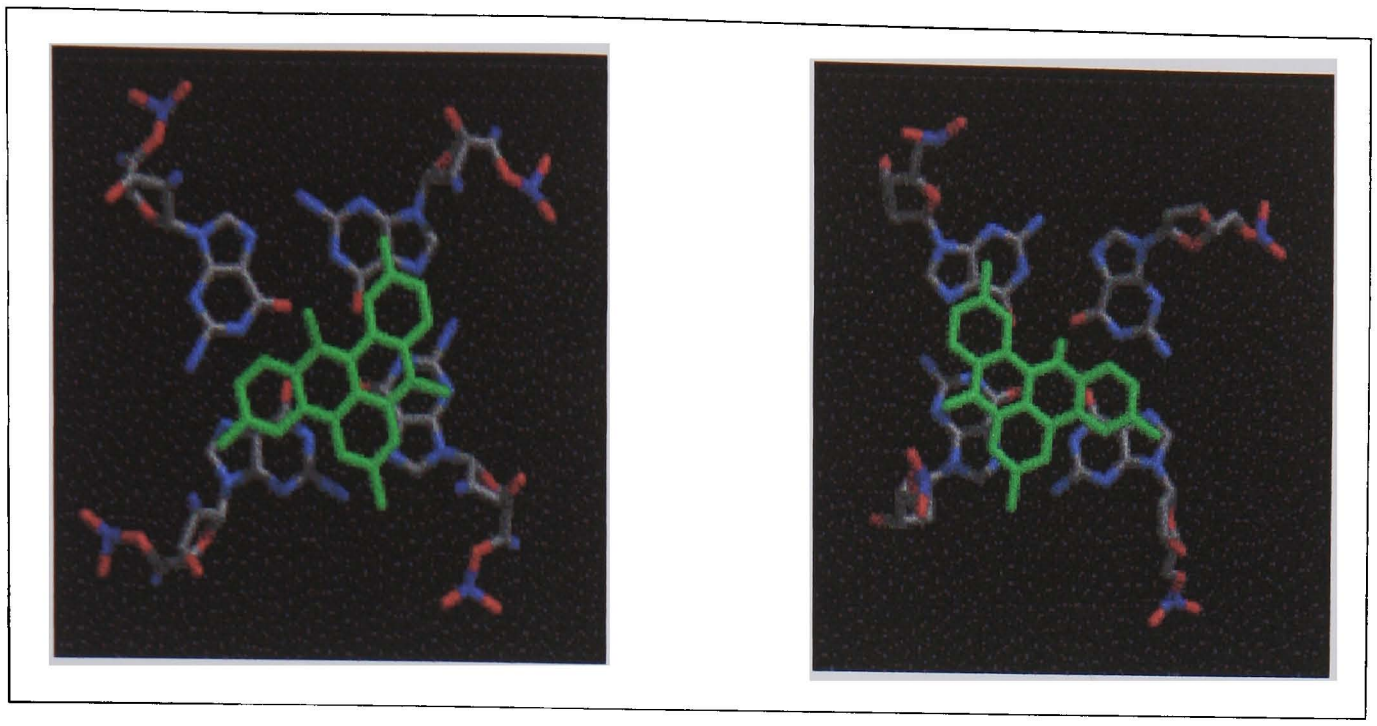


Figure 4.23 – NMR structure on left ($d(TTAGGGT)_4$ quadruplex) compared to position 6 structure on right (telomeric quadruplex).

4.7.6 Quadruplex v duplex

As we know from experimental data, RHPS4 binds selectively to quadruplex over duplex DNA, but do our modeling results show this? Below is table 4.5, which shows the results from the most favoured binding positions for quadruplex and duplex DNA.

DNA type	ΔE_{bind} (kcal/mol)	LIE (kcal/mol)	E_{stack} (kcal/mol)
Quadruplex (6)	-19.54	-13.67	-245.23
Duplex (2)	-23.72	-15.34	-197.05

Table 4.5 – Comparisons of quadruplex and duplex results for RHPS4.

We can see from these results that both the ΔE_{bind} and LIE predict that RHPS4 binds selectively to duplex over quadruplex DNA. Only the stacking interaction method predicts the correct trend in selectivity for quadruplex over duplex DNA, although this method would be expected to give the correct selectivity because the stacking interaction for quadruplex will always be larger than for duplex as there is more for the drug to interact with. There are six bases (four bases from the quartet and two from the base pair) for interaction with the drug in quadruplex, whereas only four in duplex DNA.

As only the stacking interactions get the selectivity right, which only look at simple interactions between drug and quartet/duplex and not at the energetics of the whole system, does this mean that our model is wrong? Could it be that we are trying to push the drug into a binding site that it does not want to form, suggesting that the energy needed to create the binding site within the quadruplex DNA (ΔG_{conf}) is too big?

4.7.7 RHPS3 results

Once we had the RHPS4 results we could use the most favoured positions (6 for quadruplex, 2 for duplex) to generate models to simulate the binding of RHPS3 in quadruplex and duplex DNA. The simulations were carried out as before and analysed using the same methods – full evaluation of ΔG, LIE and stacking interactions. The results are shown in Table 4.6, with the ΔE_{bind} values shown for the evaluation of ΔG, as the entropy values followed the same trend as for RHPS4.

DNA type	ΔE_{bind} (kcal/mol)	LIE (kcal/mol)	E_{stack} (kcal/mol)
Quadruplex	-28.31	-13.61	-224.76
Duplex	-16.90	-14.73	-187.32

Table 4.6 – Results from RHPS3 simulations.

The results show the correct trend in binding to quadruplex over duplex DNA for ΔE_{bind} and the stacking interaction method, but the wrong trend for LIE. Although as earlier mentioned the stacking interaction method should always predict the correct trend in selectivity, it is promising to note that on this occasion one of the other methods has predicted the correct trend also.

4.7.8 RHPS4 v RHPS3

Although the trend for binding to quadruplex over duplex DNA has not been predicted particularly well, we would also like to be able to predict the other aspect from the experimental studies, better binding by RHPS4 to quadruplex

and poorer binding to duplex DNA than RHPS3. Table 4.7 shows these comparisons.

(a) Quadruplex results

Drug	ΔE_{bind} (kcal/mol)	LIE (kcal/mol)	E_{stack} (kcal/mol)
RHPS4	-19.54	-13.67	-245.23
RHPS3	-28.31	-13.60	-224.76

(b) Duplex results

Drug	ΔE_{bind} (kcal/mol)	LIE (kcal/mol)	E_{stack} (kcal/mol)
RHPS4	-23.72	-15.34	-197.05
RHPS3	-16.90	-14.73	-187.32

Table 4.7 – Comparisons of results for RHPS4 and RHPS3 in both quadruplex and duplex DNA.

For the quadruplex systems all but the ΔE_{bind} results show that RHPS4 binds better than RHPS3, which is in agreement with experiment. The duplex results however all show the opposite trend to experiment, by predicting that RHPS4 binds better than RHPS3.

Although the $T\Delta S_{bind}$ values (not shown for RHPS3) were again found to be inconsistent and unreliable for use in calculating ΔG_{bind}, the differences between the quadruplex and duplex values can give a representation of selectivity. The $T\Delta\Delta S_{bind}$ values are both positive, 5.5 kcal/mol and 15.0 kcal/mol for RHPS4 and RHPS3 respectively, showing that both drugs are selective towards quadruplex DNA, although the results show that RHPS3 is more selective (the experimental results show RHPS4 is more selective to quadruplex DNA).

4.8 Summary

We have carried out analysis on RHPS4 and RHPS3 bound to quadruplex and duplex DNA with the aim of reproducing the experimental data and therefore being able to direct future syntheses of drugs within the series.

Whilst overall we are unable to reproduce all aspects of the experimental data, we can obtain some conclusions about the binding of these drugs.

Firstly we carried out analysis on the RHPS4 systems to find optimal structures with which to take further into our studies on RHPS3. We used the four methods: full evaluation of ΔG_{bind}; evaluation of ΔG_{bind} by LIE; stacking interactions and MIP. We reached the limits of the MIP method before the results converged and therefore discounted this method and did not use it further in the studies.

We found that the most favoured positions for RHPS4 were positions 2 and 6 for quadruplex and position 2 for duplex. We chose position 6 for quadruplex over position 2, as it resembled the NMR data more closely. Once we had these favoured positions, simulations and analysis on RHPS3 were carried out and comparisons made to the RHPS4 results. According to experimental data, RHPS4 should be a stronger binder to quadruplex DNA and a poorer binder to duplex DNA than RHPS3. Both RHPS4 and RHPS3 should show selectivity towards quadruplex DNA with RHPS4 having the greater selectivity.

The ΔE_{bind} results show the opposite trend to experiment with better binding by RHPS3 to quadruplex and better binding by RHPS4 to duplex DNA. The selectivity for quadruplex DNA is seen with RHPS3 but RHPS4 shows selectivity towards duplex DNA. The LIE results show that for quadruplex DNA, RHPS4 does bind slightly better than RHPS3 but it also binds better to duplex DNA. As for selectivity, both show the wrong trend in that they are both selective to duplex DNA. The stacking interaction method performs the best of the three methods. For quadruplex DNA RHPS4 is a better binder although results for duplex show the wrong trend. Both RHPS4 and RHPS3 show selectivity for quadruplex DNA, though as discussed earlier this is to be expected. Selectivity for quadruplex DNA can be seen from $T\Delta\Delta S_{bind}$ results for both RHPS4 and RHPS3, although the greater selectivity is for RHPS3.

As was discussed earlier, since this study began a new X-ray crystal structure has been determined for telomeric DNA (Parkinson *et al*, 2002) which differs significantly in topology to the NMR structure we have been using. The G-quartet core has its strands in a parallel orientation, while the TTA loops run diagonally between them, as opposed to the NMR structure which is anti-parallel with TTA loops at the top and bottom (see Figure 4.7, earlier in chapter).

While debate continues as to the relevance of this new structure, it does have a number of attractive features and could explain some of our results. If the model we are using is not correct we cannot be expected to obtain the correct results. One attractive feature of the X-ray structure is that the drug is expected to sit between two quadruplexes and not in the loops as we have modelled (Figure 4.24). If this were the case then we can predict that our results would be quite different. We predict that if the X-ray structure were used there would be a stronger interaction between the drug and the DNA, as it would have a quartet on both sides, which would prove beneficial to LIE analysis, possibly giving us the correct trends in binding. Also, the energy required to form the binding site (ΔG_{conf}), as discussed earlier (in section 4.7.6), may be smaller resulting in quadruplex verses duplex binding trends closer to those from experiment for both LIE and the full evaluation of ΔG_{bind}.

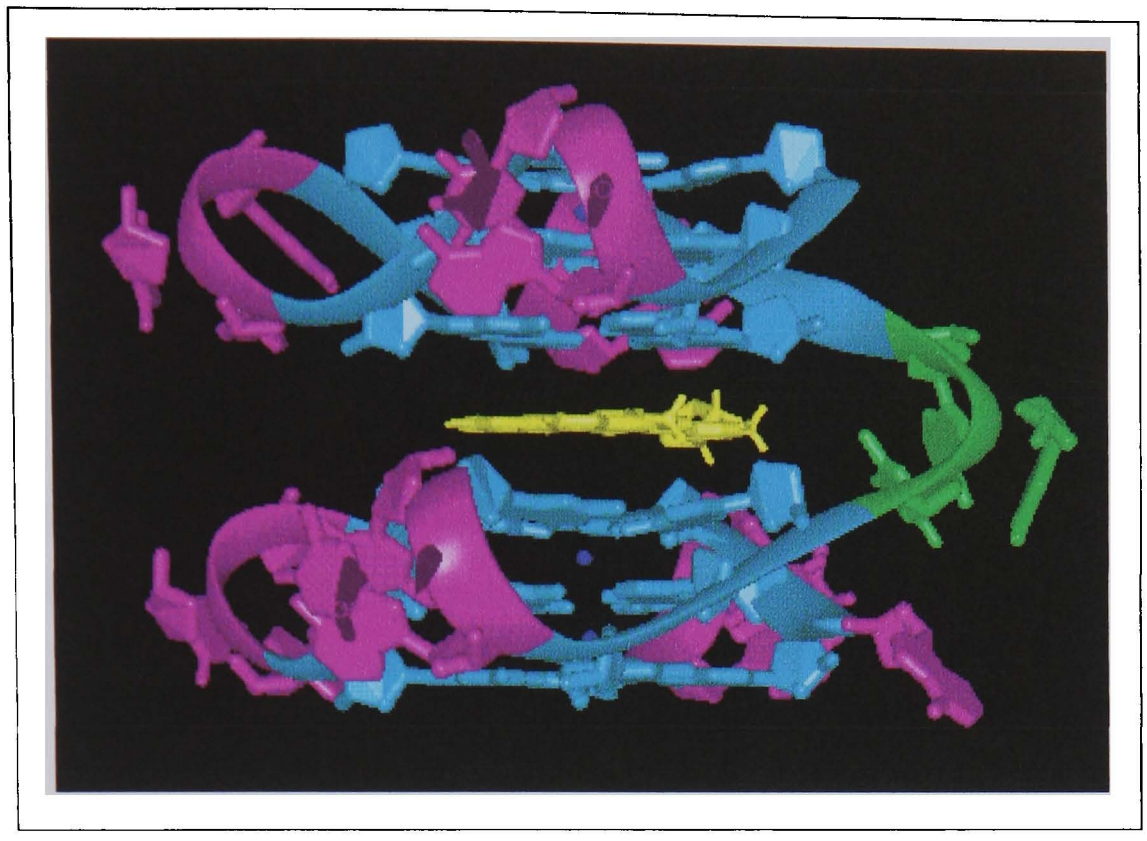


Figure 4.24 – Structure showing where a drug is expected to sit between two X-ray crystal structures (Figure courtesy of Charles Laughton).

Overall we can conclude that predicting the selectivity and binding trends for quadruplex stabilising drugs is obviously challenging due to a variety of factors, including electrostatic and vdW interactions, DNA flexibility and the energy required to form a binding site within the models used.

4.9 References

Anantha N.V, Azam M, Sheardy R.D, (1998), *Biochemistry,* **37**, 2709-2714.

Aqvist J, Medina C, Samuelsson J-E, (1994), *Protein Eng,* **7** (3), 385-391.

Bayly C.I, Cieplak P, Cornell W.D, Kollman P.A, (1993), *J. Phys. Chem,* **97**, 10269-10280.

Blackburn E.H, (1991), *Nature,* **350**, 569-573.

Blackburn E.H, (2000), *Nature,* **408**, 53-55.

Brummendorf T.H, Holyoake T.L, Rufer N, Barnett M.J, Schulzer M, Eaves C.J, Eaves A.C, Lansdrop P.M, (2000), *Blood,* **95**, 1883-1890.

Cairns D, Michalitsi E, Jenkins T.C, Mackay S.P, (2002), *Bioorg. Med. Chem,* **10**, 803-807.

Case D.A, Pearlman D.A, Caldwell J.W, Cheatham T.E. III, Ross W.S, Simmerling C.L, Darden T.L, Merz K.M, Stanton R.V, Cheng A.L, Vincent J.J, Crowley M, Tsui V, Radmaer R.J, Duan Y, Pitera J, Massova I, Seibel G.L, Singh U.C, Weiner P.K, Kollman P.A, (1999), *AMBER 6*, University of California, San Francisco.

Chaires J.B, (1998), *Biopolymers,* **44**, 201-215.

Chaires J.B, (2001), *Methods Enzymol,* **340**, 3-22.

Corey D.R, (2000), *Chem. Res. Toxicol,* **13**, 957-960.

Davis T.M, Wilson W.D, (2001), *Methods Enzymol,* **340**, 22-51.

Fedoroff O.Y, Salazar M, Han H, Chemeris V.V, Kerwin S.M, Hurley L.H, (1998), *Biochemistry,* **37**, 12367-12374.

Feng J, Funk W.D, Wang S-S, Weinrich S.L, Avilion A.A, Chiu C-P, Adams R.R, Chang E, Allsopp R.C, Yu J, Le S, West M.D, Harley C.B, Andrews W.H, Greider C.W, Villeponteau B, (1995), *Science,* **269**, 1236-1241.

Ferrin T.E, Huang L.E, Jarvis L.E, Langridge R, (1988), *J. Mol. Graphics,* **6**, 13-27.

Gavathiotis E, Heald R.A, Stevens M.F.G, Searle M.S, (2001), *Angew. Chem. Int. Ed,* **40** (24), 4749-4751.

Gowan S.M, Heald R.A, Stevens M.F.G, Kelland L.R, (2001), *Mol. Pharmacol,* **60** (5), 981-988.

Guo Q, Lu M, Marky L.A, Kallenbach N.R, (1992), *Biochemistry,* **31**, 2451-2455.

Hamilton S.E, Simmons C.G, Kathiriya I.S, Corey D.R, (1999), *Chem. Biol,* **6**, 343-351.

Han H, Cliff C.L, Hurley L.H, (1999), *Biochemistry,* **38**, 6981-6986.

Han H, Langley D.R, Rangan A, Hurley L.H, (2001), *J. Am. Chem. Soc,* **123**, 8902-8913.

Haq I, Ladbury J.E, Chowdhry B.Z, Jenkins T.C, Chaires J.B, (1997), *J. Mol. Biol,* **271**, 244-257.

Haq I, Trent J.O, Chowdhry B.Z, Jenkins T.C, (1999), *J. Am. Chem. Soc,* **121**, 1768-1779.

Haq I, Ladbury J, (2000), *J. Mol. Recog,* **13**, 188-197.

Haq I, (2002), *Archives Biochem. Biophys,* **403**, 1-15.

Hardin C.C, Watson T, Corregan M, Bailey C, (1992), *Biochemistry,* **31**, 833-841.

Harley C.B, Futcher A.B, Greider C.W, (1990), *Nature,* **345**, 458-460.

Harrison R.J, Gowan S.M, Kelland L.R, Neidle S, (1999), *Bioorg. Med. Chem. Lett,* **9**, 2463-2468.

Hayflick L, (1961), *Exp. Cell Res,* **25**, 585-621.

(a) Heald R.A, Modi C, Cookson J.C, Hutchinson I, Laughton C.A, Gowan S.M, Kelland L.R, Stevens M.F.G, (2002), *J. Med. Chem,* **45** (3), 590-597.

(b) Heald R.A, (2002), *The design and synthesis of 8,13-dimethyl-8H-quino[4,3,2-k]acridinium salts: Potent telomerase inhibitors and potential anticancer drugs.* Thesis, University of Nottingham.

Herbert B.S, Pitts A.E, Baker S.I, Hamilton S.E, Wright W.E, Shay J.W, Corey D.R, (1999), *Proc. Natl. Acad. Sci. USA,* **96**, 14276-14281.

Humphrey W, Dalke A, Schulten K, (1996), *J. Molec. Graphics,* **4.1**, 33-38.

(a) Izbicka E, Nishioka D, Marcell V, Raymond E, Davidson K, Lawrence R.A, Wheelhouse R.T, Hurley L.H, Wu R.S, Von Hoff D.D, (1999), *Anti-Cancer Drug Design,* **14**, 355-365.

(b) Izbicka E, Wheelhouse R.T, Raymond R, Davidson K.K, Lawrence R.A, Sun D, Windle B.E, Hurley L.H, Von Hoff D.D, (1999), *Cancer Research,* **59**, 639-644.

Kelland L.R, (2000), *Anti-Cancer Drugs,* **11**, 503-513.

Kerwin, S.M, (2000), *Curr. Pharm. Design,* **6**, 441-471.

Kettani A, Kumar A.R, Patel D.J, (1995), *J. Mol. Biol,* **254**, 638-656.

Kim M-Y, Gleason-Guzman M, Izbicka E, Nishioka D, Hurley L.H, (2003), *Cancer Research,* **63**, 3247-3256.

Kim N.W, Piatysek M.A, Prowse K.R, Harley C.B, West M.D, Ho P.L.C, Coriello G.M, Wright W.E, Weinrich S.L, Shay J.W, (1994), *Science,* **266**, 2011-2015.

Leong C.O, Seow H.F, (2001), *Clin. Biochemist. Rev,* **23**, 39-48.

Macaya R.F, Schultze P, Smith F.W, Roe J.A, Feigon J, (1993), *Proc. Natl. Acad. Sci. USA*, **90**, 3745-3749.

Mergny J-L, Mailliet P, Lavelle F, Riou J-F, Laoui A, Helene C, (1999), *Anti-Cancer Drug Design,* **14**, 327-339.

Modi C, (2002), *Structure selective DNA recognition by a novel class of polycyclic acridine derivatives.* Thesis, University of Nottingham.

Neidle S, Kelland L.R, (1999), *Anti-Cancer Drug Design,* **14**, 341-347.

Neidle S, Harrison R.J, Reszka A.P, Read M.A, (2000), *Pharmacol. Therap,* **85**, 133-139.

Neidle S, Parkinson G.N, (2003), *Curr. Opin. Struct. Biol,* **13**, 275-283.

Parkinson G.N, Lee M.P.H, Neidle S, (2002), *Nature,* **417**, 876-880.

(a) Perry P.J, Gowan S.M, Reszka A.P, Polucci P, Jenkins T.C, Kelland L.R, Neidle S, (1998), *J. Med. Chem,* **41**, 3253-3260.

(b) Perry P.J, Reszka A.P, Wood A.A, Read M.A, Gowan S.M, Dosanjh H.S, Trent J.O, Jenkins T.C, Kelland L.R, Neidle S, (1998), *J. Med. Chem,* **41**, 4873-4884.

Perry P.J, Jenkins T.C, (1999), *Exp. Opin. Invest. Drugs,* **8** (12), 1981-2008.

Rangan A, Federoff O.Y, Hurley L.H, (2001), *J. Biol. Chem,* **276**, 4640-4646.

Read M.A, Wood A.A, Harrison R.J, Gowan S.M, Kelland L.R, Dosanjh H.S, Neidle S, (1999), *J. Med. Chem,* **42**, 4538-4546.

Read M.A, Neidle S, (2000), *Biochemistry,* **39**, 13422-13432.

Read M.A, Harrison R.J, Romagnoli B, Tanious F.A, Gowan S.H, Reszka A.P, Wilson W.D, Kelland L.R, Neidle S, (2001), *Proc. Natl. Acad. Sci. USA,* **98** (9), 4844-4849.

Ross P.D, Subramanian S, (1981), *Biochemistry,* **20**, 3096-3102.

Shay J.W, Wright W.E, (1999), *Science,* **286**, 2284-2285.

Shin-ya K, Wierzba K, Matsuo K, Ohtani T, Yamanda Y, Furihata K, Hayakawa Y, Seto H, (2001), *J. Am. Chem. Soc,* **123**, 1262-126

Simonsson T, (2001), *Biol. Chem,* **382**, 621-628.

Spackova N, Berger I, Sponer J, (1999), *J. Am. Chem. Soc,* **121**, 5519-5534.

Sun D, Thompson B, Cathers B.E, Salazar M, Kerwin S.M, Trent J.O, Jenkins T.C, Neidle S, Hurley L.H, (1997), *J. Med. Chem,* **40** (14), 2113-2116.

Wang Y, Patel D.J, (1993), *Structure,* **1**, 263-282.

Watson J.D, (1972), *Nature New Biol,* **239**, 197-201.

Website 1 – www.stereographics.com - accessed September 2003.

Wheelhouse R.T, Sun D, Han H, Han F.X, Hurley L.H, (1998), *J. Am. Chem. Soc,* **120**, 3261-3262.

White L.K, Wright W.E, Shay J.W, (2001), *Trends in Biotech,* **19** (3), 114-120.

Wright W.E, & Shay J.W, (1992), *Exp. Gerontol,* **27**, 383-389.

Williamson J.R, (1994), *Annu. Rev. Biophys. Biomol. Struct,* **23**, 703-730.

Zahler A.M, Williamson J.R, Cech T.R, Prescott D.M, (1991), *Nature,* **350**, 718-720.

CHAPTER 5 - CONCLUSIONS AND FUTURE WORK

This thesis has involved two studies of drug-DNA recognition with the aim of discovering if the computational methods involved could reproduce experimental data. Only if this is possible can these methods be used to predict further information in the areas concerned. The two case studies show different perspectives of reproducing experimental data. Firstly, the LAMMPS case study was carried out to validate whether a new parallel molecular dynamics code could reproduce co-operative binding of the minor groove binder Hoechst 33258, as had already been seen by the established AMBER code. Secondly, the telomerase case study was carried out to see if it was possible to use modelling techniques to reproduce experimental data for polycyclic acridines binding to telomeric DNA. In both cases, if the experimental data could not be satisfactorily reproduced then there would be no point taking the projects further.

5.1 LAMMPS case study

It has been shown that by implementing the AMBER98 forcefield into the LAMMPS code we can produce stable nanosecond molecular dynamics trajectories of DNA, whose analysis is in general agreement with that of trajectories produced using the AMBER code and forcefield. Although the analysis is in agreement, the results from LAMMPS can never be expected to be exactly the same as that of AMBER due to differences in calculations within the codes. There are also slight differences in the way the trajectories were run which also means that the results will never be identical.

The AMBER trajectories were run on one processor and therefore the use of SHAKE to constrain bond lengths could be used on all bonds. All the LAMMPS trajectories were run on multiple processors which does not allow the use of SHAKE on all bonds, only on bonds to hydrogen. This therefore means that although the LAMMPS trajectories run using SHAKE are a closer

approximation to the AMBER trajectories than those using RESPA or no speed up algorithms, they still cannot be classed as identical.

Further work would need to be carried out to more rigorously test out this validation. Runs would need to be carried out which use, as closely as possible, the same parameters. As AMBER does not support the use of RESPA, direct comparisons cannot be made between AMBER and LAMMPS using RESPA. AMBER runs needs to be carried out in parallel, with no SHAKE and with SHAKE on bonds to hydrogen only, to compare directly with the runs carried out with LAMMPS.

As for scaling issues, it is clear that on higher numbers of processors LAMMPS outperforms AMBER whatever parameters are used.

5.2 Telomerase case study

One of the original aims of this study was to see if we could direct the synthesis of new and more potent drugs within the series, as this would be beneficial to the chemists involved in the project. The main stumbling block to this case study is that we do not have an NMR structure of a drug binding to telomeric DNA and because of this we have no real proof of the correct binding site nor the correct quadruplex structure we should be using. When this project began the only known structure of telomeric DNA was the Patel NMR structure and it was postulated that the most favoured binding site was in the top loop region of this structure, but confusingly when the project was well underway, an entirely different crystal structure was proposed.

Molecular dynamics trajectories were run and analysis carried out on the lead compound (RHPS4) in the series of polycyclic acridine compounds synthesised and analysed within our labs. This analysis was to try to find the most energetically favoured position for this drug to sit within the top loop of telomeric quadruplex DNA. Subsequently, when this position was elucidated,

analysis was carried out on trajectories of another compound from the series, RHPS3.

The analysis consisted of four different ways of calculating the energetics of drug binding. Although none of the methods were able to consistently reproduce the selectivity nor the binding trends seen in the experimental data we were able to conclude that the binding is a result of many factors including electrostatic and vdW interactions, DNA flexibility and the energy required to form a binding site.

Whilst these factors are obviously challenging enough, there is also great debate about which one of the two models presented for human telomeric DNA is the correct one. Until the "correct" structure is elucidated it remains a hard task for the modeller to study the binding of drugs to this sequence, as both models have to be taken into account. The structure of the second model, the Neidle X-ray structure, was published when this study was well underway and therefore there was no time to study the binding of RHPS3 and RHPS4 to this structure. This further work is now underway to see if the results of analysis of our drugs binding to the Neidle structure are in better agreement with experimental studies. It is suggested in the summary of Chapter 4 that our analysis is likely to favour the Neidle structure due to the likelihood of a smaller energy penalty when creating a binding site.

Unfortunately we have been unable to help the chemists in directing their research because of our inaccurate results and the amount of time taken to carry out the simulations required for a full and complete analysis. If there had been more time available, then the method of Free Energy Perturbation (FEP) could have been used to look at the binding abilities of several drugs. Using this method the practitioner can rank a series of closely related compounds in order of their binding ability. A calibration would have been required using several known compounds to check we could rank them theoretically in the same order as experimentally and if successful we could

have then gone on to test out new compounds before they were made in the lab.

Even though work is now being carried out on the binding of our drugs to the Neidle crystal structure, we still don't know if either the NMR or crystal structures are correct. The method of radio probing has been suggested as a way of determining the correct quadruplex structure and which site intercalating drugs are most likely to bind to. This method involves making a radioactive version of the drug, binding this drug to telomeric DNA and then using the radioactivity of the drug to determine its binding site and from that data hopefully the structure of telomeric DNA. Funding has recently been secured for this work and with a bit of luck we may have a conclusive structure in the near future.