Analysis of intrinsic DNA curvature in the TP53 tumour suppressor gene using atomic force microscopy.

Bayliss, Sion

Swansea University
Prifysgol Abertawe

ANALYSIS OF INTRINSIC DNA CURVATURE IN THE *TP53* TUMOUR SUPPRESSOR GENE USING ATOMIC FORCE MICROSCOPY

Sion Bayliss *B.Sc.* (Hons)

Submitted to the School of Medicine, Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy

September, 2012

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

The research described in this thesis aimed to evaluate the intrinsic DNA curvature of the region of the *TP53* tumour suppressor gene that codes for the sequence-specific DNA-binding domain of the p53 protein, a key protein that protects the cell from chemical insults and tumourogenesis. There have been no previous attempts to experimentally investigate the intrinsic DNA curvature within *TP53* or its relation to the functional or structural properties of the gene, such as DNA repair and nucleosomal architecture. The present study used theoretical models of *TP53* in concert with an atomic force microscopy based experimental investigation of *TP53* DNA molecules to analyse intrinsic DNA curvature within the gene. This was achieved by developing a novel software platform for the atomic force microscopy based investigation of DNA curvature, named ADIPAS. Dinucleotide wedge models of DNA curvature were used to model *TP53* in order to investigate the relationship between intrinsic DNA curvature and the structure and function of the gene. ADIPAS was applied to atomic force microscopy images of *TP53* DNA molecules immobilised on a mica surface in order to experimentally measure intrinsic DNA curvature. The experimental findings were compared to theoretical models of intrinsic curvature in *TP53*. The resulting intrinsic curvature profiles showed that exons exhibited significantly lower intrinsic DNA curvature than introns within *TP53*, this was also shown to be true for regions of slow DNA repair. This indicated that DNA curvature may play a role in *TP53* as a controlling factor for nucleosomal architecture to facilitate open chromatin and active DNA transcription. The evolutionary selection for intrinsic curvature may have played a role in the development of exons with low intrinsic DNA curvature. Low intrinsic curvature in exon position has also been implicated in the reduced efficiency of DNA repair in a number of cancer specific mutation hotspots.

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date .....23/02/2013.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .... (candidate)

Date .....23/02/2013.....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .. .. (candidate)

Date .....23/02/2013.....

Contents

Acknowledgements

I would like to thank my family for the love and support that they have shown me in all of my endeavours. My special thanks go to my parents for the understanding that they have shown during my postgraduate studies. I would like to thank my brother and grandparents for their support, guidance and chocolate.

I would like to thank my supervisor Dr. Paul Lewis and members of my group, past and present. Their influence led me to develop bioinformatics and computing skills that will undoubtedly benefit me later in my career. I would not be here had it not been for their contributions to my research.

A special mention goes to all the people who have worked with me in the ILS. Many great times have been had, innumerable cups of tea imbibed and numerous cake days enjoyed. Thanks to Georgina Menzies, Richard Charlton, Aaran Lewis, Kit Lucas, Nat DeMello, Paul Lewis, Adam Thomas, Lizzy McAdam, Owen Bodger, George Johnson, Oliver Lyttleton, Yasmin Freidmann, and Ricardo Del Sol. Special thanks go to Stephanie Hinder who always kept my mug of tea topped up. I would also like to thank Dr Josie Parker who provided me with practical guidance and always treated me with patience. Thanks also go to Owen Bodger who provided me with statistical guidance and provided distracting conversion.

List of Figures and Illustrations

The following table lists the figures and illustrations used throughout the thesis. The page on which each one is defined or first used is also given.

List of Tables

The following table lists the figures and illustrations used throughout the thesis. The page on which each one is defined or first used is also given.

List of Abbreviations and Symbols

The following table describes the significance of various abbreviations and acronyms used throughout the thesis. The page on which each one is defined or first used is also given.

Abbreviation	Meaning	Page
μm	micrometres	66
μM	micromoles	64
3'	three prime	15
5'	five prime	15
A	adenine	15
Å	angstroms	128
ADIPAS	AFM DNA image processing and analysis software	80
AFM	atomic force microscopy	27
bp	base pair	23
BPDE	benzo pyrene diol epoxide	15
C	cytosine	15
^{0}	degrees	20
DNA	deoxyribonucleic acid	15
EM	electron microscopy	27
FF	fragment flipping	41
G	guanine	15
GUI	general user interface	78
IARC	international agency for research on cancer	57
IQR	interquartile range	174
kb	kilobases	252
MD	molecular dynamics	23
Mg^{2+}	magnesium cation	33
$MgCl_2$	magnesium chloride	59
N/m	newton meters per second	66
NER	nucleotide excision repair	23
Ni^{2+}	nickel cation	33
nm	nanometres	19
NMR	nuclear magnetic resonance	26
NXS	nucleosome exclusion site	154
p53	tumour protein 53	54
PCR	polymerase chain reaction	59
rads	radians	20
RMSE	root mean square error	93
RNA	ribonucleic acid	15
ROC	radius of curvature	66
T	thymine	15
TP53	tumour protein 53 gene	54
UV	ultraviolet	57
WLC	worm-like chain	48
θ	angle	45
ξ	persistence length	25

CHAPTER 1: GENERAL INTRODUCTION

1.1 The Structure of Deoxyribonucleic Acid

Deoxyribonucleic acid (DNA) is a duplex of two polymers each individually constructed of nucleotide subunits. The duplex is connected by complementary base pairing between individual nucleotides. Nucleotides are further subdivided into chemical residues. The structure of DNA has been detailed below (Section 1.1.1.-1.1.4.) on a range of scales: the primary structure of DNA (nucleotides), the interactions that create the dipolymeric chains typically found *in vivo* (complementary base pairing) and the secondary structure of DNA (the double helix).

1.1.1 Primary Structure – Mononucleotides

Each nucleotide subunit consists of a phosphate residue, a sugar moiety and one of four nucleobases, often referred to as bases (Figure 1.1.).

The nucleobases are heterocyclic aromatic organic nitrogen-containing compounds. There are four nucleobases commonly found in DNA: adenine (A), guanine (G), thymine (T) and cytosine (C). There are two fundamental types of nitrogenous bases found in DNA: the purine bases, adenine and guanine, and the pyrimidine bases, thymine and cytosine. Nucleobases provide the molecular structure required for hydrogen bonding and complementary base pairing that gives rise to the dipolymeric structure of DNA found *in vivo*, discussed later.

The sugar residues in DNA are universally the pentose sugar monosaccharide, 2-deoxyribose, with the formula H-(C=O)-(CH_2)-($CHOH)_3$-H. This distinguishes DNA from ribonucleic acid (RNA), another important biological nucleic acid, in that RNA contains ribose rather than 2-deoxyribose. Nucleobases are connected to the sugar residue via N-glycosidic linkages that involve base ring nitrogens, N-9 for purines or N-1 for pyrimidines connected to the C-1 of the pentose sugar. The sugar and base together are called a nucleoside.

Each nucleoside is connected to the next via a phosphodiester bond between a phosphate residue at the third and fifth carbon atoms of adjacent nucleosides. Nucleosides with phosphate residue bound at the 5' terminus of the sugar ring are referred to as nucleotides. The asymmetric phosphodiester bonds give the DNA its directionality. Repeating sugar and phosphate groups form the 'sugar-phosphate backbone' of the DNA molecule. The labels 5-prime (5') and 3-prime (3') are assigned to the ends of the DNA polymer that terminate with a phosphate group and hydroxyl group respectively.

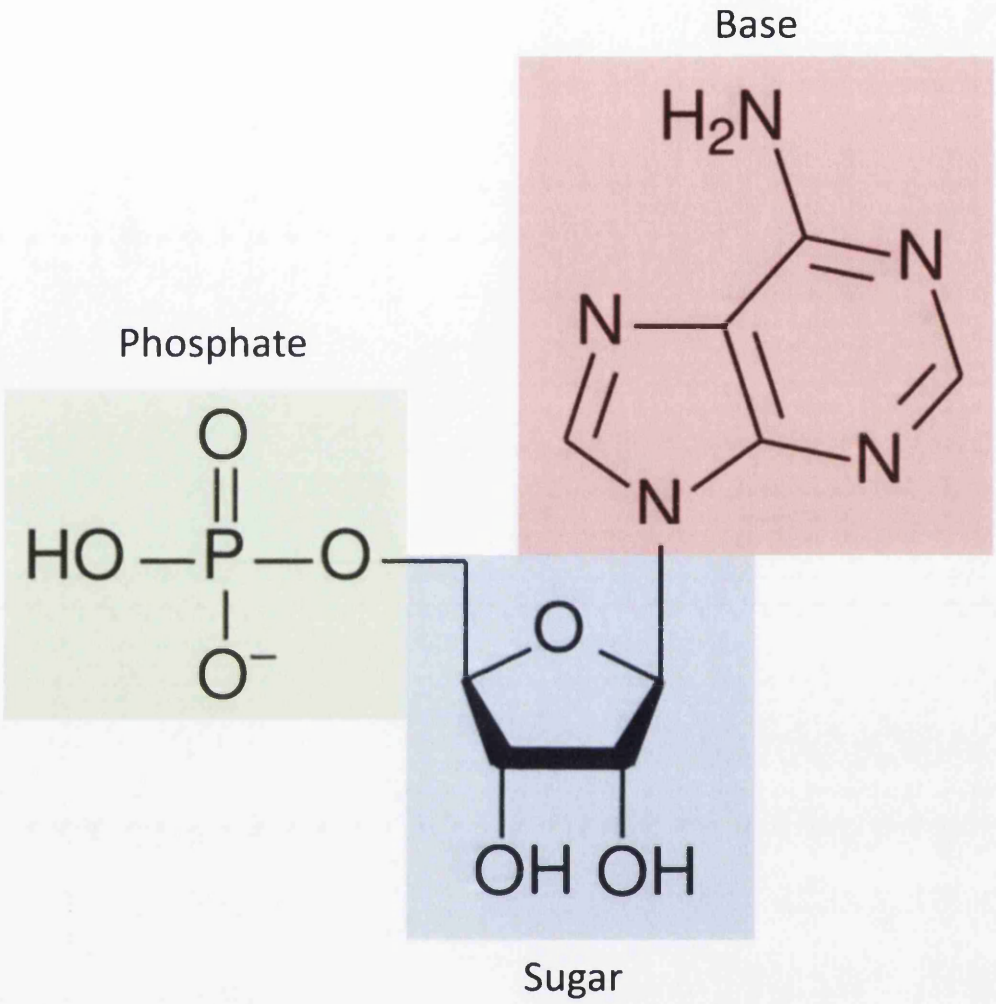

Figure 1.1. – An example of the adenine mononucleotide. The different chemical residues are indicated: phosphate (green), sugar (blue) and base (red). All mononucleotides follow this basic structure but will have a different base.

1.1.2 Complementary Base Pairing

DNA *in vivo* does not exist as a single chain. Rather, it is found as two polymer chains bound to one another by hydrogen bonds between nucleobases. Two bound nucleotides are referred to as a base pair (bp). The DNA complementary base pairing principle operates because specific geometrical requirements exist in the formation of hydrogen bonds between the heterocyclic amines. This leads to optimal geometries between complementary purines and pyrimidines. In canonical base pairing, guanine forms a base pair with cytosine while adenine forms a base pair with thymine (Figure 1.2.). Adenine and thymine form complementary base pairs via two hydrogen bonds between their respective bases. Cytosine and guanine form complementary base pairs via three hydrogen bonds between their respective bases. Therefore, the secondary strand, often called the complementary strand, has an opposite and complementary nucleotide sequence to the primary strand *e.g.* the complementary base pair sequence for ACTG would be TGAC. Conventionally, the primary strand is written in a 5' to 3' direction and the complementary strand in the 3' to 5' direction.

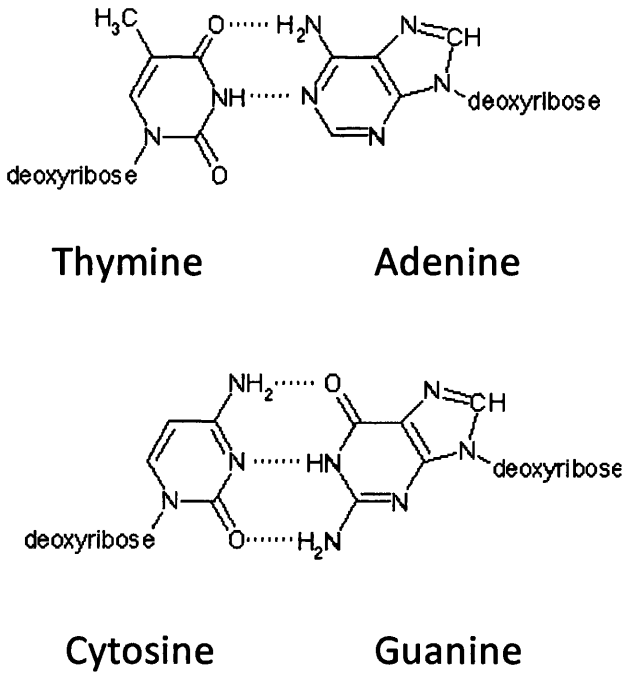

Figure 1.2. – Complementary base pairing schematic. Hydrogen bonds between base pairs are indicated by a broken dotted line. Only the bases involved in forming complementary hydrogen bonds are shown, the point where the base pair joins to the deoxyribose is indicated.

1.1.3 Secondary Structure – The Double Helix

The two complementary strands of DNA form the ultimate biological unit of DNA, the 'double helix' (Figure 1.3.). The double helix can be imagined as a vastly long rope ladder twisted about its central axis, with the sugar-phosphate backbone forming the outer 'rope' support and the base paired nucleobases forming the rungs. The phosphate 'backbone' of DNA has a strongly negative charge. In the most commonly observed form of the double helix, the B-form, the helical nature of DNA causes the nucleotides to spiral around the central axis and form two grooves within the phosphate backbone. These grooves are repeated along the double helix. The minor groove occurs when backbones are in close proximity and the major groove when they are far apart. Many sequence-specific DNA binding proteins will preferentially bind in the major groove as it displays more base identifying chemical groups than the minor groove (Xiong and Muttaiya, 2001). The most common class of eukaryotic DNA-binding transcription factors are zinc-coordinating proteins, which interact with the major groove of DNA. However, there are also a number of minor groove binding proteins, such as the TATA-box binding protein which is involved in the initiation of transcription by eukaryotic organisms (Bewley *et al.*, 1998).

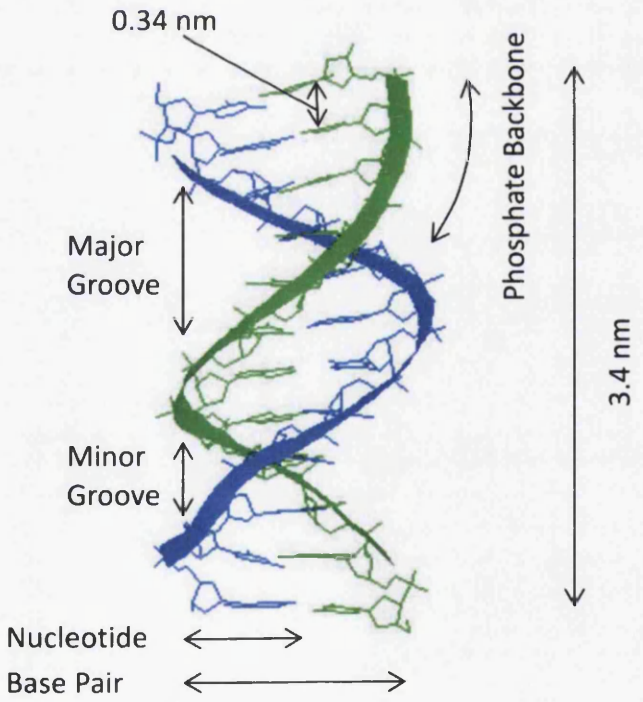

Figure 1.3. – The general structure of the DNA double helix in B-DNA. The distances in nanometres represent standard measurements. The major features of the B-form helix are indicated (Baumann *et al.*, 1997).

1.1.4 Helical Polymorphisms

The DNA helix was described as having a radius of 1 nanometre (nm) and pitch of 3.4 nm, by its discoverers James Watson and Francis Crick, with one complete turn about its axis every 10.5 bp (Watson and Crick, 1953). Whilst the true dimensions of DNA vary when in different ionic solutions, these values act as an excellent rule of thumb (Baumann *et al.*, 1997). This structure of DNA, later termed B-form DNA or B-DNA, was found to be one of many related helical structures that DNA can adopt. However, the B-form of the double helix is by far the most predominant of the possible helical structures that DNA adopts within the cell (Richmond and Davey, 2003). A generalised structure of B-DNA is presented in Figure 1.3.

While the B-form of DNA is thought to predominate in nature there are a number of different forms that DNA will adopt under both artificial and physiological conditions (Richmond and Davey, 2003). Of these possible forms only A-DNA and Z-DNA have been proposed to occur naturally (Figure 1.4.). Shortly after the discovery of B-form DNA by Watson and Crick the A-form of DNA was discovered by Franklin and Gosling (Franklin and Gosling, 1953). A-DNA has a shorter, broader helix when compared to B-form DNA with a helical turn of 11 bp in comparison to the 10-10.5 bp of B-DNA (Basham *et al.*, 1995). The formation of A-DNA is thought to have a role in transcriptional regulation (Llewellyn *et al.*, 2009) and may also form when DNA is bound by a ligand (Lu *et al.*, 2000). The propensity of DNA to adopt the A-form is sequence dependent, the major determinant for the formation of A-DNA is the hydration of phosphates along the backbone (Lu *et al.*, 2000).

Whereas A-DNA has a shorter, squatter structure compared to B-DNA, Z-DNA is quite the opposite. Z-DNA adopts a long left handed helical structure that repeats every 2 base pairs with 12 base pairs per turn (Dickerson *et al.*, 1982). The major and minor grooves in Z-DNA show little difference in width. Z-DNA has typically been difficult to study as it is only transiently formed under certain biological conditions (Zhang *et al.*, 2006). A variety of conditions have been shown to promote the formation of Z-DNA including high salt conditions, multiple repeats of the GC dinucleotide and negative supercoiling. While there is no definitive role for Z-DNA in the cell it has been hypothesised that Z-DNA forms to provide torsional relief for supercoiled DNA during transcription as the propensity for Z-DNA formation is found in regions of high transcription (Champ *et al.*, 2004).

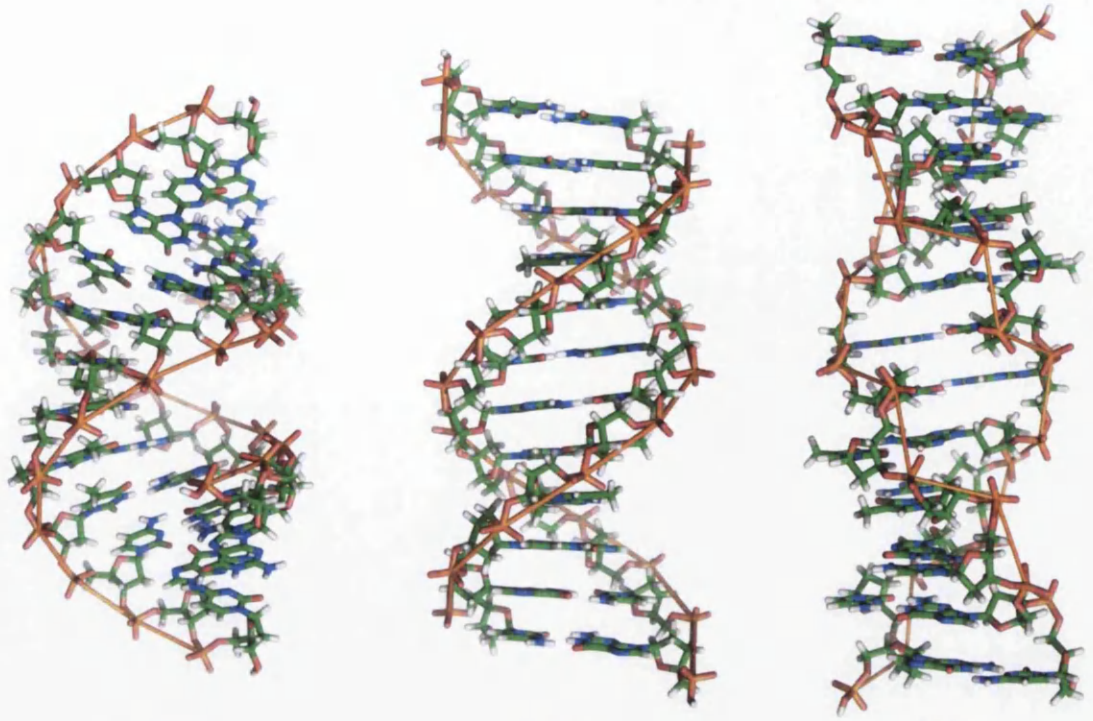

Figure 1.4. Generalised structures of A-, B- and Z- DNA. The image is owned by Richard Wheeler (www.richardwheeler.net).

1.2 Intrinsic DNA Curvature

The discovery and characterisation of DNA curvature was an incremental process contributed to by multiple researchers. The quantification of DNA curvature alongside efforts to fully characterise and model its occurrence are still ongoing. The discovery of sequence-specific DNA curvature had a profound influence on biologists studying DNA packaging, recognition and transcription. The idea of DNA as a featureless repeating polymer has long been dispelled. Asymmetrical kinks and bends are known to be caused by the binding of proteins and chemical ligands (Xiong and Muttaiya, 2001; Cassina *et al.*, 2011). However, external distortion is not necessary for local structural polymorphisms within DNA. DNA that is free of bound proteins displays heterogeneity in structure that is entirely dependent upon local DNA sequence and, to a still disputed degree, long-range sequence context. This local heterogeneity manifests as a smooth curvature over a number of helical turns that is dependent upon the local DNA sequence. DNA curvature and bending can be quantified as angles in degrees (°) or radians (rads) between two base pairs or, on a larger scale, between helical turns.

1.2.1 The Discovery of DNA Curvature

The influence of sequence on the angle between base pairs was first hinted at by DNA X-ray fiber diagrams twenty years after the discovery of the B-form of DNA (Bram, 1973). The role of local DNA sequence in the generation of DNA curvature was later confirmed and further elucidated by other experimental techniques, such as gel electrophoretic mobility and nucleotide digestion (Wang, 1979; Dickerson and Drew, 1981a; Wu and Crothers, 1984). The most conclusive evidence of sequence control of curvature and a huge source of information on DNA structure, was provided by the first X-ray crystal structures of DNA (Dickerson and Drew, 1981b).

The focus of research efforts over the last few decades, after confirmation that DNA intrinsic curvature had a sequence dependent component, has been mainly to identify and attempt to make quantitative measures of the influence of DNA sequence on curvature. The first attempt at this suggested that large amounts of eukaryotic DNA may be curved and that this intrinsic DNA curvature was likely to facilitate packing within the nucleosome (Trifonov and Sussman, 1980). Further experimentation confirmed that AA tracts in phase with repeats in the DNA helix cause gradual curvature (Marini *et al.*, 1982; Wu and Crothers, 1984). DNA sequences other than AA-TT repeats generate curvature to a greater or lesser extent. Larger scale structural context and environmental conditions, such as the amount and type of ions within solution, also plays a role (Haran *et al.*, 1994). The discovery that divalent cations induce curvature in DNA explained discrepancies between experiments in solution and X-ray crystallography data and effectively settled the debate concerning DNA sequence dependent curvature (Brukner *et al.*, 1994). The occurrence of intrinsic, sequence-dependent DNA curvature is now widely accepted.

1.2.2 Biological Roles for DNA Curvature

Intrinsic DNA curvature has been confirmed to be involved in a number of biological processes and has been implicated in many more. A selection of these have been presented below:

1.2.2.1 Protein Binding

As proteins are the ultimate effectors of processes involved in the transcription, replication and repair of DNA the effects of curvature and flexibility on protein binding have important biological implications. A number of DNA binding proteins introduce a local deformation, sometimes called a kink or bend, on binding (Luscombe *et al.*, 2000). Still other proteins recognise regions of DNA that are sufficiently curved either intrinsically or due to environmental or chemical factors (Missura *et al.*, 2001). It is likely that the structure of DNA,

either large scale curvature or localised kinking, is used by proteins to distinguish between members of the same DNA-binding protein family (Rohs *et al.*, 2010). Therefore, DNA curvature may have been selected during evolution for at least three major reasons: to facilitate histone binding and chromatin remodelling (Anselmi *et al.*, 2000; Cairns, 2009), as a feature recognised by a number of specialised proteins (Missura *et al.*, 2001; Rohs *et al.*, 2010) and a facilitator of DNA binding by ligands by reducing the mechanical cost of deformation due to intrinsic DNA curvature or flexibility. This provides a role for DNA curvature in a number of biological processes which will be detailed below.

A classic example of a protein with activity influenced by DNA curvature is DNase I. DNase I requires access to the minor groove of DNA in order to function (Suck, 1994). The activity of DNase I becomes markedly higher in intrinsically bent DNA that more often presents access to the minor groove; similarly, highly flexible DNA will also provide access to the minor groove more frequently than rigid DNA. This property of DNase I has been taken advantage of by researchers to study the curvature and flexibility of experimental DNA sequences (Brukner *et al.*, 1995a).

1.2.2.2 Nucleosome Affinity and Chromatin Structure

One of the first roles discovered for DNA curvature was its involvement in the nucleosome affinity of DNA sequences (Satchwell *et al.*, 1986). Histones are the proteins that package DNA within eukaryotic cells. Histones wind DNA around a number of nucleosomal proteins. The DNA thus packaged is called chromatin. The involvement of intrinsic DNA curvature in nucleosome affinity has not been fully explained and is only one factor that determines nucleosome affinity (Nair, 2010). However, it has been observed that nucleosome formation favours DNA with low flexibility and high curvature, via two mechanisms: decreasing the free energy of DNA distortion by nucleosomes and by increasing the energy cost that the corresponding DNA free form spends to release a part of the spine of water displaced by histone interactions (Anselmi *et al.*, 1999, 2000). The intrinsic curvature of DNA has also been implicated in the process of chromatin remodelling necessary for DNA transcription and replication (Cairns, 2009).

1.2.2.3 Transcription

DNA curvature plays a multitude of roles in DNA transcription. Highly curved DNA is present in the promoter region in prokaryotic organisms (Asayama and Ohyama, 2000). This motif is so prevalent in prokaryotes that DNA curvature, alongside other physio-chemical properties of DNA, has been used to identify and characterise different promoter regions (Jauregui, 2003). It has also been hypothesised that DNA curvature plays a role in the

termination of transcription in prokaryotes (Kozobay-Avraham *et al.*, 2006). Additionally, proteins introduce sharp bends that regulate the propagation of supercoiling in prokaryotic DNA, indicating another role for DNA curvature in prokaryotic transcription (Leng and McMacken, 2002).

Prokaryotic genomes only contain three different promoter elements (-10, -35 promoters and upstream elements) whereas eukaryotic genomes contain a wide variety of promoter elements (Struhl, 1999). Therefore, the involvement of DNA curvature in transcriptional regulation of eukaryotic organisms is less clear, due to its increased complexity. However, a number of curved DNA motifs are found clustered in and around promoter regions in eukaryotic genomes (Ohyama, 2005). This has led researchers to propose a number of functions for DNA curvature in eukaryotic transcription including: as a structural feature recognised by transcription factors, regulation of transcription in association with transcription-factor-induced bending of DNA and as an organising factor for local chromatin. Theoretical DNA curvature measurements have also been incorporated into efforts to identify novel promoters in eukaryotes (Abeel *et al.*, 2008).

1.2.2.4 DNA Damage and Repair

In many DNA damage pathways DNA damage is recognised due to the conformational effect on DNA such as double strand breaks and single strand nicks. In the case of the nucleotide excision repair (NER) pathway, damage is recognised by local bends formed by chemical adducts (Missura *et al.*, 2001). The key damage recognition proteins involved in NER pathway, XPA and RPA, have been shown to detect damage not by identification of adducted bases but by the conformational irregularities that they produce (Missura *et al.*, 2001). However, both XPA and RPA recognise different aspects of conformational change. XPA was shown to have a high affinity for sharply and rigidly bent sections of the duplex DNA, often caused by bulky DNA adducts, while RPA recognises single strand DNA loops, mainly formed due to mismatches.

While this alone indicates a role for DNA curvature in NER, more compelling evidence has recently been published. The local DNA sequence bordering a bulky chemical adduct was shown to have a measurable effect on the repair efficiency of the NER pathway (Cai *et al.*, 2009, 2010). These studies indicated that the role for DNA curvature and flexibility is that of a destabilising or stabilising factor in the presentation of DNA adducts for repair. Gel electrophoretic experiments and molecular dynamic (MD) simulations have indicated that rigidly bent DNA presents a wider minor groove leading to more efficient excision and repair of DNA lesions. The bulky adduct under investigation for these studies was benzo[a]pyrene diol epoxide (BPDE), a chemical carcinogen heavily involved in the initiation and progression of

lung cancer (Hecht, 2002; Kometani *et al.*, 2009). DNA curvature has also been identified as a possible protective factor for the protection of prokaryotic chromosomes from viruses (Abel and Mrázek, 2012). On an additional note, the tertiary structure of DNA has been shown to effect the rates of adduct formation (Raney *et al.*, 1993).

1.3 Defining DNA Curvature

It is necessary to distinguish between *curvature* and *local bending.* Local bending is the deviation from an ideal straight helix over a fraction of a helical turn, whereas DNA curvature is the conformation of a DNA tract measured over a number of helical turns (Goodsell and Dickerson, 1994). Curvature therefore discounts local writhe within the helix whilst giving a measure of how curved a sequence is on a macro scale (Figure 1.5.). A section of DNA with high curvature may be constructed from many locally straight DNA sections and include only a few bent sequences if the majority of the bent sequences are curved in the same direction. Similarly, a section of DNA with a high degree of non-uniform local bending, or writhe, may have functionally no curvature over a number of helical turns.

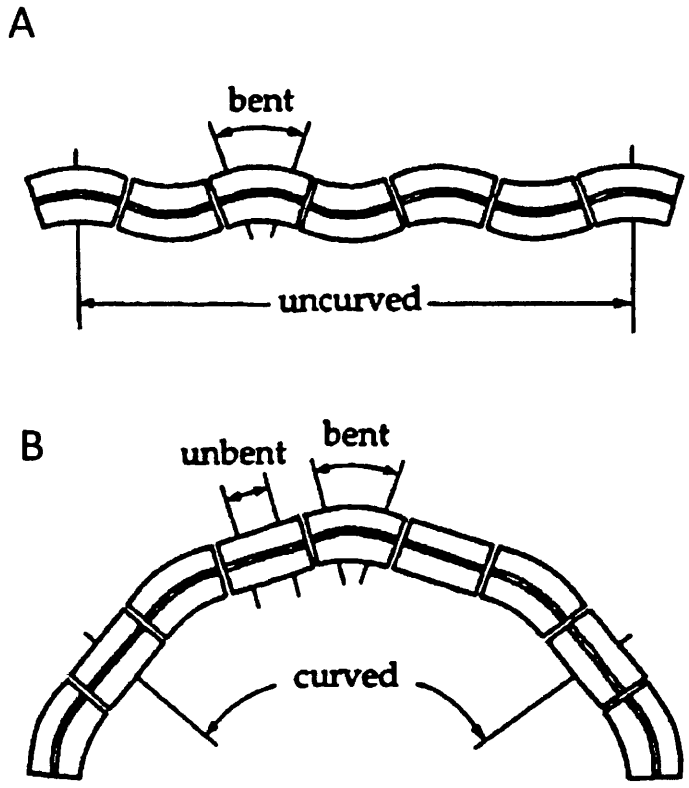

Figure 1.5. - Comparison of local bending and curvature in curved and uncurved section of DNA. A) A section of DNA with a large degree of local bending (writhe) which has low overall curvature. B) A section of DNA with a low degree of local bending interspersed with sections of straight DNA which have a large degree of curvature. Each individual section represents a helical turn or short series of base pairs. The figure is adapted from Goodsell and Dickerson, 1994.

1.4 The Experimental Investigation of DNA Curvature and Flexibility

With the discovery of the propensity of DNA sequences to intrinsically bend it became clear that there was a need to measure and quantify curvature. There have been a number of different methodologies developed for measuring the curvature of DNA sequences. One of the obstacles in quantifying DNA curvature has been that DNA is a naturally dynamic polymer and has a degree of sequence-specific flexibility (Hagerman, 1988). In many experiments this has required the considerations of the static (curvature) and dynamic (flexibility) contributions to DNA curvature. As DNA curvature and flexibility are both determined by DNA sequence context the study of curvature often goes hand-in-hand with the study of flexibility. Some studies were unable to determine which parameter, curvature or flexibility, contributed to the observed changes in DNA conformation and instead amalgamated both parameters into a single metric of DNA deformability (Brukner *et al.*, 1995b). There are also individual measures of DNA flexibility; for example, DNA persistence length (ξ) measures the distance over which DNA retains its original trajectory (Bednar *et al.*, 1995). A number of different methodologies for investigating DNA curvature and flexibility have been detailed below.

1.4.1 Gel Mobility

One of the simplest and most commonly used experiment procedures for the measurement of curvature has been polyacrylamide gel electrophoresis. Curved DNA sequences have been observed with unexpected gel mobility (*e.g.* Dlakic & Harrington, 1998a; Marini *et al.*, 1982; Zinkel & Crothers, 1990). A number of models have been proposed that relate intrinsic curvature to the mobility of DNA in gel electrophoresis. Some of these models have formed the basis of popular dinucleotide models of curvature (*e.g.* Trifonov and Sussman, 1980; Bolshoy *et al.*, 1991; De Santis *et al.*, 1988; Ulanovsky and Trifonov, 1987). The models explain the results of the gel mobility experiments and typically only consider the intrinsic curvature of a DNA tract. DNA flexibility is not considered as a large component of these models.

1.4.2 Bendability Experiments

The physical characteristics that influence the affinity of DNA for a number of proteins have been exploited by researchers studying DNA curvature in a series of related experiments. The fractional occurrence of DNA sequences in chromatin taken from chicken erythrocyte cells has been used to generate an index of the bendability of DNA sequences (Satchwell *et al.*, 1986). This experiment exploited the affinity of nucleoproteins for curved DNA. The activity of DNase I, a DNA degradation enzyme, is dependent upon access to the minor groove of DNA (Brukner *et al.*, 1995a). Flexible or intrinsically curved DNA presented the minor groove at a

higher rate than uncurved and inflexible sequences. The propensity of sequences to be cut by DNase I was taken as a measure of DNA curvature and flexibility.

1.4.3 DNA Cyclisation Kinetics

Another methodology developed for the study of DNA curvature is DNA cyclisation kinetics. This typically involves measuring the ratio of linear DNA molecules that circularise in a solution containing DNA ligase (Shore and Baldwin, 1983). The probability of forming a closed circle is related to the persistence length, a measure of DNA flexibility, of the DNA molecule. Researchers have worked on a large pool of DNA sequences to develop theoretical and computer generated models of cyclisation kinetics (*e.g.* Shore and Baldwin, 1983; Shimada and Yamakawa, 1984; De Santis *et al.*, 1996; Merlitz *et al.*, 1998; Levene and Crothers, 1986). The resulting models consider both the curvature and flexibility of experimental DNA sequences in the resulting model.

1.4.4 X-ray Crystallography and Nuclear Magnetic Resonance Imaging

Other techniques that have been invaluable for the study of DNA curvature are X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. X-ray fibre diagrams gave the first indications that DNA had intrinsic curvature (Bram, 1973). X-ray crystallography was the first tool available for the elucidation of representative roll, tilt and twist parameters for oligonucleotides. However, there are a number of possible conformations for an individual DNA molecule of which a crystal structure represents only one (El Hassan and Calladine, 1996). Additionally, crystal packing can have a large effect on the resulting structure and it is necessary to study multiple structures from different crystallisation environments in order to produce a representative picture of DNA structure (Dickerson *et al.*, 1994). In some cases the outcome of multiple experiments has produced diametrically opposite results (Crothers *et al.*, 1990; Dickerson *et al.*, 1994; Goodsell *et al.*, 1994). X-ray crystallography was the basis for a theoretical model that considered the contribution of both DNA curvature and flexibility (Olson *et al.*, 1998).

NMR spectroscopy of DNA is a powerful tool for studying DNA structure (Young *et al.*, 1995; Dornberger *et al.*, 1998; Travers, 2004). NMR results have been shown to be significantly better at predicting curvature in experimental DNA than X-ray crystallography (Gabrielian and Pongor, 1996). NMR also allows for the investigation of structures in solution, something which X-ray crystallography is unable to provide. However, NMR is unable to provide long range information about structures under investigation (Young *et al.*, 1995). The recent application of small angle X-ray diffraction "fingerprinting" to DNA structures in solution could

provide a bridge between the two methodologies and a useful tool for evaluating the results of molecular dynamics (MD) simulations (Zuo *et al.*, 2006).

1.4.5 Molecular Dynamic Simulations

Molecular dynamics simulations exist on the line between experimental and theoretical methodologies and allow researchers to test hypothesis in a meaningful way. MD simulations are computer models of the movement of atoms and molecules and the interaction of inter-atomic forces. MD simulations of DNA have shown, at least in the general features, to resemble NMR or X-ray crystallography data (Young *et al.*, 1995; Dixit *et al.*, 2004). However, variances in methodological and simulation parameters imply a level of uncertainty in the outcome of MD simulations. The results of MD simulations have also been shown to deviate from expected experimental outcomes, especially within AT rich DNA sequences (Cheatham and Young, 2000; Zuo *et al.*, 2006). MD simulations lack a sufficiently complete library of molecular structures to provide a comprehensive answer to the question of their adherence to experimental data (Beveridge *et al.*, 2004). Promising tools for checking the veracity of MD simulations, such as small angle X-ray diffraction, have recently been developed (Zuo *et al.*, 2006). Molecular dynamics can provide useful information on local DNA properties and will become more accurate and powerful as increasing amounts of experimental data become available with which to refine the method.

1.4.6 Atomic Force and Electron Microscopy

Atomic force microscopy (AFM), also called scanning force microscopy, and electron microscopy (EM) provide additional information in the form of measurements of the contour length of individual DNA molecules and ensemble population of molecules (Bednar *et al.*, 1995; Rivetti *et al.*, 1996). AFM has often been preferentially used for investigations of DNA curvature as the sample preparation procedures are simpler than those required for EM. In the preparation of DNA for EM it is necessary to treat DNA with heavy metals. AFM imaging of DNA can be performed using a range of different buffers in either air or liquid. Early AFM and EM experiments often dealt with the dynamic contribution of sequence to curvature. The works of Scipioni and colleagues gave researchers solid theoretical grounds for the separation of the effects of sequence on intrinsic DNA curvature and flexibility (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a). The authors showed that by accounting for both the direction and the magnitude of DNA curvature by AFM imaging it was possible to measure the static and dynamic contributions to curvature.

1.4.7 Experimental Separation of DNA Curvature and DNA Flexibility

In order to evaluate both the static and dynamic contributions to DNA curvature, *i.e.* curvature and flexibility, they must be experimentally separated. The first experiments that probed sequence-specific DNA flexibility used DNA tracts designed to have anomalous curvature or flexibility. Examples of such experiments include: controlling curvature with in-phase (Rivetti *et al.*, 1998) and out of phase A-tracts (Bednar *et al.*, 1995), base pair mismatches (Kahn *et al.*, 1994; Grove *et al.*, 1996), a single nick in the DNA backbone (Le Cam *et al.*, 1994), single-stranded sections in the DNA sequence (Rivetti *et al.*, 1998), asymmetric charge neutralizations of the phosphate backbone (Hardwidge, 2002) and double-stranded linker regions between two tracts of triple-helix DNA (Akiyama and Hogan, 1997). AFM was the method of investigation in the first experiments that could conclusively claim to separate the contributions of intrinsic curvature and flexibility in 'real' DNA sequences, as opposed to constructed test DNA sequences (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a). AFM has also been used to compare theoretical models of DNA flexibility to experimental DNA tracts (Marilley *et al.*, 2005) and has shown that on short scales DNA is more flexible than predicted by classical models of DNA curvature (Wiggins *et al.*, 2006).

1.5 Atomic Force Microscopy as a Tool for Studying DNA

The atomic force microscope was developed by Binnig, Quate and Gerber in 1986 (Binnig *et al.*, 1986). The precursor to the AFM was the scanning tunnelling microscope which earned Binnig and Rohrer the Nobel Prize in Physics (Binnig and Rohrer, 1993). The AFM functions by measuring the interaction between a sample surface and a nanoscale size probe. As this interaction is mechanical, not optical, it can take measurements of a surface on scales much smaller than the optical diffraction limit. The resolution of AFM images is on a nanometre scale. Measurements of interaction forces between the tip and the sample are also possible and routine. The nanoscale probe is typically a flexible cantilever on which is mounted a very small, sharp tip.

The tip is moved over the surface, much like the stylus of a record player. Deflections in the movement of the cantilever are detected by a laser coupled to a photodiode that reflects off the back of the cantilever. Nanoscale movements between the tip and surface can be precisely controlled by a piezoelectric element in the scanner head or the motor stage on which the sample is mounted (Figure 1.6.). AFM has become one of the most widely used tools for investigations in biology at the nanoscale. Another advantage of the physical nature of AFM is that it can be performed in both ambient (air) and solution (liquid) environments. This has important implications for studies investigating both naked DNA and DNA-protein interactions.

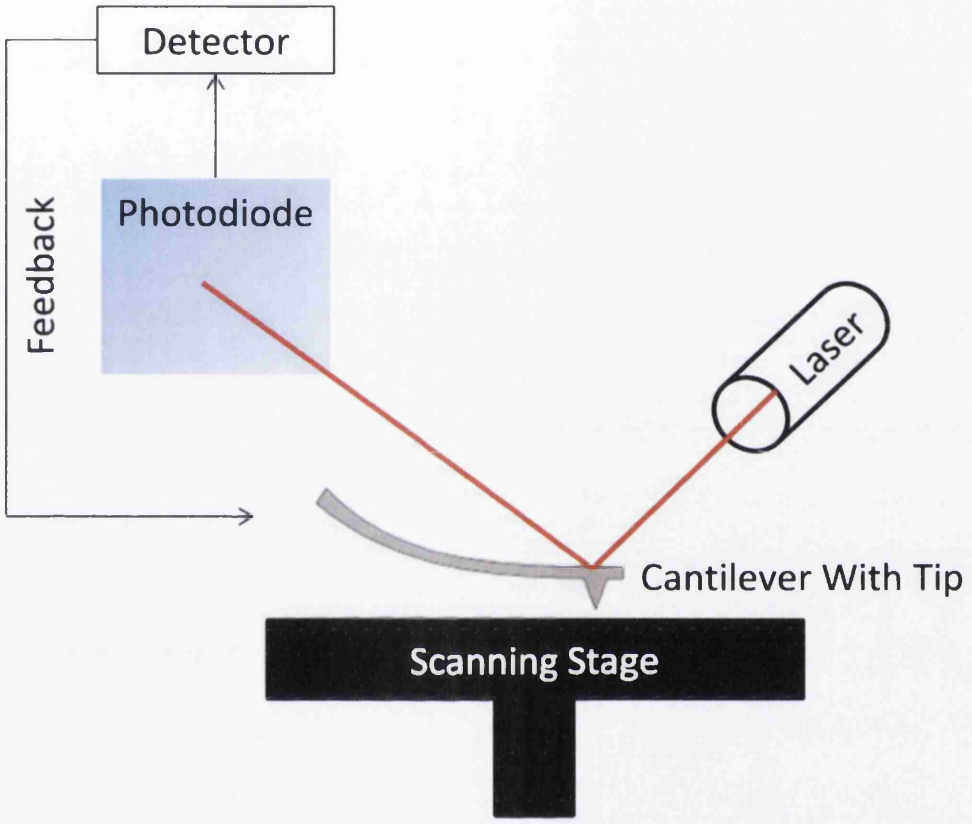

Figure 1.6. - Schematic representation of an atomic force microscope. The red line represents the laser used to track the oscillation of the cantilever.

1.5.1 AFM Imaging Modes

A number of different imaging modes have been developed for AFM imaging. In contact mode a tip is passed in close proximity to a sample surface. Deflections in the cantilever relate to deformations on the experimental surface. Contact mode is not typically considered a suitable technique for imaging DNA as there is a large possibility of the tip breaking or perturbing DNA during scanning (Hansma *et al.*, 1994). An offshoot of this mode is the direct manipulation of DNA on the surface using the AFM probe. This can be used for nano-dissection of DNA samples (An *et al.*, 2005) and has been used for highly accurate screening and selection of fluorescently labelled DNA from individual cells (Di Bucchianico *et al.*, 2011).

Intermittent contact mode, often called tapping mode, is typically used in AFM studies of DNA. This mode overcomes some of the most problematic imaging factors such as surface adhesion, friction and electrostatic forces by only bringing the tip into intermittent contact with the surface. The tip is oscillated near the resonant frequency of the cantilever. The tip is not dragged along the surface as in contact mode. Instead the oscillating tip is brought into light contact with the surface. The oscillation of the tip is maintained at a constant level by a feedback loop. As the tip interacts with the surface the oscillation is dampened. The reduction in oscillation amplitude is used to map features on the experimental surface. This is highly effective for imaging DNA, where shear forces can cause damage to the sample (Hansma *et al.*, 1994).

1.5.2 A Brief History of AFM and DNA Imaging

The first commercially available AFM was released in 1989. It was not long before the power of the system was applied to biological problems. The first published reproducible AFM images of naked DNA were produced in 1992 (Hansma *et al.*, 1992). For the first decade after the development of the AFM many DNA researchers applied themselves to methodological problems with imaging DNA under AFM. A huge amount of methodological literature was published during this time, for example: researchers explored different ways of preparing DNA for imaging (Allison *et al.*, 1992; Bezanilla *et al.*, 1995; Thomson *et al.*, 1996), the resolution limits of AFM imaging (Mou *et al.*, 1995), imaging DNA in liquid (Hansma *et al.*, 1992; Lyubchenko, 1993), enzymatic reactions and degradation of DNA (Bezanilla *et al.*, 1994), humidity effects on the height of DNA on mica (Vesenka *et al.*, 1993), imaging of DNA–protein complexes (Allen *et al.*, 1992) and, importantly, the application of intermittent contact mode to imaging DNA (Hansma *et al.*, 1994).

1.5.3 DNA Dynamics

AFM allows the researcher not just to visualise immobilised DNA but also DNA freely able to move on a surface in liquid buffers. This property makes AFM one of the only available tools for direct visualisation of DNA dynamics. It has been used for a number of different studies: the visualisation of DNA transiently forming non-B-form tertiary structures (Tiner *et al.*, 2001), formation and recognition of stem-loop structures (Lonskaya *et al.*, 2005), the formation of cruciform structures (Mikheikin *et al.*, 2006), manipulation of DNA structure by protein interaction (Jiao *et al.*, 2001), the gradual melting of replication origins (Marilley *et al.*, 2007a) and the dynamic movement of nucleosomes (Shlyakhtenko *et al.*, 2009). Additionally, researchers have begun to incorporate time-lapse imaging of individual DNA molecules into studies of DNA curvature (Marilley *et al.*, 2005). The recent development of reliable high speed AFM imaging techniques has expanded the scope of studies investigating DNA dynamics. The ability to image DNA behaviour or interactions over millisecond time-scales has begun to yield promising results and will continue to do so over the next few years (Lyubchenko *et al.*, 2011).

1.5.4 Mechanical Measurements of DNA by AFM.

AFM has also been used to measure the mechanical elastic forces in the DNA duplex (Bustamante *et al.*, 2000). One end of a single DNA molecule was attached to the AFM tip and the other end attached to the surface. By pulling the tip at a constant force away from the surface and measuring the deflection the elasticity of the DNA molecule can be measured. A related application has been applied to measure the energy required to unzip double stranded DNA (Krautbauer *et al.*, 2003). One complementary sequence is bound to an AFM tip and another to a sample surface, they are brought into contact, allowed to hybridise and then pulled apart. This has been a valuable source of information for theoretical models of inter-helical forces (Cocco *et al.*, 2002). The interaction forces between proteins and DNA can be measured in a similar way (Bartels *et al.*, 2007).

1.5.5 DNA-Ligand Interactions

AFM has obvious utility for visualising DNA-protein interactions. Some of the first applications of AFM to DNA imaged the interaction between DNA and nucleoproteins (Lyubchenko *et al.*, 1995). The deformation of DNA caused by the binding of proteins can be observed in air and liquid conditions (Yoshimura *et al.*, 2000; Lysetska *et al.*, 2002). This has proved invaluable in efforts to understand DNA transcription (Hamon *et al.*, 2007), repair (Yaneva *et al.*, 1997; Wang *et al.*, 2003; Jiang and Marszalek, 2011) and replication (Yoshimura *et al.*, 2000; Lysetska *et al.*, 2002). AFM has become a routinely utilised tool in studies of DNA-protein interactions.

AFM has also been used to assess chemical interactions and the structural perturbations caused by bulky adducts. For example, the chemical carcinogen BPDE, caused a local bend of at least 30 % in supercoiled plasmid DNA (Pietrasanta *et al.*, 2000). AFM can measure the changes in molecule length and persistence length attributed to intercalation by a number of chemical agents (Pastré *et al.*, 2005). Sufficiently large numbers of measurements over a range of different concentrations of the chemical allowed for a measurement of intercalating efficiency and estimation of the number of intercalating molecules (Cassina *et al.*, 2011).

1.5.6 AFM for the Analysis of DNA Curvature and Flexibility

The first steps in the analysis of macromolecular structure of DNA by direct imaging were performed using EM. These initial experiments utilised protein end-labels to orient DNA molecules to generate profiles of the curvature of plasmid DNA (Muzard *et al.*, 1990). These experiments were used as a basis for the first experiments that used AFM to probe intrinsic DNA curvature. The first examples of the analysis of DNA bending by AFM studied molecules designed with sections of artificially modified flexibility or conformational changes induced by proteins (Rivetti *et al.*, 1998; Cam *et al.*, 1999). It was not until the work of Scipioni and colleagues that a solid mathematical underpinning was developed for application to AFM images of DNA (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a, 2002b). These works proved that the contributions of both intrinsic curvature and flexibility to DNA conformation could be individually determined for an ensemble of DNA molecules by AFM analysis. These studies used real DNA sequences as opposed to sequences constructed with anomalous regions of curvature or flexibility.

Other studies have generated novel mechanisms for DNA molecule orientation and have studied profiles of curvature from a number of well characterised DNA sequences (Ficarra *et al.*, 2005b; Milani *et al.*, 2007; Buzio *et al.*, 2012). More recent studies have begun to investigate the functional role of DNA curvature, such as characterising curvature profiles at the origin of replication (Marilley, 2000; Marilley *et al.*, 2007a, 2007b) and the role of DNA curvature in activating the interleukin 2 receptor alpha gene (Milani *et al.*, 2011). One of the most exciting advances of recent years is the detection of the conformational changes induced by single nucleotide polymorphisms in the human osteopontin gene by AFM analysis (Buzio *et al.*, 2012). The detection of conformational changes induced by such minor modifications to the DNA sequences suggests an exciting future of AFM based analysis of intrinsic DNA curvature.

1.5.7 Adhesion of DNA to an Imaging Substrate

A major experimental consideration for AFM analysis of DNA is the selection of an appropriate imaging substrate and buffer. There are few suitable atomically flat substrates that will bind or can be caused to bind DNA. One of the most popular methods of DNA preparation is the use of a mica substrate and a divalent cation containing buffer. The mica surface has a negative charge which is unsuitable for binding DNA. However, by using positively charged divalent cations, that have an affinity for both the negatively charged DNA phosphate backbone and the mica surface, a cationic bridge is formed. One of the benefits of this method is that cationic radius has been shown to influence the strength of adhesion, therefore the strength of DNA binding can be varied by changing the constituent divalent cations and concentration of the buffer (Hansma and Laney, 1996). Chemical modification by spermidine, 1-(3-aminopropyl)silatrane and 3-aminopropyltriethoxy silane is also routinely used to make the mica surface suitable for DNA binding (Lyubchenko *et al.*, 2011). The routine use of highly oriented pyrolytic graphite is complicated by its hydrophobic nature (Oliveira Brett and Chiorcea Paquim, 2005). The typical substrate used in AFM analyses of DNA curvature is mica and the buffer used is typically a divalent cation containing buffer of either Mg^{2+} or Ni^{2+} (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Buzio *et al.*, 2012). The use of Mg^{2+} cations has been shown to produce a weak bond between the DNA and mica which allows for the DNA to equilibrate on the surface and adopt its preferred conformation (Rivetti *et al.*, 1996).

1.6 Image Processing of AFM Images of DNA

For an AFM-based study of intrinsic DNA curvature the image processing steps applied to the resulting images are of central importance. There are a number of methodological considerations to consider which are likely to influence the output of the study. AFM is a relatively high throughput technology. The imaging of naked DNA is free from high contrast topographic features and so requires only minimal oversight from the user. Large amounts of images can be captured in an automated manner on most commercial AFMs using proprietary software or in-house code. There are two major bottlenecks to consider when gathering sufficient data for a study of DNA curvature: image capture speed and image processing speed. The time it takes to capture a typical AFM image is significant. A number of factors effect the speed of imaging and include: image size, resolution (pixels per line), the size of the AFM probe and the quality of cantilever tuning in intermittent contact mode. With the advent of commercially available, high speed AFM imaging (Schitter *et al.*, 2007) with automatic image quality control (Kaemmer, 2011), this bottleneck will soon be overcome. The second bottleneck is the time it takes to process large amounts of AFM images. A number of authors

have detailed image processing workflows for automating the collection of AFM images (Sanchez-Sevilla *et al.*, 2002; Masotti *et al.*, 2004; Ficarra *et al.*, 2005a, 2005b). There are a limited number of commercially and freely available options of varying application and versatility (Collins, 2007; Horcas *et al.*, 2007; Barret, 2008; Nečas and Klapetek, 2011).

1.6.1 Image Processing Software Categorised by Level of Automation

The area of image processing for AFM-DNA images can be broadly divided into three categories: manual methods, semi-automated methods and fully automated methods. Manual methods require a user to 'draw' the backbone line of the DNA molecule using the mouse cursor (Rivetti *et al.*, 1996). These methodologies have been described as 'tedious and time consuming' (Wang *et al.*, 2007). Semi interactive measurements of DNA contour length require the user to specify a number of points along each DNA molecule in an image after which the software performs an algorithmically 'guided walk' to find the highest foreground pixels between the specific points (Marek *et al.*, 2005). Fully automated methods typically have both an unsupervised thresholding algorithm for the identification of foreground pixels and automated algorithms for the removal of imaging artefacts (*e.g.* Ficarra *et al.*, 2005a, 2005b; Fang *et al.*, 1998; Spisz *et al.*, 1998; Sanchez-Sevilla *et al.*, 2002; Wiggins *et al.*, 2006). In some notable publications the authors have used combinations of semi- and fully automated algorithms (Ficarra *et al.*, 2005a, 2005b). In these cases the authors have compared both methodologies or used the automated method only for processing computer simulated DNA images. The semi- and fully automated methodologies have been found to be largely comparable (Ficarra *et al.*, 2005b). The authors concluded with the statement 'the semi-automated procedure can be very effective for selecting molecules of interest because of the ability of the human-eye to distinguish molecules from background noise or artefacts' (pg. 2082, Ficarra *et al.*, 2005b).

1.6.2 Image Processing Toolboxes

A small number of programs for the platform specific toolboxes have been developed by research groups for the analysis of AFM images of DNA. The ALEX toolbox for MATLAB was the first image analysis platform for this purpose (Rivetti *et al.*, 1996). It has been used in a number of publications by members and associates of Dr. Rivetti's group (Zuccheri *et al.*, 2001a, 2001b; Scipioni *et al.*, 2002b). However, the ALEX toolbox has not been updated since its publication, it has limited user documentation and is not freely available to download. A similar application, named Scanning Adventure, has also been developed (Sanchez-Sevilla *et al.*, 2002). This software has been used by a number of authors associated with the original research group (Marilley *et al.*, 2005, 2007b; Milani *et al.*, 2011).

A number of freely or commercially available image analysis platforms offer some application to AFM images of DNA. For example, flexible image processing platforms such as ImageJ (Collins, 2007) have a number of packages that can be adapted for AFM imaging. Other AFM specific solutions include Image SXM (Barret, 2008), WSxM (Horcas *et al.*, 2007) and Gwyddion (Nečas and Klapetek, 2011). Many of these software platforms will allow the user to view and manipulate AFM images from a number of major AFM manufacturers. However, there is very little or no possible customisation available on such software platforms; making complex analysis time-consuming and impractical.

1.6.3 Common Image Processing Steps used on AFM Images of DNA

The aims of an image analysis package for AFM images of DNA are to take an input AFM image, identify DNA molecules, extract their orientation and output the DNA contour using a meaningful coordinate system. There are a number of confounding factors that necessarily have to be understood and appreciated in order to achieve this, which have been discussed in later sections. While there has been no definitive workflow for this type of analysis, the workflow published by Ficarra *et al.*, is debatably the most complete and detailed currently available (Ficarra *et al.*, 2005a). There are a number of common steps that have been adopted by researchers over the last 15 years: a single or multiple plane fitting step, removal or reduction of noise, extraction of foreground objects, repeated erosion of foreground objects to one pixel thinness and removal or erroneous (also called 'spurious') branches from foreground objects to leave the backbone of the DNA contour (Figure 1.7.).

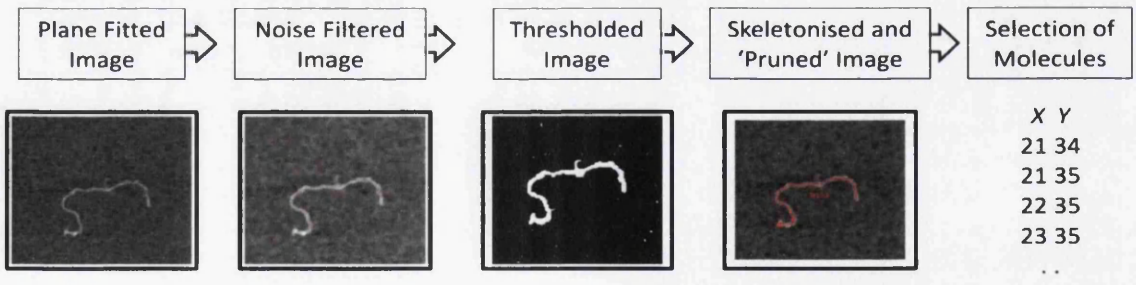

Figure 1.7. – Common steps in image processing toolboxes for AFM images of DNA.

1.6.3.1 Plane Fitting

This step has also been called flattening and is typically the first step in image analysis. AFM imaging usually produces a large amount of Z-height variation on a line-by-line basis. This variation is reduced or removed by fitting and subtracting a variable degree polynomial to each line of the image to produce a flat image. Many authors will fit and subtract a two or three degree polynomial to the image data (Bustamante and Rivetti, 1996). It is not unusual to fit multiple polynomials to an image to ensure that it is uniformly flat (Sanchez-Sevilla *et al.*, 2002). Plane fitting should not effect the resulting analysis unless the Z-height of the image is of considerable interest.

1.6.3.2 Noise Reduction/Noise Filtering

There are a number of different sources of imaging noise that can effect AFM imaging such as acoustic, electrical, vibration, surface interaction and cantilever tuning. Correction for these noises in an automated manner is impractical. They can be classed as impulsive noise and filtered using a 3x3 median filter which is effective at removing impulsive noise (Ficarra *et al.*, 2005b). Other filters have been used by researchers in AFM image analysis on an image to image basis, such as the Weiner, Gaussian (Ficarra *et al.*, 2005b) and average filters (Spisz *et al.*, 1998). The 3x3 median filter is the most often used by researchers (Sanchez-Sevilla *et al.*, 2002; Ficarra *et al.*, 2005a, 2005b). A 5x5 median filter has also been used (Sundstrom, 2008). There are reports of applications that do not use an image filter (Rivetti *et al.*, 1996; Rivetti and Codeluppi, 2001). Any filter that will increase the signal to noise ratio of the target image without causing distortion to the image is suitable for application.

1.6.3.3 Thresholding

This step is sometimes called *Image Segmentation* and should not be confused with *Molecule Extraction*. It is the separation of the foreground and background pixels. Manual methods do not need this step as the DNA contour is interactively selected by the user (Rivetti and Codeluppi, 2001). A number of automated and semi-automated methods have been employed: slider based interactive selection of a single (upper) or double (upper and lower) level thresholding (Marek *et al.*, 2005), treating the background and foreground pixel intensities as two separate distributions and fitting Gaussian curves (Fang *et al.*, 1998), manually chosen threshold values (Sanchez-Sevilla *et al.*, 2002), algorithmically calculated thresholds such as the Ridler and Otsu threshold (Ficarra *et al.*, 2005a) and custom methodologies (Rivetti *et al.*, 1996).

1.6.3.4 Skeletonisation/Thinning/Erosion

This step ideally results in the transformation of thick foreground DNA molecules into 'backbone' contours of one pixel thickness. Thinning involves an algorithm that iteratively removes (erodes) connected pixels on the outside of the binary object. There are a set of constraints, the most important of which is that thinning cannot 'break' an image object into multiple image objects *i.e.* it must stay connected. A number of algorithms have been used for skeletonising DNA molecules: the algorithm of Zhang and Suen which required the addition of a 'corner removal' step for pixels connected in an L pattern (Zhang and Suen, 1984; Spisz *et al.*, 1998), the need to remove corner pixels has been circumvented by using the thinning algorithm of Brugal and Chassery (Brugal and Chassery, 1977; Sanchez-Sevilla *et al.*, 2002) and utilising custom binary image masks (Ficarra *et al.*, 2005a). Comprehensive reviews of thinning algorithms are available (Lam *et al.*, 1992). An optional step, called end-point retrieval, recovers pixels removed during thinning that have Z-heights above the image threshold value that could be considered important for the continuity of the DNA backbone contour at the end of the molecule (Spisz *et al.*, 1998; Ficarra *et al.*, 2004, 2005b).

1.6.3.5 Removal of Image Artefacts/Critical Molecule Removal

During this step obvious image artefacts or erroneous or unsuitable molecules are removed. This includes molecules that are in contact with the image boundaries as the extent of these molecules is unknown, the removal of two molecules that overlap, the removal of 'blobs' below a certain threshold size in pixels and the removal of self-circularised or self-overlapping molecules (Spisz *et al.*, 1997; Ficarra *et al.*, 2005b). The removal of molecules with obviously erroneous contour lengths has also been applied after DNA contour identification and *xyz* coordinate extraction (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Marek *et al.*, 2005).

1.6.3.6 Removal of Spurious Branches or 'Pruning'

This has been identified as the most computationally intensive step for automated methods during image processing (Ficarra *et al.*, 2005a). The analogy most often used in the literature for a thinned DNA molecule is that of a tree trunk with a number of 'spurious branches' that protrude from the 'trunk' of the DNA backbone contour. These branches are introduced by the thinning procedure. Manual and semi-automated methods typically do not need this step as the DNA contours are identified interactively by the user on a pixel-by-pixel or section-by-section basis (Marek *et al.*, 2005). A section-by-section approach will include an algorithm for identifying the most likely intervening pixels and will not create branches. A number of methods have been previously employed for branch removal: the use of image masks for the identification of 'branches' (Ficarra *et al.*, 2005a), considering the problem in

terms of graph optimisation where the longest path is considered as the true DNA contour (Cirrone, 2007) and undocumented methods (Spisz *et al.*, 1998).

1.6.3.7 Molecular Extraction

In experiments where only the length of the DNA molecule is considered it is only necessary to return the position of each pixel in relation to its neighbouring pixels (*i.e.* diagonal, horizontal or vertical) or as a custom 'chain code' (Spisz *et al.*, 1998; Rivetti and Codeluppi, 2001). For the study of DNA curvature each pixel position is recorded in Cartesian coordinates (Ficarra *et al.*, 2005a). This must be performed iteratively from one end to the other in order to preserve the order of the coordinates for further analysis to be possible.

1.6.4 Data Processing and Analysis of DNA Molecules

After image processing there are a number of different physical descriptors that can be estimated from the resulting data. These include contour length, intrinsic DNA curvature, DNA flexibility, DNA persistence length and other customised analyses. The approaches and methods used by previous researchers are discussed below.

1.6.4.1 DNA Contour Length Calculation

A number of different methods for calculation of the length of DNA molecules from AFM images have arisen over the last decade. Digitisation of DNA contours has the effect of smoothing out small structural features below the pixel resolution of the DNA image or, alternatively, pixelising an otherwise smooth DNA contour. Therefore, it is not enough to simply measure the length of the digitised line but also to reconstruct an estimation of the true contour length. This problem has been identified for some time and a comprehensive review of different binary length estimation algorithms and their application to DNA contours is available (Rivetti and Codeluppi, 2001). Some of the most common length estimators are presented include the:

Freeman Estimator/Euclidean Distance - The simplest of the length estimators is the Freeman estimator (Freeman *et al.*, 1970; Spisz *et al.*, 1998). Pixel orientation is considered in terms of Euclidean distance *i.e.* to be in one of two states; either a single horizontal or vertical move or a 'knights' move of one up/down and one left/right. A value of 1.0 is assigned to the horizontal/vertical move and 1.4 to the 'knights' move. The sum of these values is multiplied by a correction factor based upon the resolution of the image in nanometres (*i.e.* size of image in nm divided by number of pixels) and the outcome is considered the length of the DNA molecule. This approach can lead to a length overestimation of as much as 8 % for something as simple as a digitised straight line (Sanchez-Sevilla *et al.*, 2002).

Kulpa Estimator - The Kulpa estimator is a simple modification of the Freeman estimator. It substitutes the values 0.948 and 1.343 for 1.0 and 1.4 respectively (Kulpa, 1977). This has the benefit of being both simple to implement and giving a good estimation of reconstructed DNA contour length (Rivetti and Codeluppi, 2001). It has been used by a number of authors (Rivetti and Codeluppi, 2001; Marek *et al.*, 2005).

Adjustment of Pixel Values by Weighted Average - Ficarra *et al.,* applied an *ad hoc* methodology to the problem of length estimation (Ficarra *et al.*, 2005a). The first step of the method is to transform each pixel coordinate into a weighted average of the surrounding pixel coordinates. The weight to use for the average is experimentally determined for each dataset. The reconstructed length is then calculated as for the Freeman estimator. The reasoning behind this approach is that DNA has continuous curvature and the position of each base pair is dependent upon the preceding and succeeding base pair. As a digitised line is a rough approximation of a curve a smoothing step is necessary prior to calculating curvature. The authors reported a more accurate length calculation than any of the methods previously detailed (Ficarra *et al.*, 2004).

Signal Processing Method – This method treated the DNA contour as a signal processing problem and applied a Fast-Fourier transform of the coordinate data, followed by Gaussian filtering and normalisation (Sanchez-Sevilla *et al.*, 2002). This method produced an estimated length more in-line with expectation than the Freeman estimator alone.

It is clear that there is no consensus method for length estimation within the current literature. Since the work of Rivetti and Codeluppi in 2001 there has not been a systematic attempt to compare any of the more recently developed contour length estimators (Rivetti and Codeluppi, 2001). Individual researchers are free to select a suitably accurate method from those available from the literature based upon their own criteria.

1.6.4.2 Persistence Length

The most widely used experimentally determined measure of polymer flexibility is persistence length, sometimes denoted as ξ or P. Persistence length is a measure of the 'persistence' of the memory of the initial chain direction. It is considered a measure of polymer rigidity, rather than flexibility, as it measures the distance over which a polymer maintains its original orientation. The persistence length of a polymer is defined as the "the length over which the average deflection of the polymer axis caused by thermal agitation is 1 rad." (pg. 67, Virstedt *et al.*, 2004). Although persistence length is a measure of rigidity, it is determined by both a static (curvature) and a dynamic (thermal fluctuations or flexibility) component (Bednar *et al.*, 1995).

The persistence length of DNA has been investigated using a number of techniques including rotational diffusion (Elias and Eden, 1981), light scattering (Sobel and Harpst, 1991), DNA cyclisation (Crothers *et al.*, 1992) and single molecule extension (Baumann *et al.*, 1997). These experiments determined a persistence length for DNA of ~140–180 bp (48-61 nm) and the consensus persistence length of ~50 nm is usually used for B-DNA (Hagerman, 1988). Persistence length has been used as an important global indicator of DNA equilibration on sample substrates (Rivetti *et al.*, 1996) and as a measure of chemical intercalation and adduct formation (Pastré *et al.*, 2005; Cassina *et al.*, 2011).

Persistence length is relatively simple to measure from AFM images of DNA. The two requirements for calculating persistence length from DNA images are accurate measurements of the position of the DNA contour along its length and a large enough sample size of DNA molecules. This end-to-end measure is then compared to predictions made by the WLC model of DNA flexibility (Cassina *et al.*, 2011). The persistence length of DNA estimated from AFM images varies depending on the buffer and adhesion conditions used (Rivetti *et al.*, 1996). Using a Mg^{2+} containing buffer the persistence length of DNA measured by AFM is often cited as being ~50 nm (Rivetti *et al.*, 1996). Considerable deviation from this consensus value has been reported from as low as 36 nm (Lysetska *et al.*, 2002) to as high as 56 nm (Podestà *et al.*, 2005). As an example of the effects of differing buffer conditions on persistence length, a Ni^{2+} containing buffer produces a persistence length of between 30-36 nm on a mica surface (Hansma *et al.*, 1997; Lysetska *et al.*, 2002).

1.6.4.3 Calculation of DNA Curvature and Flexibility from AFM Images

A standard methodology for calculating intrinsic DNA curvature from AFM images has been well detailed within the literature (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b). The first step is to select a set number of points along each DNA molecule at regular intervals. This separates each molecule into a number of comparable line sections or vectors. Direct comparison between points requires the assumption that all the molecules under investigation are complete DNA molecules of identical 'real' length in base pairs.

In order to achieve this, interpolants are fitted to each molecule. This step is sometimes considered to be a smoothing step. There is no consensus method for interpolation and the methods used have included a variable degree polynomial that fitted a number of points below a user defined threshold value (Ficarra *et al.*, 2005b), a number of variable degree cubic splines over a window of 5 pixels (Sundstrom, 2008) and simply standardising for length (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a). There is very little consideration for the effect smoothing or interpolant type may have on curvature angle calculation within the current literature.

The curvature of a line in space is calculated as the derivative of the tangent vector along the line. The modulus of curvature is the inverse of the curvature radius and its direction is the main normal of the curve (Timoshenko and Goodier, 1986). In the case of DNA, the helical axis corresponds to a line and the curvature is the vectorial product of the DNA sequence. Curvature represents the angular deviation between the local helical axis at *n* and *n+1* base pairs, *n* being a point (base pair, pixel or other contour length measure) in the sequence.

DNA can be considered in terms of first order elasticity due to its relatively high rigidity (Scipioni *et al.*, 2002a). The contribution of thermal noise imposing local variations of the structure of DNA is considered zero over a sufficiently large sample size (Ficarra *et al.*, 2005b). Under these assumptions the intrinsic curvature is calculated as the mean angle value of a sufficiently large population of DNA molecules. The flexibility of the DNA sequence at point *n* is the standard deviation of the assembled curvature angles at point *n*. The output of this methodology has been shown to be comparable to several theoretical dinucleotide wedge models of curvature in a number of publications (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Buzio *et al.*, 2012).

1.6.5 Experimental Orientation of DNA for AFM Imaging

For researchers to probe the site of interaction of proteins with DNA or to investigate sequence-specific curvature it is necessary to orient the DNA experimentally. The first attempts at orientation of DNA molecules used 5 nm colloidal gold spheres to label one end of a linear DNA molecule (Shaiu *et al.*, 1993). Researchers also identified enzymatic 'nicks' using biotin-streptavidin probes (Murray *et al.*, 1993). The use of repeating dimeric DNA sequences removed the need for end-labelling in particular experiments (Zuccheri *et al.*, 2001b). This experimental method involved creating a DNA molecule that was symmetrical around a central point, *i.e.* both halves of the DNA molecules have identical sequences oriented in different directions.

There are several post-imaging processing methodologies for molecule orientation. The fragment flipping (FF) algorithm has been well detailed in a pair of related publications (Masotti *et al.*, 2004; Ficarra *et al.*, 2005b). Simplistically, it is a method for orientation of a large population of molecules based upon their intrinsic curvature. Each molecule is 'flipped' into one of its four possible orientations on the flat surface. The orientation of each DNA molecule that reduces the mean column variance within the entire dataset is retained and the algorithm iterates upon every molecule within the dataset. This continues until the dataset meets a minimum optimal objective function, the objective function being the mean of the column variance of the dataset. This method has been shown to be very effective for

theoretical molecules and to give good experimental fit with the De Santis model of curvature. However, a valid concern has been raised that the algorithm relies upon a 'hill-climbing' optimisation routine that is sensitive to local minima (Buzio *et al.*, 2012).

The method described by Buzio *et al.,* is also a post-processing method that takes a very different approach than the FF algorithm (Buzio *et al.*, 2012). This method generates a profile for each molecule. This profile is the ratio of the curvature at symmetrical points along a single molecule. This chain descriptor, averaged over a suitable number of molecules, give an individual profile for a DNA sequence. This is not a typical orientation method as it does not identify either end of an individual molecule. The authors showed that the resulting profile is sensitive enough to identify single nucleotide polymorphisms between two sequences within the human osteopontin gene.

Another method for post-image processing orientation of DNA molecules uses the theoretical pitch of DNA (Milani *et al.*, 2011). The pitch was calculated using 3D theoretical models projected on to a 2D surface. The Z-height measured from experimental images is recorded and the traces are aligned with the theoretical pitch.

1.7 Theoretical Models of DNA Curvature

Theoretical models of intrinsic DNA curvature have great utility in the study of DNA curvature. In order to calculate theoretical DNA curvature, it is first necessary to model the structure of a DNA tract in three dimensions (3D). DNA dinucleotide structure can be characterised by six dinucleotide parameters: slide, shift, rise, tilt, roll and twist. These base pair geometries define the position of each base pair relative to the preceding nucleotide. Tilt and roll define bending angles between spatially adjacent base pairs. Twist is a rotation angle between two base pairs. Rise is the vertical displacement between two base pairs. Shift and slide are in-plane dislocations between base pairs. A simple schematic illustrating base pair geometries is presented in Figure 1.8. There are a number of other parameters, such as propeller twist, the rotation of one base pair in relation to the next, that describe inter- or intra-base pair geometries that have either a negligible or no net effect on the macro structure of DNA and will not be discussed further.

The 3D positions of each base pair within a DNA tract can be calculated by placing the first base pair at the origin of a Cartesian coordinate system (i.e. x,y,z co-ordinates) and then calculating the position of the next base pair using the parameters of roll, tilt and twist and translating them using a rise parameter (Vlahovicek and Pongor, 2000). Either a model-specific rise value is used for each base pair step or a constant value is chosen to reflect the ideal form of DNA, *e.g.* 3.4 Å for B-DNA (Saenger, 1984). Relevant dinucleotide wedge models provide the

dinucleotide step parameters necessary for modelling curvature (Ulanovsky and Trifonov, 1987; De Santis *et al.*, 1988; Bolshoy *et al.*, 1991; Olson *et al.*, 1998)

The calculation of the deviation in the helical axis at each base pair step can be a complex task, requiring the application of intricate matrix algebra (Ulanovsky and Trifonov, 1987). However, for B-DNA a simpler approach is often adopted by calculating the vector normals between each base pair using the base pair parameters. This has been used to great effect in programs such as CURVATURE for high speed and high throughput curvature analysis of DNA sequences (Goodsell and Dickerson, 1994).

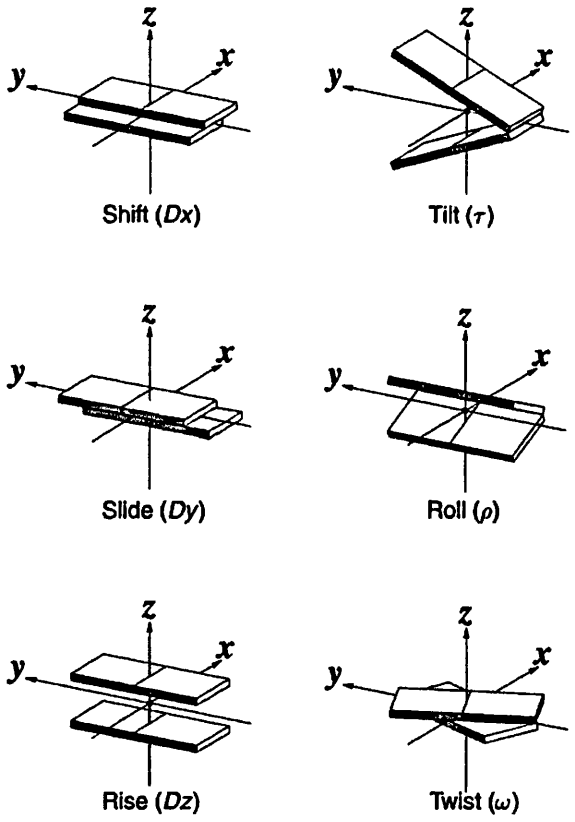

Figure 1.8. - Base pair geometry parameters of slide, shift, rise, tilt, roll and twist. The influence of each parameter on dinucleotide positions is indicated by a coordinate system. The base pair reference frame is constructed such that the x-axis points away from the minor groove edge of a base or base pair and the y-axis points toward the sequence strand. Adapted from El Hassan and Calladine, 1997.

1.7.1 Formative Models of DNA Curvature

Experimental curvature models attempt to extract base pair geometry parameters, be they roll, tilt, twist or relative scales of bendability, from experimental data. Researchers over the last few decades have used a variety of different experimental datasets and mathematical tools in attempts to estimate geometric base pair parameters from synthetic DNA oligonucleotides. Underlying these attempts and allowing for a context in which to interpret experimental results are the formative models of intrinsic DNA curvature. These underlying models are distinct from the individual experimental models of curvature described later; underlying models form a context by which di-, tri- and tetra-nucleotide models can be interpreted. There are a number of formative models that attempt to explain the occurrence of DNA curvature. Only two of the most popular are discussed below.

1.7.1.1 The Junction Model

The Junction model in its first incarnation was developed to explain observation of angles forming at the junction between A- and B-DNA (Selsing *et al.*, 1979). The structure of DNA, *i.e.* base stacking and hydrogen bonding, was preserved at the junction. The premise of the Junction model is that at the intersection, or junction, between normal DNA and an adenine-rich tract there is a change in the direction of the helical axis and a bend is formed (Figure 1.9.A). The Junction model considers distant AT/TA base pairs to have a significant effect on the angle between AA-TT dinucleotides *i.e.* that long-range influences of DNA sequence are considered relevant.

1.7.1.2 The Wedge Model

The Wedge model was first proposed to explain the correlation of sequence repeats within chromatin DNA sequences and was limited to considering the periodic repeat of AA-TT dinucleotides (Trifonov and Sussman, 1980). The original premise of the Wedge model was that non-parallel dinucleotides, *i.e.* dinucleotides forming a bent wedge, would cause unidirectional curvature in the helical axis (Figure 1.9.B). This model has since been refined and considers bending to occur primarily in AT rich sequences but also, to a lesser magnitude, in other DNA sequences (Cooper and Hagerman, 1987; Ulanovsky and Trifonov, 1987). The outcome of this model is smooth bending across the DNA sequence made up of incremental additive wedges with the dinucleotide as the unit of curvature. The model does not consider long range influences of DNA sequence on curvature.

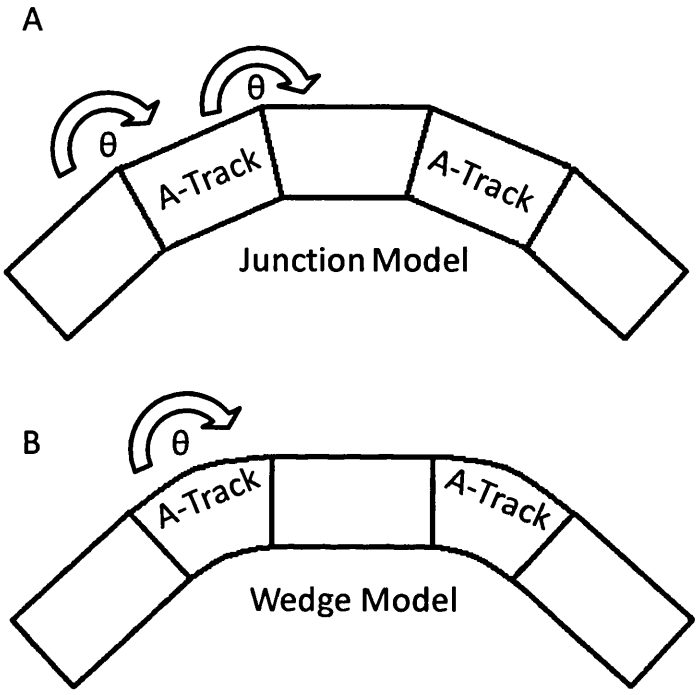

Figure 1.9. - Schematic representation of the Junction (A) and Wedge (B) models. In the Junction model the helix axis deflection (θ) occurs at the interface of B-DNA and A-tracts (B'-DNA). In the Wedge model the A-tracts are curved (θ). However, the Wedge model does not necessarily consider the general-sequence B-DNA between A-tracts to be straight.

1.7.2 Di-, Tri- And Tetra- Nucleotide Models of DNA Curvature

There is no consensus about which formative model most accurately describes sequence-specific DNA curvature. However, the base pair geometric parameters that are extracted from experimental data are often comparable even if the magnitudes of the reported parameters differ (Kanhere, 2003).

1.7.2.1 Dinucleotide Models

Dinucleotide models generate dinucleotide roll and tilt parameters from experimental datasets. While a full review and comparison is outside the purview of this study a selection of different models pertinent to AFM imaging of DNA are detailed below. The Calladine and Drew model was put forward to explain anomalous gel migration experiments in a number of test DNA sequences (Calladine *et al.*, 1988). The dinucleotide parameters were inferred from gel retardation experiments and compared against X-ray diffraction experiments. The Bolshoy model parameters were chosen to explain DNA circularisation and gel retardation experiments (Bolshoy *et al.*, 1991). This model has the largest deviations from the consensus twist and tilt angles of other models. The De Santis model calculated roll, twist and tilt angles from conformation energy calculations of dinucleotide steps (De Santis *et al.*, 1988). The resulting parameters were compared to the result of gel mobility experiments of 62 different synthetic oligonucleotides. The Olson model extracted dinucleotide parameters from a large set of DNA-protein X-ray crystal complexes. In addition to roll, twist and tilt angles the model also incorporates translational parameters of shift, slide and rise. Flexibility of the dinucleotide steps was also estimated from dispersion values of crystal complex data (Olson *et al.*, 1998). The De Santis, Bolshoy and Olson models are considered Wedge models and the Calladine and Drew model is considered a Junction model.

1.7.2.2 Trinucleotide Models

The bending parameters have also been assessed for models of curvature on a scale larger than dinucleotide scale. The following trinucleotide models have also been called *Bendability* models. The experimental results rely upon the propensity of an oligonucleotide to be deformed as a measure of curvature and flexibility. The details of the two major trinucleotide experiments have been described in Section 1.4.2. The results of these experiments have been amalgamated into the Consensus Bendability model (Gabrielian and Pongor, 1996). Trinucleotide models have been shown to be improvements over dinucleotide models (Brukner *et al.*, 1995b; Gabrielian and Pongor, 1996). However, trinucleotide models show little correlation between individual trinucleotide parameters and do not provide

accessible three dimensional parameter values for researchers wishing to simulate DNA in 3D space (Brukner *et al.*, 1995b).

1.7.2.3 Tetranucleotide Models

There are a number of different context specific instances where nearest neighbour interactions have been shown to influence individual dinucleotide angles (Brukner *et al.*, 1995b; Lankaš *et al.*, 2003). In the case of A-tracts even longer range effects have been observed (Burkhoff and Tullius, 1987). Tetranucleotide models can be considered improvements over di- and tri- nucleotide models as they begin to address some of these issues. A collaborative effort by the Ascona B-DNA Consortium has provided a library of MD simulations of all 136 tetranucleotides (Lavery *et al.*, 2010). However, they are based upon the outcome of MD simulations and, as previously mentioned (Section 1.4.5.), there are some valid concerns about the comparability of MD to experimental data.

1.7.3 Comparison of Theoretical Models

Tri- and tetra-nucleotide models have been considered improvements over dinucleotide models due to the ability to encompass more locally derived sequence fluctuations than dinucleotide models (Goodsell and Dickerson, 1994; Gabrielian and Pongor, 1996; Dlakic and Harrington, 1998b). Gel retardation experiments on phased repeat sequences have concluded that trinucleotide models are an improvement over dinucleotide models (Brukner *et al.*, 1995a; Dlakic and Harrington, 1998b). However, a study using a large experimental dataset of NMR measurements of DNA in solution observed that trinucleotide models failed to predict curvature in many of the most extensively studied experimental sequences (Kanhere, 2003). The study concluded that trinucleotide models, as combinations of measurements of both intrinsic curvature and flexibility, made poor predictions of DNA curvature. By contrast dinucleotide models showed good predictive power for all sequences under investigation with the exception of a phased GGGCCC motif. Both Wedge and Junction models typically have weak predictive power for certain GC rich sequence motifs, sometimes generating predicted curvature with the wrong direction to that observed in experimental data (Brukner *et al.*, 1994). It should be noted that curvature within the GGGCCC motif has only been observed when in solution containing divalent cations, so the low predictions could be due to experimental design (Brukner *et al.*, 1994). A quantitative assessment of tetranucleotide models and their predictive power in comparison to di- and tri- nucleotide models has yet to be made.

1.7.4 Comparison of Theoretical Models to Experimental AFM Studies

A number of theoretical dinucleotide models have been used in AFM based studies of DNA. The two most often utilised by researchers are the De Santis and the Bolshoy models. The De Santis model has been shown to provide a good estimation of real DNA curvature measured by AFM in air in a number of studies (Anselmi *et al.*, 1999; Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Buzio *et al.*, 2012). The same holds true for the Bolshoy model for liquid and air imaging, although the majority of research using the Bolshoy model has been in liquid (Sanchez-Sevilla *et al.*, 2002; Milani *et al.*, 2007, 2011; Buzio *et al.*, 2012). Both of these models have been compared by previous authors and were found to be comparable in the prediction of the position, but not magnitude, of curvature peaks (Buzio *et al.*, 2012). This was in agreement with a statistical analysis of the power of dinucleotide models to predict curvature in X-ray crystallography data that concluded that each dinucleotide model was as good a choice as any other for the prediction of intrinsic curvature (Crothers, 1998). The Olson model has not been the subject of critical comparison to curvature profiles in any available publication. However, it has shown to produce a good prediction of DNA flexibility in liquid (Marilley *et al.*, 2005).

The worm-like chain (WLC) model of semi-flexible polymers provides a good framework for generating theoretical persistence length measurements of DNA (Bustamante *et al.*, 1994). It has been shown to provide a good fit to most experimental AFM studies measuring persistence length of DNA (Bednar *et al.*, 1995; Rivetti *et al.*, 1996; Pastré *et al.*, 2005; Cassina *et al.*, 2011; Buzio *et al.*, 2012). It should be noted that on short scales DNA has been shown to be more flexible than predicted by the WLC model in AFM imaging (Wiggins *et al.*, 2006).

1.8 Theoretical Measurements of Curvature in AFM Imaging

Theoretical estimation of a number of different physical DNA parameters have been performed in AFM studies for over a decade. The creation of computer simulated DNA molecules has been important for estimating the error implicit in image analysis methods, for hypothesis generation and hypothesis testing.

One of the first recorded instances, to this authors knowledge, of a comparison of theoretical predictions of DNA curvature to the experimental DNA curvature computed from physical scanning methodologies was the comparison of multiple theoretical models to the curvature of a linearised pBR322 plasmid (Muzard *et al.*, 1990). This research was carried out using electron microscopy, but the images generated are qualitatively comparable to AFM based techniques.

The work of Scipioni and co-workers created a strong mathematical foundation for later researchers to study DNA curvature and flexibility (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a). They showed that there was a mathematical basis for the separation of the DNA curvature and DNA flexibility using AFM imaging. They also compared dinucleotide wedge models of curvature to experimental AFM measurements of DNA curvature and flexibility, showing good correlation between the results (Scipioni *et al.*, 2002a). They further extended this work to dynamic time-lapse images of DNA molecules (Scipioni *et al.*, 2002b).

Rivetti *et al.*, put forward the first standardised workflow for the generation of computer-simulated AFM images (Rivetti and Codeluppi, 2001). By generating simulated AFM images it allowed the authors to assess various contour length estimators. This approach has been adopted by many other researchers in a complete or modified form (Ficarra *et al.*, 2005a, 2005b; Marek *et al.*, 2005; Wiggins *et al.*, 2006; Buzio *et al.*, 2012). Later researchers added curvature measurements generated from dinucleotide models of DNA curvature to the computer-simulated AFM images for hypothesis and method testing (Ficarra *et al.*, 2005b; Buzio *et al.*, 2012).

Other theoretical measures have been used for comparison to AFM images including comparing the theoretical phase of DNA to experimentally determined contour height (Milani *et al.*, 2011), curvature ratio profiles for base pair sequences (Buzio *et al.*, 2012), the prediction of promoter regions in AFM images (Marilley *et al.*, 2007b) and the flexibility of DNA molecules (Scipioni *et al.*, 2002a; Marilley *et al.*, 2005; Wiggins *et al.*, 2006).

1.8.1 Programs for Analysis of Intrinsic DNA Curvature

There are a number of freely available programs for the analysis of structural and physiochemical properties of DNA. One of the oldest programs for the analysis of DNA curvature is BEND (Goodsell and Dickerson, 1994). This work described the first attempt to distinguish between local bending and intrinsic curvature. The program CURVATURE also calculates curvature for a number of popular dinucleotide models of DNA in the same way as BEND (Shpigelman *et al.*, 1993). DNAlive is a web application able to calculate a wide ranging number of structural and chemical measurements such as bendability, flexibility, nucleosome occupancy and a variety of different curvature models from DNA sequence (Goñi *et al.*, 2008). DNAlive also feeds directly into the Human Genome Browser allowing for the annotation of an input sequence with a wide range of published data (Kent *et al.*, 2002).

For many researchers it is often necessary to reconstruct and visually assess 3D DNA structure. There are number of different tools available for this purpose that do not require a high degree of molecular modelling knowledge such as: DIAMOD (Dlakic and Harrington, 1998a), FREEHELIX (Dickerson and Chiu, 1997), Model.it (Vlahovicek and Pongor, 2000),

Curves/Curves+ (Lavery *et al.*, 2009), Madbend (Strahs and Schlick, 2000) and 3DNA (Lu and Olson, 2003). While many of these tools sometimes offer overlapping utility there are many complementary novel features. For instance, Curves+ includes an in-built tool for mapping the helical axis of DNA and 3DNA allows for the reconstruction of very large scale (<2000 bp) DNA molecules that other programs struggle with or simply do not allow. Many of these programs also have web servers hosting a range of applications. In particular 3DNA has a variety of molecular modelling applications with a large degree of flexibility for the advanced user. It also has a web application, w3DNA, which provides a simple user interface for the main features of 3DNA (Zheng *et al.*, 2009).

1.8.2 Other Theoretical Estimators of Physical DNA Parameters

There has long been an effort to model the occupancy of nucleosomes in the genomes of eukaryotic organisms based on DNA sequence. Many computational approaches have been developed for the prediction of nucleosome occupancy or exclusion. A number of these tools are available on-line as web servers, for example: NuPop uses a Hidden Markov model to predict nucleosome occupancy from *S. cerevisiae* genome data (Xi *et al.*, 2010), NXsensor identifies a number of nucleosome exclusion sequences from the literature and identifies those regions that are within less than 147 bp of two exclusion sites as a region of nucleosome exclusion (Luykx *et al.*, 2006) and NuScore calculates nucleosome affinity based upon the estimation of the energy cost of the structural deformation imposed on DNA within the nucleosome core particle (Tolstorukov *et al.*, 2008).

1.8.3 Computer Generated AFM Images

In order to make predictions from dinucleotide wedge models that are comparable to the output of AFM based curvature analysis, computer simulated AFM images of DNA are often used (Cognet *et al.*, 1999; Rivetti and Codeluppi, 2001; Ficarra *et al.*, 2005b; Buzio *et al.*, 2012). The first time simulated images of DNA had been used by researchers; Cognet *et al.*, used simulated DNA as a comparison for electron microscopy of DNA when probing persistence length in 1999 (Cognet *et al.*, 1999). It was only after the work of Rivetti *et al.*, that computer simulated AFM images became commonly used (Rivetti *et al.*, 1996). Simulated images of DNA have been used in a number of studies; for generating a 'base line' length estimate for testing digitised line estimators (Rivetti and Codeluppi, 2001) , testing of the FF algorithm (Ficarra *et al.*, 2005b), hypothesis testing for novel algorithms (Buzio *et al.*, 2012), for comparison of persistence length (Bednar *et al.*, 1995) and as controls for automated image processing and analysis packages (Ficarra *et al.*, 2005a).

There were a number of common steps that the previously mentioned studies shared. These steps were standardised by Rivetti and Codeluppi and have been used in an approximately similar form by later studies (Rivetti and Codeluppi, 2001). The DNA 'chain' was considered, at a basic level, to be made up of a series of rigid rods each representing a single base pair. The rods themselves were assigned a size of 0.34 nm, which is the consensus size of B-form DNA (Rivetti and Codeluppi, 2001). The position of the next rod in the chain was considered a function of the flexibility of the polymer and the curvilinear length of the rod. The flexibility was calculated based upon the persistence length of the polymer, approximately 53 nm was taken as a typical value for B-DNA persistence length (Rivetti *et al.*, 1996). This allowed for the construction of a Gaussian probability curve of angles from which an appropriate angle was selected at random per base pair step. Plotting each rod in the chain with a random angle from the Gaussian distribution produced a 2D image of a DNA polymer. This was made more comparable to a real AFM image by applying a 'grid' at the resolution of the desired image and setting each 'pixel' within the grid that contained DNA to 1. Additional sources of noise or variation found in AFM images such as Gaussian noise or tip convolution were added to user specification. The Z-height can be set by the operator but is more typically left as binary 1 and 0 measurements. This approach allows for the estimation of the effects of digitalisation on a DNA strand, the effects of additive impulsive noise and regular tip convolution.

As a further step some authors have added a curvature value to the simulated images (Ficarra *et al.*, 2005b; Buzio *et al.*, 2012). A curvature measurement is calculated for each base pair step and is used as the mean value of the Gaussian flexibility distribution. Dinucleotide wedge models have effective synergy with this method and it is not hard to see why they are so popular with researchers.

In order to convert dinucleotide wedge models to dinucleotide angles that can be used to simulate DNA molecules in 2D there are a number of important criteria that must be met (Buzio *et al.*, 2012). Firstly, a theoretical framework must be identified which is able to produce 3D models of DNA. Secondly, a method for extrapolation of 3D models to their preferred conformation in two dimensions must be identified. Finally, the 2D models can be used as angle values to create computer simulated AFM images.

1.8.4 Modelling DNA in 3D

On the first point, the generation of robust 3D models of curvature, there are a number of well tested options available to the researcher presented earlier (Section 1.7.4.). The preferred methodology within the literature is the use of the nearest-neighbour, static dinucleotide wedge models. These models treat each base pair as an individual section with a number of experimentally determined parameters that describe the geometric position of one base pair in relation to the preceding base pair. There are a number of different sets of parameters that have been experimentally determined by different research groups using different technologies and experimental conditions (Section 1.7.2.1.).

The dinucleotide parameters can be used to generate a series of *xyz* coordinates based upon tilt roll and twist parameters and their translations: shift, slide and rise (Figure 1.8.). Bend angles can be calculated between two consecutive base pairs (a dinucleotide step). However bend angles are typically unsigned as the idea of a positive or negative angle is meaningless without a reference frame. While these theoretical absolute angles are valuable for investigations of DNA bending they are of less use in reconstructing curvature in two dimensions.

1.8.5 Modelling DNA in 2D

The understanding of the deposition of DNA on to a flat surface is a complex undertaking and not within the purview of this project. In order to simulate curvature in 2D there are two available approaches. The first approach uses the underlying assumption that the curvature modulus (magnitude) of a DNA tract will stay the same when the DNA tract is deposited on a 2D surface while the phase of curvature (direction) will adapt to the changes in the DNA conformation (Scipioni *et al.*, 2002a). This allowed the authors to simulate the resulting curvature and infer a positive or negative curvature for base pair steps.

The second approach, proposed by Buzio *et al.,* has been termed Geometric Deposition within this study (Buzio *et al.*, 2012). The methodology flattens a 3D model of a DNA tract to simulate deposition (Figure 1.10). The method separates the 3D model into a number of sections. A 2D plane of best fit is calculated for each section and the *xyz* coordinates are projected onto each plane to give a representation of the 3D model in 2D. This model assumes that the transformation from 3D to 2D will do so with a minimum number of twists in the DNA backbone. Consequently this implies a minimum energy increase in conformational energy during the flattening process, which, as long as reasonable restraints are applied to the plane fitting process, is in line with mean field models of DNA deposition (Sushko *et al.*, 2006).

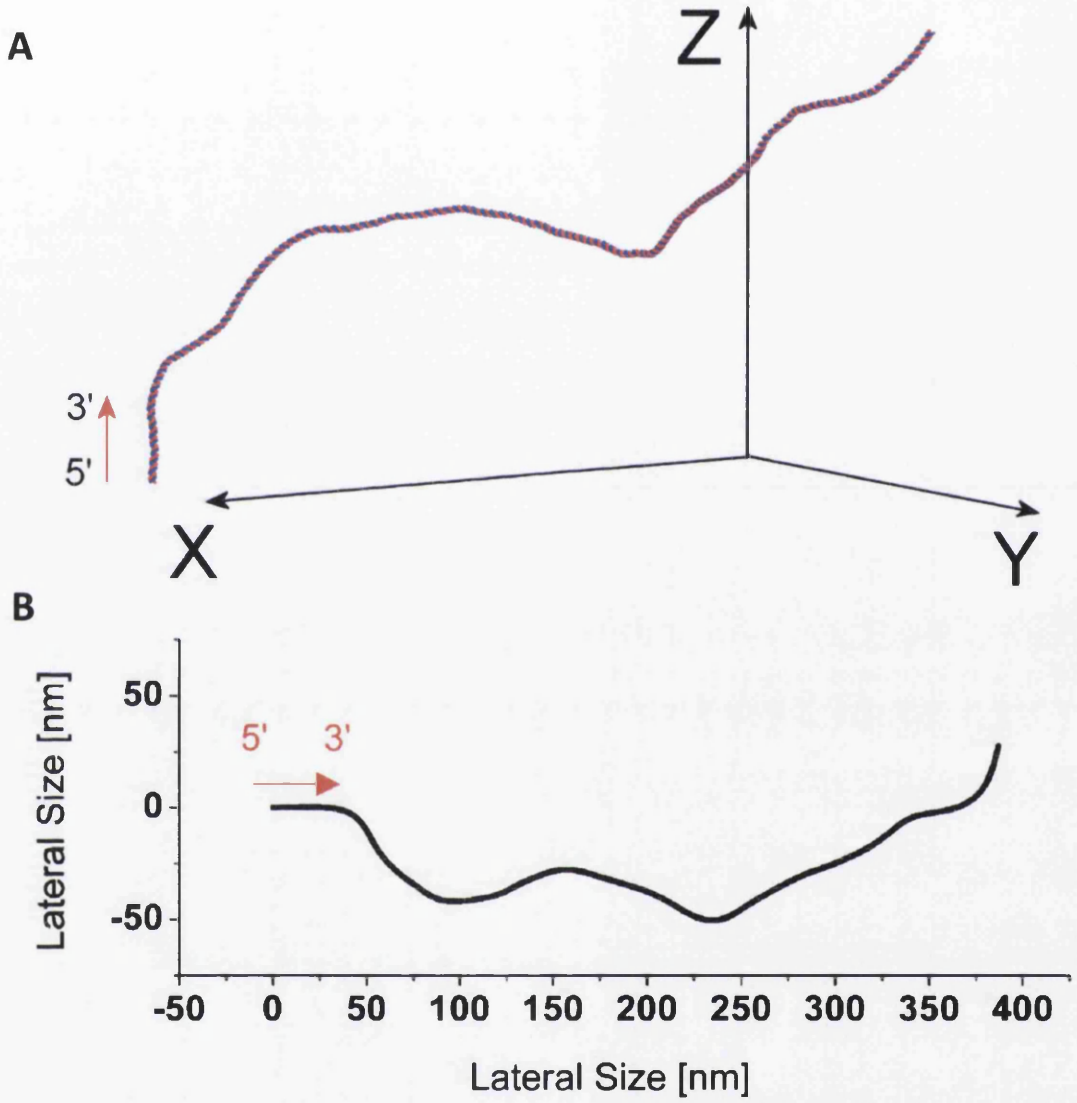

Figure 1.10. - Examples of a 3D DNA molecule projected onto two dimensions by Geometric Deposition. (A) 3D representation of the human osteopontin gene. (B) 2D projection of the human osteopontin gene. The figure was adapted from Buzio *et al.*, 2012.

1.9 *TP53* - The Tumour Protein 53 Gene

The gene of interest in this study is Tumour Protein 53 (*TP53)*. The protein encoded by *TP53*, named protein 53 (p53) due to its apparent molecular mass of 53 kilodalton on a sodium dodecyl sulfate polyacrylamide gel, was discovered in 1979 (Lane and Crawford, 1979). It wasn't until 1991 that p53 was identified as a tumour suppressor gene (Levine *et al.*, 1991). It is an important regulator of the cell cycle and has a key role in regulating cellular responses to genotoxic insults by its influence on programmed cell death, DNA repair and synthesis, senescence, transcription and genomic plasticity (Vogelstein and Kinzler, 1992).

1.9.1 The Role of p53 in the Cell

The protein encoded by *TP53* is part of a family of genes involved in regulating cellular stress alongside its paralogs p63 and p73 (Hollstein and Hainaut, 2010). The p53 protein is present at a constant low level within healthy cells and is upregulated under stress or DNA damage. The protein acts as a transcription factor as well as forming complexes with other regulatory proteins in the cell (Whibley *et al.*, 2009). As a transcription factor, p53 protects the cell against tumour growth and carcinogenesis by binding to response elements in a host of key genes. The genes regulated by p53 form the front line of defence against cellular stress and genotoxic insult and include genes that control cell cycle arrest, maintenance of genetic integrity, inhibition of angiogenesis, cellular senescence and apoptosis. The p53 protein also protects the cell through roles other than that of a transcription factor, for example p53 translocates to the mitochondria on cues from death stimuli (Mihara *et al.*, 2003). This translocation leads to cellular apoptosis.

1.9.2 Structure of *TP53*

TP53 is located on chromosome 17 (17p13). It is composed of 11 exons (protein coding regions). There is a notably large intronic (non-coding) region between exon 1 and 2. There are a number of functional domains within the p53 protein itself (Figure 1.11.). The transactivation region (Exon 2-4) is involved in activating other genes as part of the response to cellular stress, the sequence-specific DNA-binding region (Exon 5-8) is the active site of the protein involved in the recognition of DNA motifs, the nuclear localization and oligomerisation regions (Exon 9-11) have roles in localising p53 and formation of the final functional p53 tetramer. *TP53* is heavily transcriptionally regulated and there are at least 10 identified isoforms of p53 due to a number of multiple splice sites within the gene (Hollstein and Hainaut, 2010).

The p53 protein arose early in evolutionary history and *TP53* has remarkable evolutionary conservation between species (Lane *et al.*, 2010). The p53 protein in Placozoans, the simplest of free living multi-cellular organisms containing only four types of cells, has the

same key features and role shared by p53 in humans (Figure 1.12.). This conservation of p53 highlights its importance to cellular processes. The most highly conserved regions of the gene, considered exon 5 to exon 8, are also the most mutated in sporadic somatic cancers and up to 95% of all mutations occur within these highly conserved regions (Hollstein *et al.*, 1991). The exons that lie in the most highly conserved region of *TP53* code for the DNA sequence-specific binding domain of p53. This domain is required for the correct functioning of p53 as a transcription factor.

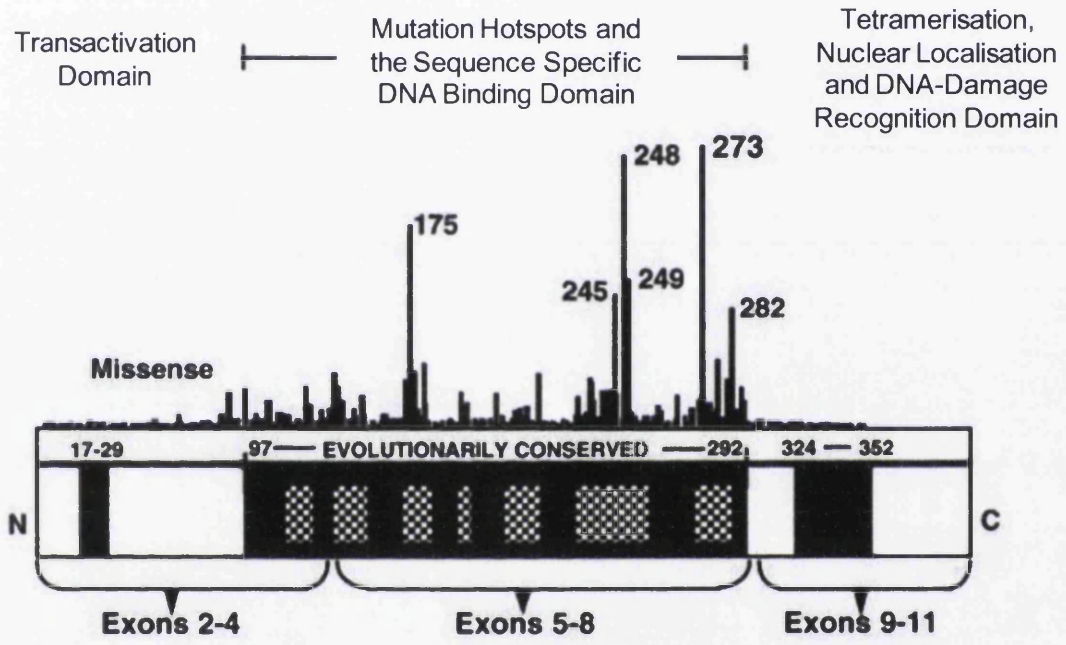

Figure 1.11. - Schematic of the p53 gene. The p53 protein consists of 393 amino acids with functional domains, evolutionarily conserved domains and regions designated as mutational hotspots. Functional domains include the transactivation region (amino acids 20–42), sequence-specific DNA-binding region (amino acids 100–293), nuclear localization sequence (amino acids 316–325), and oligomerisation region (amino acids 319–360). Evolutionarily conserved domains are indicated as black areas (amino acids 17–29, 97–292, and 324–352). Seven mutational hotspot regions within the large conserved domain are identified: amino acids 130–142, 151–164, 171–181, 193–200, 213–223, 234–258, and 270–286 (chequered blocks). Vertical lines above the schematic are missense mutations, the height of the bar represents the relative frequency of the mutations and locations of particularly prevalent mutation hotspots are labelled. The figure was adapted from Hussain and Harris, 1999.

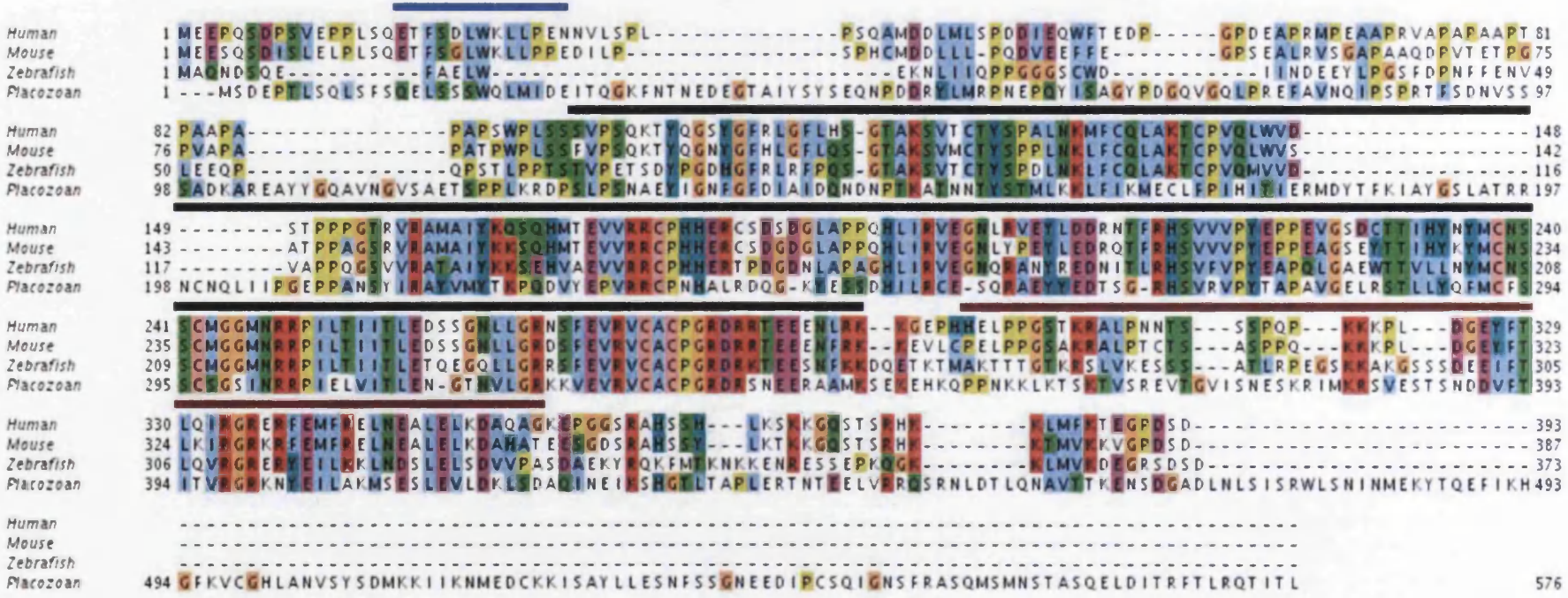

Figure 1.12. - Protein sequence alignment of the full-length p53 sequences from Human, Mouse, Zebrafish and the Placozoan *Trichoplax* genome. The bars above the sequence represent the various identified domains in human: blue denotes the N-terminal binding region, black denotes the DNA-binding domain, brown denotes the oligomerisation domain. The figure is adapted from Lane *et al.*, 2010.

1.9.3 Mutations in *TP53* and the Role of *TP53* in Carcinogenesis

TP53 has been the subject of intensive study as a model for understanding cell growth and cancer progression. *TP53* is very important for the growth and regulation of cells; mutations that cause incorrect functioning in p53 give insight into these biological processes. *TP53* is one of the most intensively studied cancer genes and there are online resources that collate the available literature on *TP53* mutation such as the International Agency for Research on Cancer (IARC) p53 database (Hernandez-Boussard *et al.*, 1999).

TP53 is mutated in 50 % of all human cancers (Greenblatt *et al.*, 1994). This percentage varies with cancer type, for example mutation frequency has been estimated to be approximately 60 % in lung cancer and 50 % in skin cancer (Biesalski *et al.*, 1998; Rigel, 2008). The acquisition of *TP53* mutations is a multi-stage process, where mutations can be picked up at an early or late stage in carcinogenesis. Tumours that contain *TP53* mutations have been shown to be more aggressive, in general, than those not carrying *TP53* mutation (Harris and Hollstein, 1993).

The etiologies of multiple types of cancer are specific to both the tissue type and mutagen involved in the initiation of carcinogenesis. There are specific patterns of somatic mutation hotspots that arise during cancer progression. Codons 175, 248 and 273 are the most frequently mutated hotspots in many cancers with the exception of lung, skin, larynx, bladder and liver carcinomas (Petitjean *et al.*, 2007). Many mutation hotspots have been linked to specific mutagens or are selected during carcinogenesis due to pro-carcinogenic properties. Many cancer types are associated with specific pathogens for the initiation of carcinogenesis. Chemical carcinogens from cigarette smoke are implicated in about 90 % of all lung cancers (Biesalski *et al.*, 1998). The mutation hotspots in codons 157, 158, 175, 245, 248, and 273 have been linked to chemical carcinogens such as benzo[a]pyrene diol epoxide (BPDE) in cigarette smoke (Pfeifer *et al.*, 2002). Approximately 90 % of skin cancers are caused by exposure to ultraviolet (UV) radiation (Rigel, 2008). Hotspots at codons 151, 177, 196, 245, 248, 278, 286 and 294 are considered to be caused by UV light (Drouin and Therrien, 1997). The patterns of mutations that arise in *TP53* are as varied as the cancers themselves. The vast majority of mutation hotspots occur in the exons within conserved regions that code for the sequence-specific DNA-binding domain of the p53 protein.

The downstream effects of mutations are varied because *TP53* regulates and is regulated by a great many genes. Missense mutations in exons 5-8 often prolong the half-life of the mutant protein (Pfeifer *et al.*, 2002). Mutations can also cause p53 to gain new functions leading to oncogenic properties (Petitjean *et al.*, 2007). For example mutations in codon 175 that lead to conversion of argenine to histidine always cause p53 to gain oncogenic functions.

Environmentally caused mutations within *TP53* are not the only causative factor for impaired p53 function. Continued infection with human papillomas virus can lead to cervical cancer by inhibiting p53 function (Klug *et al.*, 2001). Germ-line mutations in *TP53* can cause a high predisposition in individuals to the occurrence of multiple types of cancer at a young age, such as in Li–Fraumeni syndrome (Hisada *et al.*, 1998). Common polymorphisms localised to different parts of the world can also predispose individuals or whole populations to certain cancer types (Olivier *et al.*, 2010).

1.9.4 Regions of Slow DNA Repair in *TP53*

DNA repair mechanisms have been shown to have heterogeneous repair efficiency throughout the human genome (Bohr, 1987). For example, actively transcribed genes are preferentially repaired over other parts of the genome by transcription coupled repair mechanisms (Bohr, 1987; Surrallés *et al.*, 2002). This has been observed to consistently hold true for *TP53* (Denissenko *et al.*, 1998). Of particular interest is that DNA repair speed in *TP53* is sequence-specific and mutation hotspots within *TP53* genes are also regions of slow repair (Tornaletti and Pfeifer, 1994; Denissenko *et al.*, 1998; Zhu, 2000). Regions of slow repair were observed in a number of cancer specific hotspots within these studies, such as codons 157, 248 and 273 which are mutation hotspots in lung cancer (Denissenko *et al.*, 1998) and codons 177, 196 and 278 which are mutation hotspots in skin cancer (Tornaletti and Pfeifer, 1994). The mechanism underlying preferential sequence-specific repair is little understood and has been attributed to the accessibility of the DNA due to the local chromatin structure (Bohr, 1987). As nucleosome affinity is largely attributable to DNA sequence-specific curvature and flexibility this certainly provides grounds for hypothesising that there is a role for DNA curvature in the recognition and repair of DNA damage.

1.10 Aims and Objectives

The research described in this thesis aims to evaluate the intrinsic curvature of the region of *TP53* that codes for the sequence-specific DNA-binding domain of the p53 protein. This region contains exons 5, 6, 7 and 8 all of which are commonly mutated during carcinogenesis. This region is critical for the correct functioning of the p53 protein, which in turn regulates the main cellular defences against chemical insults and protects against tumourogenesis. Understanding the processes of DNA mutation and repair in *TP53* is of paramount importance to efforts to understand the cause and progression of cancer. The highly conserved region of *TP53* contains regions of slow repair which are currently poorly understood. Intrinsic DNA curvature has previously been identified as a factor involved in a number of biological processes, such as: DNA repair, nucleosome positioning and DNA transcription. Due to this involvement there was reason to believe that DNA curvature could influence the activity of DNA repair proteins at sites of slow DNA repair. The initial hypothesis was that the intrinsic DNA curvature located at, or flanking, regions of *TP53* that contain mutation hotspots or regions of slow repair would exhibit different curvature patterns to other regions.

In order to achieve these aims it was necessary to accurately measure the macromolecular conformation of *TP53*. AFM was selected as a suitable primary method of investigation alongside well established theoretical models of DNA curvature. *A TP53* DNA sequence that contained exons 5, 6, 7, 8 and 9 was identified. The polymerase chain reaction (PCR) was applied to generate experimental DNA molecules (Chapter 2). Two methods of DNA orientation were identified from the literature as being applicable to the large DNA sequence under investigation; the FF algorithm and end-labelling of *TP53* with streptavidin. These were applied to two overlapping PCR products of the *TP53* DNA sequence of interest in order to evaluate the reproducibility of intrinsic curvature measurements by AFM. A binding buffer of magnesium chloride ($MgCl_2$) was identified from the literature as providing suitably weak binding to a mica surface for intrinsic curvature measurements of DNA to be possible.

Due to the lack of software for the analysis of AFM images the first objective was to create software with the capability of processing AFM images of DNA to representative binary DNA contours. Additionally, experimental considerations lacking from the literature such as selection of interpolatory techniques and choice base pair intervals over which to calculate curvature angles were considered. To this end, a number of tools for image processing and DNA analysis were developed and encompassed within a user interface for ease of use (Chapter 3).

The second objective was to explore the available theoretical models of DNA curvature applicable to *TP53* sequences. The need for generating computer simulated AFM images of *TP53* for comparison to real AFM images was identified. These simulated AFM images allowed for statistical and hypothetical testing on ideal images of *TP53* DNA before application to real AFM images. The relationships between exons, regions of slow repair and DNA curvature were investigated using dinucleotide wedge models of DNA curvature. This approach allowed for the additional hypothesis that exon positions within *TP53* exhibit significantly lower curvature when compared to intron positions (Chapter 4).

The third objective was to evaluate two methods of DNA molecule orientation for application to *TP53*. The first methodology applied was the only protein-label free technique applicable to large *TP53* DNA molecules, the FF algorithm. Application of the FF algorithm avoided any possible interaction between DNA and protein end-labels. The algorithm was initially tested on computer simulated AFM images of *TP53*. It was then applied to AFM images of real *TP53* DNA (Chapter 5). The second orientation methodology used streptavidin end-labels attached to biotinylated PCR products of *TP53* for orientation (Chapter 6). The resulting curvature profiles were compared to theoretical predictions and statistically analysed.

CHAPTER 2: GENERAL MATERIALS AND METHODS

2.1 Design and Preparation of an Experimental DNA Template for *TP53*

2.1.1 *TP53* Sequence and PCR Primer Design

The *TP53* sequences used for this study were taken from the consensus sequence presented in the IARC *TP53* database, a compilation of *TP53* sequences taken from human population studies (Hernandez-Boussard *et al.*, 1999). Two sequences were investigated: a 1855 bp sequence covering exons 5-7 (11828 to 13682 in IARC Database notation) and a 2500 bp molecule covering exons 5-9 (11828 to 14328 in IARC Database notation). Both sequences had the same start point and the 2500 bp sequence fully overlapped the 1855 bp sequence. The full sequence is presented in Figure 2.1.

The sequences were designed with two key objectives in mind. Firstly, to assess the curvature within the section of DNA coding for the DNA sequence-specific binding region (Hollstein *et al.*, 1991). This region was of particular interest as 95 % of *TP53* mutations have been observed to occur within this DNA tract. The nature of the overlapping sections also allowed for the evaluation of the inter-experiment variation in AFM measurements of curvature. The PCR product containing exons 5 through to 7 (1855 bp) was named 'Exon 5-7' in the main text. The PCR product containing exons 5 through to 9 (2500 bp) was named 'Exon 5-9' in the main text. Capitalisation of 'Exon' within the main text indicates a reference to one of these experimental DNA sequences or molecules.

The oligonucleotide PCR primers for these two DNA sequences are presented below:

TP53 e5-7/9F	=	CATCTCTCCTGGGGATGCA
TP53 e5-7R (1855 bp)	=	TCTACTCCCAACCACCCTTG
(Reverse Complement)	=	CAAGGGTGGTTGGGAGTAGA
TP53 e5-9R (2500 bp)	=	CAGGCAAAGTCATAGAACCA
(Reverse Complement)	=	TGGTTCTATGACTTTGCCTG

```
   1 CATCTCTCCT GGGGATGCAG AACTTTTCTT TTTCTTCATC CACGTGTATT CCTTGGCTTT
  61 TGAAAATAAG CTCCTGACCA GGCTTGGTGG CTCACACCTG CAATCCCAGC ACTCTCAAAG
 121 AGGCCAAGGC AGGCAGATCA CCTGAGCCCA GGAGTTCAAG ACCAGCCTGG GTAACATGAT
 181 GAAACCTCGT CTCTACAAAA AAATACAAAA AATTAGCCAG GCATGGTGGT GCACACCTAT
 241 AGTCCCAGCC ACTTAGGAGG CTGAGGTGGG AAGATCACTT GAGGCCAGGA GATGGAGGCT
 301 GCAGTGAGCT GTGATCACAC CACTGTGCTC CAGCCTGAGT GACAGAGCAA GACCCTATCT
 361 CAAAAAAAAA AAAAAAAAAG AAAAGCTCCT GAGGTGTAGA CGCCAACTCT CTCTAGCTCG
 421 CTAGTGGGTT GCAGGAGGTG CTTACGCATG TTTGTTTCTT TGCTGCCGTC TTCCAGTTGC
 481 TTTATCTGTT CACTTGTGCC CTGACTTTCA ACTCTGTCTC CTTCCTCTTC CTACAGTACT
 541 CCCCTGCCCT CAACAAGATG TTTTGCCAAC TGGCCAAGAC CTGCCCTGTG CAGCTGTGGG
 601 TTGATTCCAC ACCCCCGCCC GGCACCCGCG TCCGCGCCAT GGCCATCTAC AAGCAGTCAC
 661 AGCACATGAC GGAGGTTGTG AGGCGCTGCC CCCACCATGA GCGCTGCTCA GATAGCGATG
 721 GTGAGCAGCT GGGGCTGGAG AGACGACAGG GCTGGTTGCC CAGGGTCCCC AGGCCTCTGA
 781 TTCCTCACTG ATTGCTCTTA GGTCTGGCCC CTCCTCAGCA TCTTATCCGA GTGGAAGGAA
 841 ATTTGCGTGT GGAGTATTTG GATGACAGAA ACACTTTTCG ACATAGTGTG GTGGTGCCCT
 901 ATGAGCCGCC TGAGGTCTGG TTTGCAACTG GGGTCTCTGG GAGGAGGGGT TAAGGGTGGT
 961 TGTCAGTGGC CCTCCAGGTG AGCAGTAGGG GGGCTTTCTC CTGCTGCTTA TTTGACCTCC
1021 CTATAACCCC ATGAGATGTG CAAAGTAAAT GGGTTTAACT ATTGCACAGT TGAAAAAACT
1081 GAAGCTTACA GAGGCTAAGG GCCTCCCCTG CTTGGCTGGG CGCAGTGGCT CATGCCTGTA
1141 ATCCCAGCAC TTTGGGAGGC CAAGGCAGGC GGATCACGAG GTTGGGAGAT CGAGACCATC
1201 CTGGCTAACG GTGAAACCCC GTCTCTACTG AAAAATACAA AAAAAAATTA GCCGGGCGTG
1261 GTGCTGGGCA CCTGTAGTCC CAGCTACTCG GGAGGCTGAG GAAGGAGAAT GGCGTGAACC
1321 TGGGCGGTGG AGCTTGCAGT GAGCTGAGAT CACGCCACTG CACTCCAGCC TGGGCGACAG
1381 AGCGAGATTC CATCTCAAAA AAAAAAAAAA AAGGCCTCCC CTGCTTGCCA CAGGTCTCCC
1441 CAAGGCGCAC TGGCCTCATC TTGGGCCTGT GTTATCTCCT AGGTTGGCTC TGACTGTACC
1501 ACCATCCACT ACAACTACAT GTGTAACAGT TCCTGCATGG GCGGCATGAA CCGGAGGCCC
1561 ATCCTCACCA TCATCACACT GGAAGACTCC AGGTCAGGAG CCACTTGCCA CCCTGCACAC
1621 TGGCCTGCTG TGCCCCAGCC TCTGCTTGCC TCTGACCCCT GGGCCCACCT CTTACCGATT
1681 TCTTCCATAC TACTACCCAT CCACCTCTCA TCACATCCCC GGCGGGGAAT CTCCTTACTG
1741 CTCCCACTCA GTTTTCTTTT CTCTGGCTTT GGGACCTCTT AACCTGTGGC TTCTCCTCCA
1801 CCTACCTGGA GCTGGAGCTT AGGCTCCAGA AAGGACAAGG GTGGTTGGGA GTAGATGGAG
1861 CCTGGTTTTT TAAATGGGAC AGGTAGGACC TGATTTCCTT ACTGCCTCTT GCTTCTCTTT
1921 TCCTATCCTG AGTAGTGGTA ATCTACTGGG ACGGAACAGC TTTGAGGTGC GTGTTTGTGC
1981 CTGTCCTGGG AGAGACCGGC GCACAGAGGA AGAGAATCTC CGCAAGAAAG GGGAGCCTCA
2041 CCACGAGCTG CCCCCAGGGA GCACTAAGCG AGGTAAGCAA GCAGGACAAG AAGCGGTGGA
2101 GGAGACCAAG GGTGCAGTTA TGCCTCAGAT TCACTTTTAT CACCTTTCCT TGCCTCTTTC
2161 CTAGCACTGC CCAACAACAC CAGCTCCTCT CCCCAGCCAA AGAAGAAACC ACTGGATGGA
2221 GAATATTTCA CCCTTCAGGT ACTAAGTCTT GGGACCTCTT ATCAAGTGGA AAGTTTCCAG
2281 TCTAACACTC AAAATGCCGT TTTCTTCTTG ACTGTTTTAC CTGCAATTGG GGCATTTGCC
2341 ATCAGGGGGC AGTGATGCCT CAAAGACAAT GGCTCCTGGT TGTAGCTAAC TAACTTCAGA
2401 ACACCAACTT ATACCATAAT ATATATTTTA AAGGACCAGA CCAGCTTTCA AAAAGAAAAT
2461 TGTTAAAGAG AGCATGAAAA TGGTTCTATG ACTTTGCCTG
```

Exon 5

Exon 6

Exon 7

Exon 8

Exon 9

Figure 2.1. – *TP53* consensus sequence from the IARC Database. Exon positions are indicated in red. The shared forward primer is indicated in yellow. The reverse Exon 5-7 primer is indicated in blue and the reverse Exon 5-9 primer in green. Base pair numbering begins at the start of the experimental DNA sequence.

2.1.2 PCR of Template *TP53* DNA

PCR amplifications were prepared in 50 µl of ddH_20. The reaction mixture included 10 µl of reaction buffer (Promega, Cat.# M791A) for a final concentration of 1.5 mM $MgCl_2$, 0.2 mM of each nucleotide (Promega, Cat.# C1141), 0.4 µM of both upstream and downstream primers, 2.5 u of Expand High Fidelity^{PLUS} PCR System (Roche, Cat.# 04743725001), and less than 50 µg of template DNA. The template DNA was human genomic DNA (Promega, Cat.# G1471). Primers were purchased from Eurofins (Eurofins MWG Operon). The primers for relevant reactions contained a 5' biotin end-label (Chapter 6).

The initial reaction for all products was a PCR amplification using Exon 5-9 primers. A second amplification step was introduced for the generation of all experimental reactions. The nested amplification from the product of the initial reaction ensured the fidelity of the final product in the case of any non-specific amplification products.

2.1.3 Reaction Conditions

Optimised reaction conditions were as follows: hot start of 5 min at 95 °C, denaturation step of 95 °C for 30 s, annealing step of 60 °C for 30 s, extension step of 72 °C for 60 s / 90 s (Exon 5-7 / Exon 5-9), the previous three steps are repeated 35 times. A final extension step of 72 °C for 5 min and a hot stop step of 95 °C for 5 min were introduced to prevent the dimerisation of primers and amplification products. Primer optimisation has been presented in Appendix 1.

2.1.4 Agarose Gel Electrophoresis

PCR products were checked on a 1 % agarose gel stained with EtBr imaged under UV. Agarose gels were run for an appropriate time in order to resolve and distinguish ladder and sample bands (30-45 mins, 100 V). PCR products were purified using a QIAquick PCR Purification Kit (QIAGEN, Cat # 28104) and re-eluted in double distilled H_20 or imaging buffer for storage at 4 °C/20 °C. Products were rechecked after purification on a 1% agarose gel. Product quantity and quality was established using a spectrophotometer (NanoDrop/NanoDrop Lite, Thermo Scientific). The final amplification product after spin column purification was sequenced (Source BioScience, Nottingham, UK). This allowed for any deviation from the theoretical DNA sequence to be identified (Appendix 1.).

2.2 Streptavidin End-Labelling of Biotinylated DNA

2.2.1 Dot Blot Analysis of 5' Biotinylated DNA

It was necessary to ensure that the biotin end-label remained viable after PCR amplification and purification. This was performed using a dot blotting analysis. The dot blot apparatus is prewashed thoroughly with 1% SDS and then with sterile water prior to use. Hybond-N nylon membrane and Whatman 3MM filter paper was cut to cover the required number of wells. Both the nylon membrane and the filter paper were pre-soaked in 2X SSC buffer (30 mM sodium citrate, pH 7.0, 0.3 M, NaCl) for 5 minutes. DNA sample of between 100 ng and 250 ng of DNA were diluted to 100 µl with sterile ddH_2O and boiled for 10 minutes. Samples were quickly chilled in an ice-water bath. An equal volume of freshly made 1 M NaOH was added and incubated at room temperature for 20 minutes.

DNA solutions were applied to the apparatus according to the manufacturer's instructions and allowed to incubate with the membrane at room temperature for 10 minutes. The solution was drawn through the apparatus under vacuum. Membrane were incubated in 100 ml of neutralizing solution (1 mM EDTA, 1.5 M NaCl, 0.5 M Tris, pH 7.2) for 30 minutes. Membranes were rinsed thoroughly with 2X SSC and air-dried before the addition of hybridisation probes. Streptavidin-horseradish peroxidase mix (Vector Laboratories, Cat.# SA-5704) was diluted 1/3000 with 1 x SCC and added to the membranes and allowed to incubate overnight. SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific, Cat.# 34079) was added and the membranes were imaged using a Gel Doc EQ System (Bio-Rad, Hercules).

2.2.2 5' End-Labelled of DNA with Streptavidin for AFM Imaging.

DNA was purified using a QIAquick PCR Purification Kit (QIAGEN, Cat # 28104) to remove biotinylated primers that could compete with 5' biotin for streptavidin end label. The purification step was repeated when necessary. The DNA concentration was diluted to 10 x the concentration used in AFM imaging. A 3:1/2:1 molar ratio of streptavidin to sample DNA (*Sigma-Aldrich* UK, Cat # S4762) was incubated overnight at 4°C before AFM imaging. Labelled and control unlabelled samples were run on a 1% agarose gel (Section 2.1.4.) in order to identify end labelling before AFM imaging.

2.3 Preparation of DNA for AFM Imaging

DNA was diluted down to an appropriate concentration for imaging (~ 1 ng/µl) in the binding buffer. Binding buffer was Tris 10 mM, $MgCl_2$ 10 mM, NaCl 5 mM pH 7.5. The Mg^{2+} buffer was used for AFM imaging as it has been previously reported by authors as providing good imaging conditions for relaxed DNA conformations at various concentrations of EtBr treatment (Coury *et al.*, 1996; Pope *et al.*, 2000). The buffer has been observed to provide a weak bind to mica surfaces that was appropriate for AFM analysis of DNA curvature (Rivetti *et al.*, 1996; Scipioni *et al.*, 2002a). DNA was applied to freshly cleaved muscovite mica and allowed to incubate at room temperature for 1-3 min before washing with ddH_20 (Millipore). The mica surface was dried under vacuum before AFM imaging.

2.4 AFM Imaging Conditions

All AFM imaging was performed on a NanoWizard 2 BioScience AFM using closed loop settings (JPK, Instruments, Berlin, Germany). The instrument operated in intermittent contact mode to minimise the possible damage caused by tip-sample interactions. The cantilever of choice was an ACTA probe (AppNano, Santa Clara, USA) with a spring constant of between 25-75 N/m (nominal 40 N/m) and a ~6 nm radius of curvature (ROC).

Images were collected in a 3x3 µm square with a pixel resolution of 1024x1024. This gives a width of 5.86/2.93 nm per pixel respectively, for both image resolutions. The proportional and integral gains and scan frequencies (typically between 0.8 – 2.0 Hz per line) were optimised for each tip and image set. Large amounts of images were collected for each sample using the Experiment Planner software (JPK, Instruments, Berlin, Germany) and in-house code. Each image was offset from the previous image by the width of the image (3 µm) in either the *x* or *y* plane. The thermal drift of the scanner head was not found to lead to the collection of duplicate molecules over long experiments. Example AFM images have been presented in Appendix 2.

2.5 Generating Computer Simulated AFM Images of *TP53*

Computer simulated AFM images were created using the method detailed by Buzio *et al.*, 2012. The De Santis dinucleotide wedge model was used to create all simulated AFM images for comparison to experimental AFM images unless stated otherwise within the text (De Santis *et al.*, 1988). The simulated images were generated using a persistence length of 53 nm, Gaussian noise with a variance of 0.025 and tip convolution by a simulated tip of 6nm ROC. The images were created to be comparable to experimental DNA sequences detailed in Section 2.1.1. Simulated AFM images were processed using the image processing software detailed in Chapter 3. The full method has been detailed below.

2.5.1 3D Models of *TP53* using w3DNA

Two different parameter sets that have been previously validated for AFM measurements of DNA were selected as theoretical values; those put forward by De Santis *et al.* 1988 and Olson *et al.* 1998. These dinucleotide models will be referred to in this text by the names of their first authors (De Santis *et al.*, 1988; Olson *et al.*, 1998).

3DNA allows for the visualisation, analysis and reconstruction of DNA *in silico* (Lu and Olson, 2008). The web interface for the application, w3DNA, was used to reconstruct 3D models of *TP53* DNA using a predefined set of dinucleotide parameters detailed in Table 2.1. Non-applicable parameters (*i.e.* shift and slide for the De Santis model) were set to 0. The De Santis model of curvature contained only base pair transitions so a constant base pair rise value of 0.34 Å was selected as a consensus value from the literature (Saenger, 1984).

Olson Dinucleotide Parameters (Olson *et al.*, 1998)

Dinucleotide Step	*Twist, deg*	*Tilt, deg*	*Roll, deg*	*Shift, Å*	*Slide, Å*	*Rise, Å*
CG	36.10	0.00	5.40	0.00	0.41	3.39
CA	37.30	0.50	4.70	0.09	0.53	3.33
TA	37.80	0.00	3.30	0.00	0.05	3.42
AG	31.90	-1.70	4.50	0.09	-0.25	3.34
GG	32.90	-0.10	3.60	0.05	-0.22	3.42
AA	35.10	-1.40	0.70	-0.03	-0.08	3.27
GA	36.30	-1.50	1.90	-0.28	0.09	3.37
AT	29.30	0.00	1.10	0.00	-0.59	3.31
AC	31.50	-0.10	0.70	0.13	-0.58	3.36
GC	33.60	0.00	0.30	0.00	-0.38	3.40

De Santis Dinucleotide Parameters (Scipioni *et al.*, 2002a)

Dinucleotide Step	*Twist, deg*	*Tilt, deg*	*Roll, deg*
CG	33.50	0.00	4.60
CA	34.10	0.40	6.80
TA	34.50	0.00	8.00
AG	34.40	-1.60	1.00
GG	33.10	-0.60	1.30
AA	36.00	-0.50	-5.40
GA	34.60	-1.70	2.00
AT	35.30	0.00	-7.30
AC	33.70	-2.70	-2.50
GC	33.30	0.00	-3.70

Table 2.1. - Dinucleotide parameters used for the generation of 3D models of *TP53*. A value of 0.34 Å value was used for the De Santis model as a generally accepted rise for B-form DNA. Atomic coordinates for each base pair were averaged to give an approximation of the centre of the DNA strand.

2.5.2 Simulated Deposition of DNA on a 2D Surface – Geometric Deposition

In order to extrapolate a simplistic simulation of the deposition of DNA onto a 2D surface it is necessary to fit a series of best fit (least squares) planes (Buzio *et al.*, 2012). Constraints were placed upon the plane fitting allowing no local fluctuations in either the *x, y* or *z* directions that exceed 2 nm from the plane. This was accomplished by fitting a plane to each *xyz* coordinate sequentially beginning with the 5' (Exon 5) end of the sequence. Orthogonal regression using principal components analysis was used to find the plane of best fit (Scholkopf *et al.*, 2005). If the plane had local variation (the local error of the fit) of less than 2 nm (20 Å) then another *xyz* coordinate was added and the best fit plane recalculated. This was iterated upon until the local error exceeded 2 nm. The next plane was then fitted to the succeeding series of *xyz* coordinates in the same manner. An additional constraint was added; if there were less than 50 bp left of the sequence when the local error was greater than 2 nm then no more planes were fitted and the remaining base pairs were included when calculating the final plane.

Calculating the angle in radians of the intersection of the planes was possible using the formula:

$$\theta_{intercept} = \arccos\left(dot(N1, N2)/(norm(N1) * norm(N2)\right)$$

Where N1 was the coefficients of the normal vector of the preceding plane (Plane 1) and N2 was the coefficients of the succeeding plane (Plane 2). By rotating the *xyz* coordinates of Plane 2 along the axis of the line of intersection of both planes by the inverse of the angle of intersection we bring them into line with Plane 1 (Figure 2.2.). The rotation was applied to all coordinates following Plane 2. The plane normals for each section was recalculated after each rotation. The *xyz* coordinates were projected onto a flat *xy* plane. The projection was the fit of orthogonal regression.

Finally, a local correction was applied at the point of intersection between two planes. A linear line was fitted to the two helical turns preceding and succeeding the point of intersection. The intersection angle of the two lines was calculated and *xy* coordinates succeeding the point of intersection are rotated by the inverse of the angle. The correction was introduced between the *xyz* points of each succeeding plane due to the observation that a large angle being intermittently introduced at this intersection. As DNA was unlikely to adopt a kinked structure except under chemical or physical stress and as deposition conditions assumed that the DNA equilibrates on the surface this should result in minimal physical stress. As the true angle at the point of intersection between two planes was unknown then a constant angle of 0^{o} was chosen as a suitable compromise. The suitability of this local

correction from this method was discussed with the original authors in a private communication (Buzio *et al.*, 2012).

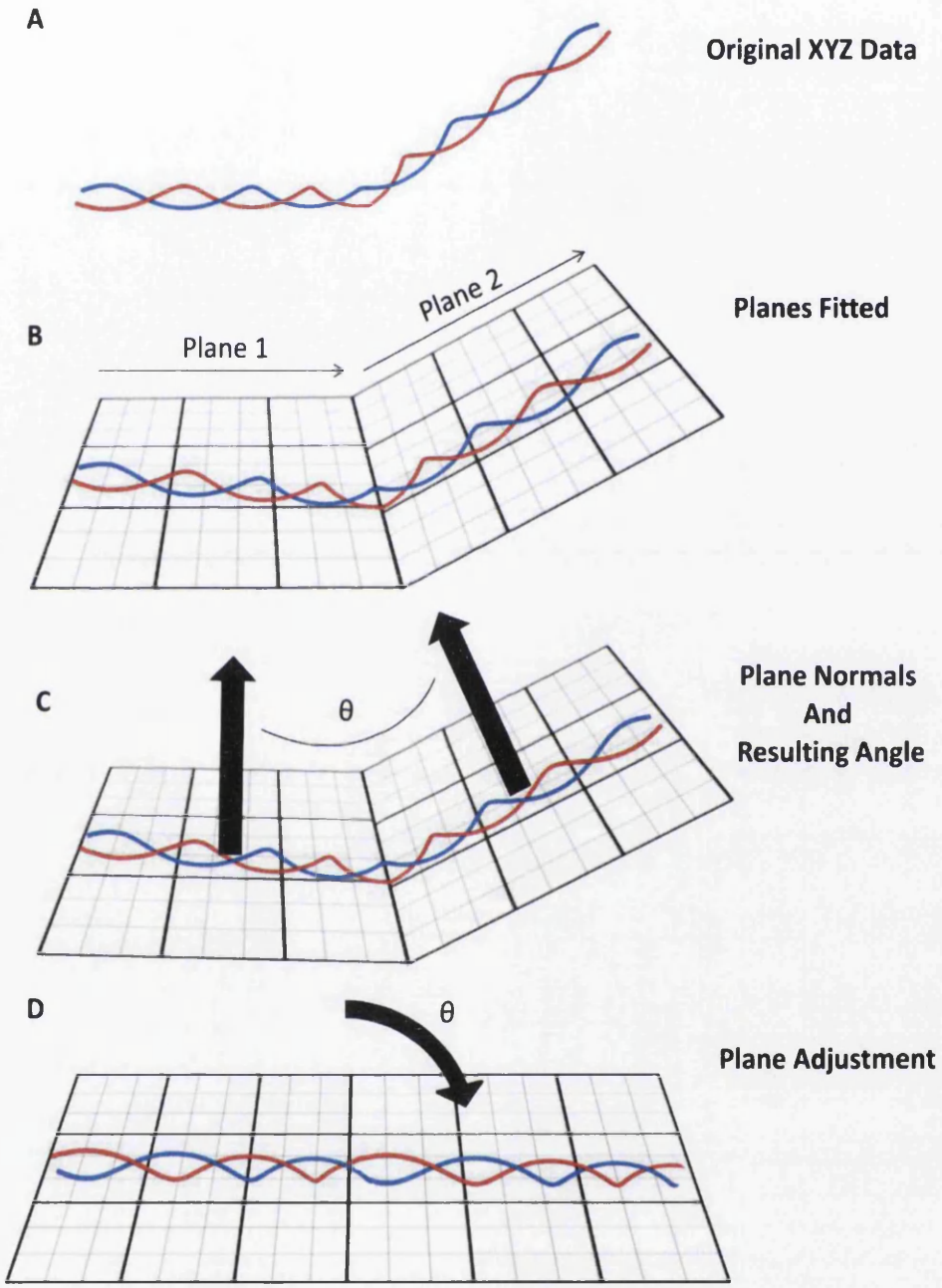

Figure 2.2. - Simple representation of the Geometric Deposition method. A) Original *xyz* coordinates for a DNA molecule. B) Two planes are fitted to the *xyz* coordinates (least squares with a maximum error of 2 nm per point). C) The angle between the plane normal vector is calculated. D) The inverse of calculated angle was used as angle of rotation along the axis of the line of interception between the two planes.

2.5.3 Simple 2D DNA Chains

The angle between each base pair *xy* coordinate from the flattened 3D model was calculated by treating coordinates as a series of vectors. The angles in radians between subsequent vectors were calculated using the formulae:

$$perpendicular\ dot\ product = -a(2) \times b(1) + a(1) \times b(2)$$
$$dot = a(1) \times b(1) + a(2) \times b(2)$$
$$\theta = arctan(perpendicular\ dot\ product, dot)$$

Where vector *a* was the product of [x(i-1) y(i-1)]-[x(i) y(i)] and vector *b* was [x(i) y(i)]-[x(i+1) y(i+1)]. The subsequent series of signed curvature angles had an angle value per base pair. Clockwise angles were denoted as positive and counter-clockwise as negative. A curvature angle was calculated for each base pair using the curvature profile as the mean of a normal Gaussian probability distribution with a standard deviation of $\sigma = \sqrt{l/\xi}$ where *l* was the length of the section of DNA, in this case 0.34 nm, ξ was the persistence length of DNA, in this case 53 nm (Rivetti and Codeluppi, 2001). A random start point was determined within a grid of user defined size in nanometres (*e.g.* 3000 nm by 3000 nm). A random trajectory was generated for the first DNA chain. Subsequent points were plotted using the formulae:

$$x(i) = x(i-1) + (l \times cos(cum\theta(i)))$$
$$y(i) = y(i-1) + (l \times sin(cum\theta(i)))$$

Where $cum\theta$ was the cumulative angle along the line segment plus the curvature value for the dinucleotide at point *i* and l were the length of a base pair in nm (0.34 nm). The final image resolution was user defined and corresponded to AFM image resolution. The *xy* coordinates were converted to AFM image coordinates *i.e.* (*xy*/size in nm) multiplied by the resolution in pixels. Values that were outside of the user defined grid area were removed. Any pixels in the final image that contained a section of DNA were set to 1 (*i.e.* a binary image of a DNA molecule is created). Examples of raw *xy* coordinates for each base pair and digitised images have been included in Figure 2.3.A and Figure 2.3.B.

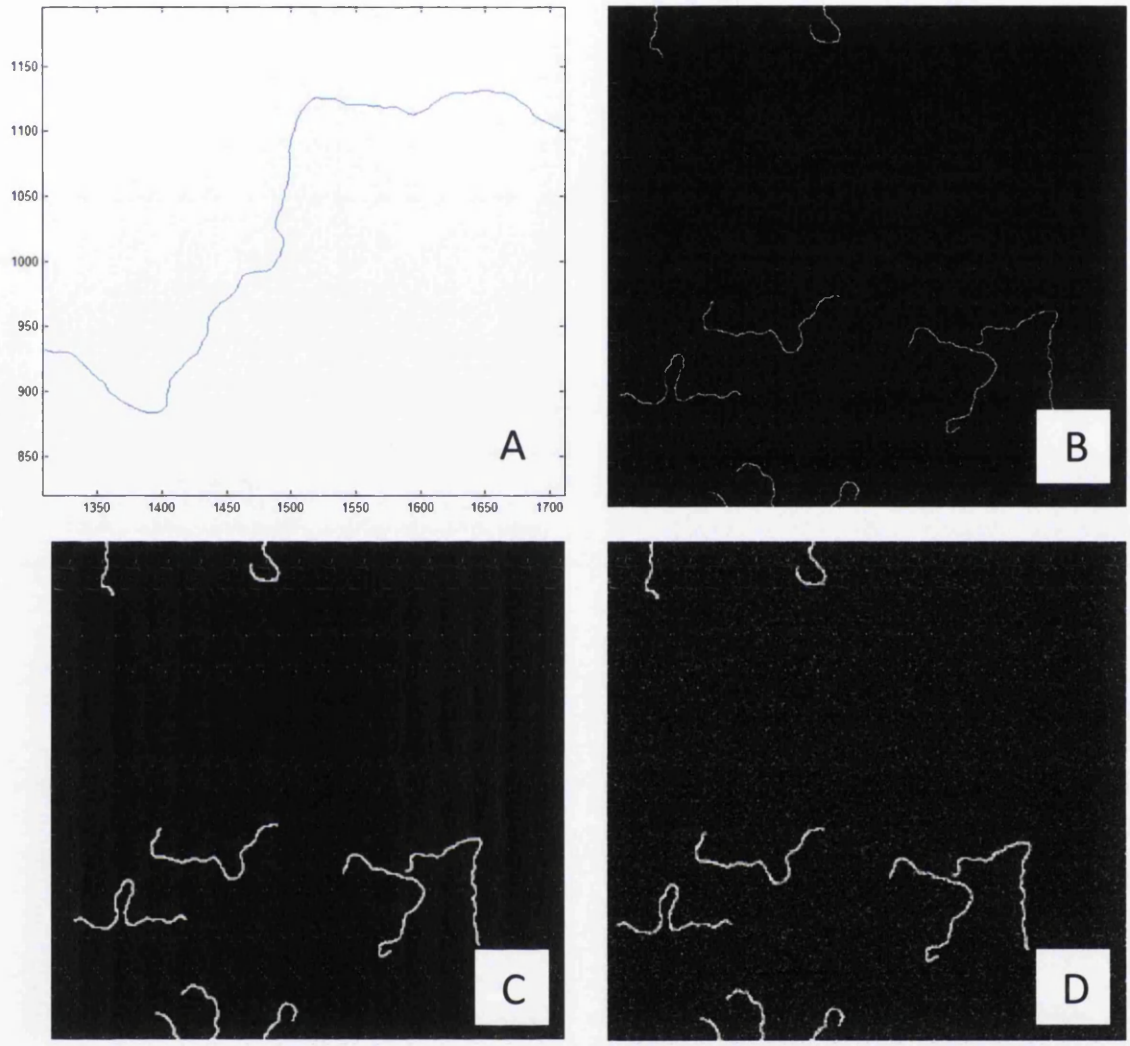

Figure 2.3. - Examples of a theoretical AFM images at each step in its production. A) *xy* coordinates for each base pair of a theoretical DNA molecule. B) Multiple molecules at the resolution of an AFM image. C) Theoretical AFM image after 3D tip convolution with a tip of 6 nm ROC. D) Final theoretical AFM image after the addition of Gaussian noise (variance = 0.025).

2.5.4 Tip Convolution

A 3D spherical function was passed over each binary image to emulate the effect of imaging DNA molecules with an AFM tip. The spherical function had a user defined radius (*r*) equal to that of the radius of curvature of an AFM tip. The function was evaluated over an *a* x *a* grid, where *a* is double the radius of the sphere in pixel resolution (*i.e.* all possible points on the image that the sphere can inhabit).

$$sphere = sqrt\ ((r\text{^}2) - (x(a_{min}{:}a_{max})\text{^}2) - (y(a_{min}{:}a_{max})\text{^}2))$$

This equation provided the Z-height of a sphere centred on a single pixel (*xy*). All values above 0 were removed to produce a half sphere. The radius of the half sphere was added to each value within the half sphere to produce a half sphere with a Z-height of zero at [*x y*]. The binary AFM image was padded with an appropriate number of zeros. The expected Z of the half sphere was then compared to the actual Z-height of the AFM image at each pixel. The discrepancy, if greater than zero, was the final recorded height at that pixel (Figure 2.4.). An example of an image after tip convolution has been included in Figure 2.3.C.

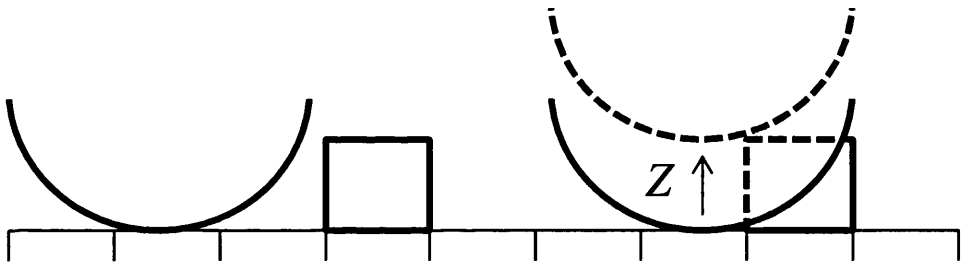

Figure 2.4. Representation of 3D spherical convolution of a binary image in 2D. A spherical function is passed over a binary image and evaluated at each pixel. If any values in the binary image are larger than the expectation of the spherical function then the value of the final image is increased by the difference (*Z*).

2.5.5 Finishing Theoretical AFM Images.

A user defined amount of Gaussian noise was added to the final images (variance = 0.025). The images were saved as uncompressed TIFF files. This file format was similar to that used by AFM manufacturer JPK. An example of a completed AFM image is presented in Figure 2.3.D.

2.5.6 Orientation of Molecules Post-Image Processing

The original *xy* coordinates of each theoretical molecule were stored in an image specific variable *i.e.* one *xy* coordinate for each base pair. The end points of each DNA molecule after image processing were compared to the first *xy* coordinate, corresponding to the 5' end of the DNA strand of the stored *xy* values. The correct endpoint was determined as having the lowest Euclidean distance between itself and the theoretical *xy* value (Figure 2.5.). The *xy* coordinates were aligned correctly, the initial *xy* coordinates being the 5' end of the molecule.

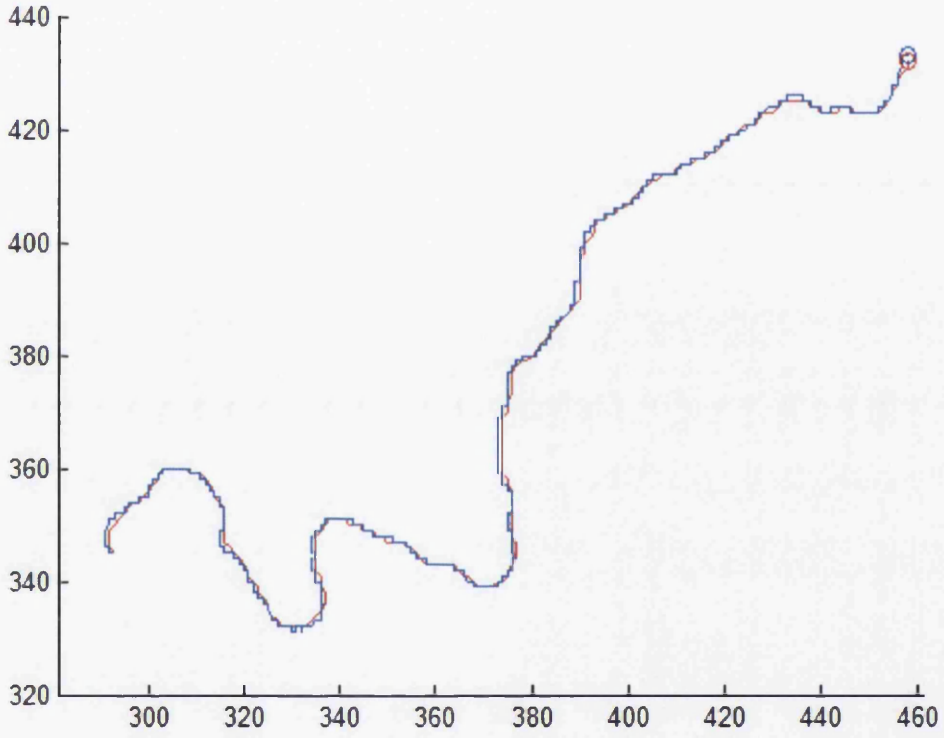

Figure 2.5. - Example of alignment of post-image processing DNA molecule (blue) to its theoretical predecessor (red). The circles (blue and red) are the 5' end of the DNA molecule. In this example both ends have been aligned.

2.6 Image Processing of AFM Images

Images were processed using the image processing software detailed in Chapter 3. All AFM images were plane fitted/flattened using a nine degree polynomial. Image processing was performed in a semi-automated manner for experimentally obtain AFM images. A median 3 x 3 filter was used to reduce noise. Other filters, such as a Gaussian 3 x 3 filter, were used where appropriate to extract the orientation of DNA contours from images with higher impulsive noise. A foreground threshold was visually identified and confirmation was provided by the user to ensure good fidelity of automated DNA identification to the DNA contour.

For simulated AFM image processing was performed in a fully automated manner. A median 3 x 3 filter was used to reduce noise. A threshold value was obtained automatically (Otsu, 1979). All DNA contours that lay within a range of ±200 nm of the theoretical size of the DNA molecule were recovered.

For appropriate experiments streptavidin end-labels were automatically identified. The Z-height of the first and last 3 pixels was compared. The end with the largest mean Z-height was designated as end-labelled with streptavidin. The presence of the streptavidin end-label was visually confirmed by the operator.

2.7 Statistical Analysis

All data processing and analysis was performed off-line using the Matlab R2007b commercial software package (MATLAB R2007b, The MathWorks Inc., Natick, MA, 2007). The normality of data was checked with a Shapiro-Wilk test (Shapiro and Wilk, 1965). The majority of the data was found to be non-normal and therefore non-parametric statistical tests were used, p-values lower than 0.05 were considered significant unless otherwise stated within the text. In order to compare intrinsic curvature or flexibility profiles the Spearman Rank Correlation coefficient was calculated (Spearman, 1904). In order to compare curvature values that occurred within exons regions to intron regions the sections of curvature profiles that corresponded to exons positions were identified from the IARC database (Hernandez-Boussard *et al.*, 1999). Exon values were compared using a Kruskal-Wallis test (Kruskal and Wallis, 1952).

2.7.1 Analysis of DNA Contours

DNA contour length was calculated using the Kulpa estimator (Kulpa, 1977). A comparable number of points were fitted to DNA contours by fitting a linear interpolant. Intrinsic DNA curvature and flexibility was calculated in the standard manner dictated from the literature (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b). The fragment flipping algorithm was instituted using the Greedy algorithm (Ficarra *et al.*, 2005b). A full analysis workflow with detailed explanations is presented in Chapter 3.

2.7.2 Curvature Peak Comparison

The largest peaks of curvature were identified within theoretical intrinsic DNA curvature profiles. The peaks that most closely corresponded to these key peaks from experimental curvature profiles were identified. The peak shift for each matching peak was calculated as a percentage value of the standardised length of the DNA sequences under investigation.

2.7.3 Visually Displaying Curvature Profiles

Two methods for calculating curvature profiles have been used within the main body of the text (Figure 2.6.). The first method is referred to within the text as unsigned curvature profiles. Unsigned curvature profiles consider curvature angle regardless of the direction of curvature on the mica surface. Unsigned curvature profiles were generated as the average of all absolute curvature angles within a dataset. This method is comparable to calculating the curvature modulus (or magnitude) and is also sometimes called 'absolute curvature' within the literature. The second method, called signed curvature profiles, consider both the magnitude and direction of curvature. Signed curvature profiles were generated as the average of all curvature angles within a dataset. Right-handed (clockwise) angles were considered positive.

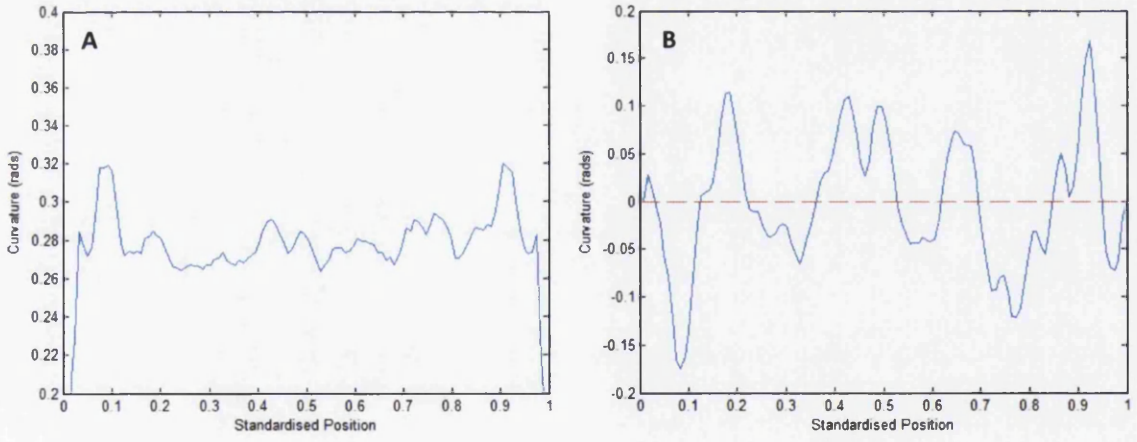

Figure 2.6. Examples of unsigned (A) and signed (B) curvature profiles. The broken red line in B represents a curvature of 0.0 radians.

CHAPTER 3: DESIGN AND IMPLEMENTATION OF THE ADIPAS IMAGE PROCESSING PLATFORM FOR THE IDENTIFICATION AND ANALYSIS OF DNA IN ATOMIC FORCE MICROSCOPY IMAGES

3.1 Introduction

3.1.1 Image Processing of AFM Images

In order to measure DNA curvature from AFM images the DNA contour must be extracted. Therefore, an image processing package must be able to accept an AFM image, identify DNA molecules, extract their orientation and output it in a meaningful and accurate co-ordinate system. In order to perform these steps the software must be able to flatten the DNA image, remove or reduce noise, extract foreground objects (*i.e.* DNA), repeatedly erode each foreground object until it is only one pixel thin and remove the erroneous branches created by erosion to leave the 'backbone' of the DNA contour (Ficarra *et al.*, 2005b). A brief summary has been provided in Figure 3.1.

There are few freely or commercially available programs that could be used for such an application, examples include ImageJ (Collins, 2007) and Gwyddion (Nečas and Klapetek, 2011). However, there is very little or no customisation possible using such software making further analysis time-consuming and impractical. For the analysis of DNA curvature and flexibility there is a pressing need to process large amounts of complex image data and perform very specific tasks. It is this need for customisation and flexibility that drives the majority of AFM researchers working with DNA to develop their own in-house software.

The first study to produce a simple image processing workflow and associated general user interface (GUI) for extracting DNA contours was the ALEX toolbox (Rivetti *et al.*, 1996). However, the ALEX toolbox has not been updated since its publication, has limited documentation for the user and is not freely available to download. The same is true for much of the research group-developed software (Sanchez-Sevilla *et al.*, 2002; Ficarra *et al.*, 2005b). Therefore, the best option available to a researcher wishing to investigate intrinsic DNA curvature by AFM is to develop their own software.

Figure 3.1. - Simplified image processing, data processing and data analysis workflow. The example AFM images are of *TP53* Exon 5-9 DNA.

3.1.2 Analysis of DNA Contours Extracted from AFM Images

There are a number of analyses that can be applied to DNA contours extracted from AFM images. The length of the DNA molecules can be reconstructed from the digitised DNA contour using a variety of estimators (Rivetti and Codeluppi, 2001). The persistence length of an ensemble of DNA molecules can be calculated (Cassina *et al.*, 2011). Of central interest to this study, intrinsic curvature and flexibility can be measured from the ensemble of DNA contours (Figure 3.1.).

In order to analyse DNA curvature each DNA contour must be fitted with a fixed number of comparable points in order to standardise the length of the molecule, an interpolant is often used to this end (Ficarra *et al.*, 2005b). The length of each DNA molecule is then assumed to be equal regardless of the measured contour length. The angle of deviation from a straight line is calculated for each consecutive point. By averaging these angles over a large population of molecules the average intrinsic DNA curvature can be calculated (Scipioni *et al.*, 2002a). Flexibility can be calculated by the variance around the average curvature values. The resulting 'curvature profile' is representative of the intrinsic curvature of the DNA sequence.

There are a number of considerations for curvature analysis of DNA by AFM that have not been clearly tackled in the current literature. There is no consensus method for selecting an interpolatory technique during length standardisation although it is likely that the choice will have an impact, however small, on the resulting curvature angles (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Marilley *et al.*, 2005; Buzio *et al.*, 2012). There is little consideration within the literature for the number of points fitted to the DNA molecule (i.e. the base pair window size) in order to calculate curvature. A typical study will fit a number of points close to the theoretical maximum resolution of the AFM image (Ficarra *et al.*, 2005b). For example, each pixel may be 2.92 nm long, the equivalent to approximately 8.5 bp, and a point would then be fitted for every 10.5 bp of the standardised length of the DNA molecule. At this low resolution there is likely to be a high degree of variation cause by digitising the DNA contour. An estimation of this variation has not been provided within the current literature and will be considered within the present study.

3.1.3 Aims and Objectives

The primary aim of this chapter was to produce a software package that could experimentally determine DNA curvature and flexibility from AFM images of DNA molecules using multiple methods available from peer reviewed literature. In order to achieve these goals the ADIPAS (AFM DNA Image Processing and Analysis Software) software was developed to provide a flexible approach to image processing and analysis suitable for multiple experiments on AFM images of DNA. This pipeline aims to be accessible to the general user and provide reliable and reproducible results.

The ADIPAS software was able to read an input AFM image, rescale the data, plane fit/flatten, filter for noise, threshold the image to identify foreground pixels (DNA molecules), skeletonise and 'prune' the resulting skeleton and finally extract *xyz* coordinates. The software calculated angles between adjacent points at user defined intervals along a DNA contour, created a curvature matrix of the resulting data and allowed either the direct creation of curvature and flexibility profiles or application of the FF algorithm. Other experimental measures were also instituted including reconstructed DNA contour length and persistence length calculation. Other experimental tools were developed during the project. This included a method for the visual identification of appropriate base pair window sizes for the calculation of curvature angles. An experiment that determined the selection of an appropriate interpolant type for the analysis pipeline has also been detailed. In order to facilitate high-throughput analysis of DNA images and usability a GUI was developed for ADIPAS. The GUI was developed to be usable with the minimum of training or knowledge of DNA studies of AFM.

3.2 Development of ADIPAS

3.2.1 Programming Platform

The AFM DNA Image Processing and Analysis Software (ADIPAS) software was developed in the Matlab 7.5.0 programming platform with the image processing and statistics toolboxes (Mathworks, Cambridge, UK). This programming environment is compatible with all major operating systems.

3.2.2 Image Processing Pipeline

3.2.2.1 Plane Fitting

Input AFM image files were read as greyscale intensity data. A nine degree polynomial was fitted to each line of the image in turn. The polynomial was subtracted from the source data in order to fit each line to a horizontal plane. This step was repeated using a polynomial fitted to the lowest seventieth percentile of Z-height data in order to ignore extreme values and smooth inconsistencies in the background. A similar method was used by previous authors and ensured that the resulting image surface was extremely flat and suitable for further image processing (Sanchez-Sevilla *et al.*, 2002). The number of iterations and the degree of the polynomial fitted could be specified by the user.

3.2.2.2 Image Filtering

While AFM has a higher signal-to-noise ratio than other comparable techniques there was still a variable level of noise in each image (Hansma and Hoh, 1994). This noise was attributable to a variety of sources: impurities in the sample, sub-optimal cantilever tuning, cantilever wear over a large number of images, surface-tip interactions, external acoustic vibration sources and poorly grounded equipment producing electrical feedback. A number of filters were implemented into the image analysis software platform. The default filter was a 3x3 median filter, used as a baseline filter for low or locally occurring noise (Ficarra *et al.*, 2005b). Other filters were utilised on a case-by-case basis and included a 3x3 Gaussian filter and a 3x3 average filter. Included in this step was a line-by-line adaptive histogram for increasing height contrast and a 3D background subtraction. These final two options were not image filters but were similarly used to improve image quality on an image-by-image basis.

3.2.2.3 Image Thresholding

The purpose of image thresholding was to separate background (sample surface) and foreground (DNA molecule) pixels. During this step the image was simplified into a binary image where background pixels were '0' and foreground pixels were '1'. The foreground contained areas of interest that were likely to be experimental DNA molecules. A threshold

value was identified which was used as a cut-off value above which all Z-height values were considered foreground pixels. This value was determined by either a visually interactive user-defined threshold value or an automatically determined threshold value. Both approaches have been incorporated into the software. For very good contrast, theoretical or low noise AFM image automatic thresholding can produce very accurate and reproducible results using the Otsu method (Otsu, 1979). For moderate to high noise images it was more suitable to visually inspect the resulting thresholded image to ensure good agreement with the original image. This was achieved using a slider that controlled the threshold value below which pixel intensity was considered background. The slider covered all values of the greyscale image.

3.2.2.4 Thinning/Skeletonisation

Successive outlying pixels were removed from the binary image until only a 'skeleton' of one pixel thickness remained. This was achieved using the default image erosion algorithm in Matlab with one hundred passes (Lam *et al.*, 1992).

3.2.2.5 Removal of Image Artefacts and Overlapping Molecules

Isolated pixels were removed from the image. Foreground pixels in contact with the image boundary were removed as it was impossible to determine how far they extended outside of the image boundary. Any molecules that were circular or formed a circular pocket were removed as it was not possible to determine which overlapping branch of the molecule was followed when extracting image coordinates. Only DNA molecules which contained a number of pixels within a user defined maxima and minima were retained for the next step. This reduced processing time for the next stage and removed clearly erroneous molecules such as fragments or DNA molecules lying end-to-end.

3.2.2.6 Removal of Spurious Branches

Skeletonisation of an image produced additional 'branches' from the backbone of the DNA contour. These branches occurred due to increases in thickness along a DNA molecule or imaging artefacts such as tip smear. In order to remove branches all spurious 'endpoint' pixels were identified and removed (Figure 3.2.). These were any pixels that were only adjacent to one other pixel in an immediate 3x3 pixel grid area. This was repeated until only two endpoints remained. During the process of removing these pixels any branches that ceased to grow were removed. The final two branches that remained were the branches that added the longest length to the molecule and formed the DNA contour or skeleton. When two branches of equal length co-existed at one end of the molecule one of the two branches was removed at random. Molecules that had more than two endpoints were not included in the final dataset.

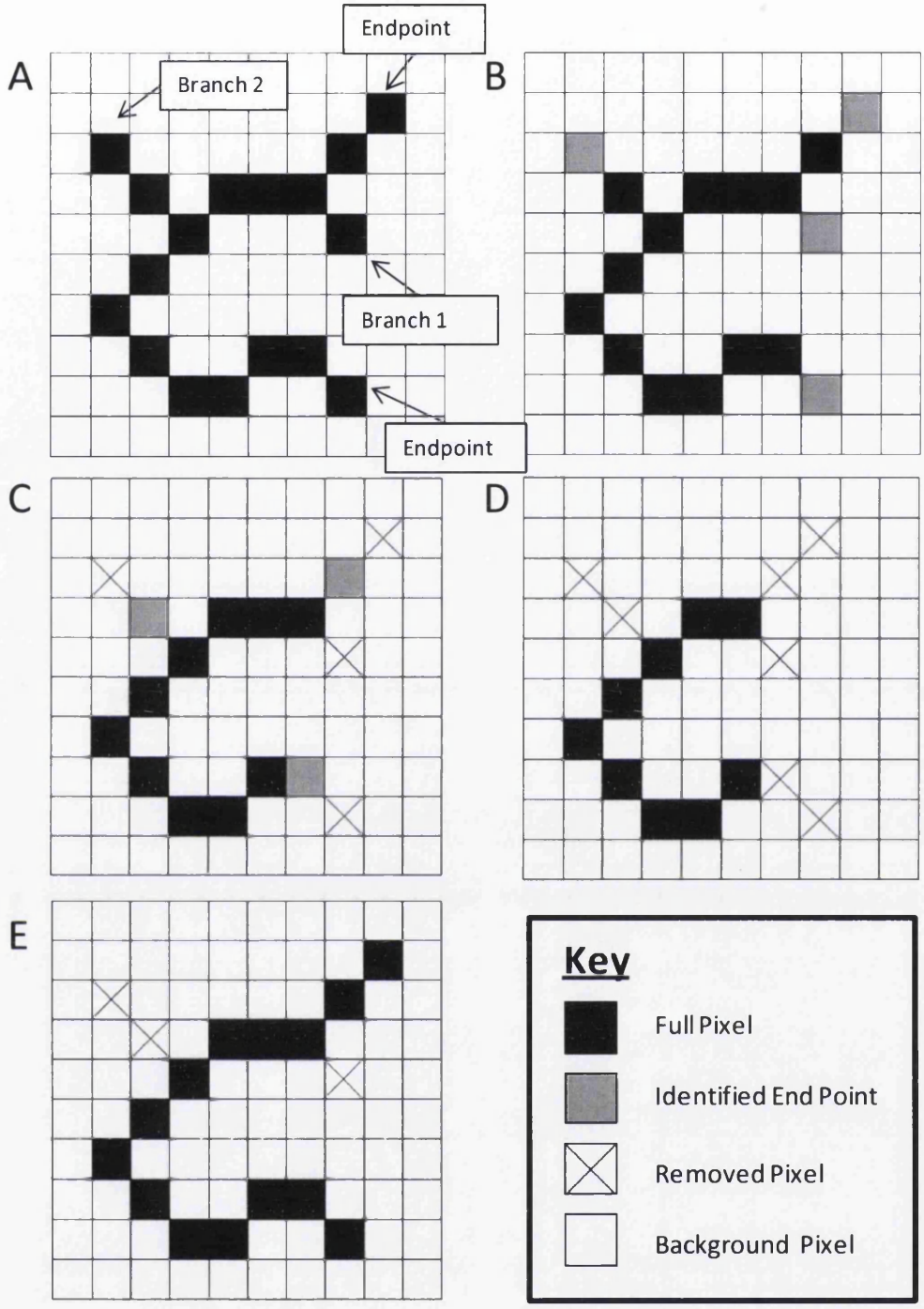

Figure 3.2. - Simple example of the algorithm for the removal of 'spurious branches' in a binary image. Black squares represent pixels that contain Z data corresponding to a skeletonised DNA molecule. Grey pixels are pixels identified as 'endpoints'. Endpoints are saved to a separate variable. White squares represent background pixels. Squares marked with an X represent pixels that have been removed. A) Initial binary image with two spurious branches. B) Endpoint pixels have been identified (grey) and stored in a separate variable. C) Endpoint identification is repeated. Branch 1 ceased to grow and is therefore removed. D) Endpoint identification is repeated. Branch 2 has ceased to continue growing and is removed. There are only two endpoints to the 'core' of the skeleton. E) The remaining branches are added back into the binary image as part of the 'skeleton core'.

3.2.2.7 Identification of Molecule of Interest

Molecules of interest were displayed visually within the ADIPAS GUI. The binary skeleton was displayed on top of the original image allowing for visual assessment of the fidelity of the image processing method. Molecules that adhered to the original image and were not sample or image artefacts were selected by the user for extraction during the next step. The end pixel that corresponded to the highest Z-height values within the original unfiltered plane fitted image was automatically identified with a red circle as the end-labelled end of the DNA molecule. In experiments without protein or small molecule end-labelling this option was disabled. The user visually confirmed endpoint tags to ensure that the program correctly identified protein end labels. Alternatively, the user could specify that the label was present at the opposite end. DNA molecules that were visually confirmed to be erroneous were removed. All DNA molecules were recovered during automated analysis of computer simulated AFM images.

3.2.2.8 Extraction of Coordinate Data

Pixel coordinates for the contours of DNA molecules identified during the previous step were extracted in sequential manner. The output was an ordered series of pixel coordinates from the first to final endpoint pixels. Corresponding Z-height values from this coordinate list were extracted from the image generated during the *Plane Fitting* step. If the experiment included visual or Z-height end-labels then the first coordinate removed was from the side of the DNA molecule that was confirmed to contain the end label. In the cases where no end labels were specified the program began extraction at a random endpoint pixel.

3.2.3 Design of a General User Interface for ADIPAS

A GUI was developed for ADIPAS that allowed the operator to visually check each stage of the image processing workflow. The purpose of the GUI was to facilitate high-throughput analysis of DNA images (Figures 3.3-3.7.).

On accessing the software the first prompt for the user was to input an image file (Figure 3.3.). This created a pop-up for file selection familiar to any user of modern operating systems. Directly below the 'Get Files' button was a textual input panel where the name of a file containing workspace variables from a previous session could be manually inputted. Image processing that was performed after the name of a valid file has been entered using 'Get Files' continued the numbering system of the previous file/dataset. Additionally, output data contained the information within the input file.

Once an image file, or series of files, had been selected the GUI displayed a plane fitted version of the AFM image file and highlighted a series of tick boxes and sliders (Figure 3.4.). The tick boxes represented image noise filters that could be applied to the displayed image by selecting the appropriate tick box followed by the 'Refresh Filter' button (Figure 3.5.). The 'Contrast' slider modified the maximum and minimum Z-height value in the displayed image and was dynamically updated. The 'Reset Image' button returned the image to its original state. Any contrast adjustment was retained in the overlaid output image.

The sliders located below the filter tick boxes controlled the upper and lower threshold value for the image. On selection of either slider the image was converted to a binary black and white image (Figure 3.6.). The threshold was dynamically updated as the slider position was changed by the user. The image could be reset to its original state by pressing the 'Reset Image' button or by applying another image filter.

On selecting the 'Done' button below the threshold sliders a pop-up window appeared during the branch removal step (Section 3.2.2.6.). As the most time consuming stage of the image processing platform (between 5-30 seconds dependent on the number of detected molecules) the pop-up window indicated the number of molecules that remained to be processed for branch removal.

After the branch removal was completed the original image appeared overlaid by the skeletonised DNA contours (Figure 3.7.). A red circle indicated the end with the largest Z-height for each molecule. This allowed for identification of end-label proteins if included as part of experimental design. A series of labelled tick boxes was highlighted to the right of the displayed image. The tick boxes corresponded to the appropriately labelled overlaid DNA molecules in the displayed image. Unchecking a tick box and selecting the 'Refresh' button removed the overlaid molecule from the image. Similarly, reselecting a tick box and pressing

'Refresh' caused the molecule to reappear on the image. A button labelled 'Flip' was located to the left of each tick box. Selecting this button when a molecule was overlaid on the displayed image caused the circle denoting the 'tagged' end of the molecule to swap endpoints. This stage also allowed for the removal of obviously erroneous DNA molecules, DNA contours that had not followed the observable DNA molecule with acceptable fidelity or otherwise undesirable DNA molecules.

On pressing the 'End' button all molecules visibly overlaid on the displayed image would have *xyz* data individually extracted in a sequential manner beginning with the end designated as 'tagged' by a red circle. The software saved the overlaid image file as a compressed jpeg. Information on the name of the source file, name of the overlaid output image and DNA contour *xyz* pixel coordinate were saved as a data file. The detected molecules in the file were labelled sequentially in ascending order beginning with zero unless a previous workspace was loaded before beginning image processing (Figure 3.3.).

The software continued to open the subsequent image file if a series of image files were selected. If there were no more files for processing the software closed. Alternatively, if the tick box located at the bottom of the screen labelled 'Reprocess Image' was checked then the software reloaded the previous file and treated it as a new image for further processing. This was useful when DNA molecules were not detected by the first pass of the software.

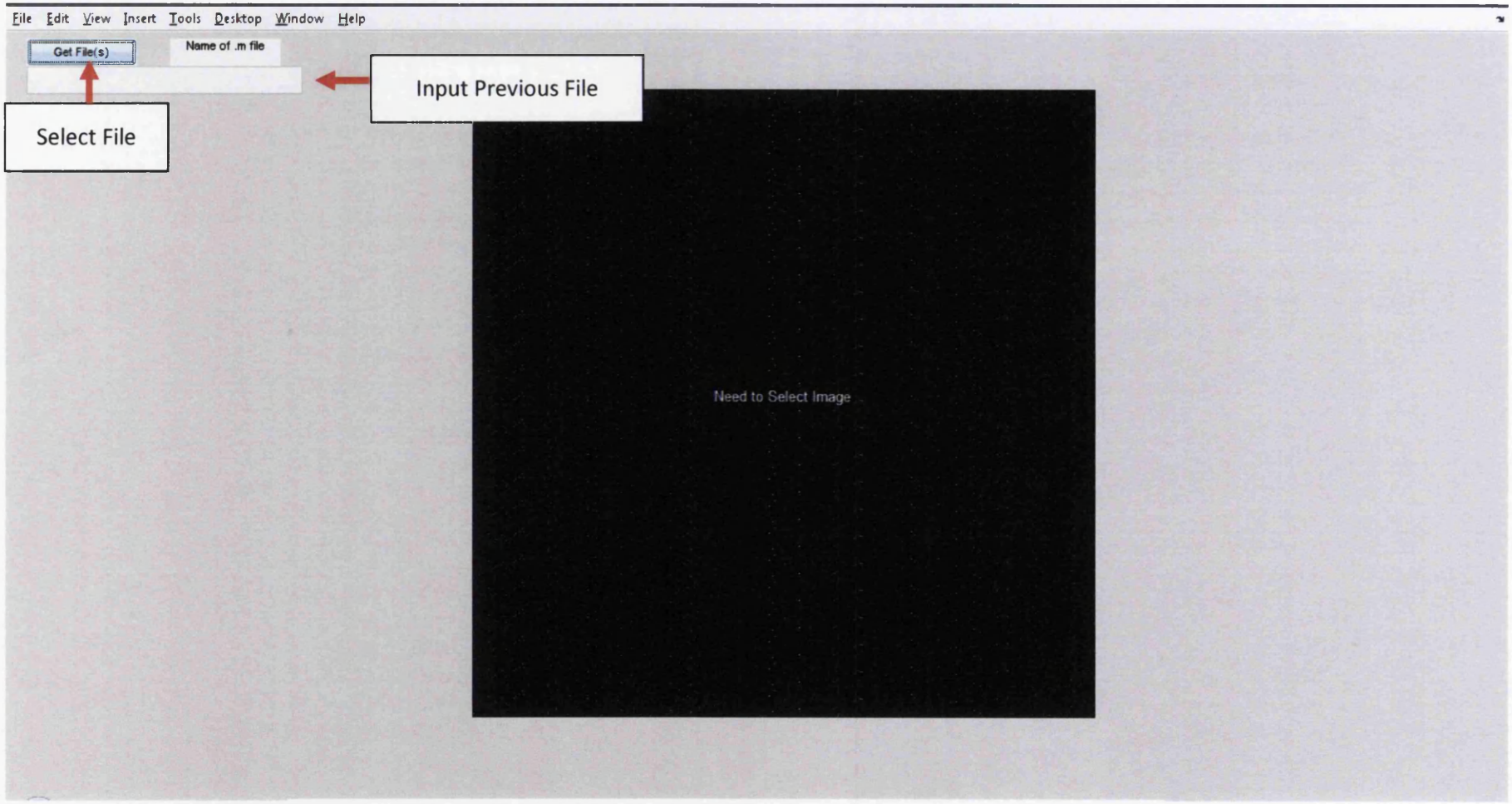

Figure 3.3. – ADIPAS GUI before image data is selected. Files can be selected using a pop-up by selecting the appropriate button. Data files containing lists of DNA molecule data from previous runs can be extended by inputting the filename in the appropriate box.

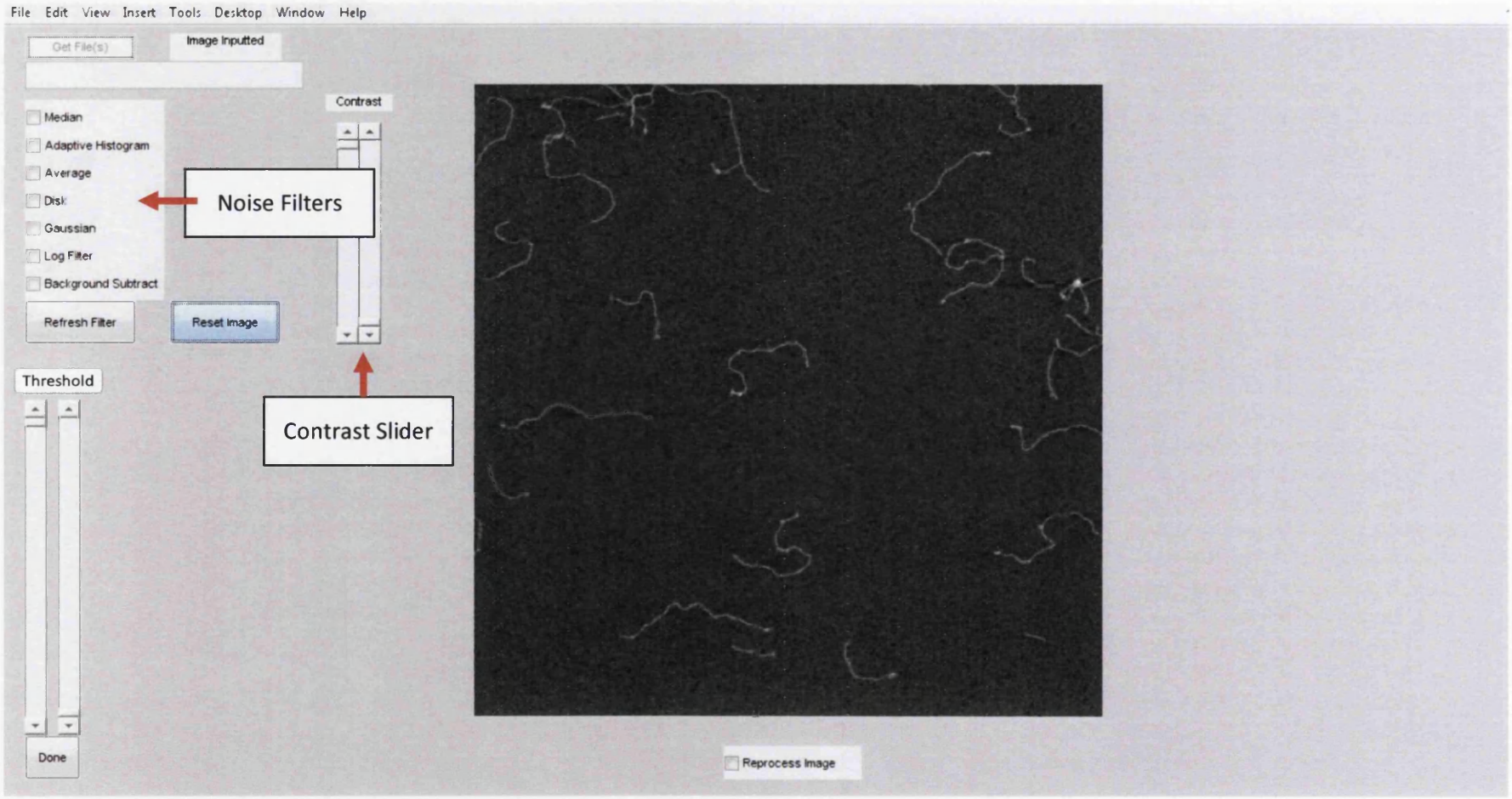

Figure 3.4. – An AFM image displayed in the ADIPAS GUI after plane fitting. A series of noise filters can be applied to the image. Additionally the contrast can be modified using the contrast sliders.

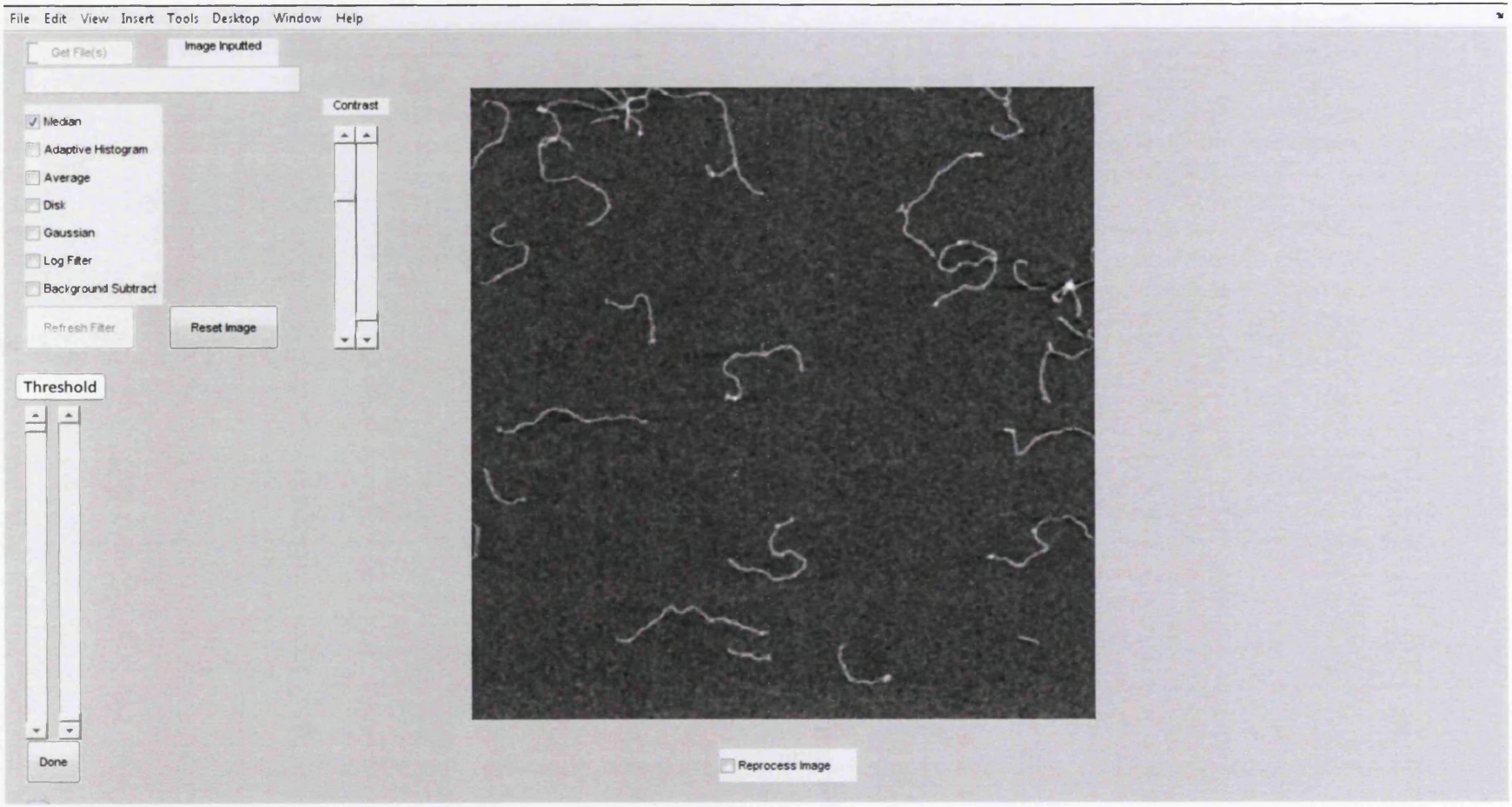

Figure 3.5. - AFM image displayed in the ADIPAS GUI after a 3 x 3 median filter has been applied. The maximum and minimum Z-height values have been adjusted to improve the contrast of the image.

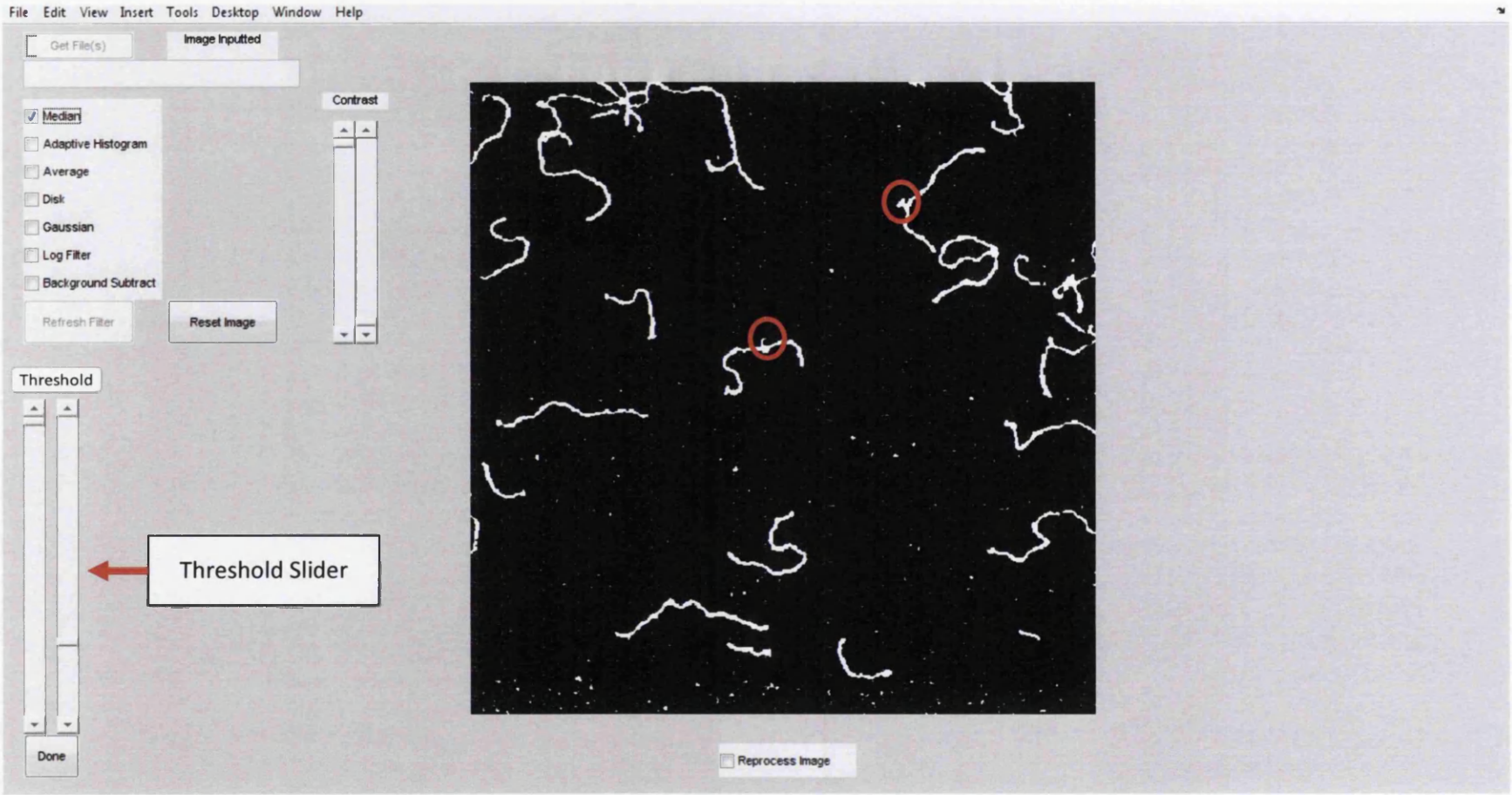

Figure 3.6. - AFM image displayed in the ADIPAS GUI after a user-defined height threshold has been applied. The resulting image is binary. White pixels are foreground DNA molecules, black pixels are background. The threshold can be dynamically modified using the sliders indicated. Note the DNA molecules that have obvious 'branches' which will be removed in the next step (red circles)

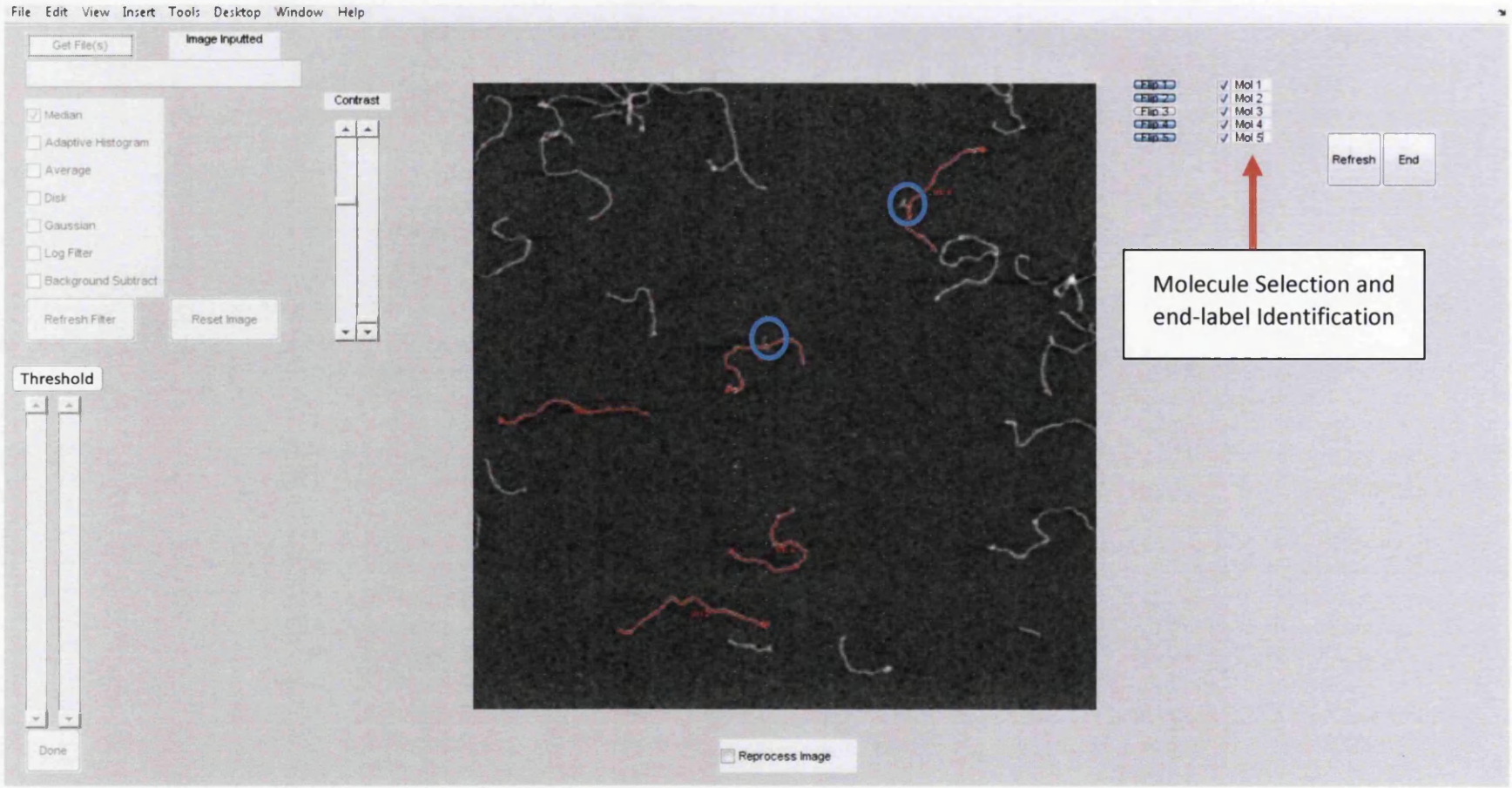

Figure 3.7. - AFM image displayed in the ADIPAS GUI overlaid with the final skeletonised and 'pruned' DNA molecules. Branches pruned by the software are indicated with blue circles. The end point of individual DNA molecule with the highest Z-height are indicated with a red circle. Molecules can be flipped or removed using the tick boxes and buttons to the right of the image. The image can be reprocessed by ticking the tick box at the bottom of the image.

3.2.4 Data Processing and Analysis Pipeline

The aim in the development of an analysis pipeline was to allow for a comprehensive investigation of physical measurements of DNA molecules extracted from AFM images. The pipeline allowed for a large amount of input *xy* coordinates from a large cohort of DNA molecules to be processed to produce measurements of reconstructed molecule length, DNA persistence length, intrinsic DNA curvature and DNA flexibility. Additionally, the FF algorithm of DNA orientation was reproduced from the original publication (Ficarra *et al.*, 2005b). The tools developed were sufficient to the task of analysing 'real' and theoretical AFM images of DNA molecules with the end result of measuring intrinsic curvature. Critically, the analysis platform enabled a full analysis of a DNA sequence, such as the *TP53* gene, with the aim of estimating DNA curvature, DNA flexibility, reconstructed length, persistence length and a full evaluation of the FF algorithm. This has been fully detailed in Chapter 5.

3.2.4.1 Calculation of DNA Contour Length

The length calculation was based upon a modified Euclidean distance measurement called the Kulpa Estimator (Kulpa, 1977). This simple estimator of distance was obtained by first calculating the Euclidean distance between each pixel using the equation below:

$$d(p,q) = \sqrt{(p^1 - q^1)^2 + (p^2 - q^2)^2}$$

In the equation above p and q are pixel *xy* coordinates. There were only two unique states for pixel orientation. This was either side-by-side in the horizontal or vertical plain or diagonal. Two pixels side by side were scored as having a Euclidean distance of 1 and those in the diagonal plane as 1.4. The Kulpa estimator used the modified values of 0.948 and 1.343 (Figure 3.8). The sum total of all the pixel distances was calculated and converted into nanometres based upon the size and resolution of the image.

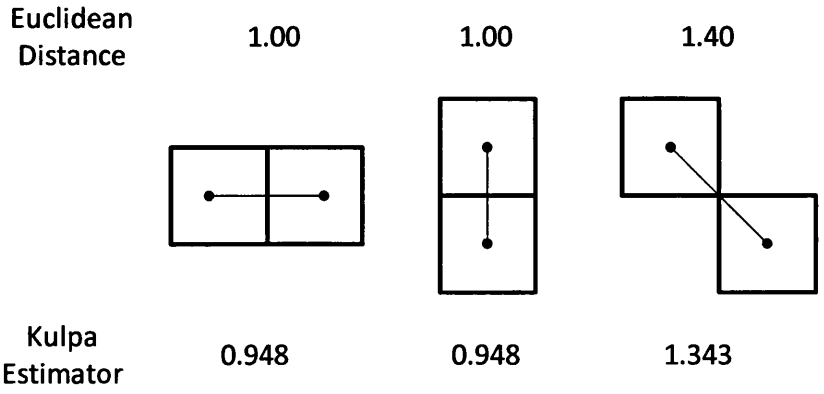

Figure 3.8. - Pixel coordinates distance as calculated for using Euclidean distance (i.e. the Freeman estimator) and the Kulpa estimator. The squares represent individual pixels and the rounded line represents the distance measured.

3.2.4.2 Persistence Length Calculation

Persistence length is an important global statistic of polymer flexibility. According to the WLC model of DNA flexibility (Rivetti *et al.*, 1996) the mean trajectory of an intrinsically straight chain in 2-D is given by the equation:

$$< R^2 >_{2D} = 4\xi L\left(1 - \frac{2}{\xi}\left(1 - e^{-\frac{L}{2\xi}}\right)\right)$$

Where R^2 is the mean curvilinear distance of the polymer in Euclidean distance, ξ is the persistence length of the polymer and L is the curvilinear distance/contour length of the polymer. Solving this equation for different values of ξ over a range of values of L allows the construction of WLC models of the end-to-end distance of DNA chains.

In order to calculate the persistence length of DNA from a set of 'real' DNA images, experimental $< R^2 >$ was calculated. The experimental $< R^2 >$ was compared to theoretical $< R^2 >$ values for a range of ξ generated using the equation above. For each ξ the value of L varied in a range of n_0 to n_L where n is an evenly spaced range of contour lengths from zero to ~300 nm. An upper limit of ~300 nm was obtained from the literature (Cassina *et al.*, 2011). This produced a prediction of $< R^2 >$ over a range of contour lengths and values of ξ.

The experimental estimation of $< R^2 >$ for a series of DNA molecules was straightforward. A linear interpolant was fitted between each pixel within each molecule in a sufficiently large dataset. The width of a pixel in nanometres was calculated (*i.e.* size of the image in nanometres divided by the number of pixels). This allowed for the selection of a series of points within the DNA molecule that were n_0 to n_L curvilinear distance from the beginning of the molecule. The end-to-end distance (R^2) between the start point (n_0) and point of interest (n_x) was then calculated in nm^2. This was repeated sequentially over an appropriate number of points (n_0-n_L) for each DNA molecule. This provided an ordered series of ascending values for the curvilinear distance along an individual DNA molecule (R^2). These values were calculated for a sufficiently large number of DNA molecules and averaged ($< R^2 >$).

In order to identify the most appropriate persistence length for experimental DNA it is necessary to identify which value of ξ produces $< R^2 >$ from the WLC model that most closely match experimental $< R^2 >$ values. A number of values of ξ are used to generate $< R^2 >$ from the WLC and the closest fit was identified using root mean square error (RMSE). The solution to the equation that best matches experimental $< R^2 >$ values provided the persistence length for experimental DNA (Figure 3.9.A). The quality of the fit can be visually confirmed by using a plot similar to that in Figure 3.9.B. If the plot significantly deviates from all theoretical predications made by the WLC model it may be necessary to assess the fit using other statistical tools.

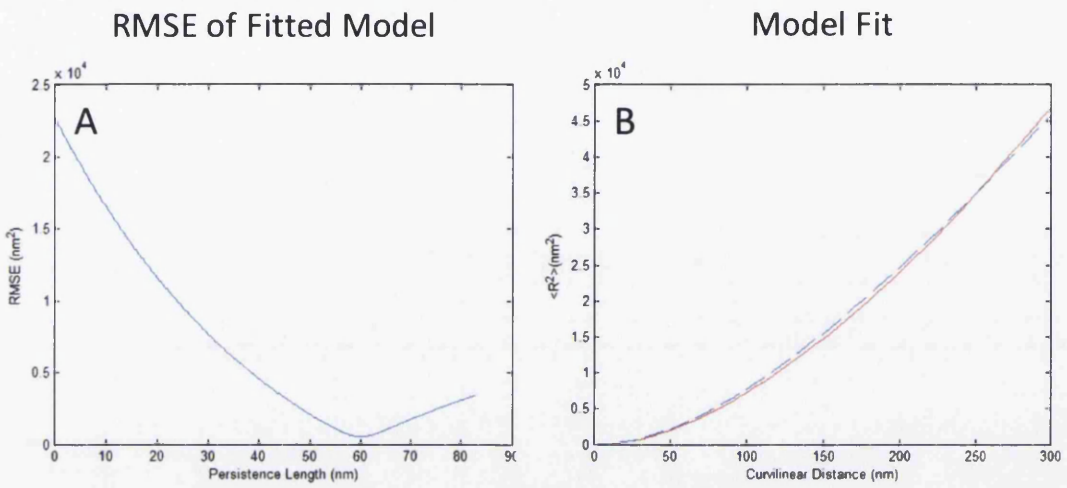

Figure 3.9. - Examples of experimentally determined DNA persistence length by comparison to theoretical values of $<R^2>$ using the WLC model. A) Plot of the RMSE fits of R^2 generated using a range of persistence length values against experimental R^2 values from the WLC model. B) Experimental R^2 values (red line) alongside R^2 values (broken blue) predicted by the WLC model for a DNA molecule of persistence length of 60 nm over a range of curvilinear distances of 0-300 nm.

3.2.4.3 Calculating Curvature Angles From DNA Molecules

3.2.4.3.1 Identification of Comparable Points in a Set of DNA Molecules

Pixel values extracted during the image processing step were fitted with a piecewise interpolation technique that passed through each point. The choice of interpolant was experimentally assessed and is presented in Section 3.2.5.1. To obtain a comparable number of data points for each molecule a suitable number of coordinates were selected at equal intervals along the DNA molecule (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a).

3.2.4.3.2 Base Pair Window Size

The number of points fitted per DNA molecule determined the base pair window size at which curvature angles were calculated. As curvature angles were calculated over three points the base pair window size was twice the distance of one fitted point. For example, a researcher wishing to calculate curvature at an interval size of 21 base pairs for a 1855 bp DNA molecule would fit a point every 10.5 bp. Therefore, 177 coordinates (1855 bp divided by 10.5 bp) would be selected from the interpolated DNA contour at regular intervals. Examples of coordinate selection over a number of base pair window sizes and the outcome angles calculated are presented in Figure 3.10.

3.2.4.3.3 Angle Calculation

In order to study curvature the angular deviation from the backbone line was calculated. Individual *xy* coordinate steps were treated as vectors. The dot product and the perpendicular dot product were used to find the angle of intersection between sequential vectors using the formulae in below (schematic in Figure 3.11):

$$theta = arctan(perpedicular\ dot\ product, dot\ product)$$

The resulting angle in radians (rads) was considered positive if it was a clockwise (right-handed) angle and negative if it was counter clockwise (left-handed). On rotation of a line around a central point the sign (+/-) of the angles will not change based upon the trajectory of the line. This is visualised using a simple series of pixel angles in Figure 3.12.

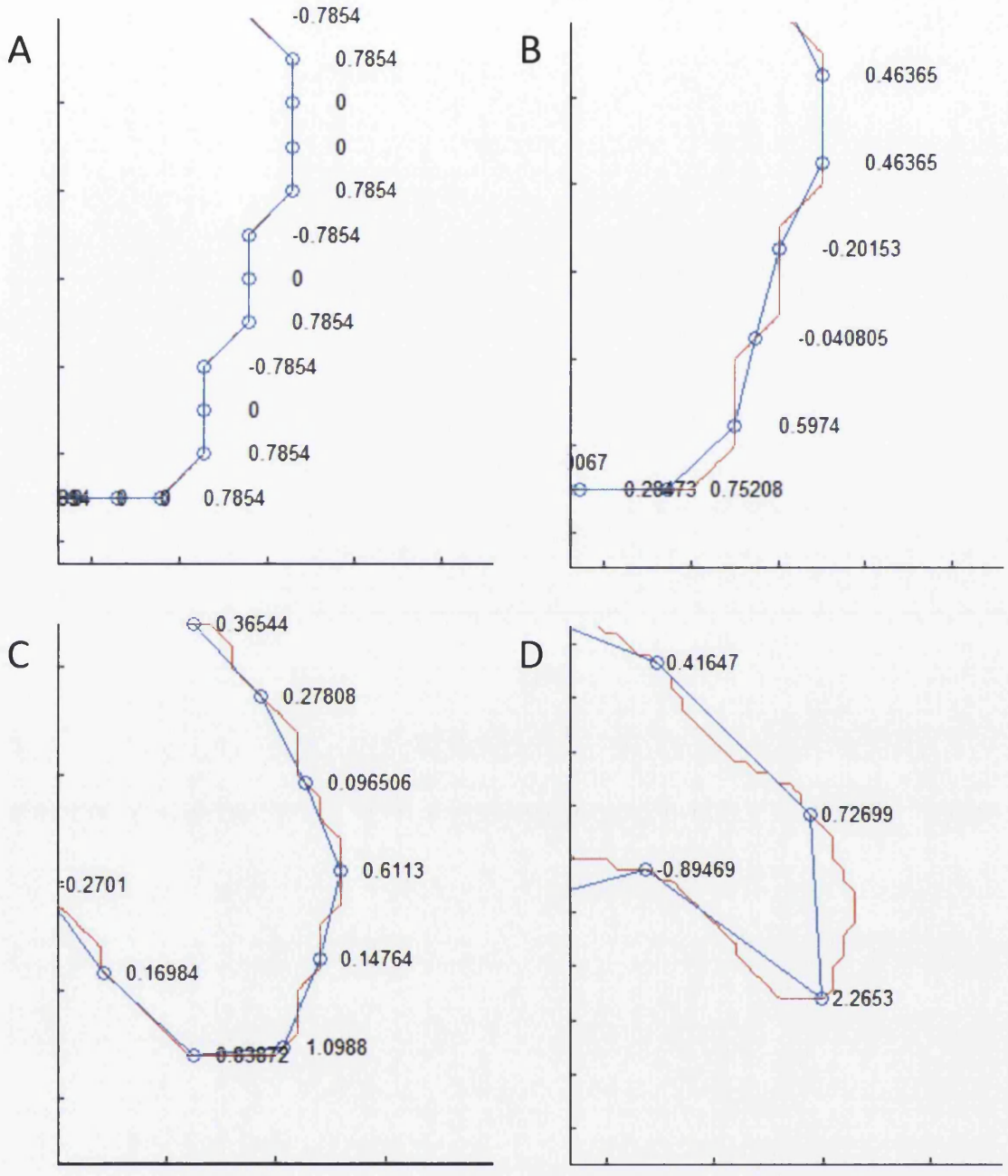

Figure 3.10. - Examples of angles calculated over four base pair window sizes. A) Base pair window size of 21 bp - 1 point fitted per pixel in original molecule. B) Base pair window size of 42 bp- 1 points fitted per 2 pixels in original molecule. C) Base pair window size of 84 bp 1 point fitted per 4 pixels in original molecule. D) Base pair window size of 400 bp - 1 point fitted per 20 pixels in original molecule. Red lines represent original data points. Blue circles represent points fitted at regularly spaced intervals of original data. Angle values were calculated as the backbone deviation from a straight line.

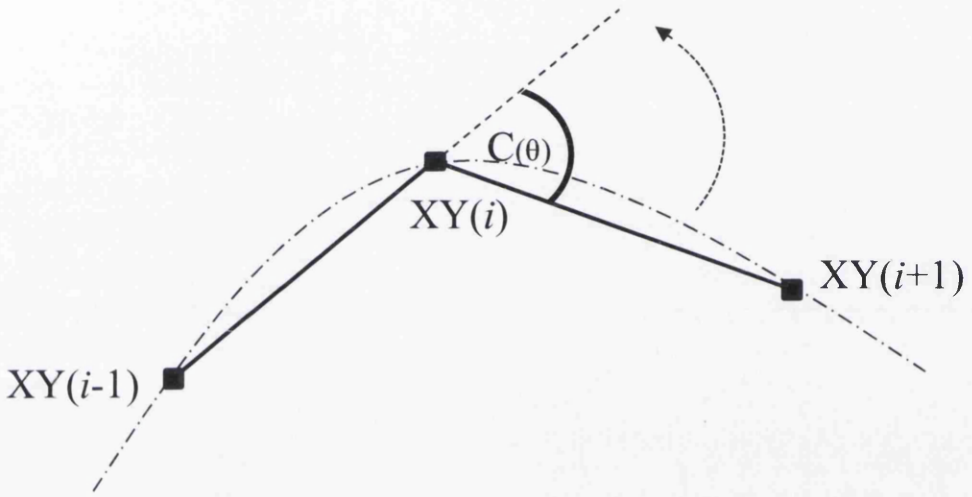

Figure 3.11. - Representation of the curvature angle at point *i*. This was calculated from the angle of intersection between both lines *XY*(*i*-1) to *XY*(*i*) and *XY*(*i*) to *XY*(*i*+1) . Note this would be a negative angle as it is counter clockwise rotation.

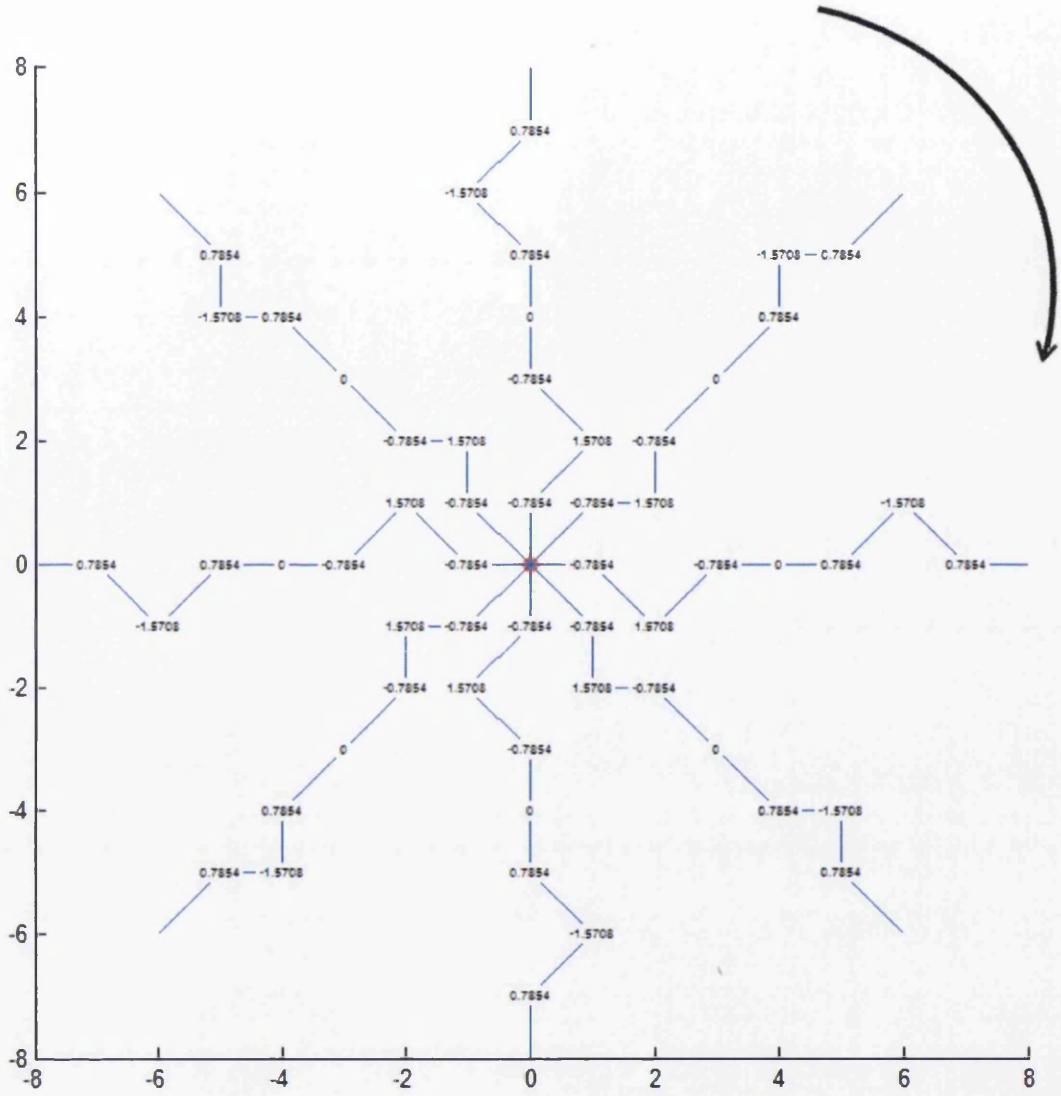

Figure 3.12. - Angle calculation for a line rotated around a central point (red circle). The line section was replicated multiple times around the central point with different orientations. All angles were calculated along the lines from the central red circle to the end of the lines. Angles were calculated at the intersection point of each section of the line. Angles along the line were identical in each rotation although the direction of the line changes. All angles calculated for a comparable point along the line were identical independent of rotation.

3.2.4.4 Calculation of Curvature Profiles.

The observable curvature of a DNA strand imaged by AFM is composed of two factors: *intrinsic curvature* (C_0) and *flexibility* (*f*). Intrinsic curvature is a product of local interactions between dinucleotides and consecutive base pair steps. Flexibility is the perturbation of DNA on interaction with the local environment. Both of these elements are products of DNA sequence. Therefore, observable curvature of a DNA sequence of base pair length *c*an be described by the equation below:

$$C(n) = C_0(n) + f(n)$$

Where *C* was the observable curvature at point *n* along the DNA sequence C_0 was the intrinsic curvature and *f* was the flexibility. Due to its relatively high rigidity DNA has been shown to follow first order elasticity theory (Scipioni *et al.*, 2002a). The contribution of thermal noise imposing local variations of the structure of DNA was considered zero over a sufficiently large sample size (Ficarra *et al.*, 2005b). Therefore averaging over a sufficiently large population of DNA molecules the intrinsic curvature at point *n* was calculated using the equation below. The flexibility parameter was characterised by the standard deviation at point *n*.

$$C_0(n) = \langle C(n) \rangle = C_0(n) + f(n)$$

In order to generate a curvature profile for an aligned set of DNA molecules of number *N* each molecule was sampled *S* number of times along its standardised length (Section 3.2.4.3.1.). This gave a matrix of curvature values *M (N x S).* Each row (*n*) of the matrix was a separate DNA molecule. Each column was a series of angle measurements at a comparable position along the length of the DNA molecule (*s*). The mean value of the rows gave the curvature profile for the dataset. The curvature profile had a length equal to *S.* The standard deviation at each point of *S* was the flexibility profile. The final curvature profile was composed of either signed or unsigned values. To produce an unsigned curvature profile (also called *absolute curvature*) all angles were made absolute before taking the average. The unsigned curvature profile took into account the magnitude of curvature and disregarded the direction of curvature. To produce a signed curvature profile both the size and direction of the curvature were considered. A sample schematic of a curvature matrix is provided in Table 3.1.

Position along the DNA Molecule in Base Pairs

	10.5 bp	21 bp	31.5 bp	42 bp	.....	End of Molecule (S)	
Molecule 1	0.165	0.073	-1.854	0.026	.....	-0.251	
Molecule 2	0.250	0.085	4.875	0.300	.....	-0.186	
Molecule 3	0.357	-0.404	0.444	-0.444	.....	-0.444	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
Molecule N	0	0.444	-0.444	0.435	-0.435	0.424	
	↓	↓	↓	↓	↓	↓	
Curvature	0.165	4.875	-1.854	0.026	-1.854	0.026	*Mean*
Flexibility	0.152	1.375	-0.894	0.563	-0.440	1.126	*Standard Deviation*

Table 3.1. Schematic of a curvature matrix of dimensions *N* x *S*. The curvature profile was the mean value of the column and flexibility profile is its standard deviation. Both profiles were of length 1 x *S*. The outcome was a signed curvature profile.

3.2.4.5 Fragment Flipping Algorithm

Each DNA molecule adopts one of four different conformations on a mica surface (Figure 3.13). The FF algorithm assumes that there is an underlying consensus curvature profile to the DNA molecules. If all molecules are oriented correctly then the objective function of the FF algorithm, the mean column variance, will be at a minimum.

The FF algorithm was instituted using a Greedy algorithm, looping throughout the curvature matrix multiple times to find an optimal solution to the objective function. The Greedy algorithm was found to be optimal compared to other well-known general-purpose heuristic solvers (Ficarra *et al.*, 2005b). A curvature matrix was constructed for a cohort of DNA molecules. The angles corresponding to the curvature of an individual molecule were then transformed into each possible orientation (*i.e.* invert sign, flip direction or both, see Figure 3.13.) and the mean of the column variances was recorded for each orientation (Figure 3.14), this value was the objective function of the algorithm (Figure 3.15). The molecule orientation that reduced the objective function the largest amount was then adopted for that molecule. This was applied to all of the molecules within the dataset. This was iterated upon for the dataset multiple times until the objective function did not significantly change, assessed by the Kolmogorov-Smirnoff test, over a user defined number of passes (default = 25). Curvature profiles were constructed from the resulting curvature matrix as previously detailed.

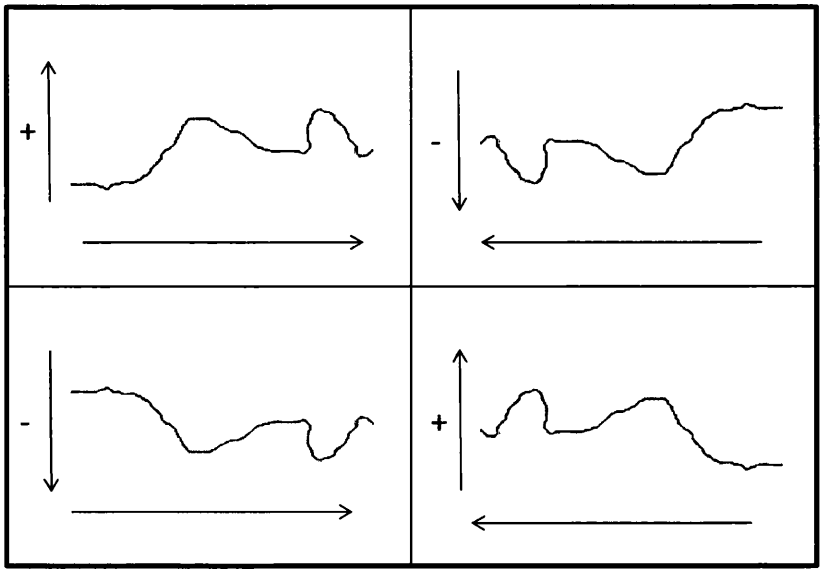

Figure 3.13. – All possible orientations of a DNA molecule on a flat surface. The shape of the molecule was the same in each orientation but the direction of the molecule changed.

Position along the DNA Molecule in Base Pairs

	10.5 bp	21 bp	31.5 bp	42 bp	52.5 bp	
Molecule 1	1	2	3	4	5	
Molecule 2	5	4	3	2	1	
Molecule 3	-1	-2	-3	-4	-5	**Mean Variance**
Molecule 4	-5	-4	-4	-2	-1	15.12

↓ **Flip Molecule 1**

	10.5 bp	21 bp	31.5 bp	42 bp	52.5 bp	
Molecule 1	-1	-2	-3	-4	-5	←
Molecule 2	5	4	3	2	1	
Molecule 3	-1	-2	-3	-4	-5	
Molecule 4	-5	-4	-4	-2	-1	11.25

↓ **Flip Molecule 2**

	10.5 bp	21 bp	31.5 bp	42 bp	52.5 bp	
Molecule 1	-1	-2	-3	-4	-5	
Molecule 2	-1	-2	-3	-4	-5	←
Molecule 3	-1	-2	-3	-4	-5	
Molecule 4	-5	-4	-4	-2	-1	2.05

⇣ **Flip Until No Significant Change in Variance for 25 Iterations**

	10.5 bp	21 bp	31.5 bp	42 bp	52.5 bp	
Molecule 1	-1	-2	-3	-4	-5	
Molecule 2	-1	-2	-3	-4	-5	**Final Mean**
Molecule 3	-1	-2	-3	-4	-5	**Variance**
Molecule 4	-1	-2	-4	-4	-5	0.05

Figure 3.14. - Demonstration of the FF algorithm on an example curvature matrix using the Greedy algorithm. Each individual DNA molecule was flipped into four different orientations. The orientation that reduced the objective function of the FF algorithm (the mean of the column variances) by the largest amount was retained. The process was repeated with all molecules in the dataset until there was no significant change in variation for a used-defined number of whole dataset iterations.

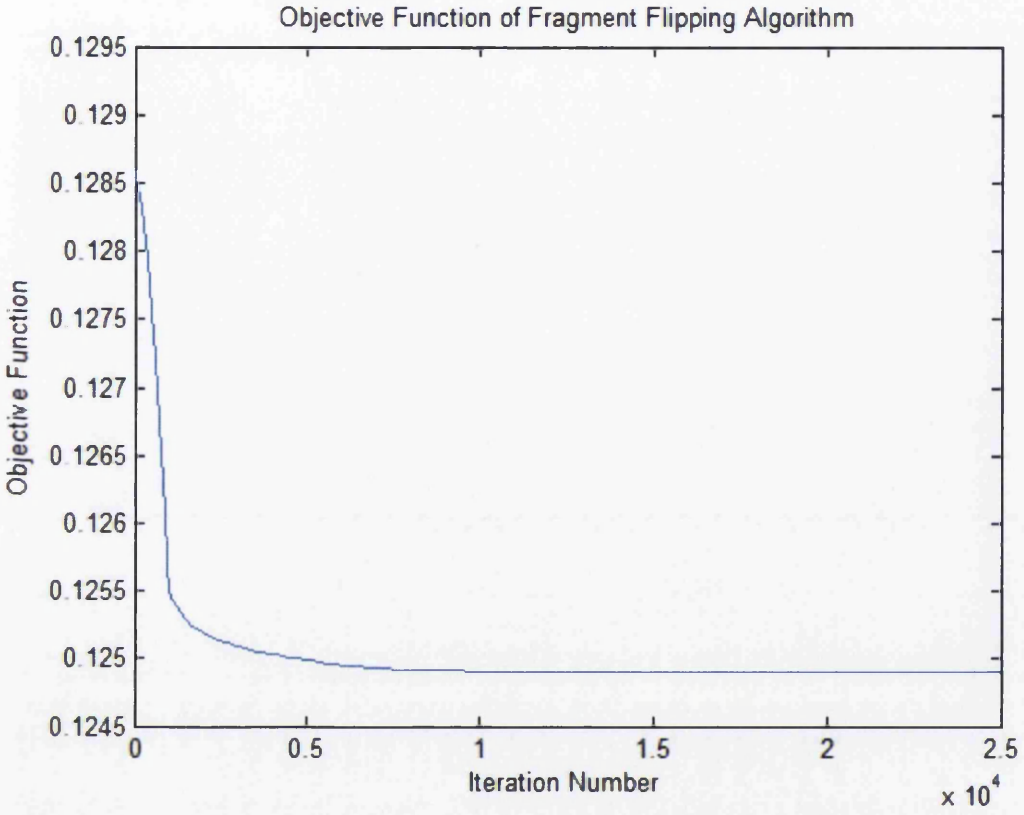

Figure 3.15. - Example of the change in the FF objective function using the Greedy algorithm. The objective function, the mean of the column variances in the curvature matrix, was recorded at the end of each iteration and plotted against the appropriate iteration number.

3.2.5 Evaluation of Methods for Calculating Interpolants and Selecting Appropriate Base Pair Window Sizes for Curvature Analysis

A lack of consideration for the choice of interpolatory method used to fit a series of regularly spaced points to DNA contours in order to calculate curvature angles was identified within the current literature. Similarly, the window size in base pairs used to select an appropriate number of points to fit to DNA contours for the calculation of curvature angles was not considered in many studies. Instead, studies typically fitted a number of points close to the number of pixels that made up individual molecules. The choice of interpolant was likely to have an influence on the curvature angles calculated on this scale. At low base pair window sizes there was likely to be an increased influence of DNA molecule variance and image noise on the calculation of curvature angles.

The following sections contain experimental work aimed at the selection of an optimal interpolation technique from those presented within the literature. Furthermore, a method of 'visual thresholding' for identification of suitable base pair windows for curvature calculation has been developed. This methodology was novel and allowed the identification of digitisation effects on AFM images of DNA and for the selection of appropriate window sizes in base pairs for the calculation of curvature angles on an experiment-by-experiment basis.

3.2.5.1 Selection of an Interpolant

In order to take a number of comparable curvature angles along DNA contours the length of each molecule had to be standardised (Rivetti and Codeluppi, 2001). A number of points were then interpolated along the standardised length. From the standardised length a suitable number of points were selected at regular intervals. The angles between these points were calculated.

There have been a number of methods used to smooth the DNA *xy* pixel coordinates by previous authors: constrained 'interpolatory splines' (Ficarra *et al.*, 2005a), piecewise fitting of polynomials every 5 coordinates (Ficarra *et al.*, 2005a) and a complex method of spline fitting (Sundstrom, 2008). Other authors did not include this step (Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a). The effect, if any, of interpolant selection on curvature measurements is not covered in any available published material. It is likely that the choice of smoothing/interpolant will effect curvature measurement at the smallest base pair windows where the effect of digitisation of the DNA contour is most pronounced.

In order to select an appropriate interpolant a number of methods were compared. Four methods of interpolation were implemented for comparison; piecewise linear interpolation, cubic spline interpolation, piecewise cubic hermite interpolation and piecewise polynomial fitting (Examples in Figure 3.16). The first three methods were available within the

Matlab programming environment. Each methodology created a series of interpolatory splines that were constrained to pass through each pixel coordinate. The linear method joined each point with a straight linear line (Figure 3.16.B). The methods used for the construction of splines for the cubic and hermite interpolants were very similar and both involved the fitting of piecewise splines (Figure 3.16.C+D). However, the hermite was typically more suitable for curved data. The end result of this was that hermite spline interpolation involved less oscillation, was less likely to overshoot in non-smooth data and typically adhered more tightly to the data (Figure 3.16.D). The final method, detailed by Ficarra *et al.*, was implemented as described by the authors (Ficarra *et al.*, 2005a). A three degree polynomial was fitted every five coordinate points (Figure 3.16.E).

A set of computer generated AFM images of *TP53* Exon 5-7 was used to test the choice of interpolant. The original *xy* coordinates per base pair were retained and used for orientation of molecules after image processing by aligning the processed molecules with their original orientation (Section 2.5.6.). The Euclidean distance between each original base pair and the comparable section after image processing was calculated. This value was used to measure the similarity between the final positions of *xy* coordinates generated using the different interpolation methods and the original DNA molecule before digitisation. The results ordered from highest similarity to lowest were; polynomial (2.83×10^3), linear (3.41×10^4), cubic (3.42×10^4) and hermite (3.42×10^4).

The clear choice of interpolant method was the polynomial, however a number of inconsistent artefacts were observed during implementation (Figure 3.17). These artefacts were not observed with the other interpolant types. The polynomial method was not implemented; it was considered that it is better to have consistently slightly poorer but predictable similarity rather than inconsistent and unpredictable artefacts. There was little difference between the remaining interpolant types. However, the linear interpolant (effectively the original digitised DNA contour) gave the best results and was implemented within the software.

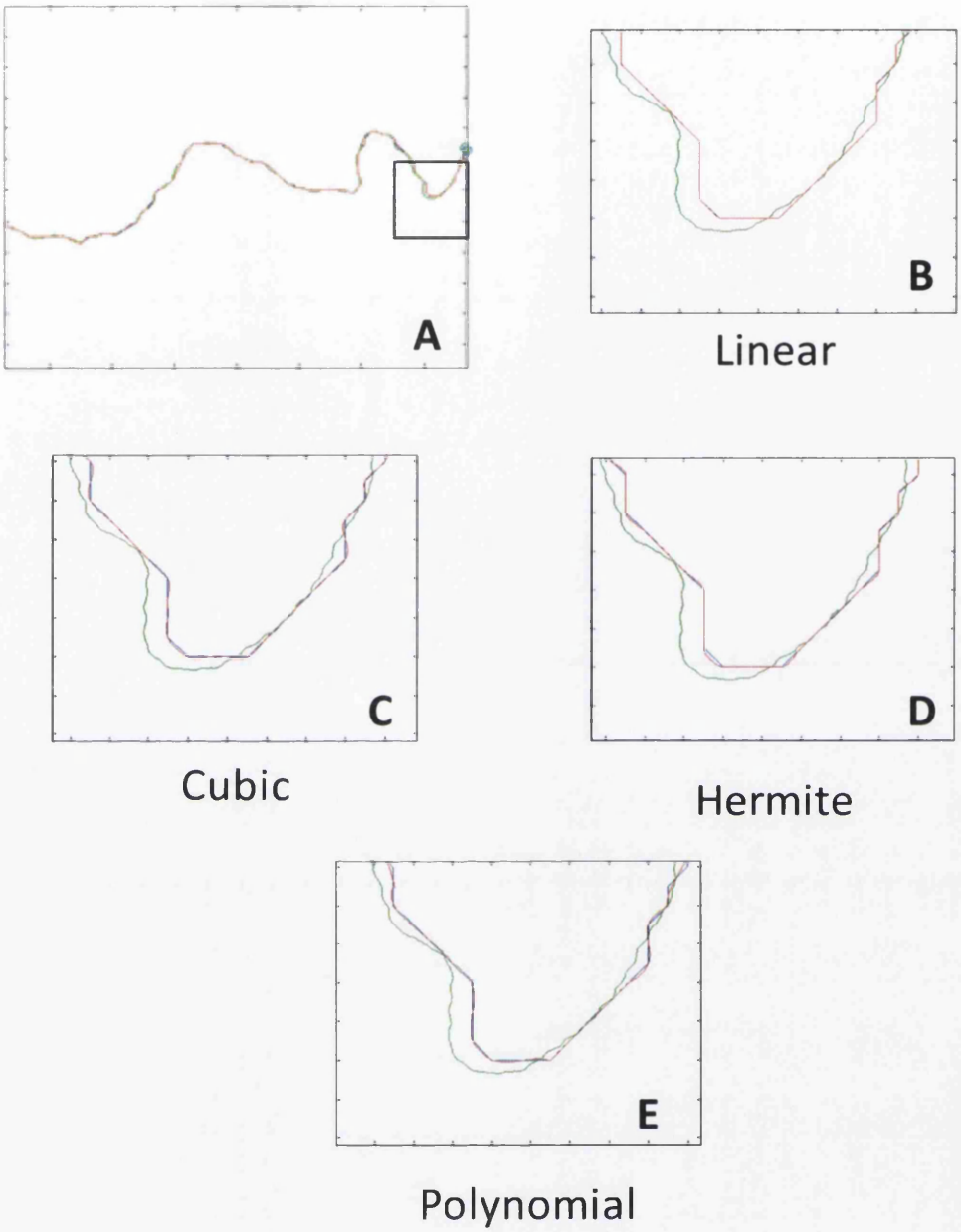

Figure 3.16. - Effect of interpolant on curvature of the DNA contour. A) Original trace a theoretical *TP53* Exon 5-7 molecule B) Piecewise linear interpolation. C) Cubic spline interpolation D) Piecewise cubic hermite interpolating polynomial. E) Polynomial interpolation (3 degree polynomial over 5 x-y coordinates). The green lines are the theoretical trace of the DNA molecule, red lines represent pixel coordinates of the DNA molecule after digitisation and the blue lines are the reconstruction of the pixel coordinates after interpolation.

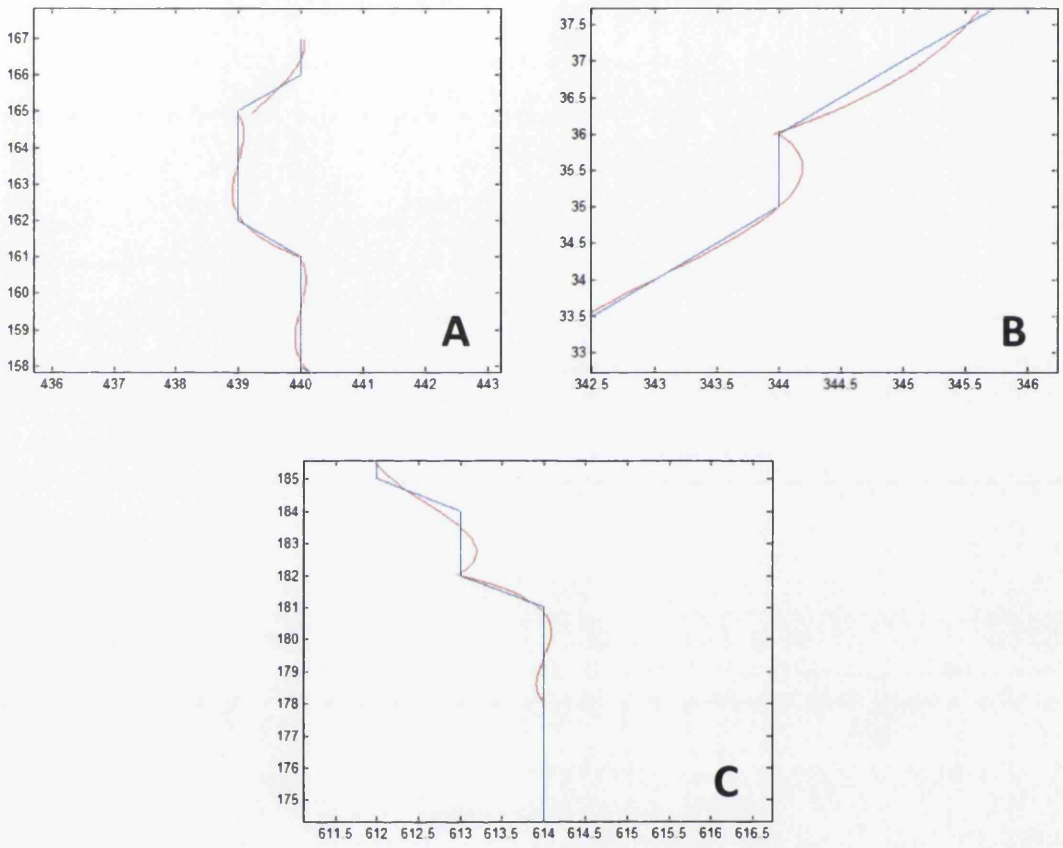

Figure 3.17. - Examples of artefacts created by the polynomial interpolation methodology. A) Contour breaks. B) Large deviations at sharp angles. C) Inconsistently following the DNA backbone. Three degree polynomials were fitted over five x-y coordinates.

3.2.5.2 Creating a Visual Threshold for Selecting Base Pair Window Size

Many researchers work at the lowest possible resolution afforded by AFM imaging. Often this resolution borders one pixel per curvature angle measurement. The following research has shown that this may not be the best resolution for a reproducible analysis, with supporting evidence described below. This study has described a method of visually selecting an appropriate window size of curvature for the AFM user alongside outputs that gives information on the quality of the data at that resolution.

3.2.5.3 Influence of Base Pair Window Size on Curvature Profiles

While there is a consensus method for the calculation of curvature from AFM images there is little consideration of how the base pair window size for curvature and flexibility will effect the observed curvature within the literature. In order to tackle this issue curvature profiles were generated for a test dataset of simulated AFM images over a wide range of base pair windows (Figure. 3.18). The images were processed using the image processing software detailed in this chapter. A number of points were fitted to the resulting *xy* coordinates at regular intervals. These intervals corresponded to the appropriate number of points for each of the base pair windows under investigation. The range used experimentally began below the limit of AFM resolution of the current experiment of 21 bp with a point fitted every 10.5 bp. The range extended up to a maximum of ten points fitted per molecule. The full details on the generation of computer simulated AFM images are presented in Chapter 4.

At low base pair window sizes (Figure 3.18 - light green lines) there was observably a greater degree of variation in absolute curvature profiles (Figure 3.18.A+B). Unsigned curvature did not consider the direction of the curvature. This contrast between peak and trough steadily increased at larger window sizes. The signed curvature profiles (Figure 3.18.C) also showed a steady increase in contrast between peaks and troughs at larger window sizes.

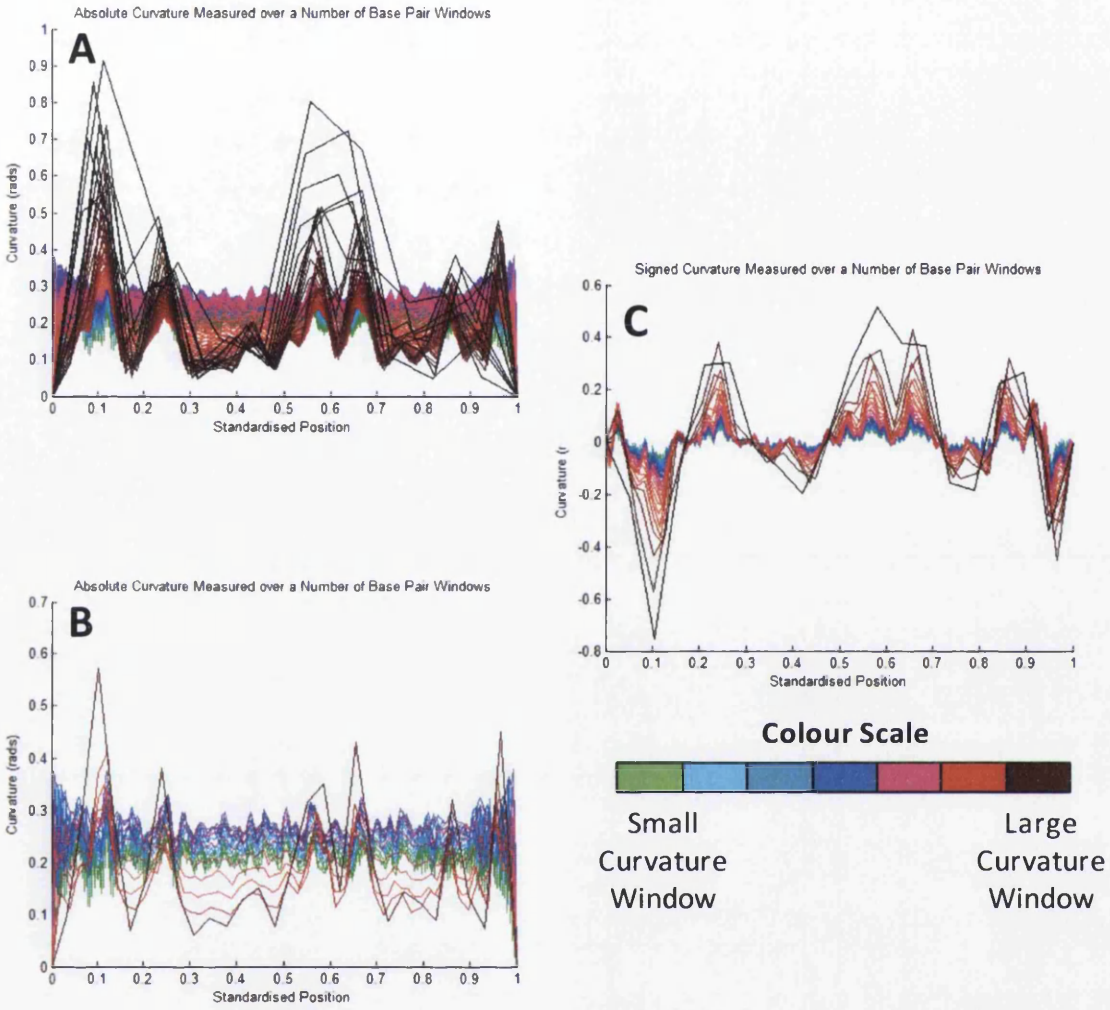

Figure 3.18. - Visualisation of the effect that the base pair sample window has on the curvature profile of *TP53* Exon 5-7 from computer simulated AFM images. A) Unsigned curvature profiles (all angles are considered positive) for window sizes from 10.5 bp to 270 bp. B) Unsigned curvature profiles for window sizes from 10.5 bp to 185 bp. C) Signed curvature profile of *TP53* Exon 5-7 (both positive and negative curvature values are considered) with window sizes from 10.5 bp to 270 bp. The position along the molecule was standardised from 0 to 1. Zero corresponds to the 5' end of the molecule. The colour scale is from 10.5 bp (light green) to ~270 bp (dark brown).

3.2.5.4 Influence of Base Pair Window Size on Mean Curvature

It was observed that when the base pair window size at which curvature was calculated was increased there was a reproducible effect on the *mean curvature* measured (Figure 3.19.). There was a peak at low window size followed by a trough and then a nearly linear increase. This effect was observed in multiple experiments. The effect was both an artefact of the data and an interesting finding.

The changes observed could be attributed to a gradation of different properties of the digitised DNA contour. The number of data points fitted at a low window size was nearly equal to average number of pixels per molecule. Where this was true there was a constrained number of different calculable angles (0, 0.78 and 1.57 radians) with a known maxima value of 1.57 radians.

If the window size was reduced below the average number of pixels per molecule then there were more points fitted than there were pixels. This led to multiple samples being taken from within one or two consecutive pixels (*i.e.* the calculated angle is 0.00 radians) and had the effect of reducing the mean curvature. This effect can be observed at window sizes below ~21 bp in Figure 3.19.A.

The trough at ~40-80 bp could be attributed to a window size that fits a point every 2-4 pixels for a large number of DNA molecules within the data set. At this scale there was still a limited number of physical orientations that the digitised DNA could conform to, however there were many more than those available at the scale of one point per pixel. While the maximum curvature could still be 1.57 radians for an individual point it was far more likely to be lower than this. Pixelated DNA could never take up a conformation of greater than 1.57 radians at a window size of 1 point every 2-4 pixels. Additional pixels would be removed during the erosion step of image processing. Additionally, within this range of window sizes some molecules would be fitted with one point for every two pixels. This had the effect of smoothing out the jagged pixelation/digitisation of the DNA contour and was likely to reduce the mean curvature.

At window sizes in excess of 60 bp the mean curvature began to rise again. At this window size the DNA was able to take up a wide range of conformations and could begin to double back upon itself (For an example see Figure 3.10.D). As the largest measurable curvature began to rise the net effect was an increase in mean curvature (Fig 3.19.A.).

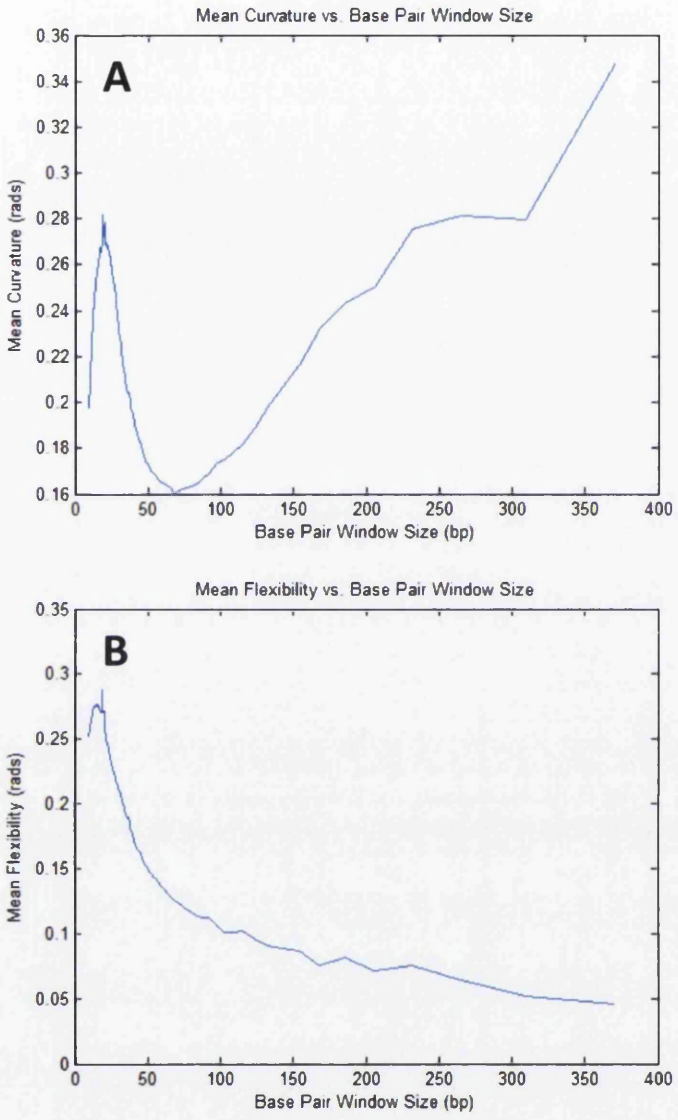

Figure 3.19. – The effect of base pair window size on unsigned mean curvature and flexibility. A) Mean curvature plotted against base pair window size. B) Mean flexibility plotted against base pair window size. Profiles were generated for a set of simulated AFM images (1171 molecules) of *TP53* Exon 5-7. Mean curvature and flexibility values were calculated as the mean value of the respective profile at each window size.

3.2.5.5 Creating a Visual Threshold for Selecting Optimal Base Pair Window Size

The observation of the mean unsigned (absolute) curvature alongside prior knowledge of how these trends occurred allowed for the segmentation of the mean curvature plot into a number of different sections. These groupings allow the researchers, on a per experiment basis, to know if the curvature measured can be attributed to digitisation of the DNA contour, local curvature or large scale curvature (Figure. 3.20.B).

There were a number of ways to approach this problem. The chosen approach was both simple to implement and visually easy to understand. The plot was smoothed (10 point average), the maximum curvature of the first peak and the minimum value of the central trough were identified. On taking the average value between these two measurements a reproducible threshold was produced. This could be applied across multiple experiments assuming the relationship between base pair window and average curvature stayed constant.

Any window sizes smaller than the window size of the peak maxima (indicated with a red circle in Figure 3.20.A) were sub-optimal as they began to sample multiple times within individual pixels. Figure 3.20 shows both the simple threshold (Figure 3.20.A) and the mean curvature with the proposed labels (Figure 3.20.B). The region from the average line and the first peak maxima was characterised as the 'pixel region', where a large proportion of molecules had a number of points fitted similar to or equal to their length in pixels. Within this region the choice of interpolant would have an effect on the curvature measured alongside the effects of DNA contour digitisation. The trough below the average line could be considered the 'local curvature' region, where the curvature measurements were free of the effects of digitisation. The minima value could be considered an optimum value. Any values that occurred within the region of window sizes larger than the average line could be considered to be curvature on the large scale, or 'gross curvature'.

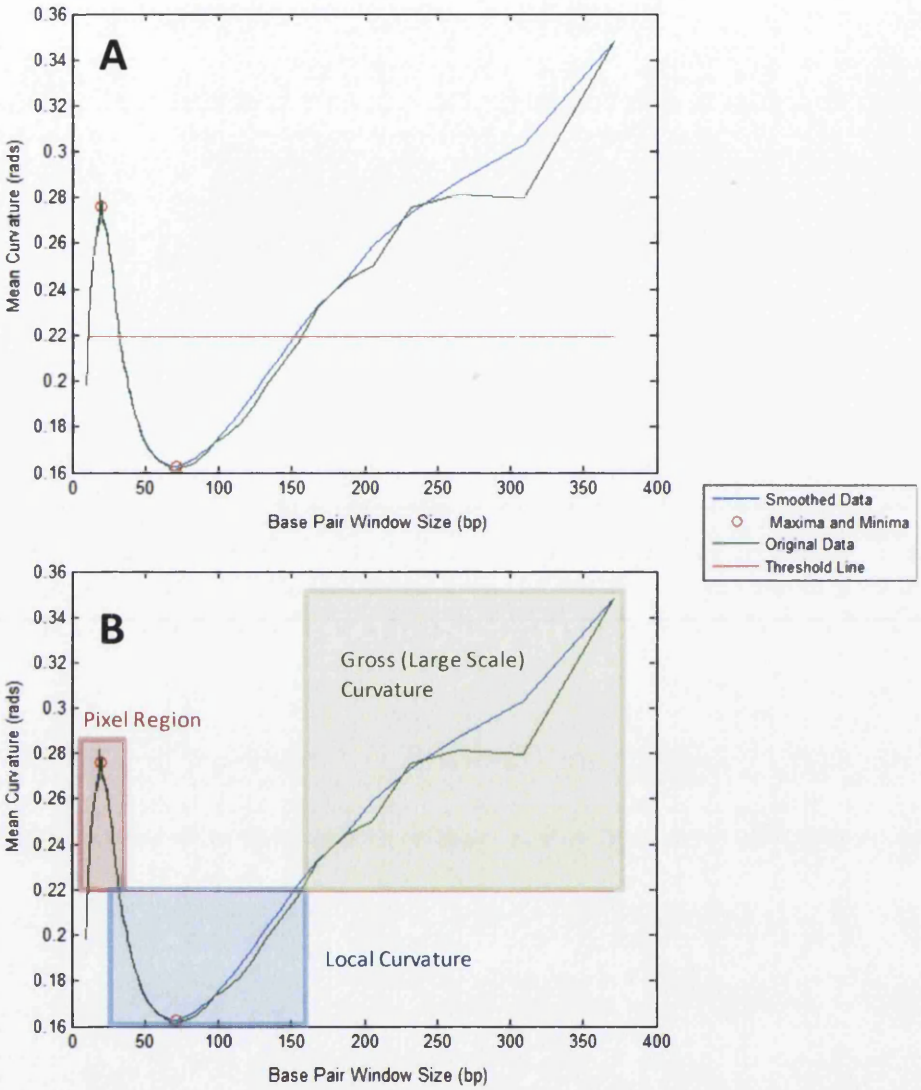

Figure 3.20. - Proposed segmentation of the mean curvature for a range of base pair windows. A) Example of the threshold model. B) Mean curvature with each section labelled and highlighted. Profiles were generated for a set of simulated AFM images (1171 molecules) of *TP53* Exon 5-7. Mean curvature values were calculated as the mean value of the curvature profile at each window size. The green line is raw data, the blue line is a ten-point smoothed average. The red circles indicate the maxima and minima of the smoothed data. The minima values are 31.44 nm and 30.48 nm Exon 5-7 and Exon 5-9 respectively. The red line represents a line drawn through the average value between the maxima and the minima. In B there are three labelled regions; the pixel region (red), local curvature (blue) and gross curvature (green).

3.2.5.6 The Effect of Base Pair Window Size on Minimum and Maximum Curvature

A number of factors were investigated to ensure that the peak in mean curvature at low base pair window size was due to the influence of digitisation on the curve. Firstly, the maximum and minimum curvature values for each dataset were calculated over a range of base pair windows (Figure 3.21.A). These measurements were the maxima and minima of angles calculated from DNA molecules (not the maxima/minima of the curvature profile). Curvature values were calculated from simulated AFM images of *TP53* Exon 5-7 (n=1171). The base pair window size was used to fit an appropriate number of linearly spaced points to each individual molecule for curvature calculations. It was observed that the matching maximum curvature values recorded within the dataset increased as base pair window size decreased (Figure 3.21.A). The graph peaked at the smallest window sizes. The value of the largest angle measured was 1.5708 radians, the angle of a right angle, which is the largest angle possible for two adjacent pixels. At this threshold window size the profile was measuring true 'pixel angles'. The highest base pair window that measures 1.5708 radians is 21 bp. It could be assumed any window size below 21 bp measured primarily pixel angles (*i.e.* 0.0, 0.78 and 1.57 radians) for all molecules. All minimum curvature measurements lay below 4.60×10^{-4} radians at each base pair window size and therefore could functionally be considered zero.

The number of individual curvature angles that matched the maxima and minima for the entire dataset were recorded at each base pair window size (Figure 3.21.B.). Occurrences of multiple maxima and minima that were non-unique began to occur regularly below the 50 bp window. It is likely that at this resolution the number of points fitted along a significant proportion of the dataset matched the number of pixels that describe individual molecules. At this window size the curvature angles calculated were non-unique: 1.5, 0.75 or 0 (only possible angles formed between three adjacent pixels). Multiple occurrences of non-unique values at such resolutions were expected due to the limited number of different conformations three adjacent pixels could assume. At larger base pair windows it was likely that points fitted would generate unique maximum curvature values as the maximum values were dictated by the local curvature over a number of points rather than the conformation of three or four pixels (which have a restricted number of different orientations). The window size with the largest matching minima and maxima value was 19 bp. This value was in good agreement with the lowest appropriate resolution of the simulated AFM images of approximately 18 bp.

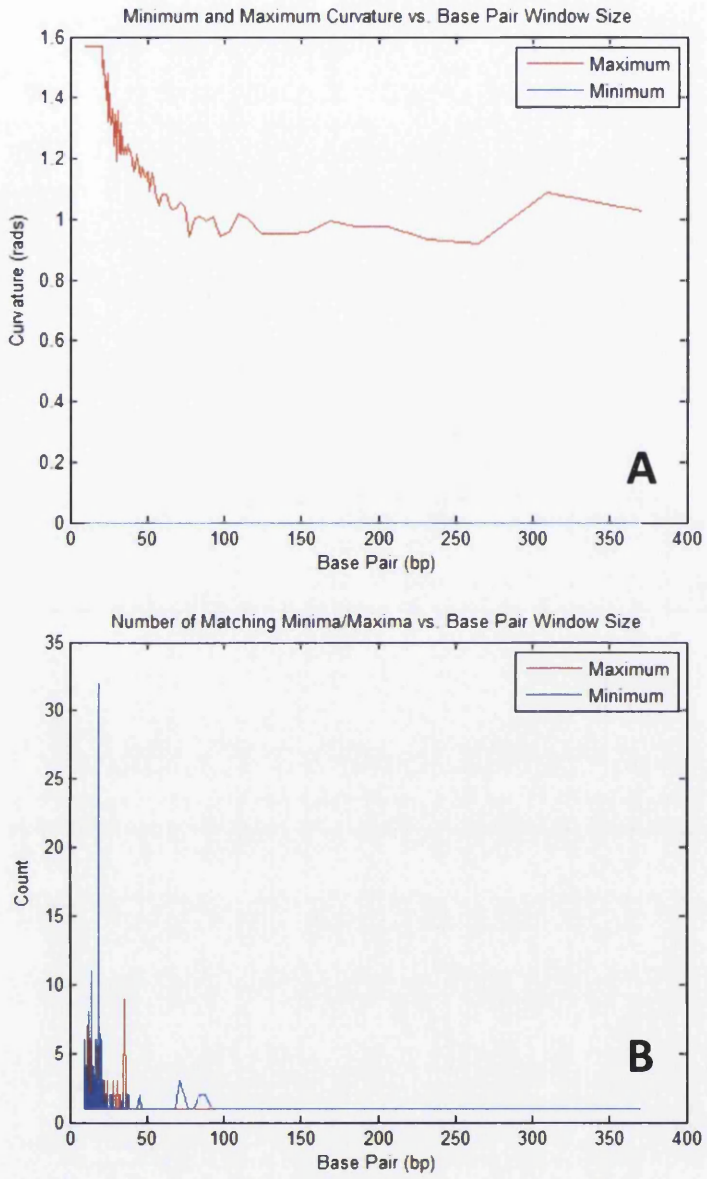

Figure 3.21. – Count of dataset extrema at various base pair windows sizes. A) Maximum and minimum curvature measurements plotted against the base pair window for simulated AFM images of *TP53* Exon 5-7. B) The number of unique angles that match the dataset maxima and minima at a range of curvature windows for simulated AFM images of *TP53* Exon 5-7. The maximum and minima were calculated for individual points within the whole dataset of simulated AFM images of *TP53* Exon 5-7 (n=1171). The base pair window size was used as a guide to fit an appropriate number of linearly spaced points to each individual molecule for curvature calculation.

3.3 Discussion

3.3.1 Publishing and Distribution

The image processing GUI and associated analysis software have been developed in the Matlab 7.5.0. programming environment. Matlab is compatible with all major operating systems (Windows, Macintosh and Linux). The ADIPAS GUI is intended for peer reviewed publication with an associated Matlab Toolbox distributed through the Mathworks website. Additionally, the software will be made available as a distributable executable file for users without Matlab using a Matlab compiler for multiple operating systems (Hernandez-Boussard *et al.*, 1999). The analysis pipeline will also be published and distributed at a later date as an extension of ADIPAS when a GUI has been finalised.

3.3.2 Comparability of ADIPAS to Previous Image Processing Pipelines

The ADIPAS image processing software described in this chapter was more comprehensive than other packages currently available for AFM image analysis. Many different techniques have been included in the software package from multiple previous studies. These include the option for automated or semi-automated image processing, a range of image filters, interactive contrast adjustment, interactive or automated thresholding, automated DNA molecule recovery and branch removal and fully annotated output images. The ADIPAS GUI allowed for operators to analyse AFM images of DNA with only minimal training. The easy access provided by the GUI, alongside the level of automation and number of available image processing options allows for quick and easy processing of large numbers of AFM images.

The steps involved in the image processing toolbox were modelled on the methods detailed by Ficarra *et al.*, with some amendments to the methodology (Ficarra *et al.*, 2005a, 2005b). A major deviation from this work was the exclusion of the *Fragment Point Recovery* step. This was initially implemented in ADIPAS and was removed due to the detrimental effect that it had on the speed of the image processing platform. The recovery of points around the DNA skeleton was computationally simple to implement. The pixels directly bordering the DNA skeleton after thinning were identified and any with Z-height above the image specific threshold were recovered and reintroduced into the DNA skeleton. This was repeated for each valid pixel. However, this step was found to introduce a large number of 'spurious branches' to each DNA molecule processed. This is likely to be less pronounced in low noise, high contrast images.

The removal of spurious branches was the most computationally intensive step within the image processing pipeline (Ficarra *et al.*, 2005a). The introduction of additional branches

dramatically reduced the speed of image processing, which was suboptimal for an image processing platform with the core aim of efficiently processing large volumes of AFM images. With the introduction of more computationally complex but faster methods of binary line tracing, such as live-wire image segmentation, the *Fragment Point Recovery* step could be reintroduced into ADIPAS without negatively impacting image processing speed (Hamarneh, 2005).

3.3.3 Identification of Image Processing Steps with Potential for Future Improvements

In terms of image processing speed the software was sufficiently fast to process large amounts of AFM images. The processing speed of a typical AFM image was between 10-30 seconds for the semi-automatic option and 5-10 seconds when fully automated. The majority of processing time was consumed with the selection of the user-defined threshold value, image filters and molecules of interest. In the case of computer-simulated AFM images these stages were entirely automated, as the level of image noise and the occurrence of erroneous molecules was considerably lower. The image filtering steps could be made redundant by implementing an automatic de-noising filter. A recent example uses the statistical features of noise sources present within an image to identify and de-noise the image (Subashini and Bharathi, 2011).

A number of automatic thresholding methods were available within the current literature. A selection of these methods have been reviewed in relation to their applicability to AFM images of DNA (Ficarra *et al.*, 2005a). Automatic thresholding using the Otsu threshold, found to be suitable by the aforementioned review, has been incorporated into the software. The Otsu threshold was found, by the present study, to be suitable for computer simulated AFM images or low-noise, high-contrast AFM images. However, real AFM images display image quality degradation over long experiments due to tip wear and other factors. This made the Otsu threshold unsuitable for the majority of real AFM images used in this study. The user-defined threshold implemented in ADIPAS allows interactive visual selection of molecules. This was found to be more suitable for the majority of real AFM images collected during this study. To improve the application of automatic thresholding to real AFM images more computationally complex automated thresholding algorithms could be incorporated into the software. A suitable algorithm exists and functions by adaptively thresholding based upon local image intensity (Gatos *et al.*, 2008). This would improve the applicability of automatic thresholding to AFM images of variable noise and contrast. The implementation of accurate automatic thresholding in addition to the previously discussed automated de-noising techniques would dramatically reduce the need for operator interaction during AFM image processing.

The method used for the removal of spurious branches in the software had the advantage of being both quick and effective for the detection of DNA contours. The algorithm iterated upon the same molecule for its execution and so it was still moderately time intensive. Methods for direct tracing of the longest path through a DNA contour, without iteration would improve the speed of this step. The current method is both quick and efficient at DNA contour detection and spurious branch removal. Without implementing another tracing algorithm into the software no quantitative comparison was possible.

3.3.4 ADIPAS General User Interface

Image processing of large volumes of AFM images was necessary for the analysis of DNA curvature and flexibility. This can be both time consuming and tedious for the user. Automation can alleviate this. However, as previously discussed, automation is only applicable to computer simulated AFM images or very high quality real AFM images. Automation applied to even moderate quality images failed to recognise DNA molecules and produced tracing errors *i.e.* false positives. Either of these eventualities would introduce bias or error into the resulting analysis.

The solution to this was a semi-automated GUI. The ADIPAS GUI allowed key decisions, such as identification of automatically traced contours as an experimental DNA molecule, to be made by the user and automates non-decision making steps. GUIs have been built for image processing packages by previous authors. For example the ALEX toolbox for Matlab had a functioning GUI for tracing plasmid DNA molecules and has been applied to linear DNA molecules (Rivetti *et al.*, 1996; Scipioni *et al.*, 2002a) The advantages of a GUI is that it allows interactive modification of visual image filters, selection of accurate threshold value and identification of DNA molecules. The current GUI was comprehensive in its level of automation and interaction. The key decisions made by the user included: choice of image noise filter, determination of image threshold value, molecule selection and necessity of image reprocessing. As previously detailed the first two key decisions could be removed if suitably efficient and accurate algorithms were implemented. The user determined selection of DNA molecules was a necessary step; it allowed for the removal of obviously erroneous DNA complexes, such as overlapping molecules overlooked by the image processing software and DNA molecules joined end-to-end.

During the design stages of ADIPAS there were two options available for the identification and removal of DNA molecules by the user. The first option was to include all DNA molecules processed by the software in the final dataset that would be edited at a later time by the operator (Ficarra *et al.*, 2005b). This had the advantage of increased automation, requiring less oversight by the user during image processing. The second option involved

selecting DNA molecules during image processing. This had two advantages. Firstly, few erroneous DNA molecules were carried through to the analysis stage. Secondly, it allowed the user to reprocess images that exhibited promising DNA molecules that were not detected by the semi-automated software during its first pass. The second option was instituted in ADIPAS as it gave higher DNA molecule recovery per image. This also had the unforeseen advantage of reducing the variable size of the analysis dataset. This was sometimes a problem as Matlab does have a maximum memory cache size which can be exceeded by very large workspace variables.

In summary, the GUI provided a balance of functionality and automation. Individual image processing time was 10-30 seconds per image, making image processing of large volumes of AFM images time consuming. However, semi-automation with operator oversight was preferable when compared to the alternatives: full automation requiring thorough removal of erroneous molecules during post-processing or the absence of a GUI requiring vast amounts of tedious operator input per image. The ability of the human eye to identify DNA molecules from background has been commented on by previous authors (Ficarra *et al.*, 2005b). The current GUI combines the benefits of automation with the decision making oversight of a human user.

3.3.5 The Analysis Pipeline

The analysis pipeline incorporated methodologies from a number of studies into one package. It provided many of the common analysis methods used by modern researchers in the field of DNA nanobiology. The analysis pipeline achieved the primary aim of the study as it was able to measure the intrinsic DNA curvature and flexibility of *TP53* or any other DNA tract of interest.

The calculation of DNA bend angles and persistence length measurement was performed using standard methodologies available from the literature (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Cassina *et al.*, 2011; Buzio *et al.*, 2012). The FF algorithm was implemented using the Greedy algorithm as recommended by the original authors (Ficarra *et al.*, 2005b). The FF algorithm was tested for accuracy and fidelity in Chapter 5.

The Kulpa DNA contour length estimator was incorporated into the software, in preference to other methodologies described in the available review of DNA length estimators (Rivetti and Codeluppi, 2001). The Kulpa estimator had the advantages of providing both an accurate estimate of DNA contour length and being simple to implement (Rivetti and Codeluppi, 2001). It produced a maximum length underestimate of -1.6 % for computer simulated AFM images and a maximum underestimation of -6.9 % on 'real' DNA images. It had the advantage over the *ad hoc* methodology described by Ficarra *et al.*, of not requiring prior

experimentation to determine a 'correction factor' for DNA length estimation (Ficarra *et al.*, 2005a). As the Kulpa estimator uses constant values for estimating contour length it is comparable between experiments. A multitude of other contour length estimators could be implemented for comparability with the Kulpa length estimator at a later date.

3.3.6 Consideration of Base Pair Window Size on Curvature

This is the first study, to the author's knowledge, that has considered the window size (in base pairs or pixels) over which to calculate curvature or flexibility and the possible downstream effect on measurements of physical parameters. The majority of research on DNA using AFM used the minimum base pair window size for calculating physical parameters based on the micrograph resolution (*e.g.* Ficarra *et al.*, 2005b; Buzio *et al.*, 2012; Cassina *et al.*, 2011; Scipioni *et al.*, 2002a, 2002b). However, at low base pair resolution there was a considerable influence of digitisation of the DNA contour on the curvature profile produced (Section 3.2.5.). The choice of interpolant influences the resulting curvature at this resolution. While there is no way to know if this had a significant impact on the results of previous studies, it was a factor that needed to be considered in order to produce the most representative estimates of intrinsic curvature and flexibility.

This study developed a method of visual thresholding to allow the user to assess the effect of base pair window size on curvature measurements on a per experiment basis (Section 3.2.5.2.). This method did not calculate an estimation of potential error or suggest a statistically optimal base pair window size. Rather, it identified a range of base pair window sizes that offered little to no interference from DNA contour digitisation and allowed the researcher to make a judgement about the selection of the experimental window size. This simple method, along with the classification system suggested, provided a foundation for future researchers to build upon. A more comprehensive study into the common effects of pixelation/digitisation on physical parameters is lacking in the current literature. This research represented progress towards facilitating accurate comparisons of curvature profiles across multiple experiments.

3.3.7 Choice of Interpolant

The identification of pixelation/digitisation effects on angle measurements resulted in the need to identify an optimal interpolant from the available literature. There have been a number of interpolation methodologies presented by various authors (Section 3.2.5.1.). Each research group presents a preferred methodology. The current literature does not provide a systematic review of interpolant methodologies or their impact on curvature measurements. This study has compared the method detailed by Ficarra *et al.*, and a simplified version of the

methodology suggested by Sundstrom (Ficarra *et al.*, 2005a; Sundstrom, 2008). The method used by Ficarra *et al.*, was taken from the paper on automated fragment sizing rather than the paper on curvature computation as the method described in the latter paper was obscurely worded (Ficarra *et al.*, 2005a, 2005b). It was assumed that the techniques used in both papers were the same; that fitting 'segmental chains' was the same as fitting polynomials.

During the analysis (Section 3.2.5.1.) an issue was raised with the method of polynomial fitting. Segmented polynomials introduced intermittent breaks and moderate deviations in the expected orientation of the DNA molecule between pixels. It is unclear if these problems were encountered by the original authors or if they corrected for these systematic errors (Ficarra *et al.*, 2005a).

When comparing the remaining three techniques there was very little difference in the curves fitted (*i.e.* linear, hermite, cubic). The three were constrained to pass through each pixel coordinate. Similar results could be predicted for any line fitting technique with rigorous constraints placed upon it, such as a series of polynomials with very small RMSE. The final choice of interpolant used in this study was the linear interpolant as it produced the least deviation from theoretical *xy* coordinates. This is similar to the methods used by previous authors for DNA contour length and persistence length estimation (Rivetti *et al.*, 1996; Rivetti and Codeluppi, 2001; Scipioni *et al.*, 2002a).

3.3.8 Proposing a GUI for the ADIPAS Analysis Pipeline

The most time and operator intensive portion of DNA curvature analysis was image processing. Therefore, image processing was prioritised over the analysis pipeline for development of a GUI. However, many of the individual steps in the pipeline produce an automated and labelled graphical output. For example, R^2 can be calculated and compared to theoretical values in order to estimate the persistence length of a set of DNA molecules with a single function. Similarly, curvature and flexibility profiles can be generated from a raw set of pixel coordinates. The steps and considerations presented in this study provide a good schematic for future nano-biologists to complete a working GUI. Further improvements could include: a number of dinucleotide wedge model parameters in the style of CURVATURE allowing comparison to experimentally produced curvature profiles (Shpigelman *et al.*, 1993), integration of novel techniques for curvature analysis (Buzio *et al.*, 2012) and multiple DNA contour length calculation methods for comparison (Rivetti and Codeluppi, 2001). Additionally, simple modifications made to the analysis pipeline would make it suitable for calculation of intrinsic DNA curvature and flexibility for time lapse experiments of DNA dynamics (Scipioni *et al.*, 2002b).

3.3.9 Limits of the Available AFM Analysis Software

None of the available software, including the software developed in this study, made any attempts to remove the effects of tip convolution from an image. The problem is widely acknowledged (Li, 2007; Sundstrom, 2008) and algorithms exist solely for the purpose of tip deconvolution of AFM images (Villarrubia, 1997). Another source of blurring, thermal drift, is also not accounted for in any of the available software. Thermal drift can cause blurring in images and algorithms suited to tackling this problem are available (Carasso, 1999). However, thermal drift is often circumvented by investigators using closed-loop settings available on most modern AFMs. Closed-loop scanning monitors the physical position of the scanner and corrects for drift introduced while driving the scanner head. Closed-loop settings have been used in this study.

The process of DNA adsorption to the mica surface is poorly understood. It has been observed that DNA sometimes undergoes a transition from B- to A- form DNA on the mica surface (Rivetti and Codeluppi, 2001). This effect has also been attributed to condensation of the DNA on interaction with the cation loaded mica surface (Sanchez-Sevilla *et al.*, 2002). Accurate models to account for this possible transition would allow for more accurate length calculations in DNA measurements. Interestingly, it has been theorised that increasingly accurate intrinsic DNA curvature calculations will allow for improved contour length estimation by modelling the predicted DNA contours as a series of arcs and straight sections (Sundstrom, 2008). The ADIPAS software would be an ideal platform for the development of such a length estimator.

3.4 Conclusions

The ADIPAS software has been developed with the primary aim of analysing intrinsic curvature and flexibility of *TP53* DNA molecules. The lack of flexible and available AFM image analysis tools was identified from the current literature and internet search engines. To this end ADIPAS was able to analyse AFM images of DNA and calculate curvature from the resulting coordinate data. The software incorporated analysis methods from a range of previous studies, allowing the use of a range of image filters, rescaling of image contrast, automated or operator interactive thresholding, automated branch removal and molecule selection. ADIPAS allowed for a more comprehensive analysis of the structural properties of DNA molecules than any other available software pipeline. It was scalable, allowing analysis of DNA molecules from a range of different AFM images sizes. ADIPAS presented the image analysis portion of its package in a GUI that would allow even unskilled operators to process AFM images of DNA after only limited training. The GUI for the analysis portion of ADIPAS will be developed in the future using the same flexible design philosophy. Other estimates of statistical and physical DNA measurements, such as DNA contour length and persistence length, were implemented into the software. The software is aimed at online distribution and publication with the hope that it will be of use to researchers within the field and also to encourage further investigation of DNA curvature by allowing other research groups to overcome the large technological hurdle of in-house software development necessary for this type of investigation.

Considerations such as choice of interpolation technique prior to curvature calculation have been investigated before implementation into the pipeline. Additionally, a novel visual method of identifying potential interference of digitisation noise in curvature calculations has been developed. These considerations have been applied to real AFM molecules and have been expanded upon in later chapters. These developments provide a strong foundation for future researchers to build upon and also represent progress towards improving accessibility to the field of DNA curvature investigation as AFM technology becomes more widespread.

CHAPTER 4: GENERATING AND EVALUATING THEORETICAL MODELS OF INTRINSIC DNA CURVATURE IN *TP53*

4.1 Introduction

The theoretical estimation of a number of different physical DNA parameters have been performed in AFM studies for over a decade. The creation of computer simulated DNA molecules has been important for estimating the error implicit in image analysis methods and for the generation and testing of hypotheses. The first standardised workflow for the generation of computer-simulated AFM images was put forward by Rivetti *et al.*, 1996. This approach has been adopted by many other researchers in a complete or modified form (Ficarra *et al.*, 2005a, 2005b; Marek *et al.*, 2005; Wiggins *et al.*, 2006; Buzio *et al.*, 2012). Intrinsic DNA curvature measurements can be added for improved hypothesis and method testing (Ficarra *et al.*, 2005b; Buzio *et al.*, 2012). No current studies have included a sequence specific flexibility parameter. Most studies rely upon a constant value of flexibility derived from the average persistence length of DNA of ~53 nm (Rivetti *et al.*, 1996).

Other theoretical measures have been used for comparison to AFM images including comparison of the theoretically determined pitch to DNA contour height (Milani *et al.*, 2011), curvature ratio profiles for base pair sequences (Buzio *et al.*, 2012), the prediction of promoter regions in AFM images (Marilley *et al.*, 2007b) and DNA flexibility (Scipioni *et al.*, 2002a; Marilley *et al.*, 2005; Wiggins *et al.*, 2006).

A number of dinucleotide wedge models have been used in AFM based studies of DNA. The two most often utilised by researchers are the De Santis and the Bolshoy models (De Santis *et al.*, 1988; Bolshoy *et al.*, 1991). The De Santis model used energy minimisation calculations to generate base pair parameters from gel electrophoresis experiments. It has been compared to real AFM measurements of DNA curvature by a number of groups and has unanimously been in good agreement under ambient (air) conditions (Anselmi *et al.*, 1999; Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Buzio *et al.*, 2012). The same holds true for the Bolshoy model, calculated from gel migration data, for liquid and air imaging (Sanchez-Sevilla *et al.*, 2002; Milani *et al.*, 2007, 2011; Buzio *et al.*, 2012). Both of these models have been compared by previous authors and were found to be comparable in the prediction of the position, but not magnitude, of curvature peaks (Buzio *et al.*, 2012). This was in agreement with a statistical analysis of the power of dinucleotide models to predict curvature in X-ray crystallography data that concluded that each dinucleotide model was as good a choice as any other for the prediction of intrinsic curvature (Crothers, 1998). The Olson model, based upon data mined from DNA–protein X-ray crystal complex experiments, has not been the subject of critical comparison to curvature profiles in any available publication (Olson *et al.*, 1998). However, it has been used to compare against flexibility measurements of DNA using AFM (Marilley *et al.*, 2005).

4.1.1 Aims and Objectives

The primary aim of the research in this chapter was to assess the theoretical curvature of the *TP53* gene that codes for the sequence-specific DNA-binding region of the p53 protein. To this end the De Santis model of curvature was used to predict the intrinsic curvature of *TP53*. This allowed for the statistical evaluation of the relationship between intrinsic DNA curvature and functional regions of the gene. Intrinsic curvature in regions of *TP53* that have been shown to exhibit slow DNA repair were considered separately. Other relevant physical theoretical measurements of *TP53,* such as nucleosome affinity, were also assessed.

The secondary aim of the research in this chapter was to generate computer simulated AFM images of *TP53* in order to make realistic predictions about intrinsic DNA curvature in real AFM images. To this end the De Santis and Olson dinucleotide wedge models were used to create computer simulated AFM images of *TP53*. The De Santis model has been compared to real AFM measurements of DNA curvature by a number of groups and has unanimously been in good agreement. The Olson model has previously been used to compare against flexibility measurements of DNA using AFM, but not curvature measurements. The De Santis model was included as a gold standard for comparability to AFM data. The Olson model was included to assess the effect the inclusion of DNA translations would have on the relevance of a model to experimental AFM measurements. Two different simulated deposition methodologies were tested. The same two overlapping *TP53* PCR product DNA sequences that were used formed the basis of the analysis described in later chapters. The resulting theoretical curvature profiles were statistically analysed to generate expectations for curvature measured from real *TP53* DNA.

4.2 Results

4.2.1 3D Model of *TP53*

3DNA allows for the visualisation, analysis and reconstruction of DNA *in silico* (Lu and Olson, 2008). The web interface for the application, w3DNA, was used to reconstruct 3D models of *TP53* DNA using a predefined set of dinucleotide parameters (Section 2.5.1.). 3D models are presented in Figure 4.1.

The resulting 3D models of *TP53* were relatively planar; the majority of both 3D models molecular structure lies in two dimensions. The Olson model (Figure 4.1.C+D) was observably less curved than the De Santis model (Figure 4.1.A+B.). The 5' sections of all of the molecules were relatively straight. The De Santis model showed regions of moderately large-scale curvature towards each end of the DNA fragment. There was a great deal more 'writhe' present in the De Santis model when compared to the Olson model.

4.2.2 Plane Fitting

In order to extrapolate a simplistic simulation of the geometric deposition of DNA onto a 2D surface it was necessary to fit a series of best fit (least squares) planes that allowed for no *xyz* coordinates to exceed a local deviation from the plane of best fit by more than 2 nm (Section 2.5.2.).

A plane was fitted for the De Santis model of *TP53* on average every 277 bp and every 416 bp for the Olson model (Figure 4.2.). The number of planes fitted was an indicator of the amount of 'writhe' within the 3D model. This confirmed the visual observations made in the previous section about the shape of each 3D model, *i.e.* that the De Santis model produced a more curved molecule with greater 'writhe' (Figure 4.2.A+B). Additionally, many of the planes fitted to the Olson model lie within a similar plane, which emphasised the planarity of molecules generated using Olson parameters (Figure 4.2.C+D).

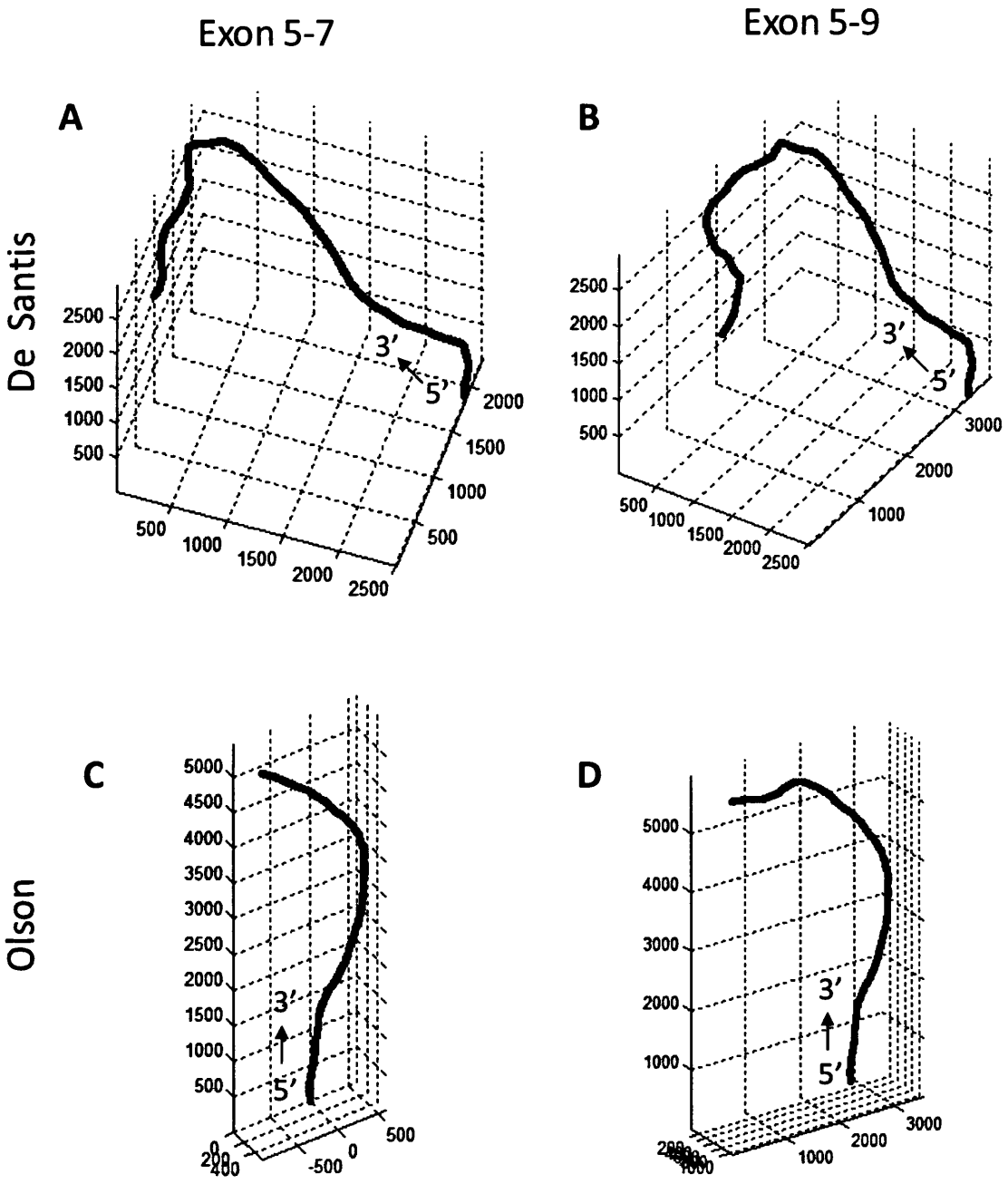

Figure 4.1. - 3D *TP53* DNA molecular orientations generated using two different dinucleotide parameter sets. Atomic coordinates were generated from w3DNA and the average *xyz* coordinate value for each base pair was plotted. Axis units are in angstroms (Å). A) *TP53* Exon 5-7 using De Santis parameters. B) *TP53* Exon 5-9 using De Santis parameters. C) *TP53* Exon 5-7 using Olson parameters. A) *TP53* Exon 5-9 using Olson parameters.

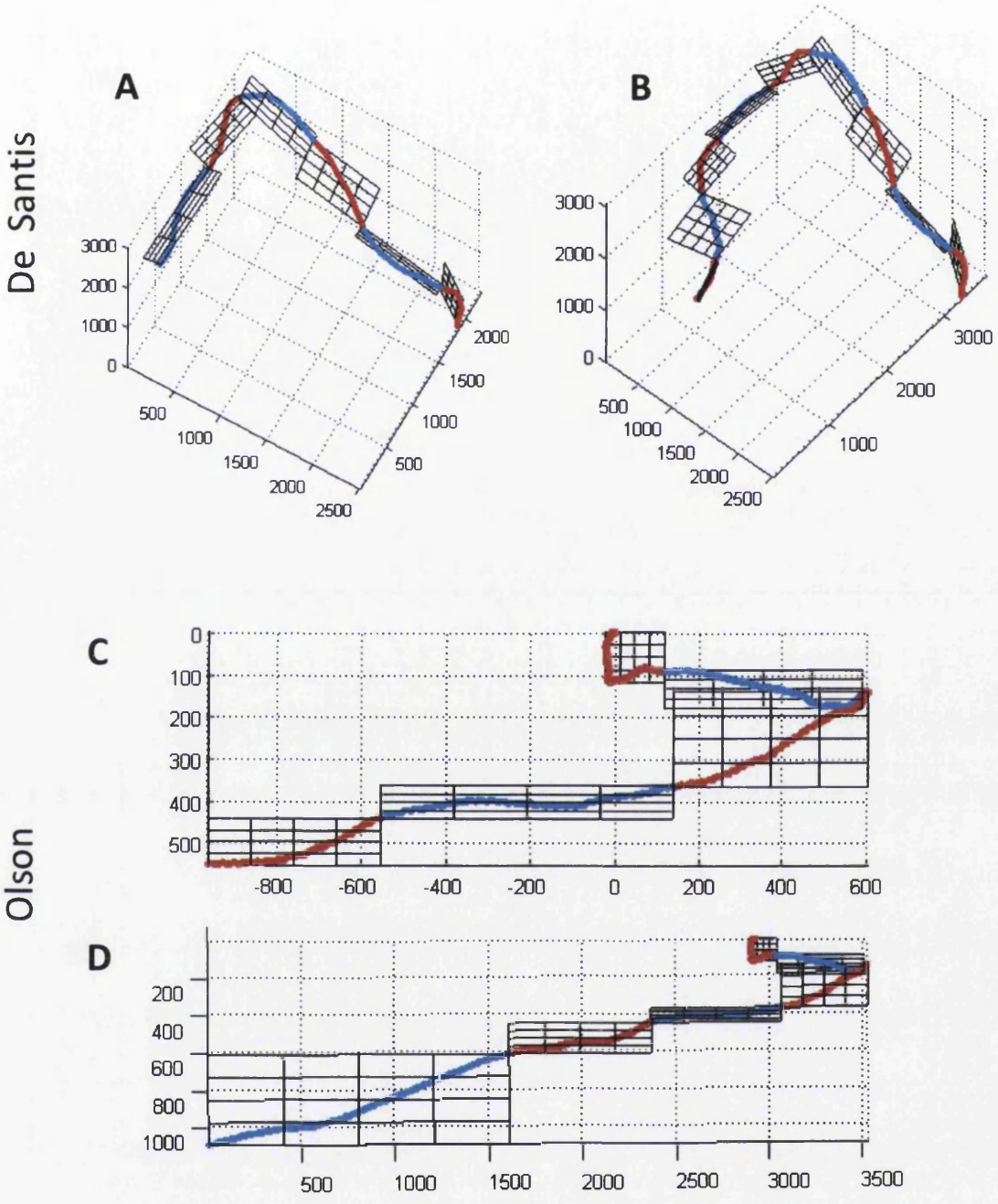

Figure 4.2. - Least squares plane fitted 3D *TP53* DNA generated using two different dinucleotide parameters sets. A) *TP53* Exon 5-7 using De Santis parameters – 6 planes. B) *TP53* Exon 5-9 using De Santis parameters – 9 planes. C) *TP53* Exon 5-7 using Olson parameters – 5 planes. D) *TP53* Exon 5-9 using Olson parameters – 6 planes. Planes are fitted with no more than 2 nm of local deviation from the plane. Atomic coordinates were generated from w3DNA. Axis units are in angstroms (Å). DNA in consecutive planes is been highlighted in red and blue.

4.2.3 Geometric Deposition of 3D Models onto a 2D surface.

The *xyz* coordinates detailed in the previous section were rotated along the line of intersection between succeeding planes until all *xyz* coordinates lay within one plane. The resulting *xyz* coordinates were projected onto a flat *xy* plane. A local correction was made at the point where two planes intersected as previously detailed (Section 2.5.2.).

It was observed that the Olson model of curvature contained a great deal less curvature than that predicted by the De Santis model. The Exon 5-7 molecule produced a consensus convex shape in both models. The projection of Exon 5-9 was in slight disagreement between both models, the final three 3' planes in the Olson model were orientated in a different direction to those in the De Santis model. This was most likely because the increased curvature in the De Santis model led to a preferential rotation of the seventh fitted plane (Figure 4.3.B). If this was not the case the De Santis model would have been expected to adopt a horse-shoe shape when deposited on a 2D surface.

The distance between each coordinate was calculated for each model and projection. The final 2D projection led to a length contraction of 2.6% for De Santis and 3.6 % for Olson model as compared to the original 3D molecule. This was a slight increase in contraction over the 1.5 % reported by previous authors on different DNA sequences (Buzio *et al.*, 2012).

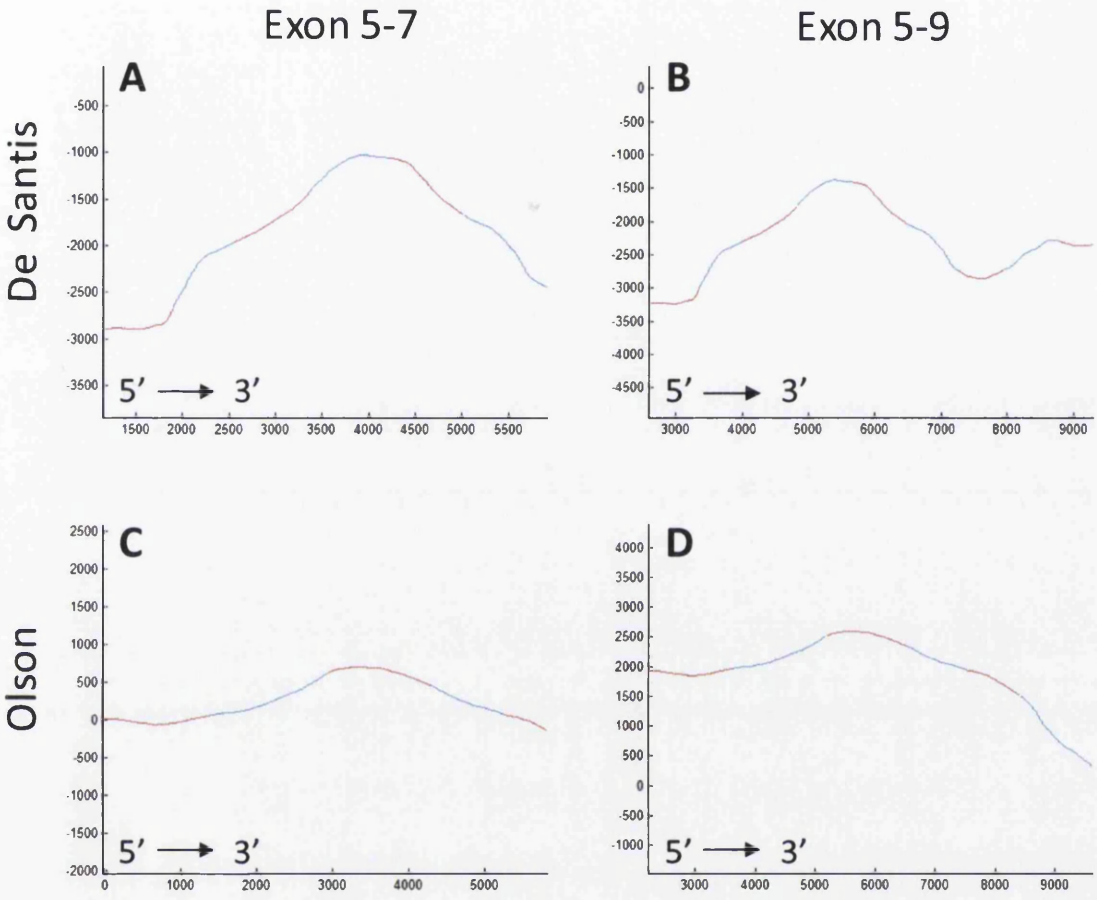

Figure 4.3. - 2D projection of 3D *TP53* DNA by plane fitting. A) *TP53* Exon 5-7 using De Santis parameters – 6 planes. B) *TP53* Exon 5-9 using De Santis parameters – 9 planes. C) *TP53* Exon 5-7 using Olson parameters – 5 planes. A) *TP53* Exon 5-9 using Olson parameters – 6 Planes. 3D coordinates were projected onto least squares fitting planes. Planes are rotated into alignment with one another. Planes were fitted with no more than 2 nm of local deviation from the plane. Atomic coordinates were generated from w3DNA. Axis units are in angstroms (Å). DNA in consecutive planes is highlighted in red and blue.

4.2.4 Creation of Computer-Simulated AFM Images of *TP53* for Curvature Analysis

Computer simulated AFM images were generated for both Exon 5-7 and Exon 5-9 using the De Santis and Olson dinucleotide wedge models. Datasets of over 1000 molecules were collected from the computer simulated AFM images for both sequences using the Geometric Deposition method. The correct orientation of each molecule was ascertained as described in Section 2.5.6. Three test datasets were generated:

- *Curvature Images* – The DNA molecules were generated using a fixed value of curvature at each base pair step identified from the De Santis and Olson dinucleotide wedge model (*i.e.* they were all identical). The only sources of image variation were the orientation of the DNA molecules, the effect of digitisation of the DNA contour and the effect of skeletonisation on the resulting AFM images.
- *Flexibility Images* - The DNA molecules were generated using a variable value of curvature at each base pair step. The mean value of the Gaussian distribution of curvature angles at each step was the same as that used in *Curvature Images*. The variation around the mean value was determined using a persistence length of 53 nm (Rivetti *et al.*, 1996). The flexibility of DNA molecules provided another source of variation in addition to that of contour digitisation.
- *Theoretical AFM Images* - The DNA molecules were generated in the same way as the *Flexibility Images* but also had both tip convolution (6 nm ROC) and Gaussian noise (variance = 0.025) added as additional sources of experimental variation. These images were the most comparable to real AFM images and had additional sources for potential variance between molecules as the images were subjected to noise filtering and automatic thresholding (*i.e.* all image processing steps of the ADIPAS software were applied).

4.2.5 Reconstructed Contour Length of Simulated DNA Molecules

The length of simulated DNA molecules in nanometres was calculated (Section 2.7.1.). Images were subjected to image noise and DNA molecule conformational flexibility (Section 4.2.4.) to investigate the effects that these conditions had on the final contour length estimates. The results are summarised in Table 4.1. and reconstructed length distributions in Figure 4.4.

Curvature images were as close to as invariable AFM images as possible. There was no tip convolution, no noise and no DNA molecule flexibility. The orientation of each molecule in 2D space led to a degree of variability based upon the digitisation of the DNA contour. The standard deviation for this set was the lowest for both samples. The standard deviation was larger for the Exon 5-9 dataset.

The idealised *Curvature* images exhibited a reconstructed length which was larger (Exon 5-7 - 3.60 %; Exon 5-9 - 3.49 %) than a theoretical value based upon 0.34 nm per base pair step. This length increase was likely due to the effects of contour digitisation. The second set of DNA molecules, *Flexibility* images, contained no noise or tip convolution but did contain a flexibility parameter *i.e.* each angle was selected at random from a Gaussian distribution with a mean angle taken from a theoretical curvature profile. An increase in the standard deviation of angles was observed and a shorter average length for both Exon 5-7 and Exon 5-9.

The final set of images, *Theoretical AFM* images, included a flexibility parameter, tip convolution (6 nm ROC) and Gaussian noise (variance = 0.025). They exhibited the smallest average reconstructed length and the largest deviation from the mean. This set also had the smallest difference between mean reconstructed length and theoretical length (0.61% and 0.43%).

Exon 5-7	Number of Molecules	Mean (nm)	Standard Deviation (nm)	Percentage Difference from Theoretical (%)
Theoretical	-	630	-	-
Curvature Images	1198	654	2.96	3.60
Flexibility Images	1253	646	4.48	2.40
Theoretical AFM Images	1171	634	5.97	0.61
Exon 5-9	**Number of Molecules**	**Mean (nm)**	**Standard Deviation (nm)**	**Percentage Difference from Theoretical (%)**
Theoretical	-	850	-	-
Curvature Images	1181	881	3.88	3.49
Flexibility Images	1046	870	5.18	2.29
Theoretical AFM Images	913	854	10.17	0.53

Table 4.1. - Summary of reconstructed length measurements of DNA molecules taken from images with various amounts of noise added. The theoretical length value was 0.34 nm per bp step. *Curvature* images were generated using angle values taken directly from the De Santis curvature profile. The *Flexibility* images had values taken from a Gaussian distribution generated using a persistence length of 53 nm. *Theoretical AFM* images were the same images as *Flexibility* images with tip convolution (6 nm ROC) and Gaussian noise (variance = 0.025) added. The De Santis dinucleotide wedge model was used as a basis for curvature measurements for each set.

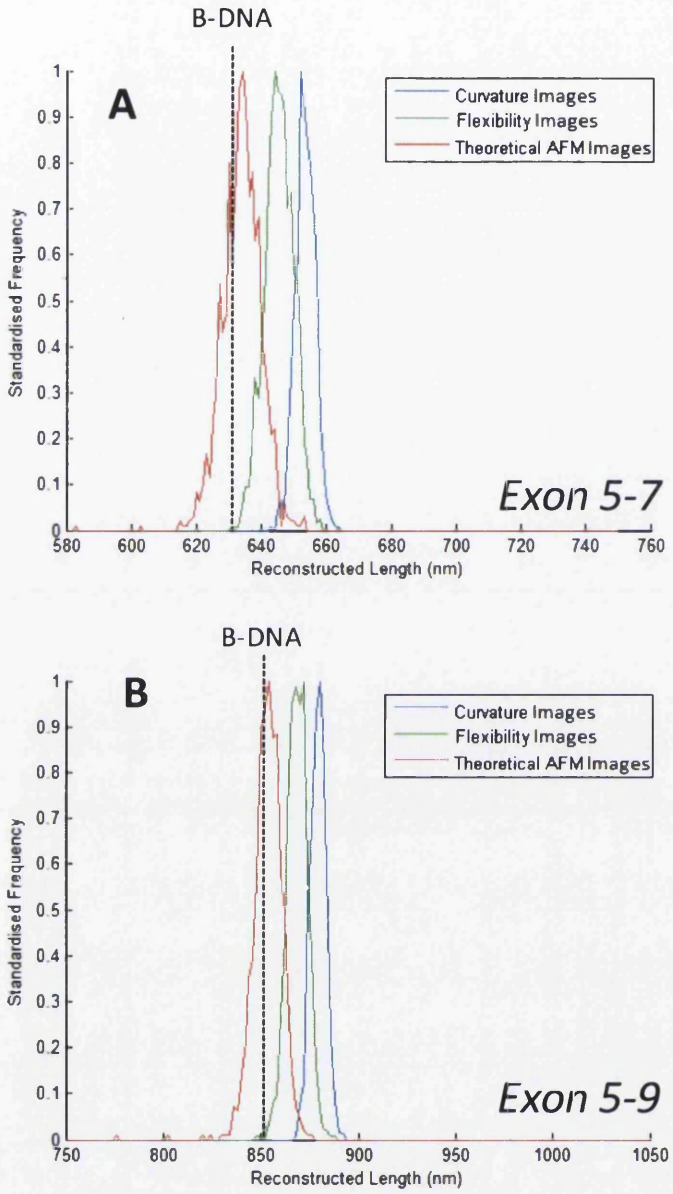

Figure 4.4. - Distribution of reconstructed length measurement of theoretical DNA molecules. A) *TP53* Exon 5-7. B) *TP53* Exon 5-9. *Curvature* images were generated using angle values taken directly from the De Santis curvature profile. The *Flexibility* images had values taken from a Gaussian distribution generated using a persistence length of 53 nm. *Theoretical AFM* images were the same images as the *Flexibility* set with tip convolution (6 nm ROC) and Gaussian noise (variance = 0.025) added. The De Santis model was used as a basis for curvature measurements for each set. The theoretical length values for B-DNA are indicated with a broken line.

4.2.6 Generation of Idealised Curvature Profiles after Image Processing

Within the literature the theoretical curvature profile has been often compared to experimental curvature profiles produced by an ensemble of AFM images. This approach is valid; however, there has been little consideration of the effects of contour digitisation or image processing on the resulting curvature profiles. By using the *Curvature* simulated *TP53* AFM images the effect of contour digitisation on curvature profiles was assessed. Only the trajectory of the DNA molecules in simulated images was different; so the only source of variation between curvature angles was caused by DNA contour digitisation. Both signed and unsigned curvature profiles were produced from the simulated images. The unsigned curvature profiles were comparable, at least in identification of peaks and troughs with the output of CURVATURE (Shpigelman *et al.*, 1993). The appropriate base pair window size was used when generating profiles in CURVATURE for comparison. The results of this comparison are summarised in Figure 4.5.

The De Santis model gave a larger average curvature value and a more curved theoretical DNA molecule (Section 4.2.2.). The general features of the simulated and theoretical curvature profiles at the 42 bp window of curvature were largely similar (Figure 4.5.B.). Many of the key peaks and troughs were retained after digitisation of the DNA molecule. There was little to no peak shift for the major peaks of curvature (this was quantitatively measured in a later section). While there were a few common features retained between the theoretical and experimental curvature profiles for the 21 bp window the lack of obviously large peaks within the theoretical profiles made visual comparison difficult (Figure 4.5.A.). Additionally, the more homogenous curvature of the 21 bp window size profile made this resolution of curvature unlikely to be suitable for further analysis.

The Olson model had lower average curvature than the De Santis model (Figure 4.5.C+D). There was little obvious similarity between the theoretical and simulated AFM profiles after contour digitisation and image processing. Peaks that occurred within the theoretical profile at ~0.4 and ~0.7 (Figure 4.5.D.) were not present in the curvature profile after contour digitisation. As there was little comparability between curvature profiles measured from simulated AFM images and theoretical curvature profiles in even these most idealised of AFM images the Olson model was not pursued further as a basis for comparison to experimental AFM images of DNA molecules.

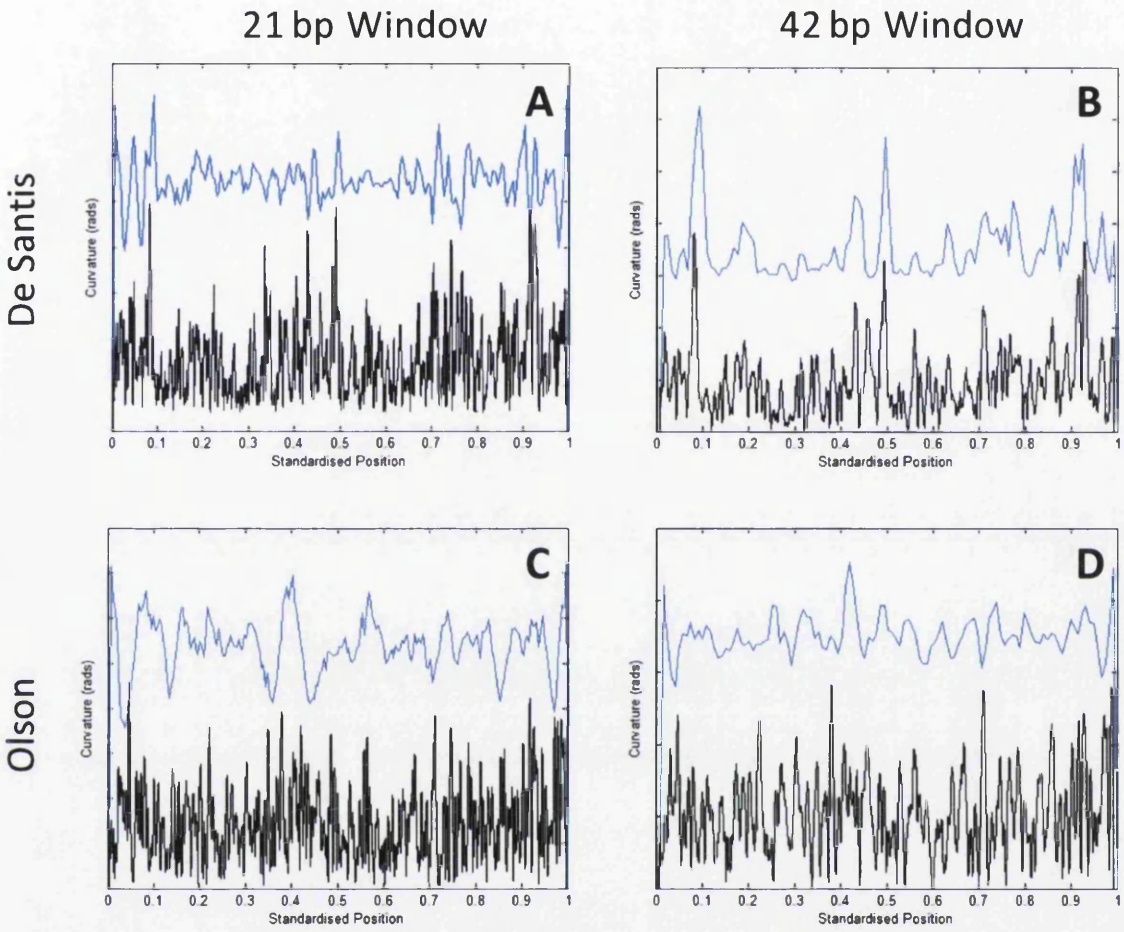

Figure 4.5. - Comparison of theoretical profiles generated in CURVATURE to curvature profiles generated from ideal computer generated AFM images of *TP53* Exon 5-9. Theoretical curvature profiles (black) were generated in CURVATURE using a 21 bp (A+C) and a 42 bp (B+D) window. Experimental profiles (blue) were produced from a large set (~1000 molecules) of AFM images containing computer generated DNA molecules with no deviation from an ideal curvature profile appropriate for the model comparison.

4.2.7 Effect of Imaging Conditions on Curvature Profiles.

The three simulated *TP53* AFM image sample sets with various degrees of molecule flexibility and image noise addition were used to assess effects of these factors on the resulting curvature profiles (Section 4.2.4.). All molecules within the datasets were correctly oriented. Curvature profiles for both signed and unsigned curvature were generated using a 42 bp window for calculating curvature angles (Figure 4.6).

In the unsigned curvature profiles (Fig 4.6.A+B) it was observed that the contrast between peaks was greater in the test set with no flexibility parameter (*Curvature* images). It was observed that the background curvature was higher in both profiles with sources of molecule flexibility or image noise (*Flexibility* and *Theoretical AFM* images) than the idealised *Curvature* images. Both *Flexibility* and *Theoretical AFM* images had approximately the same baseline curvature. As the *Theoretical AFM* images were *Flexibility* images with the addition of tip convolution and Gaussian noise then the increase in the baseline in comparison to *Curvature* images was attributed to the addition of flexibility to the simulated molecules. With the addition of flexibility, tip convolution and Gaussian noise the shape of the underlying profile was retained but the contrast between large peaks and troughs was reduced. The signed curvature profiles were observably very similar under all noise conditions and the characteristic shape of the curvature profile was retained (Fig 4.6.C+D.). A slight smoothing of the peak apex was observed indicating that there may have been a small amount of peak shift in the samples with additional sources of image noise and molecule variation (Section 4.2.13.).

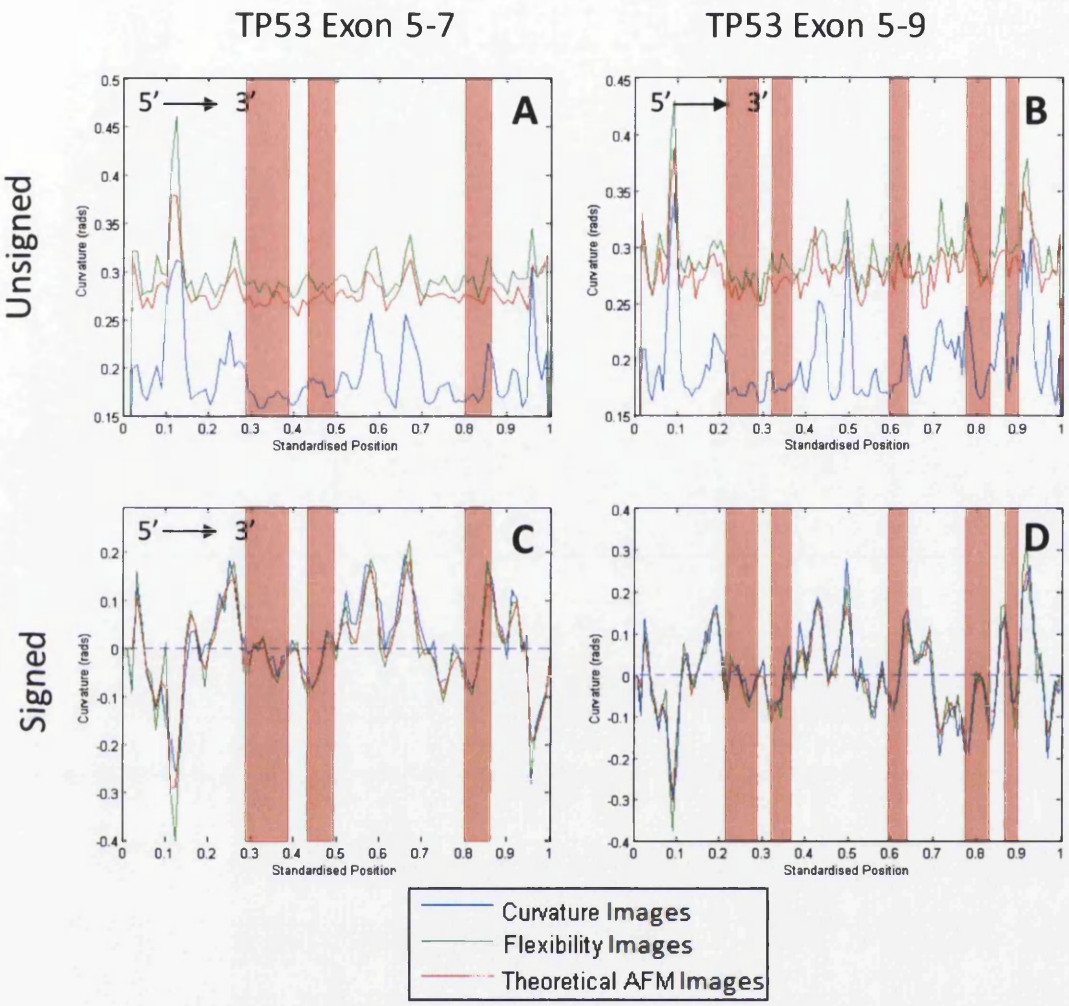

Figure 4.6. - Comparison of reconstructed curvature profiles from datasets that variously had DNA molecule flexibility and image noise. A) *TP53* Exon 5-7 unsigned curvature profiles. B) *TP53* Exon 5-9 unsigned curvature profiles. C) *TP53* Exon 5-7 signed curvature profiles. D) *TP53* Exon 5-9 signed curvature profiles. Unsigned curvature is calculated from absolute angles within the dataset. Exons are highlighted in red and read from left to right in ascending order. The window size of curvature is 42 bp.

4.2.8 Computer Generated Curvature Profiles for Comparison with Experimental AFM Images

By producing computer-generated AFM images of idealised DNA molecules that adhered perfectly to the theoretical curvature parameters of the De Santis model a realistic expectation for curvature profiles was generated for comparison with experimental data. Curvature profiles were generated at a base pair window size of 42 bp and used for comparison to experimental images in later chapters (Figure 4.7.). There were a number of prominent peaks that were expected to be retained in experimental *TP53* curvature profiles.

The unsigned curvature profiles considered all angles to be positive regardless of direction. As observed in the previous section, there was a more extreme effect of increasing variation on the unsigned curvature profiles than the signed curvature profiles. The resulting profiles exhibited reduced contrast on increasing noise. There were a number of key features retained in all the profiles that were expected to be observed in experimental profiles generated for real DNA molecules. These included: peaks of curvature preceding exon 5 and following exon 9, a number of large curvature peaks in the intronic region between exons 6 and 7 and a multitude of moderate to large peaks of curvature at the 3' end of the sequence. Perhaps the most important observation was that all exon positions occurred in regions of low curvature, with the exception of exon 7 which contained a small peak.

The signed curvature profiles of *TP53*, considered clockwise angles to be positive and anticlockwise angles to be negative. The Geometric Deposition method predicted that there would be two troughs of negative curvature at either end of the molecule and that the majority of the curvature in the rest of the molecule would be in the opposite (positive) direction to the end region curvature. Exon positions were in regions of low curvature (*i.e.* they were close to the dotted line denoting 0.0 radians of curvature).

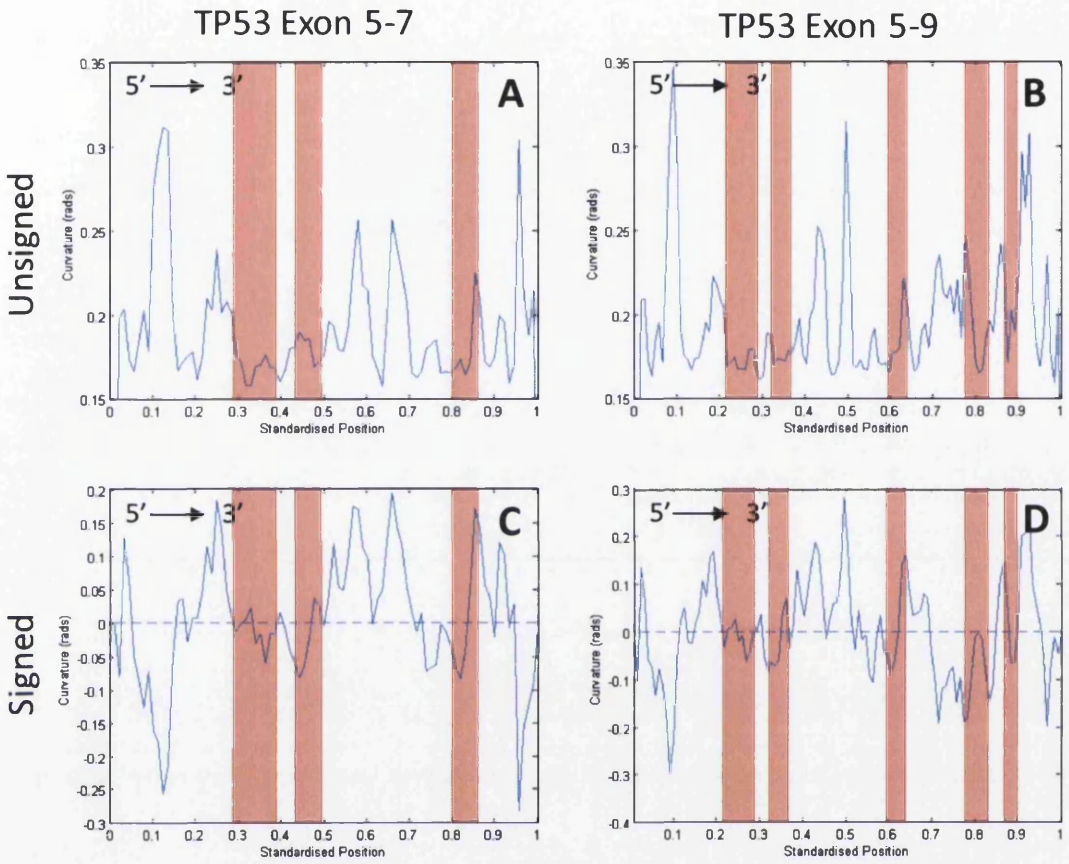

Figure 4.7. - Curvature profiles from simulated AFM images of *TP53* DNA at a 42 bp window of curvature for comparison to experimental *TP53* curvature profiles.

4.2.9 Signed Curvature Profiles Generated using the Scipioni Method

Scipioni *et al.,* estimated the likely deposition orientation of DNA molecules by mathematical methods (Scipioni *et al.*, 2002a). The same method was applied to *TP53* sequences. The underlying assumption was that the curvature modulus (magnitude) of a DNA tract would stay the same when the DNA tract is deposited on a 2D surface while the phase of curvature (direction) adapts to the changes in the DNA conformation (Scipioni *et al.*, 2002a). The curvature profiles were provided by the original authors by private communication (Figure 4.8.).

The maximum region of curvature for *TP53* Exon 5-7 was +0.16 radians at the 5' end of the sequence. The maximum region of curvature for *TP53* Exon 5-9 was ~+0.18 at the 3' end of the sequence. It was observed from both profiles that exon positions typically lay within regions of low curvature (close to the dotted line denoting 0.0 radians of curvature). All of the major peaks in curvature occurred during intronic positions.

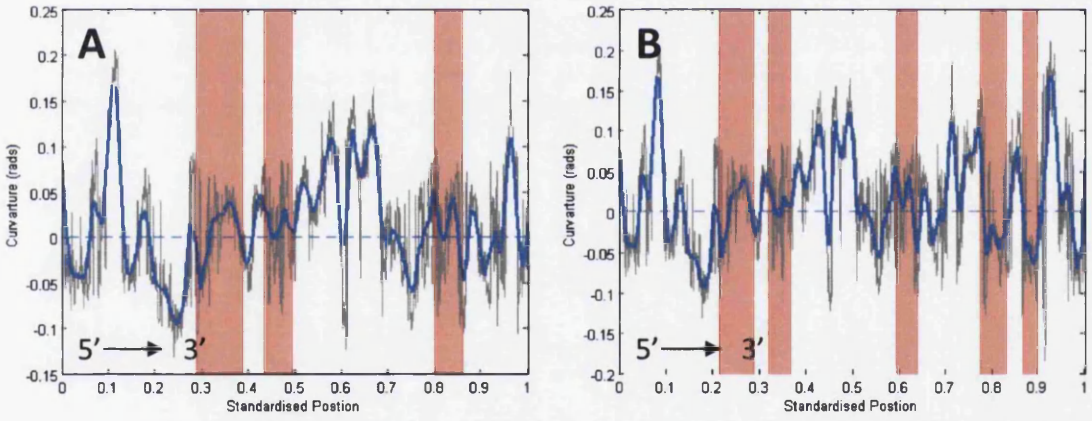

Figure 4.8. - Signed curvature profile generated using the method detailed by Scipioni *et al.,* using the De Santis model of curvature (Scipioni *et al.*, 2002a). A) *TP53* Exon 5-7. B) *TP53* Exon 5-9. Curvature is plotted against standardised position along the DNA sequence. Exon positions are indicated by shaded red areas and are read in ascending order (*e.g.* Exon 5-7 reads from left to right exon 5, 6 then 7). The grey line represents the raw data calculated for every base pair. The blue line is a smoothed profile averaged over 42 base pairs.

4.2.10 Comparison of 2D Deposition Methodologies

Both the methods detailed by Buzio *et al.,* (Buzio *et al.*, 2012) and Scipioni *et al.*, (Scipioni *et al.*, 2002a) were applied to model the likely deposition of *TP53* on a flat surface. The first method used a geometric minimisation approach to model likely deposition onto a flat surface. The second method used the phase of DNA as a guideline for the direction of DNA. Both models used the De Santis model of curvature. The comparison of the results using a 42 bp window size over which curvature angles were calculated is displayed in Figure 4.9. There were notable differences between the two predictions. Firstly, the Geometric Deposition model included the additional noise of digitisation; which had the effect of increasing the theoretical curvature at all large curvature peaks and increased the width of a number of large peaks. The models predicted opposite curvature at the terminal ends of *TP53* Exon 5-9 and the 5' end of *TP53* Exon 5-7.

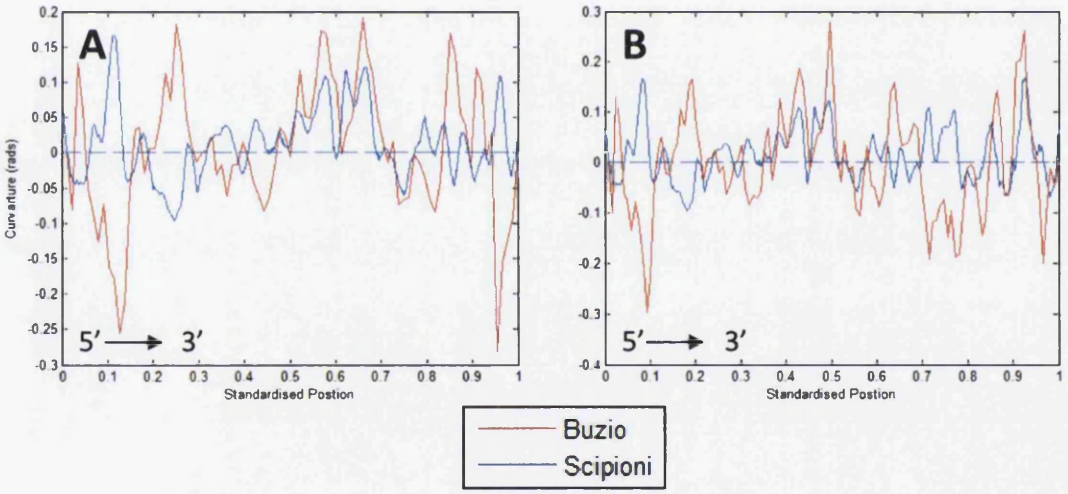

Figure 4.9. - Comparison of signed curvature profiles generated using the method detailed in Scipioni *et al.,* 2002 (blue line) and the method detailed by Buzio *et al.,* 2012 (red line) using the De Santis model of curvature. A) *TP53* Exon 5-7. B) *TP53* Exon 5-9.

4.2.11 Analysis of Correlation between Curvature Profiles of Simulated AFM Images of *TP53*

A correlation analyses was performed between curvature profiles produced by the *Curvature* and *Theoretical Image* samples. This allowed for an assessment of the effects of DNA molecule flexibility and image noise on the reproducibility and comparability of curvature profiles. This was repeated for curvature profiles calculated using different base pair window sizes for the calculation of curvature (21 bp, 42 bp and 63 bp). The majority of profiles exhibited a non-normal distribution (Shapiro-Wilks, $p = <0.05$) so a Spearman's Rank correlation test was used (Table 4.2.).

Each base pair window size exhibited significant correlation between noisy and non-noisy profiles. All correlation coefficients were positive, indicating a positive correlation. The strength of the positive trend increased on increasing base pair window size in both signed and unsigned curvature profiles. The correlation measured was weaker in the unsigned curvature profiles at all comparable window sizes. The 21 bp window size for unsigned curvature exhibited a very weak positive trend (Rho = 0.15) in comparison to other profiles.

Correlation analysis was also performed on the overlapping portions of the Exon 5-7 and Exon 5-9 (Table 4.2.). Three base pair window sizes were analysed: 21 bp, 42 bp and 63 bp. The profiles were created using 'noisy' *Theoretical AFM* images (Section 4.2.4.). Each base pair window size for both signed and unsigned profiles showed positive significant correlation of varying strength. The signed curvature profiles all exhibited a very strong positive trend (Rho = > 0.9). The unsigned profiles exhibited a positive trend that increased on increasing window size. The unsigned 21 bp profile showed the lowest significant positive correlation (Rho = 0.26) of all window sizes.

Base Pair Window Size	Correlation with and without image and measurement noise		Correlation between overlapping Exon 5-7 and Exon 5-9 profiles	
Unsigned Curvature	**Spearman's Correlation (Rho)**	**p-value**	**Spearman's Correlation (Rho)**	**p-value**
21 bp	0.15	<0.05	0.26	<0.05
42 bp	0.69	<0.05	0.50	<0.05
63 bp	0.89	<0.05	0.76	<0.05
Signed Curvature				
21 bp	0.88	<0.05	0.90	<0.05
42 bp	0.95	<0.05	0.95	<0.05
63 bp	0.98	<0.05	0.96	<0.05

Table 4.2 – Correspondence analysis using Spearman's Rank correlation applied to simulated AFM images with and without noise addition and between overlapping sections of Exon 5-7 and Exon 5-9 curvature profiles. The first (columns 3 and 4) comparison was between simulated image with and without image and measurement noise. Sources of noise were: a flexibility parameter for the prediction of DNA conformation (ξ = 53), tip convolution (ROC = 6 nm) and Gaussian noise (variance = 0.025) added to the final images. The second comparison was between overlapping sections of curvature profiles for Exon 5-7 and Exon 5-9. All profiles had image and measurement noise added as described above.

4.2.12 Comparison of Peaks Estimated from CURVATURE and Curvature Reconstructed for Computer Simulated AFM Images of *TP53*

A curvature profile was calculated for *TP53* Exon 5-9 using CURVATURE on default settings and using the De Santis model of curvature (De Santis *et al.*, 1988; Gohlke, 1994). The ten peaks with the highest curvature values were identified. This was repeated using a curvature profile reconstructed from computer simulated AFM images for *TP53* Exon 5-9 at a 42 bp window of curvature. The curvature profile reconstructed from the computer simulated AFM images used a signed profile (considered direction of curvature) in which all of the angles had been made absolute (positive); this was the most comparable the profile could be made to the output of CURVATURE (Figure 4.10).

The ten largest peaks were identified from both profiles. A comparison of peaks showed that there were notable differences. A number of peaks that occurred in the CURVATURE profile were merged into one peak (0.5 and 0.93 standardised length), which was expected considering the difference between the resolution of the curvature profiles (CURVATURE provided a resolution of one measurement per dinucleotide whereas theoretical AFM images provided a resolution of one point per 21 bp). There were a number of peaks within the reconstructed profile that were not present at comparably large magnitudes within the CURVATURE profile (1.8, 6.3 and 0.97 standardised length). Additionally, there was a peak within the CURVATURE profile that was not present within the reconstructed profile at a comparable magnitude (0.34 standardised length).

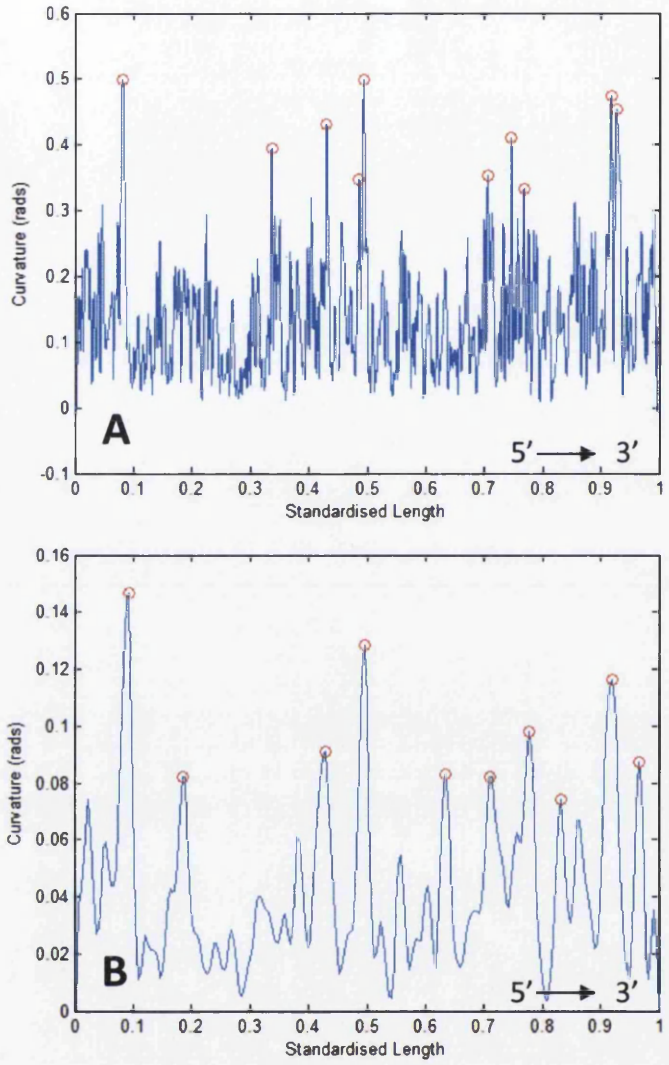

Figure 4.10. - Comparison of ten major peaks within curvature profiles generated using different approaches. A) Curvature profile produced from CURVATURE (Shpigelman *et al.*, 1993). B) Curvature profile reconstructed from computer simulated AFM images using a 42 bp window of curvature. Peaks were identified as the ten peaks with largest curvature value within a profile. Peaks are indicated with red circles.

4.2.13 Estimation of Peak Shift after Noise Addition

The ten largest peaks of curvature were identified in the intrinsic DNA curvature profiles reconstructed from an ideal set of computer simulated AFM images with no flexibility parameter (*Curvature* images – Section 4.2.4.). The peaks that most closely corresponded to these key peaks from curvature profiles produced from AFM image with a degree of molecule variation and image noise were recorded (*Theoretical AFM* images – Section 4.2.4.). The peak shift was calculated as a percentage value for three base pair window sizes of curvature: 21 bp, 42 bp and 63 bp (Figure 4.11).

The 21 bp window of curvature produced a mean peak shift of 0.63 %. However, it was necessary to apply a Savitzky-Golay smoothing filter (9 degrees, 15 points) to identify the corresponding peaks. The maximum peak shift detected was 1.2 % or 31.5 bp and visually the pattern of peaks was dissimilar to the noiseless images. Only 8 of the 10 peaks were accurately identified in the noisy image, indicating that two of the peaks had merged or been lost. The magnitude of curvature at the detected peaks was not significantly different (Paired t-test, $t = 0.75$, $p = 0.47$).

The 42 bp window of curvature produced a mean peak shift of 0.59 %. The maximum peak shift detected was 0.84 % or 21 bp. Visually, the pattern of peaks was similar between the images. The two peaks at the 3' end of the molecule in the noiseless profile had merged into one peak in the noisy image. The magnitude of curvature at the detected peaks was not significantly different (Paired t-test, $t = 0.14$, $p = 0.89$).

The 63 bp window of curvature produced a mean peak shift of 0.25 %. The maximum peak shift detected was 1.3 % or 31.5 bp. Visually the pattern of peaks was similar between the images and all peaks were identified. The magnitude of curvature at the detected peaks was significantly different between the profiles (Paired t-test, $t = 4.37$, $p = <0.05$).

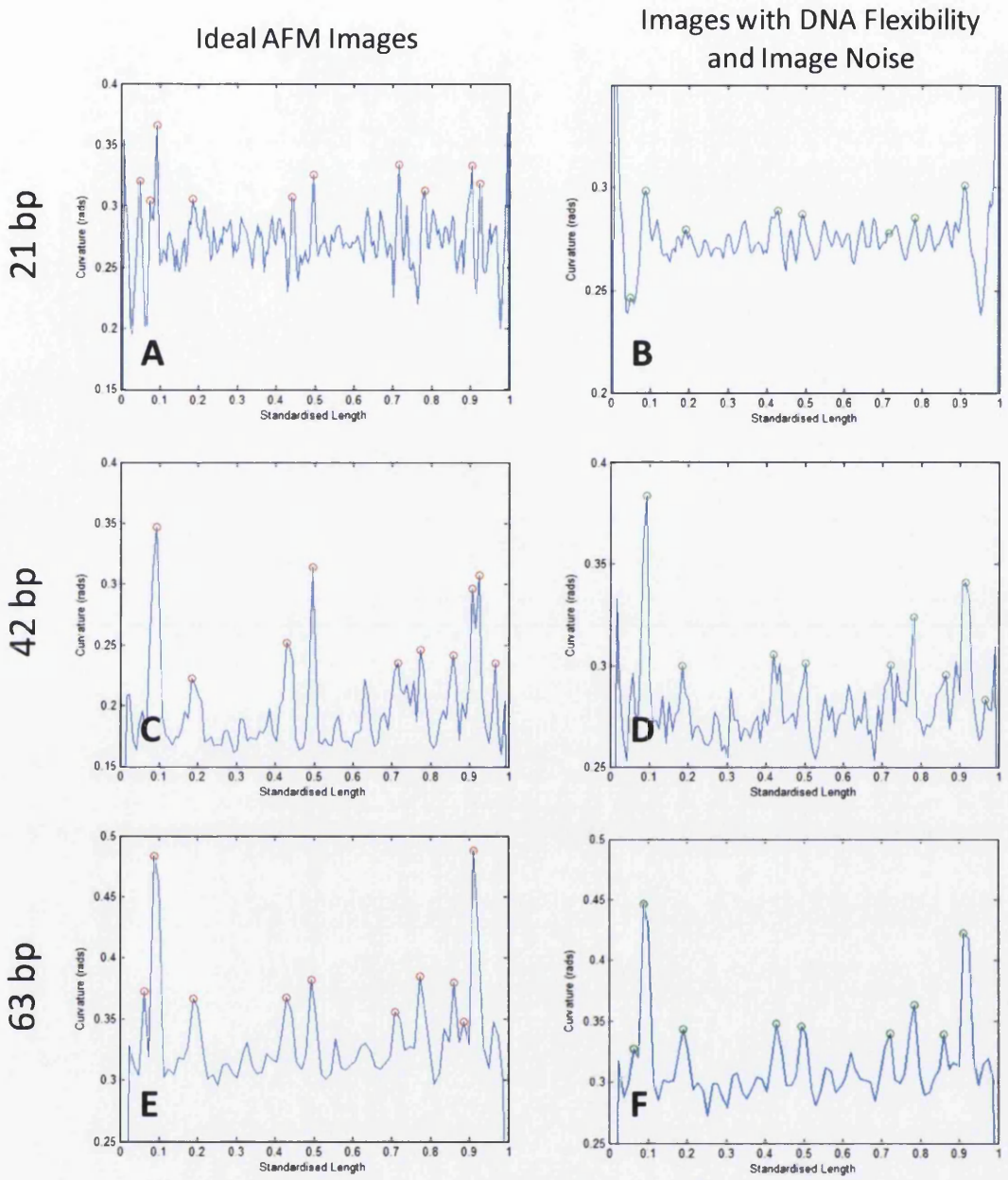

Figure 4.11. Comparison of ten major peaks from curvature profiles reconstructed from simulated AFM images of *TP53* before and after the addition of DNA molecule flexibility and image noise. Three base pair window sizes of curvature with and without sources of image noise are presented. Sources of noise were: a flexibility parameter for the prediction of DNA conformation (ξ = 53), tip convolution (ROC = 6 nm) and Gaussian noise (variance = 0.025) added to the final images. Peaks are identified with a red circle.

4.2.14 Curvature and Regions of Slow Repair.

Six regions of slow DNA repair within the *TP53* gene were identified from the available literature: codons 177, 196 and 278 in skin cancer (Tornaletti and Pfeifer, 1994) and codons 157, 248 and 273 in lung cancer (Denissenko *et al.*, 1998). These codons are also common mutation hotspots. Curvature values were calculated from CURVATURE for *TP53* Exon 5-9 using the defaults setting and the De Santis model of curvature (Shpigelman *et al.*, 1993). Curvature values for the three nucleotides in each codon position (n = 18) were statistically compared to the rest of the sequence (n = 2482) using the Kruskal-Wallis test. Regions of slow repair showed significantly lower median curvature (Kruskal-Wallis, $p = <0.05$) than the rest of the profile (Figure 4.12).

The Kruskal-Wallis test was performed on curvature values that corresponded to the regions of slow repair from curvature profiles generated from simulated AFM images of *TP53* using a 42 bp window of curvature. The curvature for region of slow repair in the signed profile was not significantly different from curvature throughout the rest of the gene (Kruskal-Wallis, $p = 0.32$). The curvature for region of slow repair in the unsigned profile was not significantly different from curvature throughout the rest of the gene (Kruskal-Wallis, $p = 0.06$).

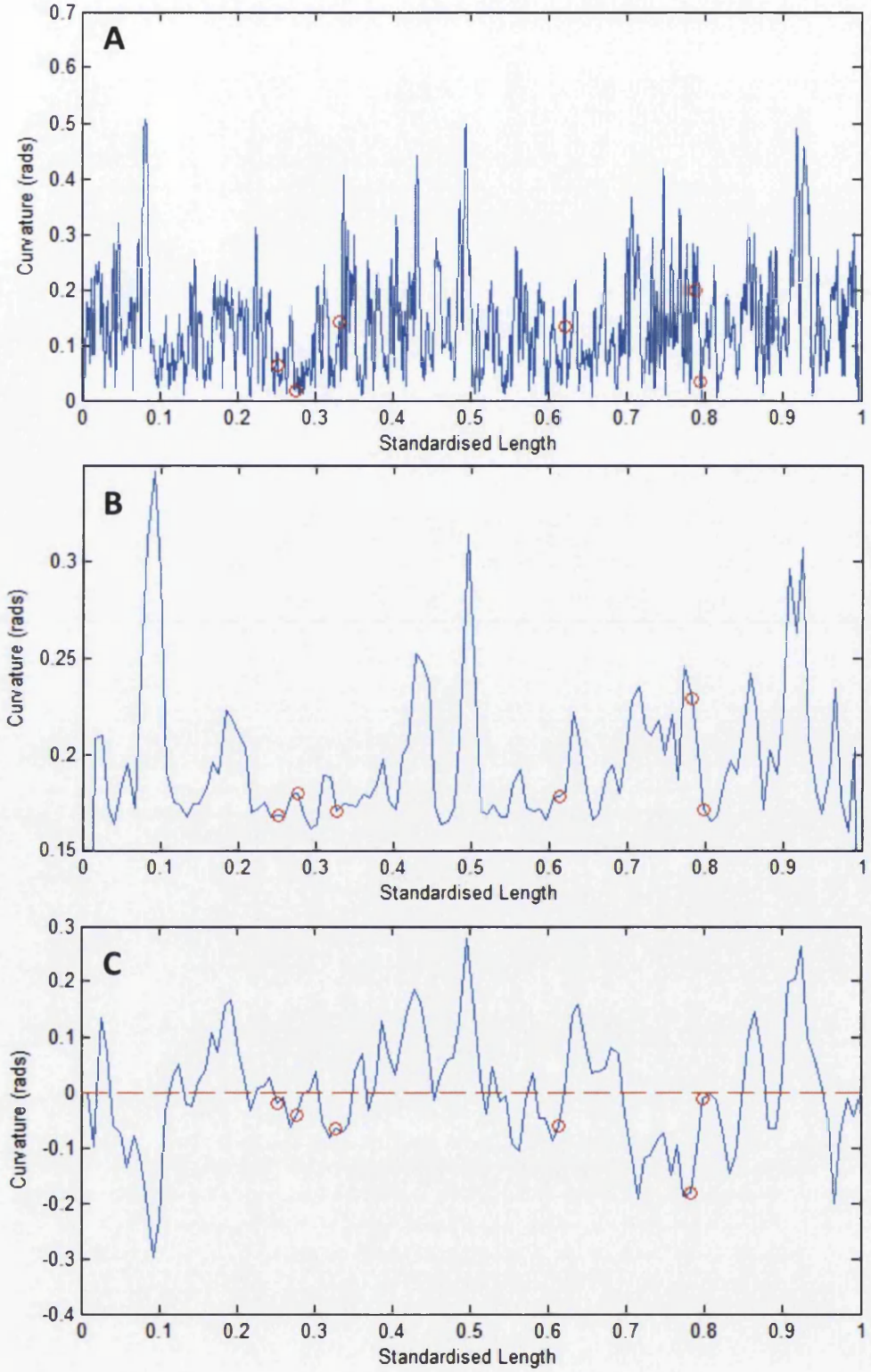

Figure 4.12. - Comparison of curvature profiles with regions of slow repair in *TP53* Exon 5-9. A) Curvature profile generated using CURVATURE (De Santis curvature, default settings) B) Unsigned curvature profile from simulated AFM images using a 42 bp window of curvature (De Santis curvature, no flexibility or noise). C) Signed curvature profile from simulated AFM images using a 42 bp window of curvature (De Santis model, no flexibility or noise). The length of the DNA sequence was standardised on a scale of zero to one; zero is 5' and 1 is 3' end of the sequence. Regions of slow repair are indicated with small red circles.

4.2.15 Statistical Comparison of Exon Curvature to Intron Curvature

Curvature values that lay within the exon positions as designated by the IARC database were statistically compared to intron positions (Hernandez-Boussard *et al.*, 1999). The distribution of data was largely non-parametric and the Kruskal-Wallis test was used to test comparisons. Each curvature value within the exon boundaries in standardised length was designated as 'exon'.

The curvature values of exon positions were pooled and compared against intron positions (Table 4.3.). Exons had significantly lower curvature in unsigned 42 bp and 63 bp profiles. Exons had significantly lower curvature in the signed 21 bp profile. A curvature profile generated using CURVATURE had significantly lower curvature in exons than introns.

The individual exons were compared to intronic regions *i.e.* all values that did not lie within an exon region (Table 4.4.). Curvature values were calculated from CURVATURE for *TP53* Exon 5-9 using the defaults setting and the De Santis model of curvature (Shpigelman *et al.*, 1993). The median values for exon and intron regions were compared. Exons 5, 6 and 7 each had significantly lower curvature than intron regions. Exon 8 and 9 did not exhibit significantly different curvature from intronic regions.

Curvature values for *TP53* Exon 5-9 were generated from simulated AFM images from the 'noisy' *Theoretical AFM Images* sample at three windows of curvature: 21 bp, 42 bp and 63 bp. Exons in signed curvature profiles had no significantly different curvature than from introns. Exon 5 had a significantly lower curvature than intron regions in the 42 bp and 63 bp windows of curvature.

		Kruskal-Wallis (p-value)		
Exon 5-9	**Window Size**	**Unsigned Curvature**	**Signed Curvature**	**CURVATURE (32 bp)**
	21 bp	0.40	<0.05	-
	42 bp	<0.05	0.25	<0.05
	63 bp	<0.05	0.24	-

Table 4.3. - Summary of the Kruskal-Wallis test applied to the pooled curvature and flexibility of exon positions to the pooled curvature and flexibility of intron positions. The distribution of data points was non-normal and not size matched; a Kruskal-Wallis test was used to test for significant differences between median values. Significant p-values are highlighted in red.

CURVATURE							
Base Pair Window (bp)		**Exon 5**	**Exon 6**	**Exon 7**	**Exon 8**	**Exon 9**	**Intron**
32 bp	*Number of Sample Points*	184	113	110	137	63	-
	Median Curvature (rads)	0.06	0.10	0.10	0.18	0.15	0.13
	Kruskal-Wallis (p)	<0.05	<0.05	<0.05	0.08	0.06	-
Absolute Curvature							
Base Pair Window (bp)		**Exon 5**	**Exon 6**	**Exon 7**	**Exon 8**	**Exon 9**	**Intron**
21 bp	*Number of Sample Points*	18	11	10	13	6	-
	Median Curvature (rads)	0.29	0.30	0.30	0.30	0.30	0.30
	Kruskal-Wallis (p)	0.06	0.99	0.69	0.99	0.36	-
42 bp	*Number of Sample Points*	8	6	6	6	3	-
	Median Curvature (rads)	0.17	0.17	0.18	0.18	0.19	0.19
	Kruskal-Wallis (p)	<0.05	0.10	0.66	0.40	0.82	-
63 bp	*Number of Sample Points*	6	3	3	4	2	-
	Median Curvature (rads)	0.31	0.31	0.33	0.32	0.34	0.32
	Kruskal-Wallis (p)	<0.05	0.08	0.92	0.49	0.30	-
Signed Curvature							
Base Pair Window (bp)		**Exon 5**	**Exon 6**	**Exon 7**	**Exon 8**	**Exon 9**	**Intron**
21 bp	*Number of Sample Points*	18	11	10	13	6	-
	Median Curvature (rads)	-0.01	-0.02	-0.02	-0.00	-0.04	-0.00
	Kruskal-Wallis (p)	0.29	0.24	0.78	0.08	0.13	-
42 bp	*Number of Sample Points*	8	6	6	6	3	-
	Median Curvature (rads)	-0.01	-0.05	-0.02	-0.05	-0.07	-
	Kruskal-Wallis (p)	0.68	0.61	0.87	0.10	0.67	-
63 bp	*Number of Sample Points*	6	3	3	4	2	-
	Median Curvature (rads)	-0.02	-0.00	-0.03	-0.05	-0.01	-0.01
	Kruskal-Wallis (p)	0.37	0.68	0.95	0.25	0.92	-

Table 4.4. - Summary of statistical analysis of comparisons between curvature measurements of exon and intron positions for *TP53* Exon 5-9. The distribution of data points was non-normal and not size matched; a Kruskal-Wallis test was used to test for significant differences between median values. Significant p-values are highlighted in red.

4.2.16 Nucleosome Positioning Using Theoretical Models

It was observed that a number of exonic regions within *TP53* exhibited very low theoretical curvature. The possible physiological significance of this observation was investigated. Low intrinsic curvature is a feature of DNA that is unlikely to be included within the nucleosome (Figure 4.13.). A number of motifs that are unfavourable for binding by histones have been compiled (Luykx *et al.*, 2006). These motifs were termed Nucleosome Exclusion Sites (NXS) by the authors. The NXSensor software package was created that identifies these exclusion sites and predicts regions that are unlikely to be bound by nucleosomes based upon the proximity of NXS. Approximately 147 base pairs are wrapped around the nucleosome core. The presence of two NXS within 147 base pairs indicates a region that is unlikely to be occupied by a nucleosome. The NXsensor software was applied to the *TP53* DNA sequences and a number of NXS were identified. Of particular interest were two NXS in close proximity to one another at the beginning and end of Exon 5 (Figure 4.13.C). This indicated that theoretically nucleosomes were unlikely to occupy Exon 5. The other NXS occurred within introns and none were close enough together to create a nucleosome exclusion region.

The NuPop algorithm was also applied to *TP53* (Xi *et al.*, 2010). This method explicitly models the nucleosome linker DNA and was trained on nucleosome positioning data from *S. cerevisiae.* The outcome of the model was a probability value for the start of nucleosome occupancy (Figure 4.13.A.) and an occupancy score (Figure 4.13.B.). The results for *TP53* showed a regular pattern of likely nucleosome occupancy regions. The major intron regions were likely to be wrapped up in the histone core. All of the exons were predicted to be occupied by nucleosomes with the exception of exon 6 which contains a central region of low occupancy. The 5' border of exon 5 also contained a region that was highly unlikely to be occupied by a nucleosome.

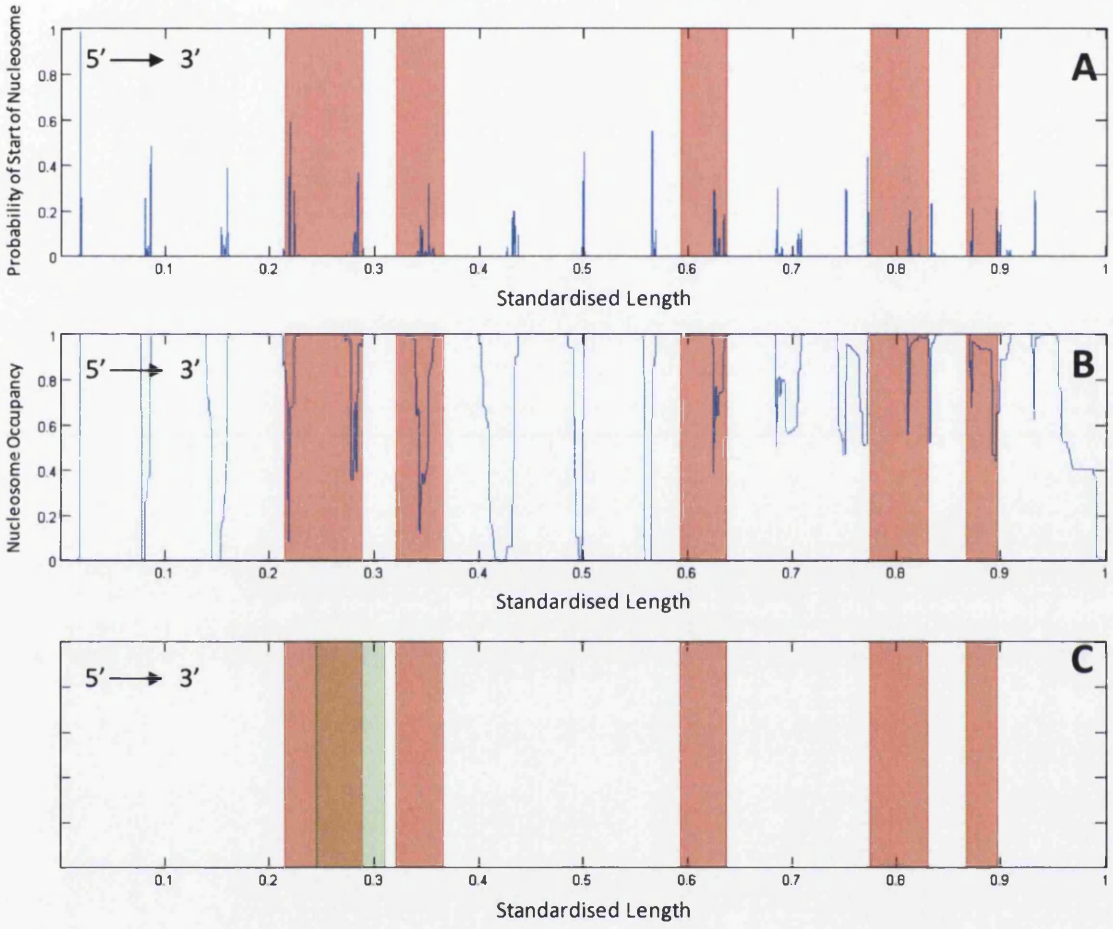

Figure 4.13. - Summary of nucleosome positioning algorithms applied to *TP53*. A) The probability of the start of nucleosome binding from NuPop. B) Nucleosome occupancy score from NuPop. C) Nucleosome exclusion sites from NXSensor (green section). The nucleosome occupancy score is from 0, low likelihood of occupancy, to 1, high likelihood of occupancy. Only the most prominent nucleosome exclusion site is shown for NXSensor, all other exclusion sites were scattered throughout intron positions. Exons 5 to 9 are highlighted in red and read in ascending order from left to right (5′ to 3′).

4.3 Discussion

4.3.1 Simulating Deposition of 3D *TP53* models onto 2D surfaces

The two dinucleotide parameter sets were used to model 3D molecules of *TP53* (Section 2.5.1.). The 3D molecules had the same shape but a considerable difference in the magnitude of curvature. This suggested that the models predicted curvature in the same regions but differed in the estimation of the magnitude of curvature. The De Santis prediction of *TP53* 3D shape was more curved than the Olson prediction and required more planes to be fitted in order to 'flatten' it into 2D (Figure 4.2 - Planes fitted every: De Santis = 277 bp; Olson = 416 bp). For comparison, in the work by Buzio *et al.,* the authors fitted a plane approximately once every 222 bp *i.e.* 6 planes were fitted to a 1332 bp sequence (Buzio *et al.*, 2012). This was shown to have implications for the simulated deposition methods discussed later and the applicability of the Olson model to the study.

There were two available methodologies for projecting 3D models of DNA onto 2D surfaces. The first was a mathematical model that assumed that during deposition of DNA the local intrinsic curvature remained the same while the curvature phase changed to accommodate modifications in DNA architecture (Scipioni *et al.*, 2002a). This was kindly generated for *TP53* sequences by the original author in a private communication (Scipioni *et al.*, 2002a). The second approach, Geometric Deposition, assumed that DNA would undergo the least possible conformational changes in order to equilibrate on the mica (Buzio *et al.*, 2012). There were benefits and limitations to both methods. The Geometric Deposition method was complex to implement, but required minimal understanding of underlying theory. The phase method required a complete understanding of the underlying mathematics of DNA curvature and flexibility. It was for this reason that phase curvature profiles were generated for this study by the original authors. This has also been the case for other studies, perhaps indicating that this methodology is too complex for common usage (Buzio *et al.*, 2012).

Although both methods have different underlying assumptions they have been observed to produce very similar results (Buzio *et al.*, 2012). Application of both methods to *TP53* indicated that there were two regions of disagreement between the models about the direction (phase) of DNA curvature: the central intron region between exon 6 and exon 7 and the region beginning during exon 8 until the end of the sequence (Section 4.2.10.). These differences were likely to be due to conformational changes introduced by Geometric Deposition to accommodate the larger scale curvature of the De Santis model. Figure 4.3 illustrates this; the Olson model, with a lesser degree of curvature, had a difference in DNA direction at the 3' end of Exon 5-9 in comparison to the De Santis model. The large degree of twist needed to flatten the 3' end section of the De Santis model is likely to account for the

discrepancy between the deposition predictions. It should be noted that both models were just predictions of possible molecule orientation and experimental results can been used to amend the predictions, such as by modifying the direction of the predicted curvature to match experimental results (Buzio *et al.*, 2012).

4.3.2 Evaluation of a Suitable Dinucleotide Parameter Set for Comparison to Experimental AFM Images

Curvature profiles produced from computer simulated AFM images using the De Santis and Olson dinucleotide wedge parameters were compared (Section 4.2.6.). It quickly became apparent that the Olson model would be unsuitable for studying *TP53*. The De Santis parameters produced an unsigned profile in good agreement with that generated by other methods. The Olson model produced a homogenous, featureless profile that did not exhibit characteristic peaks that had been predicted by CURVATURE. The curved regions predicted by the Olson model were obscured by even the minimal noise introduced by digitising the DNA contour. An initial target of this study was to obtain a sequence-specific flexibility profile for *TP53* using the Olson model. Where communication was achieved, authors were unable to provide this study with a clear methodology for calculating flexibility from dinucleotide parameters or to calculate these values for *TP53* DNA sequences (Olson *et al.*, 1998; Marilley *et al.*, 2005). It is possible that had the flexibility parameters corresponding to the Olson model been included then the model would have been more suitable for comparison to AFM imaging. The De Santis model provided a well tested and robust dinucleotide parameter set for comparison to real AFM images and was used exclusively within the current study.

4.3.3 Evaluation of the Effects of Digitisation of DNA Contour Length

The effect of digitisation alone increased the contour length of DNA as measured by the Kulpa estimator (Section 4.2.5.). The overestimation decreased on the addition of a flexibility constant to the DNA molecules and again on addition of Gaussian noise/tip convolution. The percentage difference values for the datasets with the most sources of variation was 0.61 % and 0.53 % for Exon 5-7 and Exon 5-9 respectively (Table 4.1). This presented a better agreement between reconstructed length and theory than the -1.2% underestimate observed by previous authors (Rivetti and Codeluppi, 2001). The introduction of flexibility to the DNA molecules may have led to a larger degree of local writhe within the molecules that would have been removed by digitisation of DNA contours or skeletonisation of the resulting images.

This allowed for a prediction of increased underestimation of contour length in real AFM images that have larger sources of image and measurement noise. In real AFM images

the expected underestimation using the Kulpa estimator was between -4.5 to -6.9 % (Rivetti and Codeluppi, 2001). The standard deviation of the simulated images was greater in the larger molecules; this was expected as there was a longer contour length over which variations can be introduced by digitisation. In conclusion, the Kulpa estimator provided an accurate reconstructed length of *TP53* in simulated images but may deviate to a greater extent in real AFM images.

4.3.4 Evaluation of the Effects of Molecule Variation and Image Noise on Curvature Profiles

Curvature profiles were produced with variable amounts of image noise and DNA flexibility (Section 4.2.7.). There was very good visual agreement between the major peaks in the profiles for both signed and unsigned curvature. It was observed that on addition of noise the unsigned curvature profiles had reduced peak contrast. Real AFM images have stronger sources of noise and variation. Therefore, the expectation for real AFM images was a greater reduction in peak to background contrast and loss of smaller peaks. This is not the case with the signed curvature profile. Due to this, signed curvature profiles were used where possible for experiments on real AFM images. Sources of molecule variation also had an impact on the comparability of curvature profiles at low base pair window sizes (Section 4.2.11.). The 21 bp window was near the minimum resolution of the micrograph (~18 bp). Image and molecule variation had an increased effect on curvature measurements at low base pair windows (Section 3.2.5.)

4.3.5 Evaluation of Peak Shift on the Addition of Image Noise

Peak shift in the curvature profiles of two sets of images, one containing noise and one without, was evaluated using a range of base pair windows (Section 4.2.13.). The 21 bp profile produced a peak shift of 1.2 % of the standardised length. The introduction of noise caused two of the major peaks to become unidentifiable in the noisy profile. Additionally, the peak to background contrast was poor and it was necessary to smooth the profile to identify peaks in both noisy and noiseless profiles. This base pair window of curvature was unsuitable for this sort of analysis. At larger window sizes the influence of image noise was reduced. The 42 bp and 63 bp profiles had a percentage peak shift equivalent to a single data point. Both larger window sizes were suitable for analysis of real images.

4.3.6 Identification of Suitable Base Pair Window Size for Curvature Calculation

The correlation between curvature profiles produced by overlapping DNA sequences has been assessed (Section 4.2.11.). The 21 bp window size was unsuitable for analysis of *TP53* due

to the poor comparability between the 21 base pair window size profile and the output of curvature (Section 4.2.7.), weak correlation between ideal and noisy images indicating poor reproducibility (Section 4.2.11) and the loss of key peaks (section 4.2.12.). The 21 bp window size is close to the minimum window size dictated by the resolution of the images (~18 bp) and it is likely that it is strongly influenced by the digitisation of the DNA contour. Larger window sizes provided a better peak contrast and were less influenced by sources of molecule and image variation. The 42 bp and 63 bp window size were used in the study for experimental AFM images of *TP53* for this reason.

4.3.7 Features of Curvature Profiles Lost or Accentuated after Digitisation

It was clear that a number of peaks had been merged in the simulated AFM image profiles in comparison to the output of curvature (Section 4.2.12.) All of the peaks predicted by CURVATURE had been retained with the exception of one peak that occurred in the region of exon 6. What was surprising was that two curvature peaks had been introduced in the simulated AFM image profile. These peaks occurred in the region of exon 7 and near the 3' end of the sequence. The base pair window over which curvature was calculated from the simulated AFM images was larger than the curvature profiles calculated using CURVATURE. The peaks that were unique to the simulated AFM image profiles may have represented curvature that occurred over a larger scale. This larger scale curvature would not have been detected by CURVATURE, which calculates curvature using a dinucleotide site base pair window. Alternatively, these peaks may have been introduced by the simulated deposition method. Either way, the curvature profiles produced from simulated AFM images were more comparable to curvature profiles produced from real AFM images than those produced by other methods. The peak differences highlighted the need to generate models of curvature, using computer-simulated AFM images, for comparison to AFM image real data; the curvature profile produced by CURVATURE, although based upon the same set of dinucleotide parameters (De Santis) generated a different expectation.

4.3.8 The Intrinsic DNA Curvature of Exons in *TP53*

On statistical analysis of *TP53* curvature profiles, exon positions exhibited significantly lower curvature than the introns positions in *TP53* (Section 4.2.15.). This is most evident in the curvature profiles produced by CURVATURE. The reduced curvature predicted by the De Santis dinucleotide wedge model was significantly lower in exons 5, 6 and 7. Exons 8 and 9 were not significant. However, they were bordered by regions of high curvature. This could have increased the curvature measured within the region designated as 'exon' as curvature is averaged over a bp window of 31 bp by the CURVATURE algorithm. The pooled curvature

values of all exons exhibited significantly lower than curvature in pooled introns. In order to evaluate whether this trend was likely to be observed in real AFM images of *TP53* curvature profiles from simulated AFM images were analysed. The results indicated that only exon 5 was likely to exhibit significantly reduced curvature in unsigned curvature profiles of 42 bp or above. The significance of exons 6 and 7 are likely to be lost in real AFM images. However, the statistically significance of the reduction of curvature in pooled exon to intron positions is likely to be retained in real AFM images.

This observation has interesting implications. *TP53* is heavily conserved in evolution due to its key importance to cell regulation and repair (Lane *et al.*, 2010). The reduced curvature in exons may indicate that the structural architecture of coding regions in *TP53* has been selected for during evolution. Alternatively, the low curvature could be a by-product of the accumulation of GC base pair content in coding sections of DNA throughout evolutionary time (Galtier *et al.*, 2001). If intrinsic DNA curvature is actively selected for then this is most likely due to the influence of curvature on nucleosome positioning and the maintenance of nucleosome structure (Shrader and Crothers, 1990; Virstedt *et al.*, 2004). Although curvature has been shown to influence transcription and replication, the impact of curvature is predominantly in the origins of replication and promoter regions of genes (Ohyama, 2005; Marilley *et al.*, 2007b). The sequence under investigation contains no promoters or replication origins so the role of intrinsic curvature in *TP53* is likely to be structural. Low levels of DNA curvature in genes have been linked to open chromatin and active transcription (Vinogradov, 2003). *TP53* is constantly transcribed at a low level within the cell, and its transcription is tightly regulated, so evolutionary selection for DNA architecture to enhance stable transcription is likely to be beneficial (Hollstein and Hainaut, 2010). The theory of evolutionary selection for architectural features in genes has been previously proposed and favours active selection for intrinsic curvature rather than selection for GC content leading to reduced curvature (Vinogradov, 2003).

4.3.9 Intrinsic DNA Curvature in Regions of Slow Repair in *TP53*

There are number of sites in *TP53* that have been shown to exhibit slow DNA repair of bulky chemical adducts (Tornaletti and Pfeifer, 1994; Denissenko *et al.*, 1998; Zhu, 2000). The curvature values for slow repair codons, produced in CURVATURE, were statistically analysed and found to exhibit significantly lower curvature in comparison to the remaining *TP53* sequence (Section 4.2.14.). However, regions of slow repair were localised to exons within *TP53* which independently exhibited reduced curvature (Section 4.2.8.). The possibility of low curvature in slow repair codons being caused by the localisation of the codons to exons has not been discounted.

Reduced curvature in codons of slow repair may indicate a role for curvature in DNA repair in *TP53*. The local DNA sequence bordering a chemical bulky adduct was shown to have a measurable effect on the repair efficiency via the NER pathway (Cai *et al.*, 2009, 2010). Two of the key proteins in the NER pathway, XPA and RPA, specifically recognise DNA structural deformities due to chemical adduction and are also required to deform DNA in order to function (Missura *et al.*, 2001). Studies have concluded that DNA curvature may have a role as a stabilising factor in the presentation of DNA adducts for repair (Cai *et al.*, 2009, 2010). Gel electrophoretic experiments and molecular dynamic simulations indicate that rigidly bent DNA sequences present a wider minor groove leading to more efficient excision and repair of the DNA lesion. The DNA adduct used in these studies, BPDE, was derived from benzo[a]pyrene, a chemical carcinogen heavily involved in the initiation and progression of lung cancer (Hecht, 2002; Kometani *et al.*, 2009). BPDE has also been implicated as a causative agent for the three lung cancer specific sites of slow repair used in this study (Denissenko *et al.*, 1998; Hussain *et al.*, 2001). Therefore, it can be hypothesised that the regions of slow repair in *TP53* may be due to, at least in part, the low curvature of slow repair codons causing reduced presentation of the chemical adduct for removal by the NER pathway.

Additionally, the mechanism underlying sequence-specific DNA repair has also been attributed to the accessibility of the DNA due to the local chromatin structure (Bohr, 1987). As curvature has an active role in nucleosome positioning and the maintenance of nucleosome structure it may also effect DNA repair efficiency indirectly through nucleosome positioning (Shrader and Crothers, 1990; Anselmi *et al.*, 1999).

4.3.10 Nucleosome Positioning

Nucleosome positioning algorithms were applied to *TP53* but failed to produce a consistent result (Section 4.2.16). NXsensor identified a large nucleosome exclusion site within exon 5. NuPop instead predicted that nucleosomes would be unlikely to occupy intronic regions of *TP53* with, perhaps, the exception of exon 6. Both algorithms are equally valid, but identify nucleosome occupancy/exclusion differently. NXsensor identifies sequence motifs unfavourable for nucleosome binding and NuPop explicitly models linker DNA. While the results from the different algorithms do not corroborate one another they do indicate that the nucleosome occupancy of *TP53* should be investigated further, especially in relation to DNA curvature and repair. The potential for exon 5 or 6 to be excluded from the nucleosome core has interesting implications for DNA damage models. For example, exon 5 is highly mutated in lung cancer (Denissenko *et al.*, 1996). One of the major carcinogens involved in lung cancer, BPDE, has shown preferential binding to DNA not contained in the nucleosome core (Jack and Brookes, 1982; Kurian *et al.*, 1985). Intrinsic DNA curvature has an active role in nucleosome

positioning and the maintenance of nucleosome structure (Shrader and Crothers, 1990; Anselmi *et al.*, 1999). It may therefore indirectly influence DNA damage rates and DNA repair rates, as discussed in the previous section, indirectly via control of nucleosome architecture.

4.3.11 Limitations of Theoretical Models

The results presented in this study have highlighted how valuable theoretical models are for generating and testing hypotheses. However, they do possess a number of limitations. One of the limitations of this study has been the lack of sequence-specific DNA flexibility parameters for creating computer simulated AFM images. The current study used a constant persistence length value for modelling DNA although flexibility is known to be sequence-dependent (Hagerman, 1988). A number of DNA flexibility models do exist but not all are applicable to the problem. For example, both bendability trinucleotide models, normalized melting temperatures and stacking energies offer a measure of flexibility but do not provide values that can be easily converted to a measure of persistence length (Brukner *et al.*, 1995b; Scipioni *et al.*, 2002a). The most promising method was the unavailable crystallographic deformity data (Olson *et al.*, 1998; Marilley *et al.*, 2005). Estimates of sequence-specific flexibility would allow for better evaluation of the influence of flexibility on curvature.

Another well reported source of unmodellable variation is the shortening of DNA measured in both air and buffer by AFM. The shortening of DNA has been variably attributed to a B- to A-form DNA transition (Rivetti and Codeluppi, 2001) and electrostatic interactions with the cation loaded mica surface (Sanchez-Sevilla *et al.*, 2002). The source of this shortening has yet to be conclusively identified. However, shortening is assumed to be uniformly distributed throughout the DNA molecules (Buzio *et al.*, 2012). If DNA shortening is due to a B- to A-DNA transition, then the propensity of DNA to transition has been shown to be sequence dependent (Ivanov and Minchenkova, 1995). Currently, due to limited understanding of the underlying cause of DNA shortening, the assumption of uniform condensation must be used. In the eventuality that DNA shortening is confirmed to be due to B- to A-DNA transition, efforts will be needed to model the effect on curvature measurement from individual molecules.

Other sources of experimental variation in AFM imaging cannot be accounted for by theoretical models. DNA molecules can break during DNA deposition; if the break is sufficiently close to either end then the molecule will be treated as a full length molecule in the analysis. There is no way of identifying erroneous DNA molecules that still lie within the expected reconstructed length distribution. These molecules are likely to cause a widening of curvature peaks in real DNA analysis and a shortening of average DNA contour length. Additionally, the likelihood of breakage may be sequence dependent and create a second weak overlapping curvature profile which will be introduced as an experimental source of variation.

4.4 Conclusions

Theoretical models have provided a working hypothesis for the section of the *TP53* gene that codes for the sequence-specific DNA-binding region of the p53 protein. Exon positions exhibited significantly lower curvature than intron positions. The evolution of low curvature in exons may be caused by selection for nucleosomal architecture in *TP53*. This selection for low curvature, to promote stable transcription, may have implications for DNA damage and repair in this most crucial of genes. Low DNA curvature has also been shown to be associated with regions of slow DNA repair in *TP53*, introducing another role for DNA curvature in the functioning of *TP53*. Exons 5 and 6 were predicted to be excluded from the nucleosome core by separate nucleosome positioning algorithms. The propensity to transition from B-DNA to A-DNA was also found to be lowest in exon 5. A number of factors including intrinsic DNA curvature, nucleosome positioning and propensity for structural transition may collectively contribute to a very different structure for exons within *TP53*. Exon 5 in particular was consistently found to have significant differences. This could indicate that it has a distinctly different structural architecture from other regions of the *TP53* gene.

The use of simulated AFM images allowed for a number of predictions to be made about the AFM based analysis of *TP53*. The experimentally determined contour length of *TP53* DNA molecules would be an underestimate in comparison to the theoretical estimates of B-DNA length. A number of key peaks would be retained in the curvature profiles processed using the ADIPAS software. Curvature profiles would be more reproducible at comparable window sizes for signed profiles in comparison to unsigned curvature profiles. Exon 5 would be expected to have significantly lower curvature in comparison to intron regions in the experimental curvature profiles. Finally, pooled curvature measurements of exon positions would likely be significantly lower when compared to pooled intron curvature.

The use of simulated AFM images also produced guidelines for the analysis of real AFM images of *TP53*. The 21 bp window size for calculating curvature angles was shown to be unsuitable for the analysis of curvature in *TP53*. Larger base pair window sizes were more suitable. Theoretical curvature profiles were generated at a representative base pair window size for further comparison to real AFM images.

CHAPTER 5: INTRINSIC DNA CURVATURE ANALYSIS BY APPLICATION OF THE FRAGMENT FLIPPING ALGORITHM TO EXONS 5 TO 9 OF THE *TP53* GENE

5.1 Introduction

5.1.1 Methods of DNA Orientation in Nano-Biology

The precise investigation of intrinsic DNA curvature and flexibility is important for understanding the physical interactions of DNA with other biomolecules. To this end, AFM is a very useful tool for the researcher, allowing nanoscale measurements of the physical conformation of DNA molecules. One of the problems faced by nano-biologists working with DNA is the identification of the correct orientation of DNA on a surface. Researchers have overcome this problem by using protein end-labels (Shaiu *et al.*, 1993; Marilley *et al.*, 2005). However, limited local interaction between the protein end-label and DNA have been reported (Marilley *et al.*, 2005). There are a number of other methods for the orientation of DNA molecules without end-labelling, such as using palindromic DNA dimers (Scipioni *et al.*, 2002a), using symmetrical curvature ratios (Buzio *et al.*, 2012), orientation based upon theoretical models of twist and Z-height (Milani *et al.*, 2011) and the FF algorithm which uses local curvature measurements to orient DNA molecules within a dataset (Ficarra *et al.*, 2005b). The subject of the research detailed in this chapter is the last of the methods listed, the FF algorithm (Ficarra *et al.*, 2005b).

5.1.2 The Fragment Flipping Algorithm

The FF algorithm has been well detailed in a number of related publications (Masotti *et al.*, 2004; Ficarra *et al.*, 2005b). The FF algorithm has been shown by the authors to be effective for real and simulated AFM images and the results have been in good agreement with the De Santis model of curvature. The FF algorithm has been applied to both repeat dimers and linear non-palindromic DNA. The algorithm uses the mean in variance within local curvature measurements as the objective function of a hill-climbing optimisation routine. The aim of the routine is to reduce the variation within the dataset to reach an optimum state. The intrinsic curvature can then be calculated from the orientated DNA dataset using well researched mathematical methods (Scipioni *et al.*, 2002a). Therefore, the FF algorithm is considered a post-processing method of molecule orientation. It introduces no experimental end-labels and it is this quality that makes it desirable to the researcher. The only potential perturbations to the curvature of the DNA are controlled by the researcher, such as the choice of buffer, temperature and adhesion method.

5.1.3 The Underlying Assumptions of the Fragment Flipping Algorithm

The FF algorithm makes a number of assumptions: that the intrinsic curvature is measurable within DNA molecules, that the measurable curvature is greater than background thermal perturbations (*i.e.* there is sufficiently high signal-to-noise ratio), that all the DNA molecules have a random orientation on the surface and the FF algorithm finds the global optimum configuration of the majority of DNA molecules. As to the first and second points, that DNA intrinsic curvature can be measurable using AFM techniques, it is well documented that this is the case within the available literature (Cognet *et al.*, 1999; Zuccheri *et al.*, 2001b; Scipioni *et al.*, 2002a). For the third point, that all DNA molecules have random orientation on the surface, there is evidence that this may not be the case; the face of the DNA double helix that contains the most thymine has been shown to preferentially bind to inorganic surfaces (Sampaolese *et al.*, 2002). On the final point, that the global optimum curvature is found by the FF algorithm, a valid concern with the FF algorithm has been raised (Buzio *et al.*, 2012). The authors noted that the FF algorithm is a hill-climbing optimisation algorithm and, in this respect, is sensitive to solutions that provide local minima in its objective function instead of the global minima (*i.e.* it will find a suitable solution, but that solution may not be the desired 'global' solution). The authors provided an example where this was the case. This is a well documented limitation of hill-climbing algorithms (Morris, 1993).

5.1.4 Aims and Objectives

The main aim was to evaluate whether the FF algorithm can accurately reconstruct curvature from *TP53* DNA. In order to achieve this, the FF algorithm was initially tested using computer simulated AFM images of *TP53*. Using the simulated images as a guideline, alongside guidance tools that had been previously developed, the FF algorithm was applied to 'real' *TP53* DNA molecules. The resulting intrinsic curvature profiles were compared to theoretical curvature models that had been previously generated. The fidelity and applicability of the FF algorithm to *TP53* DNA sequences was investigated and discussed.

5.2 Results

5.2.1 Testing the Fragment Flipping Algorithm using Computer Generated AFM Images.

A large number of computer simulated AFM images were generated for both Exon 5-7 and Exon 5-9 using the De Santis dinucleotide wedge model as described in Chapter 4. Datasets of over 1000 molecules were collected from the computer simulated AFM images for both sequences. The correct orientation of each molecule was ascertained as described in Section 2.5.6. Three test datasets were generated:

- *Curvature Images* – The DNA molecules were generated using a fixed value of curvature at each base pair step identified from the De Santis dinucleotide wedge model (*i.e.* they were all identical). The only sources of image variation were the orientation of the DNA molecules, the effect of digitisation of the DNA contour and the effect of skeletonisation on the resulting AFM images.
- *Flexibility Images* - The DNA molecules were generated using a variable value of curvature at each base pair step. The mean value of the Gaussian distribution of curvature angles at each step was the same as that used in *Curvature Images*. The variation around the mean value was determined using a persistence length of 53 nm. The flexibility of DNA molecules provided another source of variation in addition to that of contour digitisation.
- *Theoretical AFM Images* - The DNA molecules were generated in the same way as the *Flexibility Images* but also had both tip convolution (6 nm ROC) and Gaussian noise (variance = 0.025) added as additional sources of experimental variation. These images were the most comparable to real AFM images and had additional sources for potential variance between molecules as the images were subjected to noise filtering and automatic thresholding (*i.e.* all image processing steps of the ADIPAS software were applied).

5.2.1.1 Accuracy of the Fragment Flipping Algorithm on Increasing Image Noise and DNA Conformational Flexibility

Theoretical molecules were oriented from 5' to 3' within a curvature matrix (a table of all molecules with angles calculated at a number of comparable points). The molecules were randomly flipped into one of the four possible orientations available to them. The transformation was recorded at each step. The FF algorithm was then applied to the curvature matrix. The orientation of each molecule in comparison to its original orientation was recorded after each iteration of the algorithm. This was repeated at three different window sizes for measuring curvature angles; 21 bp, 42 bp and 84 bp. The 21 bp window was included to assess the effect on the FF algorithm at curvature window sizes close to minimum based upon resolution of AFM images. A summary of the results is provided in Table 5.1.

At the smallest bp window, 21 bp, the FF algorithm was at its lowest accuracy for all test samples. For Exon 5-7 the outcome was complete random orientation of DNA molecules. For Exon 5-9 the 21 bp window was more effective at ~50 % accuracy. For larger bp windows the FF algorithm worked at nearly 100 % accuracy for the idealised *Curvature Images*. There was a drop in accuracy with the introduction of DNA molecule flexibility and image noise to a minimum accuracy of 87.15 % in Exon 5-7 using *Theoretical AFM Images*. It should be noted there was a difference in percentage accuracy on repetition which has been investigated in the next section. An example of the resulting curvature profiles has been presented in Figure 5.1. There was less visible effect of lower FF accuracy on signed curvature profiles than unsigned curvature profiles.

Exon 5-7	No. of Molecules	21 bp Window		42 bp Window		84 bp Window	
		Max. Similarly Oriented	% Correctly Oriented	Max. Similarly Oriented	% Correctly Oriented	Max. Similarly Oriented	% Correctly Oriented
Curvature Images	1198	331	27.63 %	1197	99.92 %	1198	100%
Flexibility Images	1171	370	32.00 %	1151	98.29 %	1118	95.47 %
Theoretical AFM Images	1253	311	24.82 %	1092	87.15 %	991	97.06 %
Exon 5-9	**No. of Molecules**	**21 bp Window**		**42 bp Window**		**84 bp Window**	
		Max. Similarly Oriented	**% Correctly Oriented**	**Max. Similarly Oriented**	**% Correctly Oriented**	**Max. Similarly Oriented**	**% Correctly Oriented**
Curvature Images	1181	617	52.24 %	1181	100.00 %	1181	100.00 %
Flexibility Images	1046	680	65.09 %	984	94.07 %	971	92.93 %
Theoretical AFM Images	913	471	51.59 %	842	92.22 %	823	92.22 %

Table 5.1. - Summary of the number and percentage of molecules correctly oriented by the FF algorithm. A correctly oriented *TP53* dataset of curvature measurements from simulated DNA molecules was generated. The orientation of each DNA molecule was transformed randomly in one of the four possible orientations. The FF algorithm detailed by Ficarra *et al.*, 2005 has been applied and the maximum number of molecules in a single orientation was recorded as the percentage '*Correctly Oriented*'.

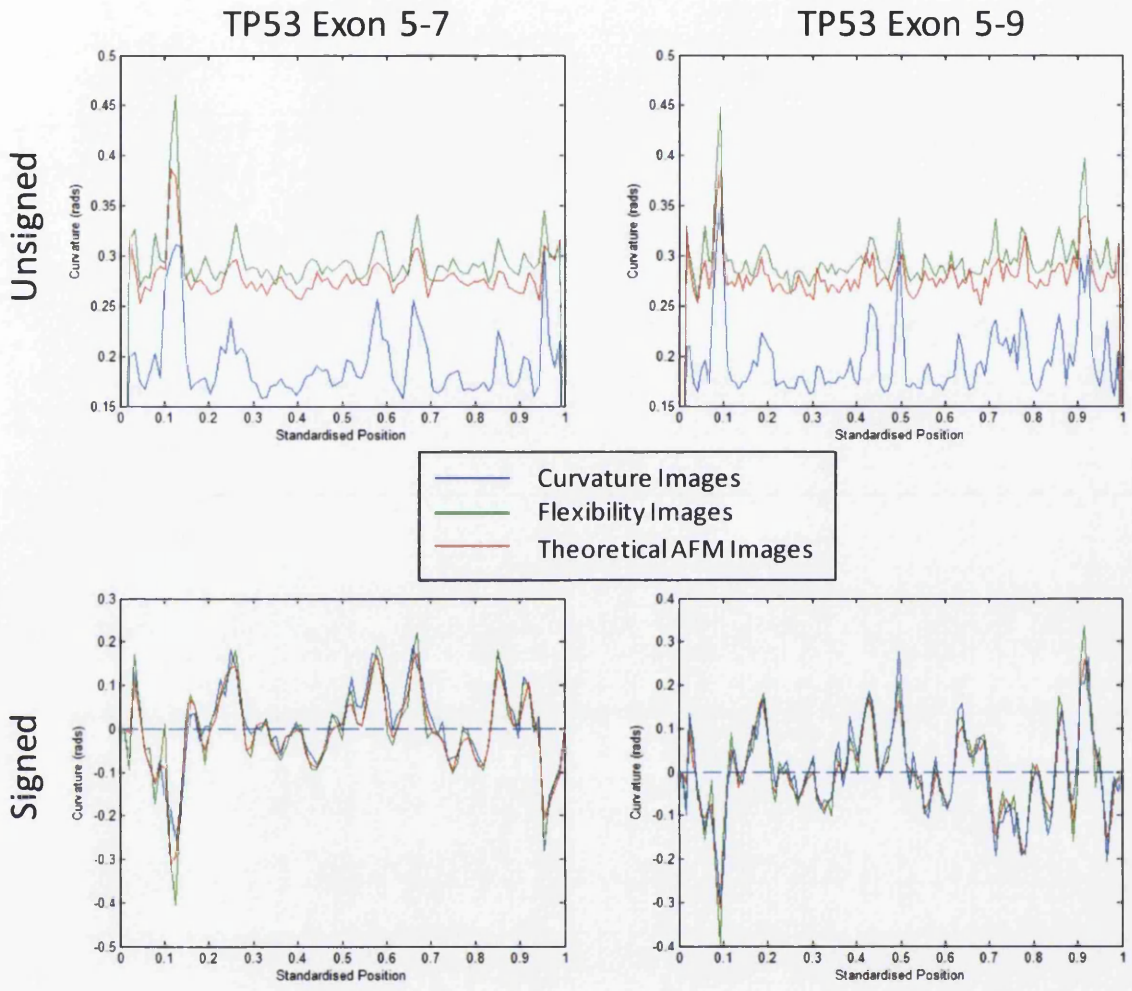

Figure 5.1. - Comparison of reconstructed curvature profiles from computer simulated AFM images with using the FF Algorithm. The computer simulated experimental sample is indicated within the central legend. The window size of curvature is 42 bp.

5.2.1.2 Effects of Base Pair Windows Size on the Accuracy of the FF Algorithm.

It was observed in the previous section that base pair window size had a measurable effect on the fidelity of the FF algorithm. By testing a number of window sizes it was possible to estimate the effect that window size had on the number of correctly oriented DNA molecules. A dataset of 1171 molecules from the *Theoretical AFM Images* sample was used.

The dataset was correctly oriented using the method described in Section 2.5.6. It was then randomised, the orientation of the molecules (in comparison to its original orientation) was recorded during randomisation, curvature angles were calculated for a number of window sizes and the FF algorithm was applied. The maximum number of DNA molecules oriented in the same direction was scored and converted into a percentage. This was repeated three times to give the average percentage accuracy. Standard deviation was used as a measure of variability at each base pair window (Figure 5.2.). The FF algorithm failed completely at the lowest window size of 21 bp; orientating only ~ 25 % of molecules in the same direction, which is nearly complete randomisation. The FF algorithm produced reproducibly good results (>85 % oriented correctly) between the window sizes of 32-84 bp. The maximum accuracy of the FF algorithm was 88.24 % at a window size of 49 base pairs.

The deviation of the accuracy of the FF algorithm was assessed at each base pair window. Poor reproducibility was observed at window sizes below 36 bp. The window size that exhibited the maximum amount of variability in the accuracy was the 34 bp widow (standard deviation = 19.47 %). The window sizes of 34-103 bp showed little to no variation in FF accuracy.

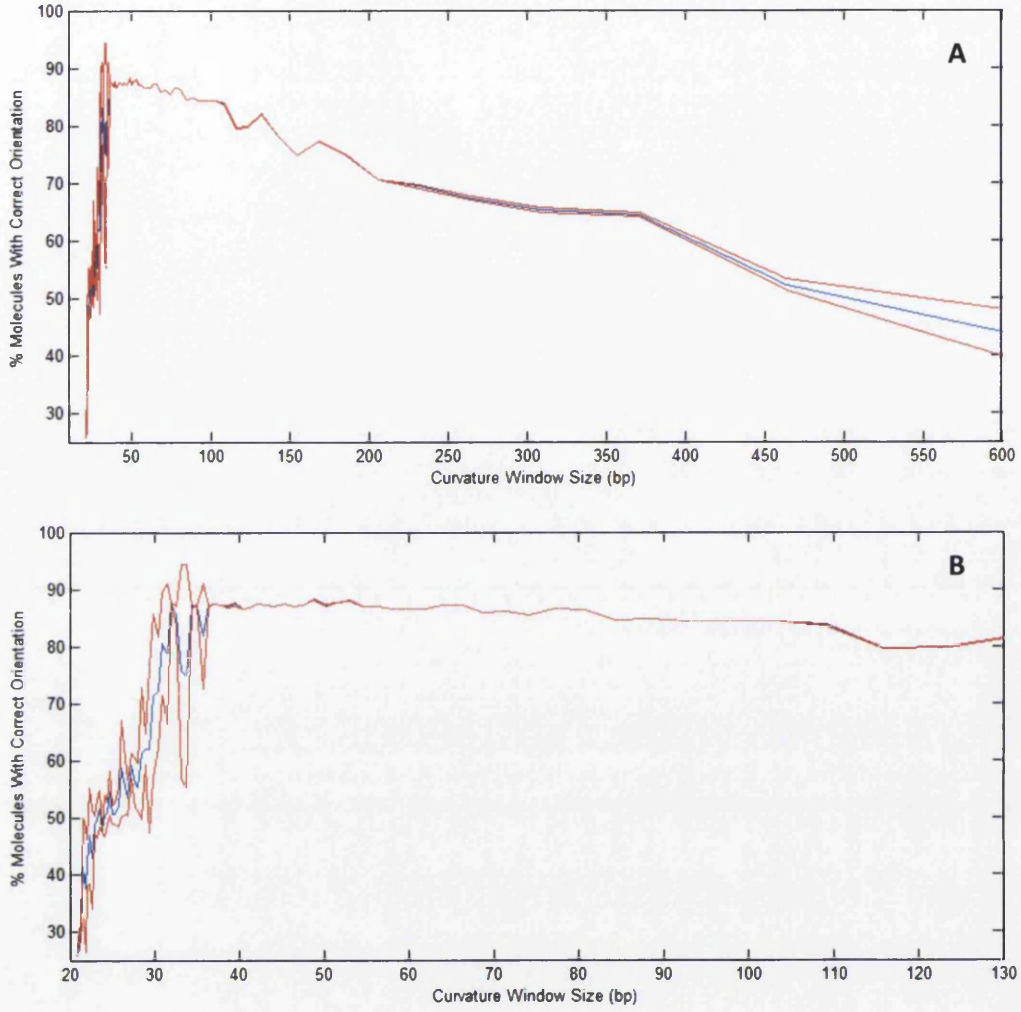

Figure 5.2. - The average percentage of correctly oriented DNA molecules within a dataset over a range of base pair window sizes. A) Base pair window size range of 21 to 600 base pairs. B) Base pair window size range of 21 to 130 base pairs. Average values (blue line) were generated from three repeats of the FF algorithm on 1171 Exon 5-7 simulated DNA molecules or random orientation. The maximum number of DNA molecules with the same orientation were scored and converted into a percentage. The variation in the percentage accuracy was characterised by the standard deviation of three repeats (red line).

5.2.2 Application of the FF Algorithm to Real AFM Images of *TP53*

5.2.2.1 Collection of AFM Images

A large number of AFM images of the PCR product of *TP53* Exon 5-7 and Exon 5-9 were collected. The AFM images were processed using the ADIPAS software detailed in Chapter 3. The DNA molecules were deposited on the mica surface in Mg^{2+} containing buffer in order to allow weak binding and surface equilibration (Hansma and Laney, 1996; Rivetti *et al.*, 1996). A dataset of more than 1000 DNA molecules was processed for both *TP53* PCR products (Table 5.2.).

5.2.2.2 Reconstructed Length Measurements

Reconstructed length measurements were calculated for both the *TP53* Exon 5-7 and Exon 5-9 datasets. *TP53* Exon 5-7 exhibited a non-normal distribution before and after log transformation (Shapiro-Wilks, $p = <0.05$). The median contour length of Exon 5-7 was 598 nm. *TP53* Exon 5-9 exhibited a non-normal before and after log transformation (Shapiro-Wilks, $p=<0.05$). The median contour length of Exon 5-9 was 835 nm. A summary of the reconstructed contour lengths of Exon 5-7 and Exon 5-9 can be found in Table 5.2.

It was observed that there were a number of molecules with contour lengths that did not lie within the main distribution and were far from the median values. It was necessary to remove these outlying molecules; this was achieved by selecting a number of molecules around the median value of the distribution for further analysis. The removal of obviously erroneous molecules has been performed in numerous studies (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Marek *et al.*, 2005). For further analysis 1000 molecules were selected around the median value. After this outlier removal the median of Exon 5-7 was 603 nm and Exon 5-9 was 840 nm. The reconstructed contour length values before and after outlier removal are presented in Figure 5.3. and Table 5.2.

***TP53* Exon 5-7**	**Number Of Molecules**	**Normality Test [Shapiro-Wilks] (p-value)**	**Median (nm)**	**IQR [Q3-Q1] (nm)**
Original Data	1433	<0.05	598	41.34
Outliers Removed	1000	<0.05	603	25.51
TP53 Exon 5-9				
Original Data	1546	<0.05	835	34.95
Outliers Removed	1000	<0.05	840	18.58

Table 5.2. - Summary of the reconstructed length of *TP53* Exon 5-7 and *TP53* Exon 5-9 datasets directly from image processing software and after outliers removal. The median and interquartile range (IQR) values were generated from the reconstructed length measurements from the appropriate datasets. The Shapiro-Wilks test for normality was performed on reconstructed length measurements from the same datasets. Significant p-values for the Shapiro-Wilks test are indicated in red.

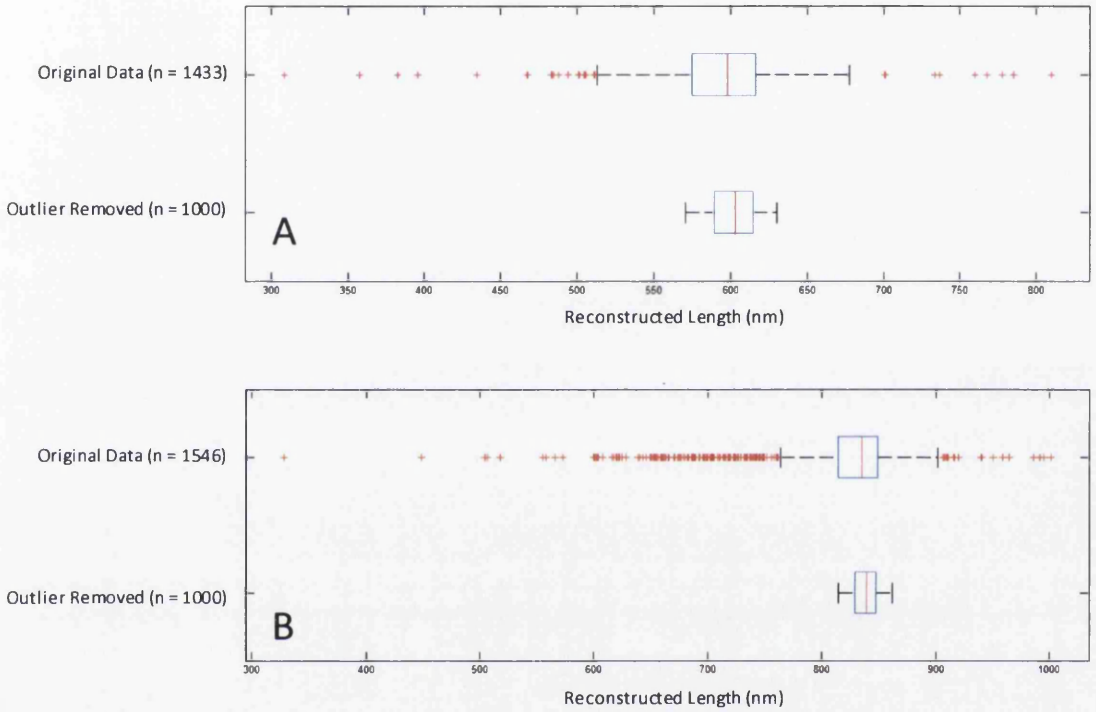

Figure 5.3. - Distributions of reconstructed contour length of *TP53* molecules before and after outlier removal. A) Boxplot of reconstructed length of Exon 5-7 and after outlier removal. B) Boxplot of reconstructed length of Exon 5-9 and after outlier removal. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as red crosses.

5.2.2.3 Persistence Length

Persistence length for both the *TP53* Exon 5-7 and Exon 5-9 datasets was calculated as detailed in Section 3.2.4.2. (Figure 5.4.). The persistence length of DNA was investigated over a curvilinear distance range of 0-400 nm. The persistence length calculated for Exon 5-7 was ξ = 52 nm and Exon 5-9 was ξ = 49 nm. Model fitting over a smaller range of contour lengths (0-300 nm) produced smaller persistence length measurements for both DNA sequences of ξ = 49 nm and ξ = 47 nm for Exon 5-7 and Exon 5-9 respectively.

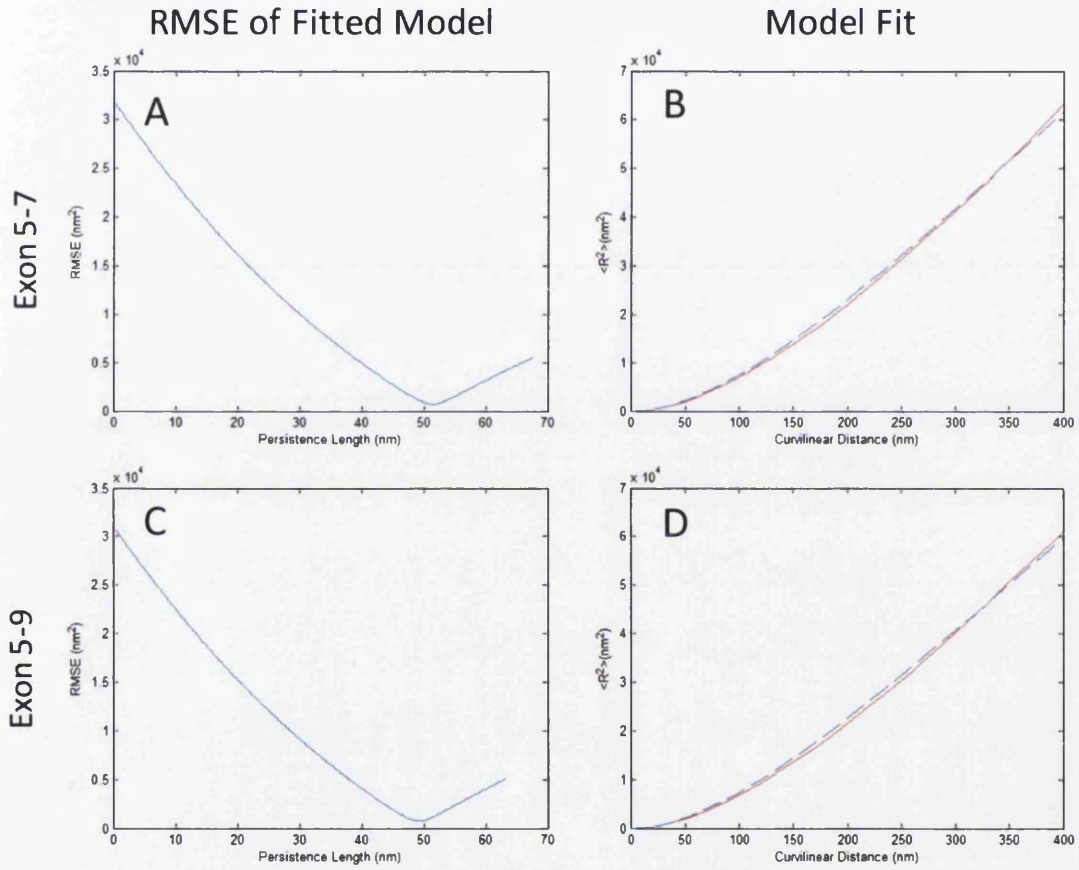

Figure 5.4. - Experimentally determined DNA persistence length for *TP53* Exon 5-7 and Exon 5-9 by comparison to theoretical values of $\langle R^2 \rangle$ from the WLC model. A) Plot of the RMSE fits of $\langle R^2 \rangle$ generated using the WLC theory using a range of persistence lengths for experimental $\langle R^2 \rangle$ of *TP53* Exon 5-7. B) Experimental $\langle R^2 \rangle$ values (red line) of *TP53* Exon 5-7 alongside predicted $\langle R^2 \rangle$ values (broken blue) for the WLC model at a persistence length of 52 nm for a range of curvilinear distances from 0-400 nm. C) Plot of the RMSE fits of $\langle R^2 \rangle$ generated using the WLC theory using a range of persistence lengths for experimental $\langle R^2 \rangle$ of *TP53* Exon 5-9. D) Experimental $\langle R^2 \rangle$ values (red line) of *TP53* Exon 5-9 alongside predicted $\langle R^2 \rangle$ values (broken blue) for the WLC model at a persistence length of 49 nm for a range of curvilinear distances from 0-400 nm.

5.2.2.4 Selection of Base Pair Window for Curvature Calculation

Having previously observed that the FF algorithm was more accurate within a certain range of base pair window sizes (Section 5.2.1.) it was necessary to identify a window size to use on experimental DNA molecules. The lowest approximate window size for the resolution of the images was ~18 bp. Typically a base pair window size close to the resolution limit of the image is used by researchers (Ficarra *et al.*, 2005b). However, the 21 bp window size was previously shown to be unsuitable for *TP53* sequence when using computer generated AFM images (Section 4.2.11.). Computer simulated AFM images are 'ideal' images and only contain a few controlled sources of noise. The failure of the FF algorithm on simulated DNA molecules indicated that it would be unlikely to work on experimental images that have greater sources of image noise and DNA molecule variance.

There were a number of sources of information available for the selection of appropriate base pair windows. Firstly, there was the experiment that evaluated the accuracy of the FF algorithm at a range of base pair window size (Section 5.2.1.2.). The window sizes with minimum variance and maximum accuracy suggested a base pair range of 34-84 bp. The Visual Threshold, developed in Section 3.2.5.2., was also applied (Figure 5.5.). This allowed for the visual assessment of curvature calculated over a range of base pair window sizes. This was used to identify the influence of digitisation of the DNA contour on curvature angle measurements. Both Exon 5-7 and Exon 5-9 followed the expected pattern. The minimum curvature values were 55 nm for Exon 5-7 and 46 bp for Exon 5-9. The Visual Threshold suggested a range of window sizes of 34-80 bp for Exon 5-7 in remarkable agreement with the experimentally determined optimum FF algorithm window sizes.

The window sizes of 42 bp and 63 bp were used for further analysis. In some instances the window size of 21 bp has been included for comparison to previous research. The window sizes of 42 and 63 bp lie within experimentally determined optimal ranges. Additionally, these window sizes have been shown to provide good curvature peak-to-background contrast in theoretical studies of *TP53* (Section 4.2.6.). Both window sizes are multiples of a helical turn (10.5 bp in B-DNA) and can be discussed in terms of a biologically relevant measure.

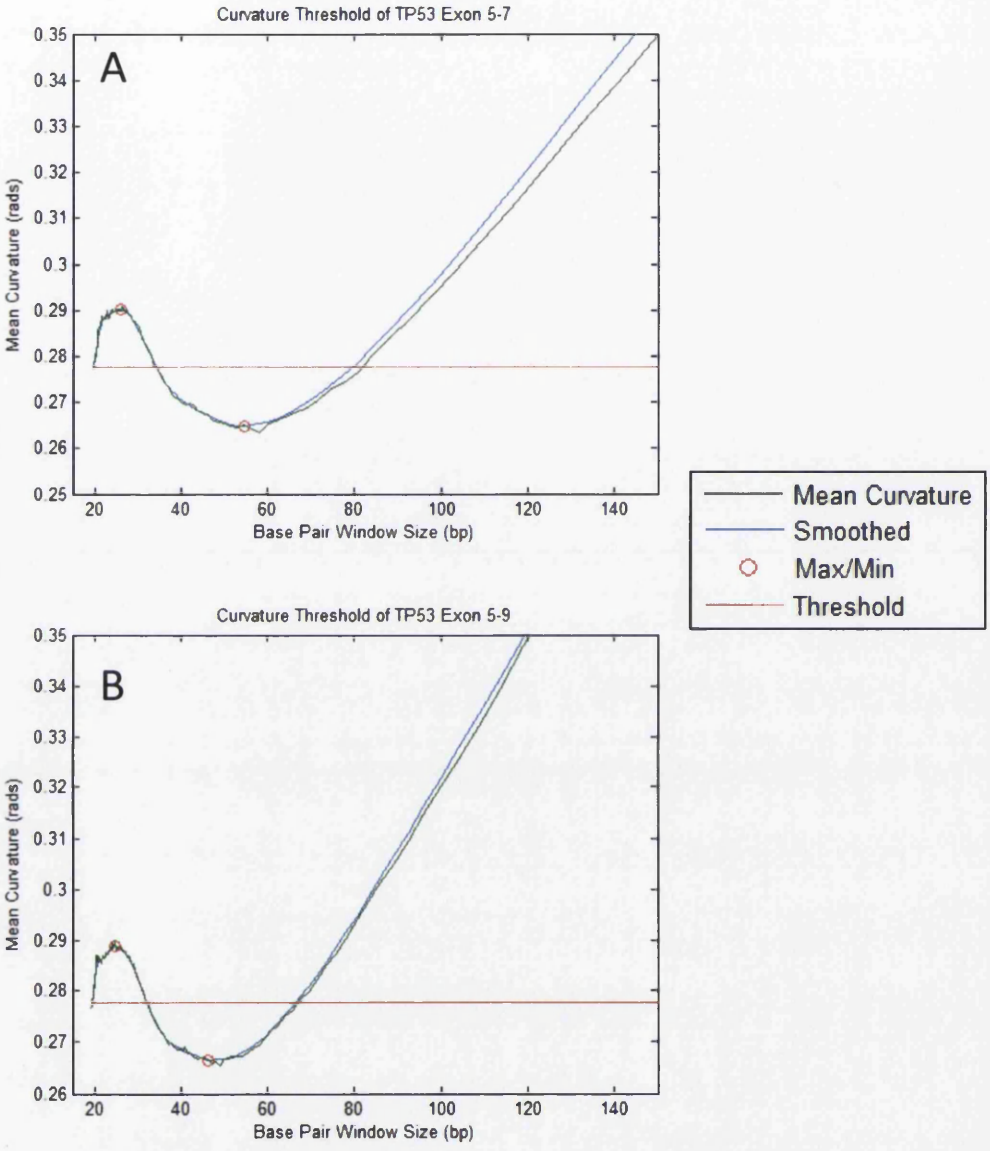

Figure 5.5. - Visual Threshold of mean curvature for *TP53*. A) *TP53* Exon 5-7 (minimum = 55 bp). B) *TP53* Exon 5-9 (minimum = 46 bp). Mean curvature is plotted as a green line, smoothed (three point moving average) as blue, the maxima and minima values are denoted as red circles and the threshold as a red line.

5.2.2.5 Curvature Profiles Generated using the Fragment Flipping Algorithm

The FF algorithm was applied to the unoriented set of DNA molecules after outlier removal. There are 1000 molecules in the final set for both Exon 5-7 and Exon 5-9. A full description of the FF algorithm can be found in Section 3.2.4.5. and in the original publication (Ficarra *et al.*, 2005b). The resulting curvature profiles are presented in Figure 5.6. Signed profiles were calculated as the mean value per base pair window interval of a dataset with both negative and positive curvature angles. Unsigned profiles were calculated as the mean value per base pair window where all angles were considered positive *i.e.* there was no direction attributed to the curvature angles. Profiles were smoothed (three point moving average) to improve the peak-to-background contrast and highlight trends in the curvature profiles.

The FF algorithm was initially applied to unoriented sets of DNA molecules. The curvature profiles before and after application of the FF algorithm were recorded. The results are presented in Figure 5.6. It was observed that the profiles before the FF algorithm was applied exhibited weak trends in curvature, perhaps indicating that not all DNA molecules were randomly oriented on the mica surface. The curvature profiles after application of the FF algorithm exhibited strong similarities to profiles produced before FF. At the 42 bp window the majority of the curvature had been flipped to one end of the molecule. The pattern of curvature for both Exon 5-7 and Exon 5-9 was in almost perfect agreement with the profile before FF. A similar effect was observed at the 63 bp window of curvature, although there was less visual agreement between the profiles. The magnitude of curvature measured after application of the FF algorithm was larger than the initial profiles before FF. However, the positions of many peaks were in good agreement.

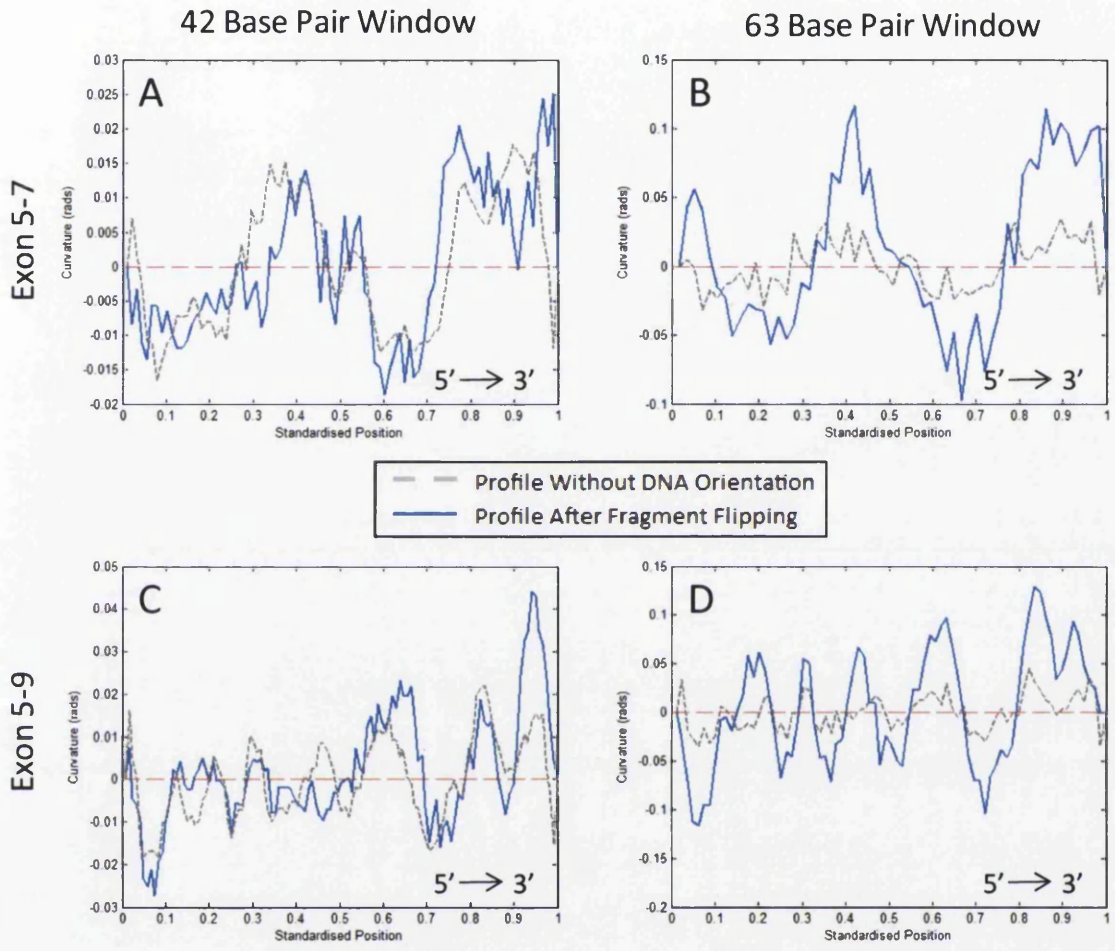

Figure 5.6. - Comparison of curvature profiles before and after application of the FF algorithm. A) *TP53* Exon 5-7 42 bp curvature window. B) *TP53* Exon 5-7 63 bp curvature window. C) *TP53* Exon 5-9 42 bp curvature window. D) *TP53* Exon 5-9 63 bp curvature window. Raw profiles are indicated with broken grey lines and profiles after FF algorithm are indicated with blue lines. Curvature is in radians and the direction of curvature is indicated (signed curvature). The position along the profile is standardised from 0-1 (5′ to 3′).

5.2.2.6 Reapplication of the Fragment Flipping Algorithm after Randomisation of DNA Orientation

The similarity between the profiles before and after FF orientation indicated that the algorithm was being influenced by the weak initial trends in the curvature of *TP53* (Section 5.2.2.5.). The assumption was made that the weak trends in the curvature profiles were providing a local solution to the objective function of the FF algorithm. This has been observed in other studies (Buzio *et al.*, 2012). In order to overcome this problem the orientation of the DNA molecules within both datasets were randomised before reapplication of the FF algorithm.

To avoid the effects of local solutions to the objective function of the FF algorithm further steps were taken. The FF algorithm was repeated ten times on randomised DNA molecules. The results were aligned and averaged. This was the equivalent of applying the FF algorithm to the results of the FF algorithm. If randomisation has the effect of modifying this local solution then sufficient randomisations and reapplication of the FF algorithms may provide a number of local or global solutions that, when averaged, would produce a consensus profile. The aim of this experiment was to establish whether the FF algorithm can provide a consensus outcome for the DNA sequences of interest.

The signed curvature after randomisation and re-application of the FF algorithm is presented in Figure 5.7 alongside theoretical curvature profiles. Curvature angles were calculated using the appropriate base pair window before application of the FF algorithm. Theoretical curvature profiles were calculated separately for each base pair window size and rescaled to allow visual comparison to experimental curvature profiles.

The outcome for Exon 5-7 at a 42 bp window provided an unclear result (Figure 5.7. A.). The magnitude of the curvature peaks measured was small and there were few clear trends within the data before and after smoothing. It was not clear if the curvature profiles produced were representative of the DNA sequence. At a larger window of curvature, 63 bp, there were more obvious trends within the data (Figure 5.7.B). There were similarities between the theoretical profiles and the experimental profiles towards the ends of the DNA sequence. These similarities were less clear within the centre of the sequence. The peak of curvature between 0.3 and 0.4 standardised length, roughly corresponded to Exon 5 and 6, was unexpectedly large relative to the rest of the profile.

The curvature profile for Exon 5-9 at a 42 bp window of curvature provided a number of peaks of curvature with which to make a comparison to theoretical models (Figure 5.7.C.). There was good visual correlation between a number of experimental peaks within the curvature profile and agreement peaks within the theoretical profile although the direction of

the peaks may differ. This was especially strong for peaks near either end of the molecule. The 63 bp window of curvature of Exon 5-9 provided even greater contrast between peaks and there was a consensus shape between the two base pair window sizes (Figure 5.7.D.). There was good visual agreement between the occurrences of peaks at the end of the DNA sequence, disregarding the sign of curvature. This visual agreement was weaker at the centre of the DNA sequence, although there were a number of corresponding peaks between the profiles.

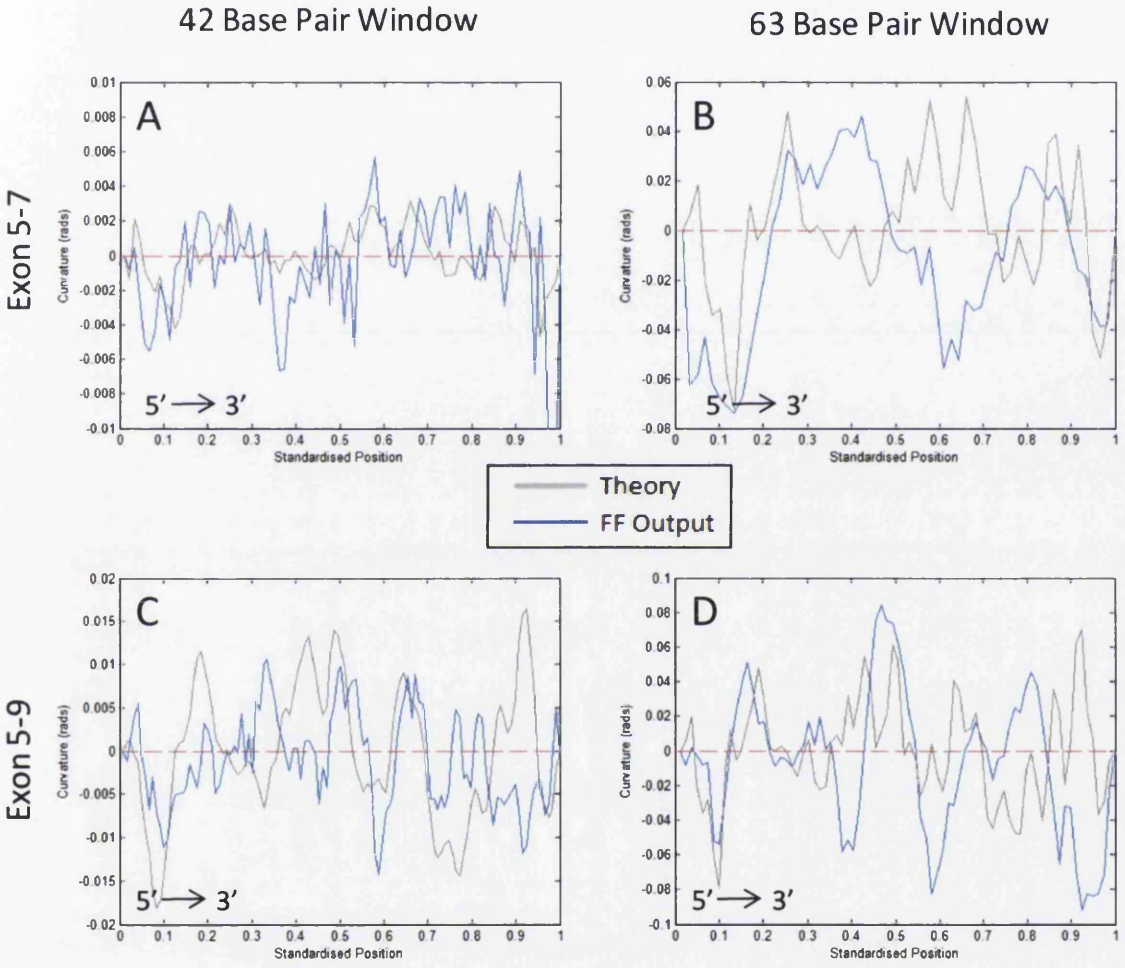

Figure 5.7. - Experimental curvature profiles for *TP53* aligned with theoretical curvature profiles. A) *TP53* Exon 5-7 42 bp curvature window. B) *TP53* Exon 5-7 63 bp curvature window. C) *TP53* Exon 5-9 42 bp curvature window. D) *TP53* Exon 5-9 63 bp curvature window. Experimental profiles are indicated in blue and theoretical in grey. Curvature is in radians and the direction of curvature is indicated. The position along the profile has been standardised from 0-1 (5′ to 3′). The 42 bp window profiles were smoothed with a 3-point average filter. Theoretical profiles have been rescaled for comparison to the experimental profiles. Theoretical curvature profiles were produced from the De Santis dinucleotide model .

5.2.2.7 Comparison of Curvature Profiles to Amended Theoretical Profiles

The Geometric Deposition model for simulating adsorption of DNA produced theoretical profiles that were comparable to the experimental profiles for Exon 5-9. This model gave an approximation to the actual average geometry of adsorbed molecules. There may be relevant differences between the projected geometry of the DNA molecules and the experimentally observed trajectory of DNA molecules. This must be considered when comparing experimental and theoretical profiles and adjustment to the theoretical must be made *a posteriori* (Figure 5.8.). An example of this was provided by the original authors of the method (Buzio *et al.*, 2012). The authors observed a preferential 180° rotation between neighbouring sections of DNA and used this to adjust the results of the deposition model. Practically, this was performed by inverting the sign of the region containing the preferential twist in the relevant theoretical curvature profiles (*i.e.* values changed from positive to negative and *vice versa*). This sort of modification of the chain architecture cannot be predicted by current theoretical models as it is a direct product of the adsorption process.

TP53 Exon 5-9 was the first to be considered as it provided the clearest comparison between experimental and theoretical peaks in curvature. Two preferential 180° twists within the DNA sequence were proposed to provide a better agreement between experimental and theoretical curvature. The first proposed twist occurred between exon 5 and 6 and continued for a short way into the intronic region between exons 6 and 7. The second twist occurred at the 3' end of the DNA sequence and incorporated exons 8 and 9. There was very good visual similarity between both the occurrence of curvature peaks and the curvature direction for both windows of curvature (Figure 5.8.C+D.).

A correlation analysis was performed between the theoretical and experimental profiles before and after a preferential twist was introduced to the theoretical data (Table 5.3.). Application of the FF algorithm at a 42 bp window showed significant weak positive correlation with the original theoretical projection (Spearman's, Rho = 0.21; p = <0.05). After amending the projection *a posteriori* the profiles exhibited an improved correlation (Spearman's, Rho = 0.49; p = <0.005). Application of the FF algorithm at a 63 bp window showed no significant correlation with the original theoretical projection (Spearman's, Rho = 0.09; p = 0.44). After amending the projection *a posteriori* the profiles exhibited a moderate significant correlation (Spearman's, Rho = 0.60; p = <0.05). Theoretical profiles were in better agreement with experimental profiles at a larger base pair window size, in agreement with predictions made from simulated AFM images of *TP53* (Section 4.2.11.).

The potential occurrence of preferential twists within the DNA sequence of Exon 5-7 was less easily accounted for due to the reduced degree of similarity between the theoretical

and experimental curvature profiles (Figure 5.8.A+B). However, assuming similar adsorption behaviour between the sequences, a 180° twist was introduced between exon 5 and 6 and continued into the intronic region between exons 6 and 7. This accounts for a small amount of the deviation between the theoretical and experimental curvature profiles although there were still dissimilarities within the centre of the profile. A correlation analysis was performed between the theoretical and experimental profiles before and after a preferential twist was applied to the theoretical data (Table 5.3.). Significant positive correlation was observed for Exon 5-7 at a 63 bp window of curvature using the amended theoretical profile (Rho = 0.31, p = <0.05).

		Spearman's Rank Correlation Coefficient			
		Original Theoretical Projection		**Amended Theoretical Projection**	
Exon 5-7	**Window Size**	**Rho**	**p-value**	**Rho**	**p-value**
	42 bp	-0.97	0.47	0.31	<0.05
	63 bp	0.10	0.36	0.06	0.60
Exon 5-9	**Window Size**	**Rho**	**p-value**	**Rho**	**p-value**
	42 bp	0.21	<0.05	0.49	<0.05
	63 bp	0.09	0.44	0.60	<0.05

Table 5.3. – Spearman's Rank correlation between experimental and theoretical curvature profiles of *TP53* before and after amending the theoretical profile using *a posteriori* knowledge. Significant p-values are indicated in red.

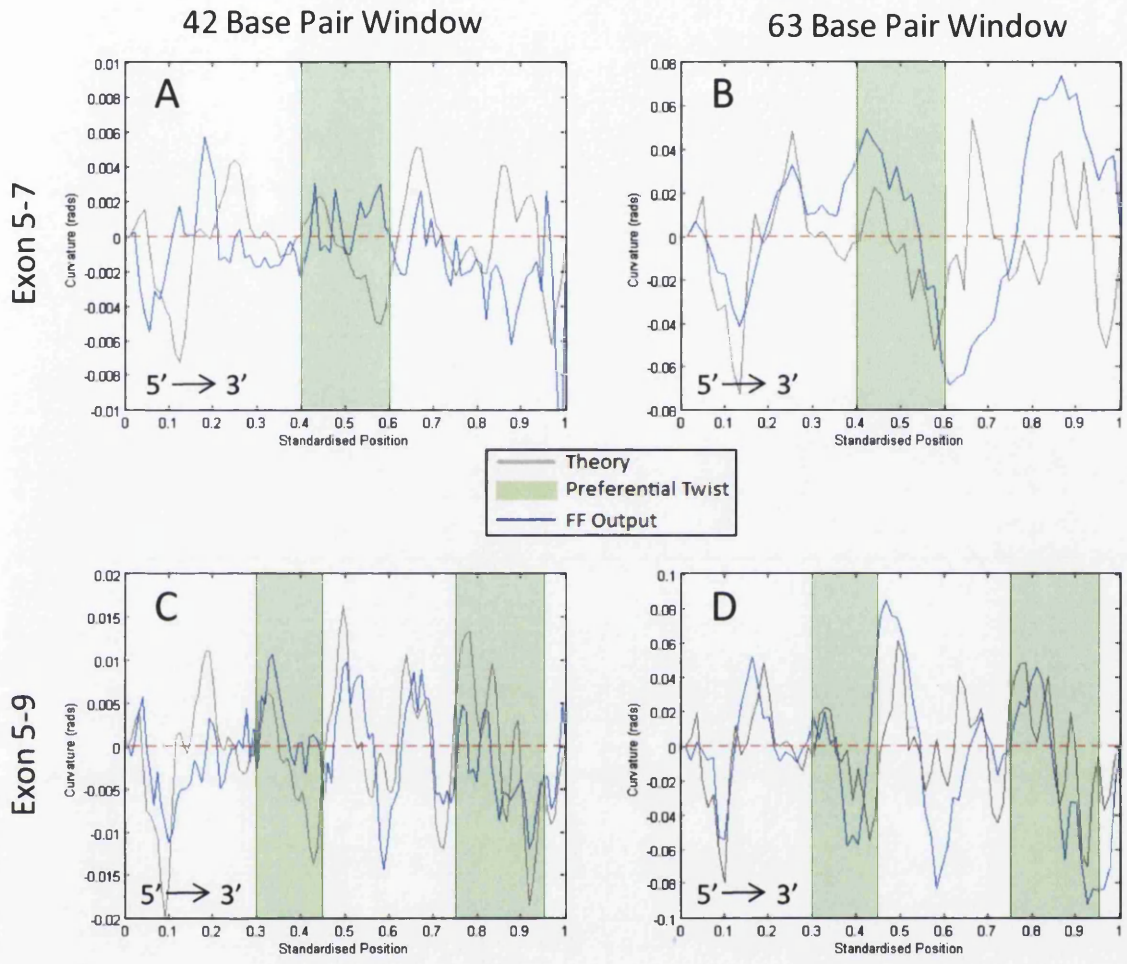

Figure 5.8. - Experimental curvature profiles for *TP53* aligned with theoretical curvature profiles with preferential twist. A) *TP53* Exon 5-7 42 bp curvature window. B) *TP53* Exon 5-7 63 bp curvature window. C) *TP53* Exon 5-9 42 bp curvature window. D) *TP53* Exon 5-9 63 bp curvature window. Experimental profiles are in blue and theoretical are in grey. The proposed preferential twist in the theoretical profile is highlighted in green. Theoretical profiles (De Santis) have been rescaled for comparison to the experimental profiles. Curvature is in radians and the direction of curvature is indicated (signed curvature). The position along the profile has been standardised from 0-1 (5′ to 3′). The 42 bp window profiles were smoothed with a 3-point average filter.

5.2.2.8 Assessing the Peak Shift of Key Curvature Peaks

Ten key peaks were identified in the theoretical profiles as the peaks with the largest magnitude of curvature. This was performed for the 42 bp and 63 bp windows of curvature. This was performed using the data for Exon 5-9 and the theoretical profiles amended with two 180° preferential twists (Section 5.2.2.7.). Peaks of curvature that corresponded to the theoretical peaks were identified in the experimental profile produced by the FF algorithm.

The 42 bp window of curvature showed a number of similarities between the occurrences of peak positions although the magnitude and sometime direction of the peaks was different. Nine of the ten key peaks were identified in the experimental profile. The only key peak not identified was likely to have merged into one peak in the experimental profile (green circle in Figure 5.9.A+C). The average peak shift of the identified peaks between prediction and experimental profiles was 1.31 % or 32.75 bp. The magnitude of corresponding curvature peak values were significantly different (Wilcoxon Rank Sum, $p = <0.00$).

The 63 bp window of curvature showed a number of similarities in the occurrence of peak positions although the magnitude of the peaks was different. Eight of the ten key peaks were identified in the experimental profile. Two key peaks were not identified. One of these key peaks was likely to have merged into one peak in the experimental profile (green circle in Figure 5.9.B+D). The other peak was missing from the experimental profile. Additionally, there was a notably large peak in the experimental profile that was not present within the theoretical profile that could have been produced from the merging of multiple peaks (red circles in Figure 5.9.B+D). The average peak shift of the eight identified peaks between prediction and experimental profiles was 2.25 % or 56.25 bp. The magnitude of corresponding curvature peak values were significantly different (Wilcoxon Rank Sum, $p = <0.00$).

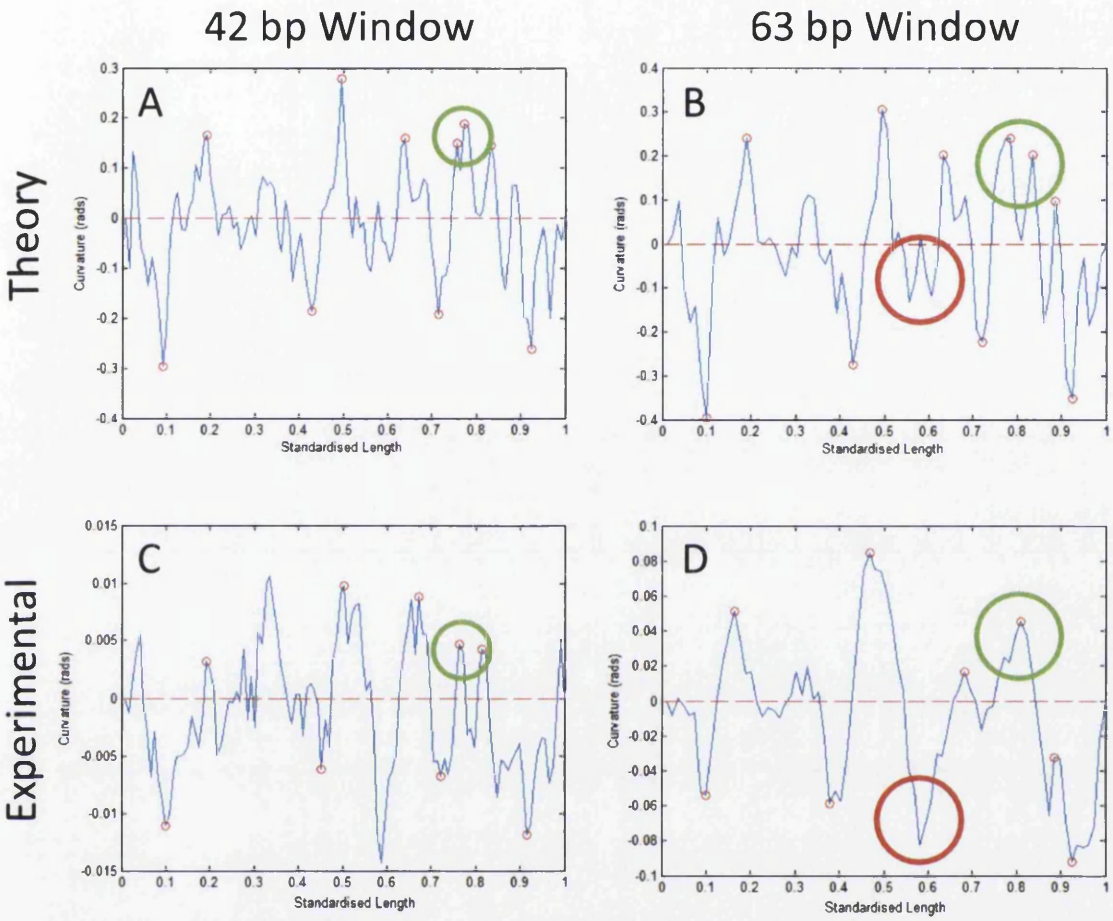

Figure 5.9. - Identification of key peaks between experimental and theoretical profiles. Key peaks were identified as those with the largest curvature values in the theoretical profiles. Key peaks are shown with small red circles. Proposed regions where peaks merged are indicated with circles (red and green).

5.2.2.9 Observations on the Final Curvature Profiles for *TP53*

The final curvature profiles for *TP53* were aligned with exon and intron positions (Figure 5.10.). There were fewer peaks in the curvature profile of Exon 5-7 at the 42 bp window and fewer observations could be made. Exon 5-7 at a 63 bp window showed gradually modulating curvature with the largest peak at the 5' end of the DNA sequence. Exon positions in Exon 5-7 at a window size of 42 bp and 63 bp occurred during small peaks in curvature. Exon 5-7 shared few visual similarities with theoretical profiles (Section 5.2.2.6.) and this provided a reason to assume that the FF algorithm had failed to reconstruct the true intrinsic curvature of the Exon 5-7 DNA sequence.

The Exon 5-9 curvature profiles were in good agreement with the theoretical profiles (Section 5.2.2.7.). Exon 5 exhibited very low curvature (close to the broken red line denoting 0.0 radians of curvature) at both window sizes. Exon 6 showed a peak in curvature within the 42 bp profile that was not present within the 63 bp profile. This peak was expected from theoretical models but may have been obscured at larger window sizes. Exon 7 occurred directly after a large peak in curvature in both window sizes. Exon 8 occurred as a trough in curvature in the 42 bp profile, which was in agreement with theoretical models, and during a peak in the 63 bp profile. The expected small trough may have been obscured by noise in the larger window size profile. Exon 9 was in full agreement with theoretical expectation as it appears as a small peak of moderate curvature within both window sizes. A statistical analysis of unsigned curvature values of exon positions compared to intron positions indicated that exon 5 had significantly reduced curvature than intronic positions at the 63 bp window of curvature (Kruskal-Wallis, $p = <0.05$). This was not the case for other exons or for exon 5 at the 42 bp window. Exon 5 was the only exon predicted to exhibit significant curvature using simulated AFM images (Section 4.3.8.).

Exon 5-9 shared good visual similarity with theoretical curvature profiles (Section 5.2.2.7.). This similarity increases on the addition of two preferential twists into the theoretical profiles. The similarity suggested that the FF algorithm was functioning correctly for this molecule and that the De Santis model of curvature was providing a good estimation of intrinsic DNA curvature.

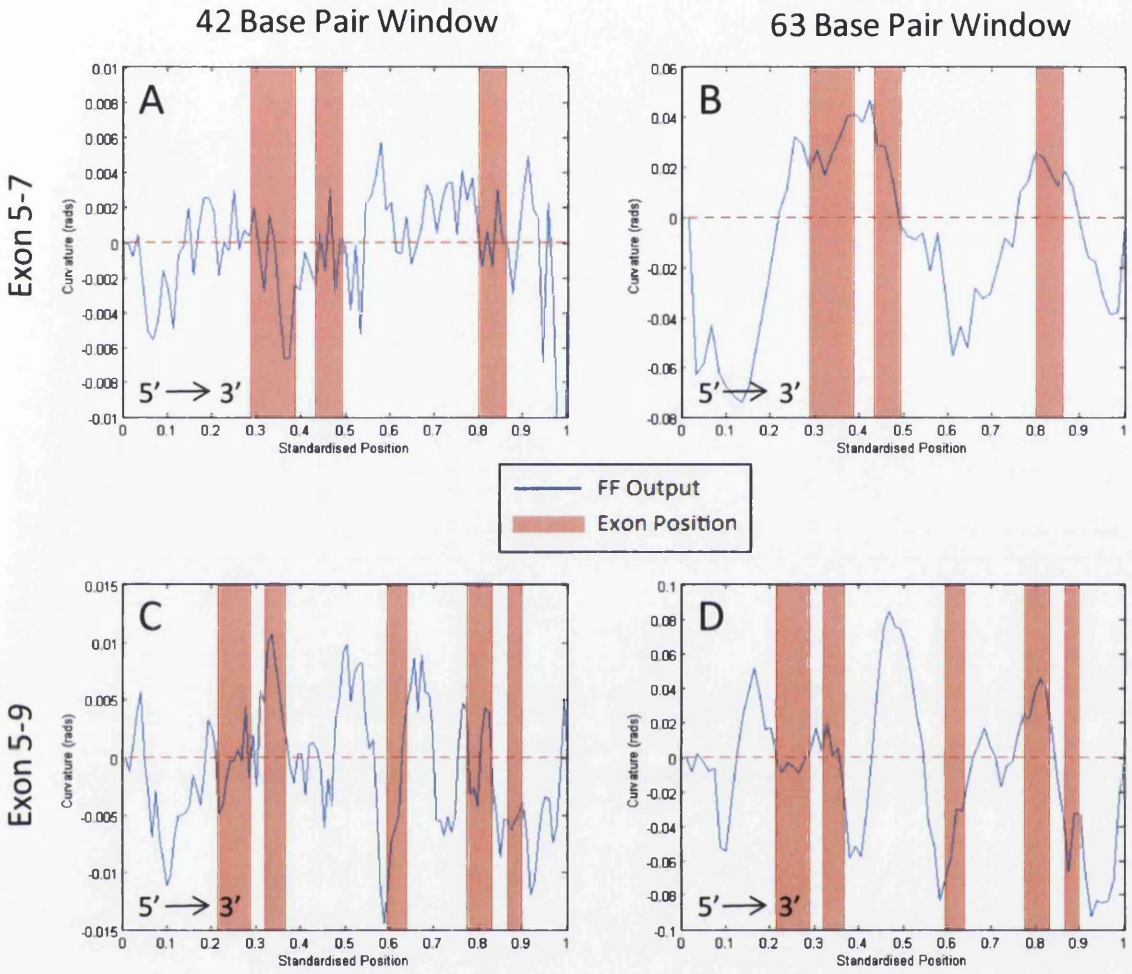

Figure 5.10. - Final experimental curvature profiles for *TP53* with exon positions highlighted. A) *TP53* Exon 5-7 42 bp curvature window. B) *TP53* Exon 5-7 63 bp curvature window. C) *TP53* Exon 5-9 42 bp curvature window. D) *TP53* Exon 5-9 63 bp curvature window. Exon positions are highlighted in red. Curvature is in radians and the direction of curvature is indicated (signed curvature). The position along the profile has been standardised from 0-1 (5′ to 3′). The 42 bp window profiles were smoothed with a 3-point average filter.

5.3 Discussion

5.3.1 The Effects of Image Noise and DNA Molecule Conformational Flexibility on FF Accuracy

The accuracy of the FF algorithm was tested using simulated images before it was applied to real AFM images of *TP53*. The accuracy of the FF algorithm was initially tested on ideal AFM images and then AFM images with the addition of DNA flexibility and image noise (Section 5.2.1.1.). The most striking observation was the failure of the FF algorithm at the 21 bp window of curvature. DNA molecules were completely randomly oriented at the 21 bp window for Exon 5-7 (25 % accuracy) and only half were correctly oriented for Exon 5-9. This window size of curvature was close to minimum resolution of the images (~18 bp). Low base pair window sizes were previously shown to be effected by digitisation of the DNA contour (Section 4.3.4.) and curvature measurements were significantly influenced by experimentally introduced variation at this window size. These two sources of variation were likely to have had a significant impact on the accuracy of the FF algorithm at the 21 bp window size. The original authors of the FF algorithm used base pair window sizes close to the maximum resolution of the AFM images (Ficarra *et al.*, 2005b). One of the theoretical samples used in the study was reported to have an accuracy of only 76.19 %. This low accuracy may have been caused by the effects of digitisation.

The FF algorithm at larger base pair windows exhibited improved accuracy in excess of 87 % (42 bp and 63 bp – Table 5.1). The FF algorithm correctly oriented 99.92-100.00 % of all DNA molecules in simulated images sets that contained inflexible, idealised DNA molecules. The addition of sources of DNA flexibility and image noise caused a decrease in the accuracy of the FF algorithm at all base pair windows. This was most pronounced in the shorter Exon 5-7 molecule at a 42 bp window size. Even with the addition of image noise the accuracy of the FF algorithm at larger base pair window sizes was greater than the accuracy reported by the original authors of the FF algorithm of between 76.19 % - 96.97 % (Ficarra *et al.*, 2005b). The improved accuracy in the present study may be attributed to the poorer resolution of simulated images used in the previous study (3.91 and 7.81 nm per pixel in comparison with the 2.92 per pixel of this study) and the considerable differences between intrinsic curvature profiles. Additionally, the original authors did not consider the potential for variation introduced by DNA contour digitisation on angle calculations; instead points were fitted to each DNA molecule close to the resolution of the images.

5.3.2 Evaluation of the Effects of Base Pair Window Size on the Accuracy of the FF Algorithm

The base pair window size used to calculate curvature had a measurable effect on the accuracy of the FF algorithm (Section 5.3.1.). At low (<34 bp) and high (>200 bp) window sizes the FF algorithm produced poor results with a large degree of variability upon repetition. The window sizes that produced the most accurate reconstruction by the FF algorithm (>85 %) were between 31 and 81 bp. It was likely that these base pair window sizes provided the best peak-to-background contrast for optimal FF accuracy. At lower base pair window sizes the influence of noise and digitisation increased, effectively reducing the ability of the FF algorithm to function with any degree of accuracy.

This analysis indicated that there were a number of factors that needed to be considered before applying the FF algorithm to real AFM images. Firstly, the choice of pixel resolution will influence curvature measurement due to pixilation/digitisation noise. Low pixel resolutions may introduce higher levels of DNA contour variance during image processing and mask small-scale curvature features that are necessary for accurate orientation by the FF algorithm. The second and more important consideration identified was the base pair window size at which to calculate comparable curvature angles before application of the FF algorithm.

5.3.3 Selection of a Base Pair Window for Application of the FF Algorithm to Real AFM Images of *TP53*

The previously discussed experiments were used as a guideline for selection of an appropriate base pair window size to calculate curvature angles before application of the FF algorithm to experimental *TP53* DNA (Section 5.2.1.). Additionally, the Visual Threshold developed in Section 3.2.6.5. was also applied to the experimental AFM molecule to ascertain which base pair window sizes were effected by digitisation noise. The window sizes suggested by both methods were in excellent agreement (Visual Threshold – 34-80 bp; window size experiment – 31-81 bp). This is a further indication that the accuracy of the FF algorithm is dependent upon peak-to-background contrast and that the window size over which to calculate curvature angle was an important consideration.

5.3.4 Reconstructed Length Measurements of AFM images of *TP53*

Reconstructed lengths were calculated for both *TP53* DNA sequences using the Kulpa estimator (Section 5.2.2.2.). Approximately 0.34 nm per base pair was used as a consensus length for B-DNA taken from X-ray crystallography experiments (Saenger, 1984). The median reconstructed length measurement of 603 nm for *TP53* Exon 5-7 slightly underestimated the theoretical measurement of B-form DNA of 631 nm by -4.36 %. A similarly small

underestimation was observed for Exon 5-9 with a theoretical value of 850 nm in comparison to the experimental measurement of 840 nm for an underestimation of only -1.20 %. These percentage differences were within the experimental boundary of less than -6.9 % previously identified for the Kulpa estimator (Rivetti and Codeluppi, 2001). The distribution of the length around the median value was higher in the smaller molecule, which was the opposite of the trend expected from theoretical measurements (Section 4.2.5.). The variation around the median value (standard deviation - Exon 5-7 = 16.17; Exon 5-9 = 12.04) was still comparable to the standard deviation values of between 22.4 and 112.7 reported by previous authors (Rivetti and Codeluppi, 2001; Scipioni *et al.*, 2002a).

5.3.5 Persistence Length Measurements of *TP53*

The persistence lengths calculated for *TP53* DNA molecules (Exon 5-7, $\xi = 52$; Exon 5-9, $\xi = 49$) were in good agreement with flexibility reported by other authors of around ~50 nm for Mg^{2+} deposited DNA (Rivetti *et al.*, 1996; Moreno-Herrero *et al.*, 2006; Wiggins *et al.*, 2006; Buzio *et al.*, 2012). These results suggested that DNA molecules deposited under these experimental conditions were thermodynamically equilibrated within two dimensions before immobilisation on the mica surface. This was the intended outcome when the Mg^{2+} buffer was selected for the present study. It allows for the estimation of curvature values from *TP53* molecules under the most minimal of surface interactions.

5.3.6 Identifying Pre-Existing Curvature Trends in *TP53* and the Effect of the FF Algorithm

Before the initial application of the FF algorithm pre-existing curvature patterns were identified in the unoriented sets of DNA molecules. After application of the FF algorithm the curvature profiles that were produced closely resembled the pattern of curvature present in the unoriented dataset (Section 5.3.6.). A similar effect had been noticed by previous authors using the FF algorithm (Buzio *et al.*, 2012). The authors illustrated that the pre-existing trends in the unoriented data led to local minima in the objective function of the FF algorithm (the mean column variance). The local minima were reached before the desired global minima. These authors produced an example DNA sequence where the FF algorithm failed to produce meaningful results.

5.3.7 Amendments to the FF Algorithm

The original authors of the FF algorithm, assumed that DNA molecules deposited on a mica surface would be unoriented (Ficarra *et al.*, 2005b). This has been shown to be, at least in terms of direction (*i.e.* up or down on the mica surface) of curvature, to be partially false. For example, the thymine rich strand of DNA preferentially binds to inorganic crystal surfaces, such

as mica (Sampaolese *et al.*, 2002). It is unclear if this is the case in this study. However, there were weak but clear trends within the unoriented data. It was assumed that this provided the FF algorithm with a 'local minima' to its objective function.

In order to remove pre-existing trends from curvature profiles the orientation of each molecule in the analysis was completely randomised before reapplication of the FF algorithm (Section 5.2.2.6.). This satisfied the assumption that DNA molecules on a mica surface would be randomly oriented. The results were repeated multiple times, aligned and averaged to avoid isolated instances of local minima effecting the outcome of the FF algorithm. Assuming there was some underlying profile discernible by the FF algorithm it should have been identified by this method. The resulting profile, effectively flipping the result of the FF algorithm, provided a consensus curvature profile for both Exon 5-7 and Exon 5-9. A similar approach has been used by previous authors (Buzio *et al.*, 2012). The approach used by these authors was also to randomise before flipping, but to repeat multiple times and select the curvature profile that presented the lowest solution to the objective function. This represented a major hurdle for easy application of the FF algorithm and may be avoided in future studies by using other optimisation algorithms less effected by existing trends.

5.3.8 Comparisons of Curvature Profiles to Theoretical Profiles of *TP53*

The consensus curvature profile for Exon 5-9 provided an excellent agreement to the theoretical prediction (Section 5.2.2.7). The visual comparison was as good as those produced by previous studies using the FF algorithm (Ficarra *et al.*, 2005b; Buzio *et al.*, 2012). The agreement was improved by introducing two 180° preferential twists to the theoretical profile at two positions. The justification for this was detailed in Section 5.2.2.7. and has been used by previous researchers (Buzio *et al.*, 2012). It was highly likely that Exon 5-9 adopted a slightly different conformation on the surface than that predicted by the Geometric Deposition method. The increased agreement between the theory and experimental profile was visible in both 42 bp and 63 bp windows of curvature. Both windows had the FF algorithm applied separately and produced a similar result. Correlation analysis after amending the theoretical curvature profile showed strong significant positive correlation between experimental profile and theory (63 bp - Rho = 0.56, $p = <0.05$). Spearman's correlation coefficient values were not as high as the prediction based on simulated AFM images of Rho = ~0.9 (Section 4.2.11.). This was likely to be due to increased sources of interference in real AFM images and also the reduced accuracy of the FF algorithm. Additionally, exon 5 showed significantly lower curvature in Exon 5-9 (Kruskal Wallis, $p = <0.05$) which was in good agreement with predictions made using simulated AFM images (Section 4.2.15.). Overall *TP53* Exon 5-9 provided excellent agreement with predictions based upon the De Santis dinucleotide wedge model.

Exon 5-7 exhibited a curvature profile that had low visual similarity to the theoretical profile. The general shape of the flipped curvature profile at the 63 bp window of curvature was in moderate agreement with the theory with the exception of the region between 0.5 and 0.7 standardised length. There was only one instance of Exon 5-7 showing significant correlation between theoretical and experimental profiles (Table 5.3 – Amended 42 bp, Rho = 0.31, p = <0.05). It seemed likely the FF algorithm failed to correctly orient the majority of the DNA molecules for Exon 5-7. One explanation may be that Exon 5-7 showed a greater underestimation in comparison to theoretical length and may have undergone partial condensation or transition to A-DNA (Rivetti and Codeluppi, 2001; Sanchez-Sevilla *et al.*, 2002). This may have been facilitated by differences in the surface charge of the mica sheets. Alternatively, there may have been a repeated curvature motif in Exon 5-7 that was not accounted for by the De Santis dinucleotide wedge model that made it unsuitable for the FF algorithm. Unfortunately, due to the nature of the FF algorithm, the true orientation of the curvature profile was unknown and a definitive comparison was not possible.

5.3.9 Evaluating the Agreement between Experimental and Theoretical Curvature by Peak Shift for Exon 5-9

As *TP53* Exon 5-9 showed good visual agreement to the theoretical De Santis curvature profile the degree of peak shift between the experimental profile and the dinucleotide model was evaluated (Section 5.2.2.8.) At a 42 bp window the peak shift was 1.31 % and at a 63 bp window the peak shift was 2.25 %. These values were slightly larger than estimations based on simulated AFM images of 0.84 % and 1.27 % (Section 4.2.13.). Increases in peak shift percentages were expected due to the increased sources of molecule variance and image noise in real AFM images in comparison to simulated images.

It should be noted that not all peaks present in the theoretical profile were identified in the FF profile. The majority of these peaks can be accounted for by the merging of nearby peaks. This also accounts for the large peak introduced by the FF algorithm. The direction of the peaks (*i.e.* positive/negative) was not considered in the peak analysis; only their presence or absence. The direction of curvature was dependent on the conformation of the DNA molecules on the surface, which was itself dependent on the process of deposition. The methods of simulated deposition provided only one estimate of curvature direction as discussed more fully in the previous section. The addition or loss of peaks by the FF algorithm has been observed by previous authors (Buzio *et al.*, 2012). Experiments using simulated DNA molecules have indicated that the FF algorithm is most likely functioning at somewhere below ~87% accuracy at the two base pair window sizes estimated (Section 5.2.1). This may be

sufficient to explain some of the discrepancies between the experimental and theoretical profiles.

5.3.10 The Problem of Orientation after Flipping

Curvature profiles after application of the amended FF algorithm were aligned to provide the best agreement to theoretical profiles generated in Chapter 4. The need for this type of orientation has been considered a flaw in the FF algorithm by previous researchers (Buzio *et al.*, 2012). This can be avoided by the use of palindromic repeat dimers of the DNA tract under investigation (Ficarra *et al.*, 2005b). This created a repeat pattern in the final curvature profiles allowing orientation of the regions of resulting curvature profiles. The present study required the use of a large DNA molecule in order to cover the entire nucleotide sequence that codes for the sequence-specific DNA-binding domain of *TP53*. This sequence would have been too large to image as a palindromic dimers (~5 kb). Although imaging of large DNA molecules (>2.5 kb) is possible using AFM, the time taken to collect a sufficiently large number of images to perform curvature analysis would have been impractical (Reed *et al.*, 2007). However, this approach may be suitable for future studies on smaller regions of interest in *TP53*, such as exon positions, or other genes.

There was no suggested final orientation to the output of the FF algorithm. Therefore, detailed theoretical models were needed for comparison. These models alone were time-consuming and complex to produce. Additionally, their applicability to the results of the FF algorithm are limited if the results are sufficiently variable, as seen for *TP53* Exon 5-7. The need for these models greatly reduces the utility of the FF algorithm to researchers. Finally, there is no internal gauge as to the accuracy of the final output of the FF algorithm other than visual agreement with theoretical models. Idealised percentage accuracy can be generated for the FF algorithm using theoretical images, as has been performed in this study. This should be performed as an experiment-by-experiment optimisation.

5.4 Conclusions

The FF algorithm produced a curvature profile in good agreement with theoretical profiles of *TP53* Exon 5-9. The results for Exon 5-9 indicated that intrinsic DNA curvature in *TP53* was accurately predicted by the De Santis model of curvature. Furthermore, the majority of predicted peaks were present within the FF reconstructed profile although the magnitude of curvature was significantly different. Exon 5 was observed to exhibit significantly reduced curvature in comparison to intronic regions as predicted by the De Santis model of curvature. The experimental result indicated that there were some disagreements between the simulated deposition model and experimental DNA conformation. These results highlighted the inaccuracy of methods for simulating the deposition of DNA molecules on a flat surface. The FF algorithm was less successful for Exon 5-7, failing to produce a curvature in good agreement with theoretical profiles.

This study has identified a number of potential pitfalls when applying the FF algorithm to real DNA that had not been discussed by the original authors. The identification of an appropriate window over which to calculate curvature has been proven to be extremely important for the accurate reconstruction of curvature by the FF algorithm. This effect was quantified and could be used as a template for other researchers wishing to use the FF algorithm. A simple method of randomisation and repetition was also proposed for profiles containing weak curvature trends before application of the FF algorithm.

CHAPTER 6: ANALYSIS OF INTRINSIC DNA CURVATURE AND FLEXIBILITY OF EXONS 5 TO 9 OF THE *TP53* GENE USING STREPTAVIDIN END-LABELLING

6.1 Introduction

6.1.1 End Labelling of DNA Molecules for Orientation by AFM Analysis

The first reproducible AFM images of naked DNA were published in 1992 (Hansma *et al.*, 1992). It was not long before researchers realised the need for the identification of the orientation of DNA, a uniform and relatively featureless polymer. The first AFM based attempt at end-labelling used a chimeric fusion protein between streptavidin and two immunoglobulin G-binding domains of staphylococcal protein A (Murray *et al.*, 1993). Later that year another successful example made use of 5 nm colloidal gold spheres as a label for linear DNA molecules (Shaiu *et al.*, 1993). These techniques had been adapted from previous research on DNA using EM which labelled DNA with an avidin-ferritin-biotin complex (Muzard *et al.*, 1990). Protein labelling has also been used to identify structural features such as enzymatic 'nicks' (Murray *et al.*, 1993), abasic sites (Sun *et al.*, 2001) and direct haplotyping of DNA sequences by AFM (Woolley *et al.*, 2000). One study used dual labelling with different size proteins to identify both structural motifs and orientation (Woolley *et al.*, 2000; Sun *et al.*, 2001). AFM analysis was used to differentiate between the proteins by width, height or visual analysis. Protein labels have been shown to be effective for the study of DNA curvature and flexibility in a number of studies (Muzard *et al.*, 1990; Cognet *et al.*, 1999; Marilley *et al.*, 2005). Examples of protein end labels used in previous studies are presented in Figure 6.1.

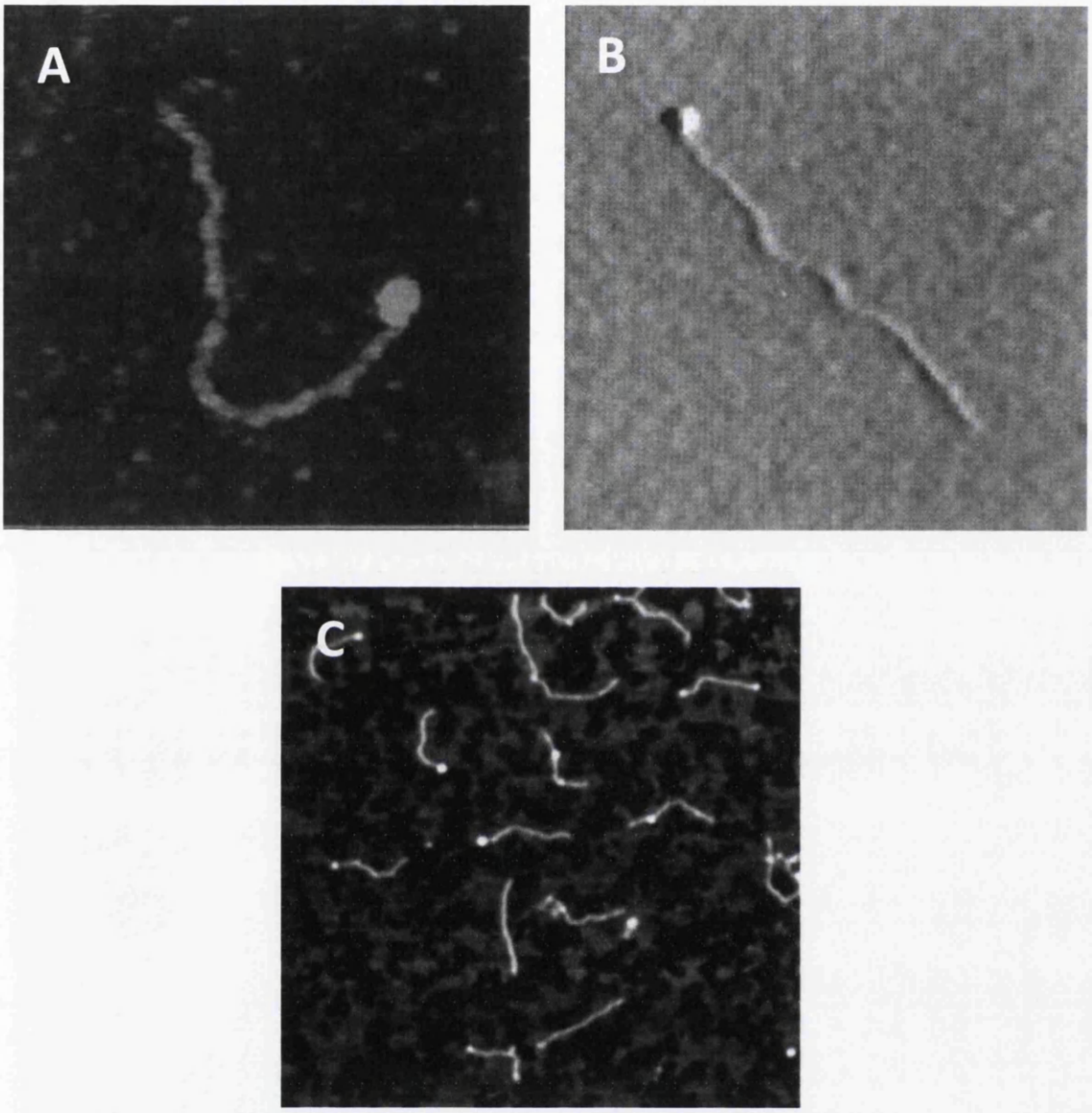

Figure 6.1. - Example images of end-labelled DNA taken from the available literature. A) End-labelling using a chimeric fusion protein between streptavidin and two immunoglobulin G-binding domains of staphylococcal protein A (Murray *et al.*, 1993). B) End-labelling using 5nm colloidal gold spheres bound to streptavidin-biotin (Shaiu *et al.*, 1993). C) End-labelling using a streptavidin-biotin complex (Marilley *et al.*, 2005).

6.1.2 Potential Conformational Effects on Local DNA Structure by Streptavidin End-Labelling

Streptavidin is a tetrameric protein with a high affinity for the vitamin biotin (Weber *et al.*, 1989). Biotin can be optionally incorporated into commercially available oligonucleotide PCR primers making the production of streptavidin-biotin end-labelled products for AFM analysis relatively simple. Some authors have commented that protein labelling could effect the DNA localised around the tag (Buzio *et al.*, 2012). This has been observed in one study that used streptavidin end-labelling of DNA imaging in air, although the nature and extent of the perturbation were not discussed (Marilley *et al.*, 2005). The reported interaction was not observed by the authors during liquid AFM imaging. However, this interaction was likely to be due to sample preparation methods as numerous authors have not reported any perturbation of the local structure of DNA when using streptavidin labelling for curvature or conformational analysis (Murray *et al.*, 1993; Rivetti *et al.*, 1996; Woolley *et al.*, 2000; Neish *et al.*, 2002; Seong *et al.*, 2002). A study specifically looking at DNA bound to mica under different conditions reported that streptavidin did not effect the ability of DNA to equilibrate onto a mica surface or have any measurable effect on DNA persistence length (Rivetti *et al.*, 1996). Other authors have called it a model ligand for DNA end-labelling (Neish *et al.*, 2002).

6.1.3 Aims and Objectives

A level of variability was observed in the curvature analysis of *TP53* by application of the FF algorithm in Chapter 5. In order to provide further corroboration of theoretical curvature measurements produced in Chapter 4 the following study utilised AFM analysis of *TP53* PCR products 5' end-labelled with a biotin molecule. The biotin molecules were conjugated to streptavidin proteins in order to provide orientation to a suitably large number of DNA molecules. Two PCR products were under investigation: one spanning exons 5 to 7 and the other exons 5 to 9. Thus oriented using the streptavidin end-label the DNA molecules were used to generate intrinsic DNA curvature and flexibility profiles for *TP53*. The experimental results were compared and contrasted with theoretical models that had been previously generated from simulated AFM images. An assessment of the curvature profiles and the relationship between curvature and exon positions was attempted alongside comparison of the two different experimental molecules used.

6.2 Results

6.2.1 Confirmation of Streptavidin End-Labelling

6.2.1.1 Identification of Streptavidin Binding using a Band Shift Assay.

After incubation of streptavidin with 5' biotinylated *TP53* DNA the product was run on a 1% agarose gel stained with ethidium bromide (Figure 6.2.). A slight but noticeable band shift was observed in the streptavidin labelled PCR product compared to the unlabelled. All bands occurred within the expected height range of 2500 bp.

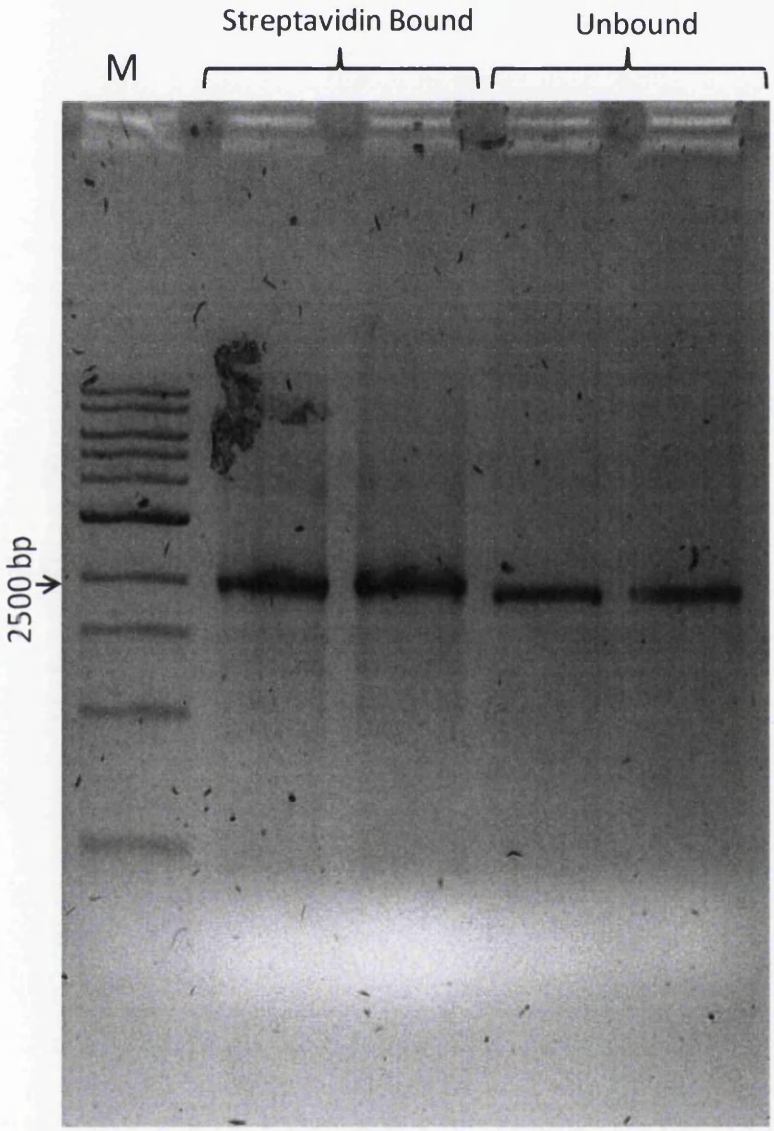

Figure 6.2. - Comparison of streptavidin bound and unbound 5' biotinylated *TP53* Exon 5-9 PCR amplification product. Lane M contained New England Biolabs 1 Kb DNA ladder. Lane 2 and 3 contained 5' biotinylated *TP53* Exon 5-9 (2500 bp) incubated with streptavidin overnight. Lane 4 and 5 contained 5' biotinylated *TP53* Exon 5-9.

6.2.1.2 Identification of Streptavidin Binding by Dot Blot Analysis

As a secondary confirmation of the efficiency of biotinylated *TP53* DNA to bind free streptavidin a Dot Blot analysis was performed. This was performed on the primers used in the PCR amplification of genomic DNA as a quality control for primer biotinylation (Figure 6.3.) and on the final PCR product for AFM analysis (Figure 6.4.). Strong banding was observed in the primer lanes. Weak to intermediate strength banding was observed in the lanes containing biotinylated PCR product. This was in line with expectation as the same weight to weight ratio of primer to PCR product contains a smaller amount of biotin molecules.

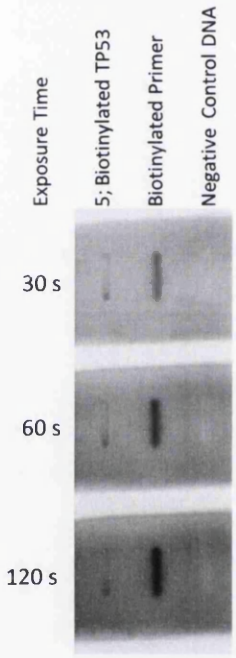

Figure 6.3. - Dot blot of biotinylated primer DNA. Lane 1 contained 100 ng of 5' biotinylated *TP53* Exon 5-9 amplification product. Lane 2 contained 100 ng of biotinylated 5' primer used in the amplification of the DNA product used in Lane 1. Lane 3 contained 100 ng of a negative control DNA that contained no biotin. Three different exposure times (30, 60 and 120 s) are shown.

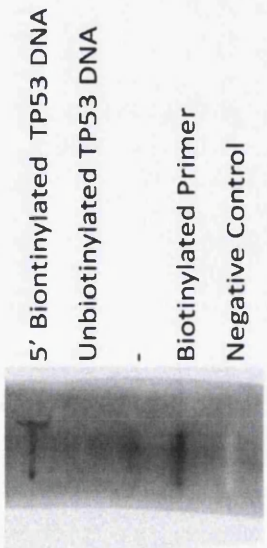

Figure 6.4. - Dot blot of *TP53* PCR product and biotinylated primer DNA. Lane 1 contains 5' biotinylated *TP53* Exon 5-7 amplification product. Lane 2 contained the unbiotinylated *TP53* Exon 5-7 amplification product. Lane 4 contained the 5' primer used in the amplification of the PCR product used in Lane 1. Lane 5 contained a negative control DNA. Each lane contains 250 ng of DNA.

6.2.2 Collection of Experimental AFM Images of 5' End-Labelled *TP53* DNA

A large number of AFM images were collected of *TP53* DNA labelled with streptavidin (examples in Figure 6.5). The final number of DNA contours extracted from the images was 1305 for Exon 5-7 and 588 for Exon 5-9.

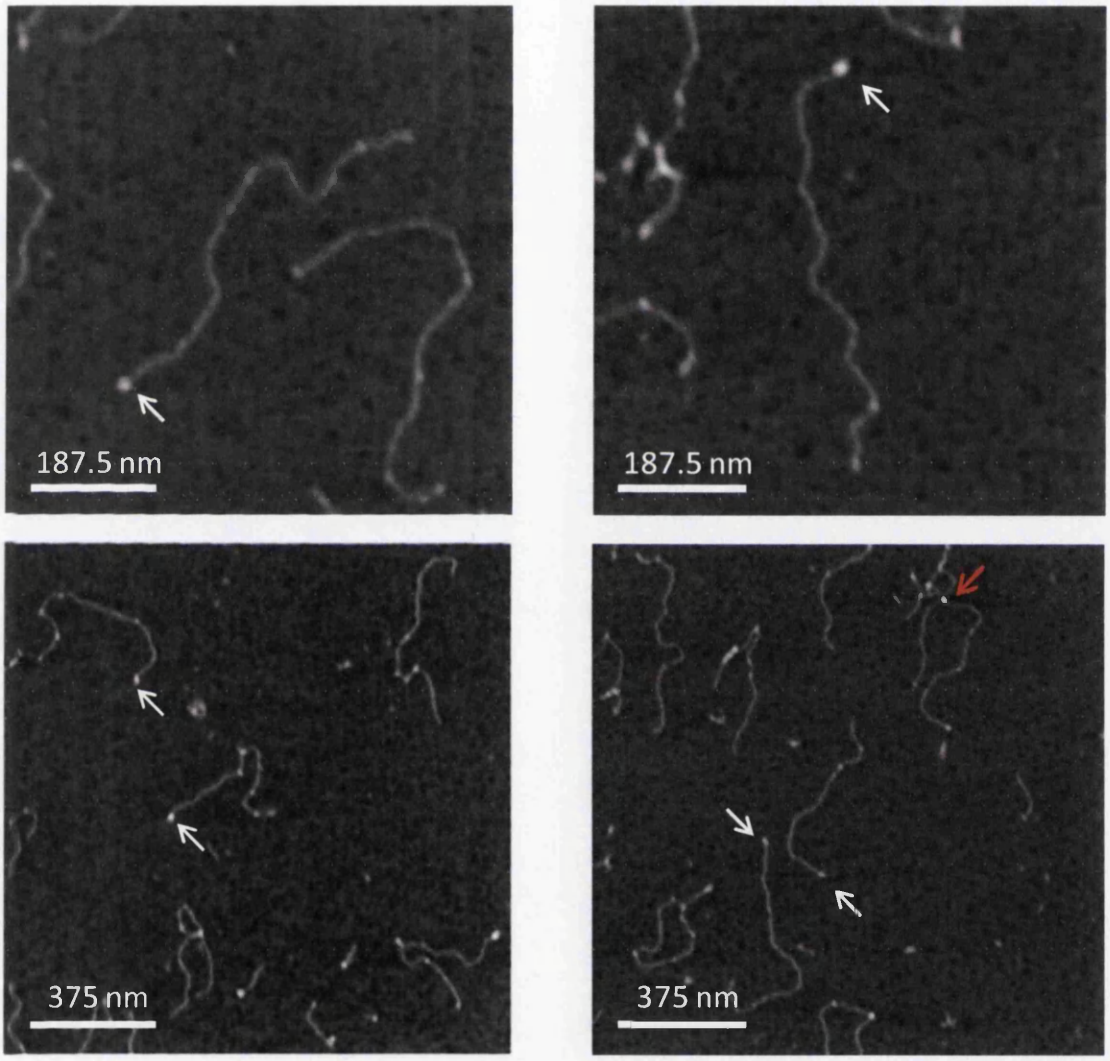

Figure 6.5. – Examples of *TP53* DNA end-labelled with streptavidin-biotin. End labels are indicated with white arrows. An example of multiple *TP53* molecules bound to one streptavidin molecules is indicated with a red arrow. Scale bars are in nanometres.

6.2.3 Removal of Unsuitable DNA Molecules by Z-Height Analysis.

During AFM image processing the Z-height values were recorded for each molecule included within *TP53* Exon 5-9 and *TP53* Exon 5-7 datasets. The values were background corrected by fitting and subtracting a 1 degree polynomial. The contour length of each molecule was standardised from 0 to 1 by interpolating a linear line between each pixel coordinate and selecting a set number of equally spaced points. The mean Z-height at each comparable point was calculated (Figure 6.6.A. and Figure 6.6.C.).

To ensure that only labelled molecules were analysed a Z-correction step was introduced. Z-height at each end of each molecule was compared. The values analysed lay within a standardised length of 0-0.05 and 0.95-1.0 (the beginning and end 5 % of the molecule) where the streptavidin end-label was observed. The expectation was that the Z-values for the labelled end of each molecule would be larger than the unlabelled end. Any molecules that were not in line with this expectation were removed from the analysis. For *TP53* Exon 5-7 a total of 365 (940 molecules remaining) molecules were removed from a set of 1305 molecules to give an error rate of 27.96 %. For *TP53* Exon 5-9 a total of 85 molecules (503 molecules remaining) were removed from a set of 588 molecules. This gave an error rate of 14.46 %. The final number of DNA molecules remaining after Z-correction was 940 for *TP53* Exon 5-7 and 503 for *TP53* Exon 5-9.

The mean Z-height was recalculated (Figure 6.6.B and Figure 6.6.D). It can be observed that the Z-height at the unlabelled end (1.0 in standardised notation) was reduced after Z correction, as per expectation. However the Exon 5-9 dataset still retained a small peak in mean Z-height at the untagged end (Figure 6.6.D. - green square).

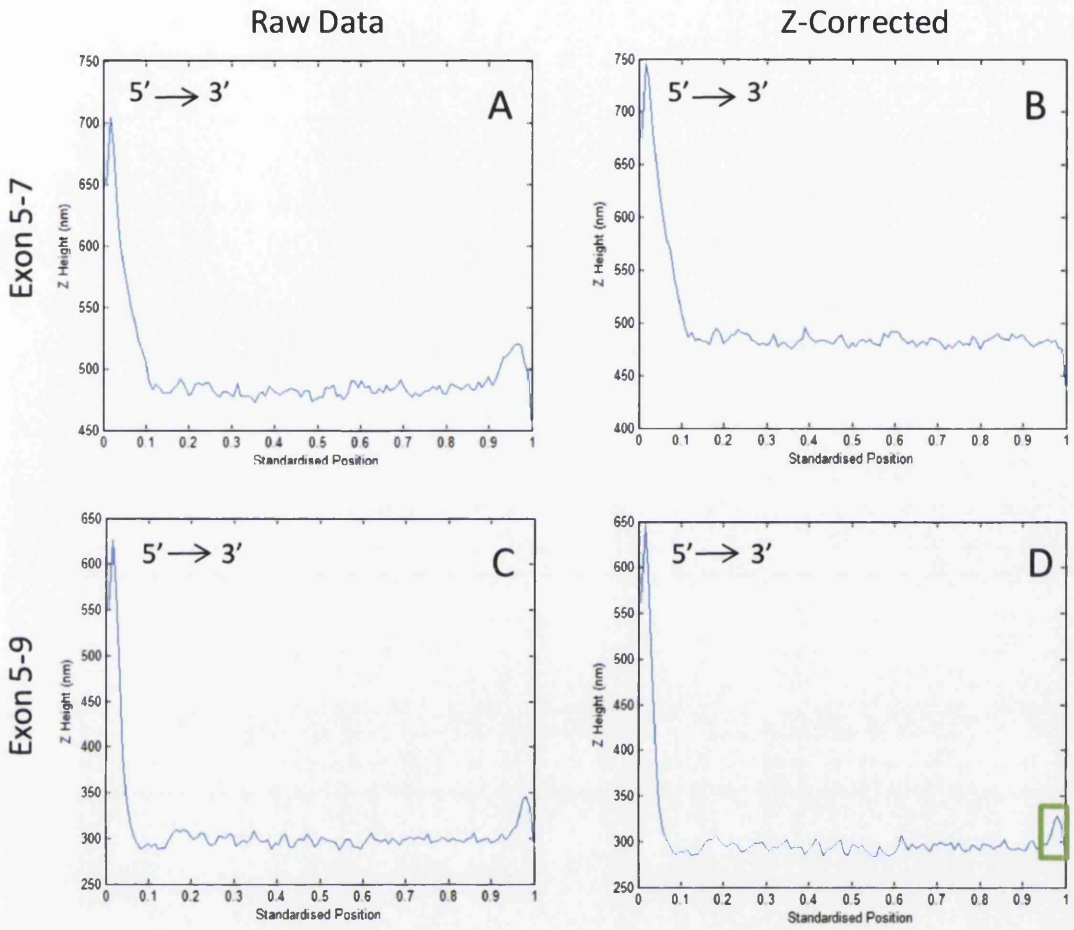

Figure 6.6. - Mean Z-height values at standardised length for *TP53* Exon 5-7 and Exon 5-9 molecules. A) Z-heights of original dataset for *TP53* Exon 5-9. B) Z-height of *TP53* Exon 5-9 after Z-correction. C) Z-height of original dataset for *TP53* Exon 5-7. D) Z-height of *TP53* Exon 5-9 after Z-correction. Z-heights were background corrected by subtracting a 1 degree polynomial and averaged for each dataset. The contour length measurements (x-axis) were standardised to a scale of 0 to 1. Z-correction was implemented by removing all molecules with a greater mean Z-height at the untagged end than the tagged end (*i.e.* comparing the first and last 5% of each molecule). A small increase in Z-height for Exon 5-7 remained after Z-correction (green square).

6.2.4 Reconstructed Length Measurements

Reconstructed length measurements were calculated for the datasets after Z-correction (Figure 6.7.). *TP53* Exon 5-7 exhibited a non-normal distribution before and after log transformation (Shapiro-Wilks, $p = <0.05$). Exon 5-7 had a median contour length value of 563 nm. *TP53* Exon 5-9 exhibited a non-normal before and after log transformation (Shapiro-Wilks, $p = <0.05$). Exon 5-9 had a median value of 767 nm. A summary of the reconstructed length of Exon 5-7 and Exon 5-9 can be found in Table 6.1 and Figure 6.7.

There were a number of molecules in the Z-corrected dataset with contour lengths that did not lie within the main distribution. It was necessary to remove these outlying molecules; this was achieved by selecting out a number of molecules around the median value of the distribution for further analysis. The removal of obviously erroneous molecules has been performed in numerous studies (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Marek *et al.*, 2005). The relevant statistics of the dataset before and after outlier removal are presented in Table 6.1.

Approximately 0.34 nm per base pair was the consensus length for B-DNA taken from X-ray crystallography experiments (Saenger, 1984). The median reconstructed length measurement of 560 nm for *TP53* Exon 5-7 underestimated the theoretical measurement of B-DNA of 631 nm (1855 bp x 0.34 nm) by 11.17 %. The same held true for Exon 5-9 with a theoretical value of 850 nm compared to the experimental measurement of 771 nm. This was a 9.24 % underestimation. A similarly large reduction in reconstructed length in comparison to theoretical length for B-DNA was not observed in unlabelled DNA (4.36 %. and 1.20 % for Exon 5-7 and Exon 5-9 respectively – Section 5.2.2.2.).

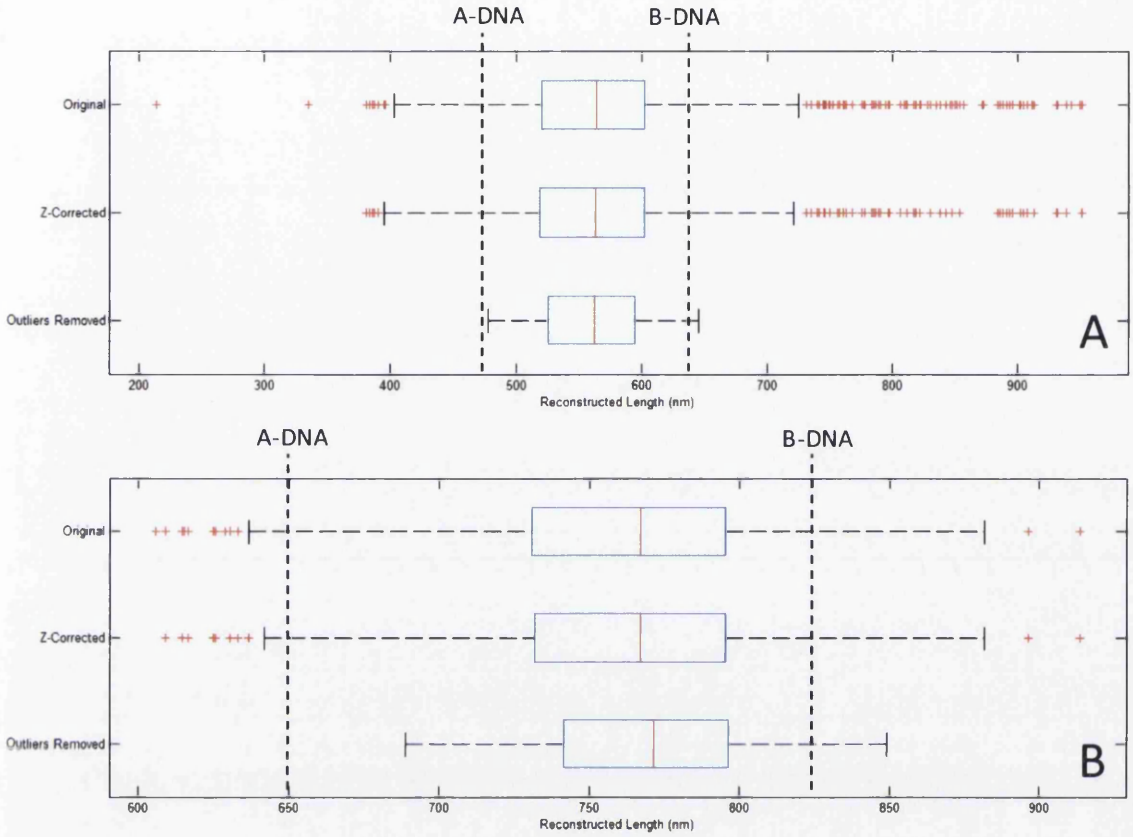

Figure 6.7. - Boxplot of reconstructed length measurements for *TP53*. A) *TP53* Exon 5-7. B) *TP53* Exon 5-9. On each box, the central red mark is the median, the edges of the blue box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as red crosses. Theoretical contour length measurements for A-DNA and B-DNA are indicated with a broken line (Saenger, 1984).

TP53 Exon 5-7	Number Of Molecules	Normality Test [Shapiro-Wilks,] (p-value)	Median (nm)	IQR [Q3-Q1] (nm)
Original Data	1305	<0.05	565	82.21
Z-Corrected	940	<0.05	563	83.58
Outliers Removed	800	<0.05	560	72.95
TP53 Exon 5-9				
Original Data	588	<0.05	767	64.66
Z-Corrected	503	<0.05	767	63.46
Outliers Removed	450	<0.05	772	54.66

Table 6.1. - Summary of the properties of *TP53* Exon 5-7 and *TP53* Exon 5-9 datasets at each stage of outlier molecule identification and removal. The median and IQR values were generated from the reconstructed length measurements from the appropriate dataset. The Shapiro-Wilks test for normality was performed on reconstructed length measurements from the same datasets. Significant p-values are indicated in red.

6.2.5 Analysis of Correlation between End-Label Z-Height and Reconstructed Length.

There have been reports of interaction between the protein end labels and experimental DNA sequences (Marilley *et al.*, 2005). While the nature of the interaction was not explicitly stated by the authors the possibility of interaction was investigated. The maximum end-label height, taken as the maximum Z-value of the first and last 12 data points of each molecule, was plotted against the reconstructed length of each molecule (Figure 6.8.). A correlation analysis was performed on each set of molecules. Exon 5-7 exhibited no correlation (Spearman's Rank, Rho = -0.05, p = 0.15) and Exon 5-9 exhibited significant weak negative correlation (Spearman's Rank , Rho = -0.40, p = <0.05).

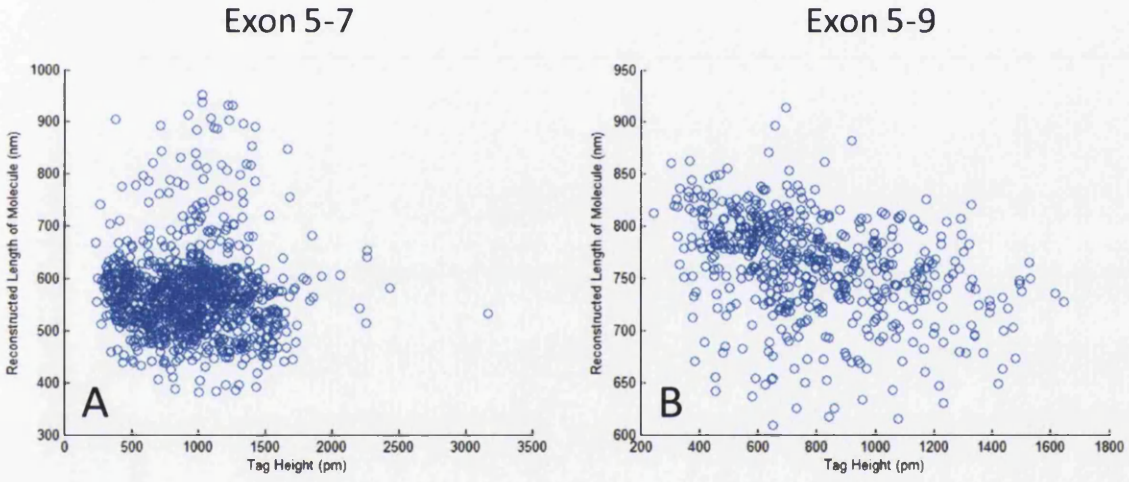

Figure 6.8. - Maximum end-label height plotted against contour length measurements of DNA molecules. A) *TP53* Exon 5-7. B) *TP53* Exon 5-9.The maximum Z-height of the first 10 % of a molecule was taken as the maximum tag height.

6.2.6 Persistence Length Measurements of *TP53*

The persistence length for both Exon 5-7 and Exon 5-9 *TP53* datasets was calculated as detailed in Section 3.2.4.2. The range of curvilinear distances investigated was 0-200 nm and 0-300nm (Figure 6.9 and Figure 6.10). The persistence length calculated for the range of 0-200 was ξ = 56 and ξ = 54 for Exons 5-7 and Exons 5-9 respectively. The persistence length calculated for the range of 0-300 was ξ = 61 and ξ = 60 for Exons 5-7 and Exons 5-9 respectively.

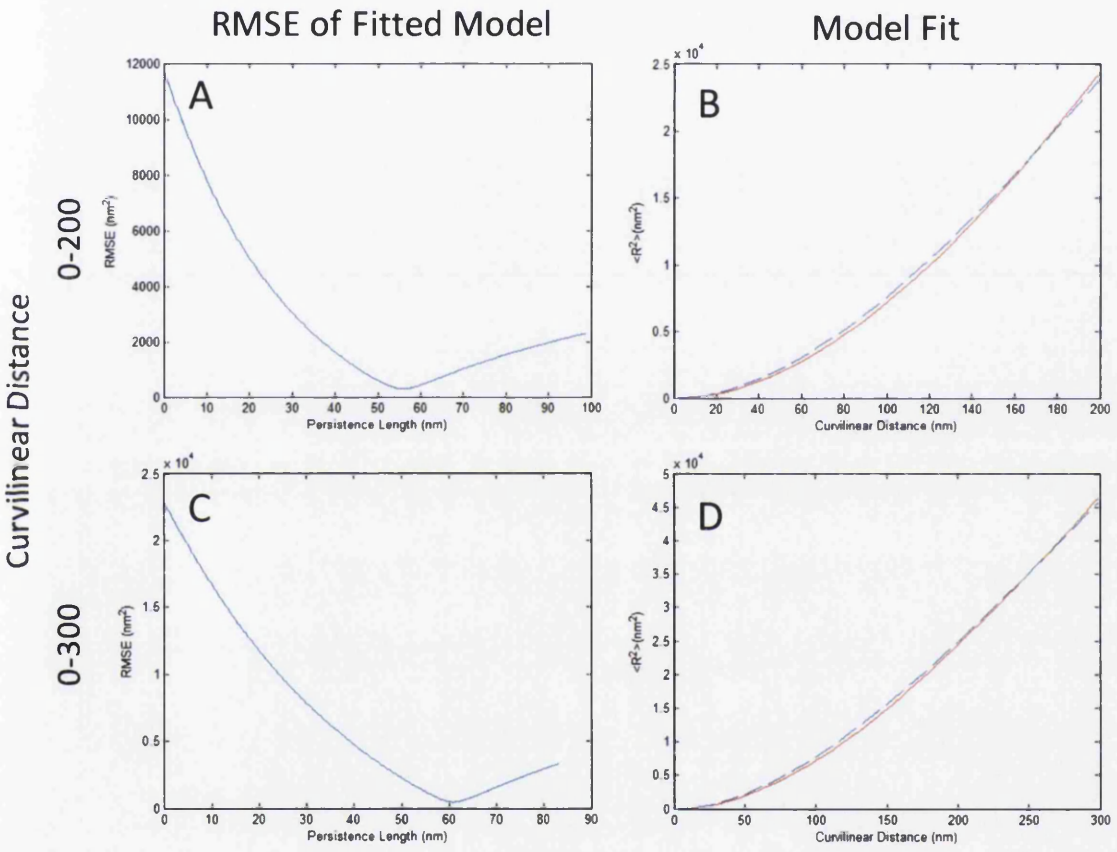

Figure 6.9. - Experimentally determined DNA persistence length for Exon 5-7 by comparison to theoretical values of $\langle R^2 \rangle$ from the WLC model. A) RMSE of fitted models generated from the WLC theory using a range of persistence lengths for experimental $\langle R^2 \rangle$ values. B) Experimental $\langle R^2 \rangle$ values (red line) alongside predicted $\langle R^2 \rangle$ values (broken blue) for the WLC model at a persistence length of 56 nm for a range of curvilinear distances from 0-200 nm. C) RMSE of fitted models generated from the WLC theory using a range of persistence lengths for experimental $\langle R^2 \rangle$ values. D) Experimental $\langle R^2 \rangle$ values (red line) alongside predicted $\langle R^2 \rangle$ values (broken blue) for the WLC model at a persistence length of 61 nm for a range of curvilinear distances from 0-300 nm.

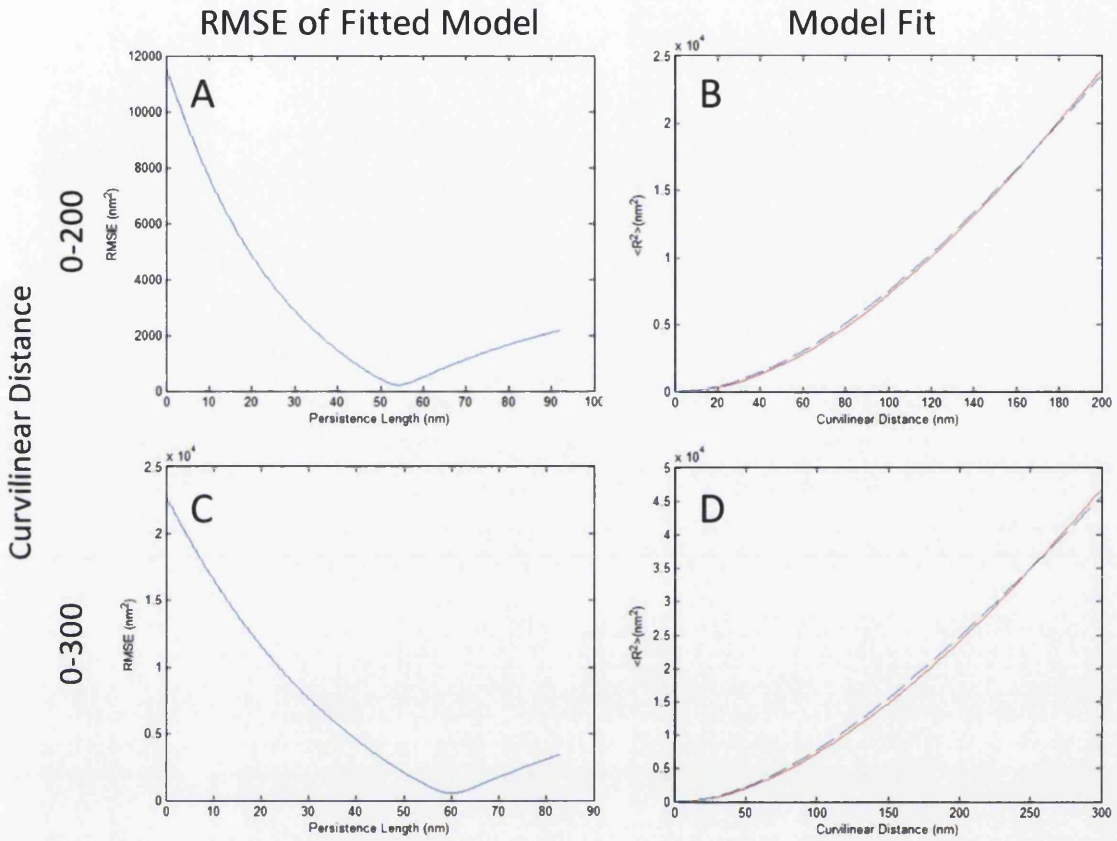

Figure 6.10. - Experimentally determined DNA persistence length for Exon 5-9 by comparison to theoretical values of $\langle R^2 \rangle$ from the WLC model. A) RMSE of fitted models generated from the WLC theory using a range of persistence lengths for experimental $\langle R^2 \rangle$ values. B) Experimental $\langle R^2 \rangle$ values (red line) alongside predicted $\langle R^2 \rangle$ values (broken blue) for the WLC model at a persistence length of 54 nm for a range of curvilinear distances from 0-200 nm. C) RMSE of fitted models generated from the WLC theory using a range of persistence lengths for experimental $\langle R^2 \rangle$ values. D) Experimental $\langle R^2 \rangle$ values (red line) alongside predicted $\langle R^2 \rangle$ values (broken blue) for the WLC model at a persistence length of 60 nm for a range of curvilinear distances from 0-300 nm.

6.2.7 Selection of Base Pair Window for Curvature Calculation

It has previously been shown that at low window sizes there was a measurable effect of digitisation on DNA contours (Section 4.2.7.). There were a number of sources of information available for the selection of appropriate base pair window sizes. The Visual Threshold was applied to *TP53* Exon 5-7 and Exon 5-9, detailed and developed in Section 3.2.5.2. This allowed for the visual assessment of curvature calculated over a range of base pair windows sizes. This was used to identify the influence of digitisation of the DNA contour on curvature angle measurements. The results of the Visual Threshold are shown in Figure 6.11. Both Exon 5-7 and Exon 5-9 followed the expected pattern. The minimum curvature window sizes were 62 nm for Exon 5-7 and 45 nm for Exon 5-9. The ranges of acceptable base pair windows suggested were 38 – 103 bp for Exon 5-7 and 27 - 74 bp for Exon 5-9.

Another method for measuring the influence of digitisation on curvature angle calculations was to look at the dataset maximum and minimum angles calculated and the number of matching occurrences of the extrema values within the dataset. The results are presented in Figure 6.12. It was observable that below a window size of 31 bp there were multiple instances of individual angles that matched the dataset extrema. This was more pronounced in Exon 5-7. This indicated that at window sizes lower than 31 bp the choice of interpolant would have an effect on curvature measurements.

The window sizes of 42 bp and 63 bp were used for further analysis. In some instances the window size of 21 bp was included for comparison to previous research. The window sizes of 42 and 63 bp lie within experimentally determined optimal ranges. Additionally, these window sizes have been shown to provide good curvature peak-to-background contrast in theoretical studies of *TP53* (Section 4.2.7.).

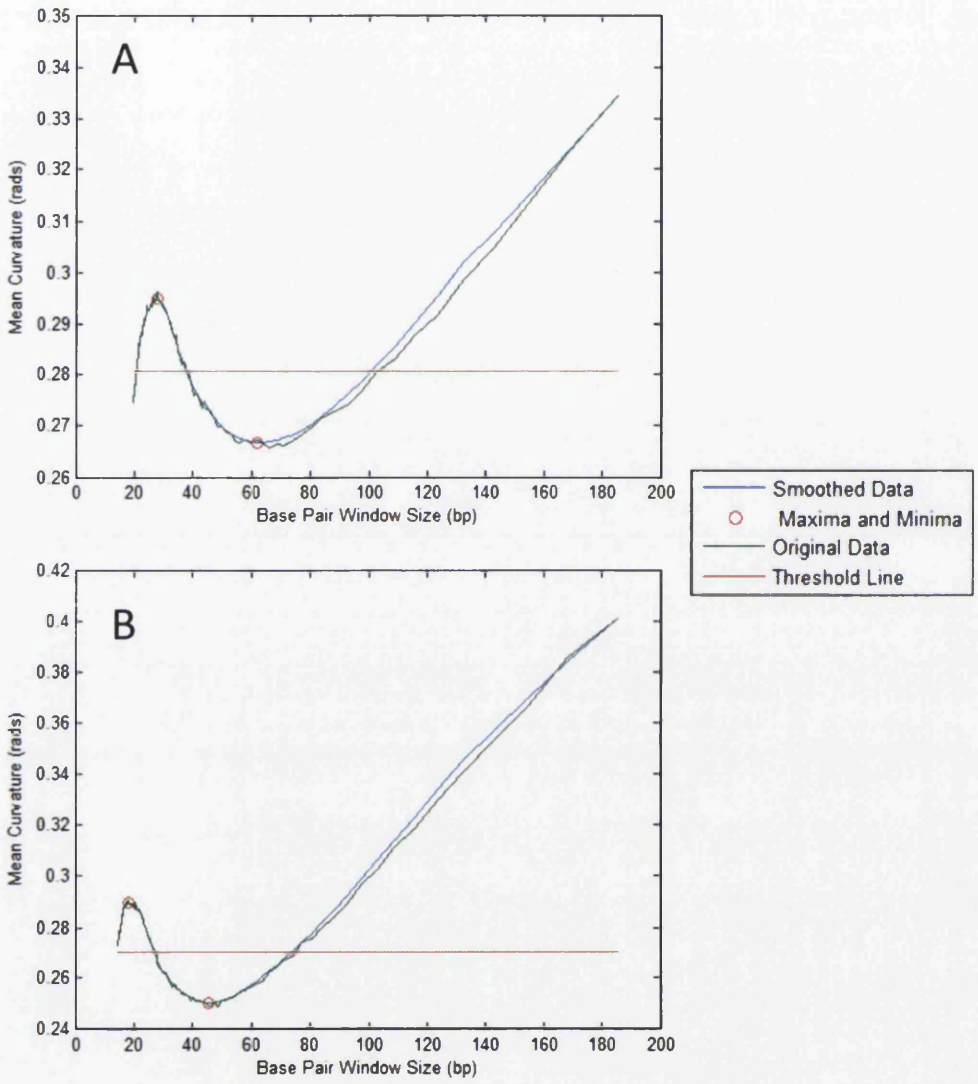

Figure 6.11. – The Visual Threshold applied to the mean curvature of *TP53*. A) *TP53* Exon 5-7 (minima = 62 bp). B) *TP53* Exon 5-9 (minima = 45 bp). Mean curvature is plotted as a green line, smoothed (three point moving average) as blue, the maxima and minima values are denoted as red circles. The threshold is denoted as a red line, acceptable windows of curvature lie below the threshold line.

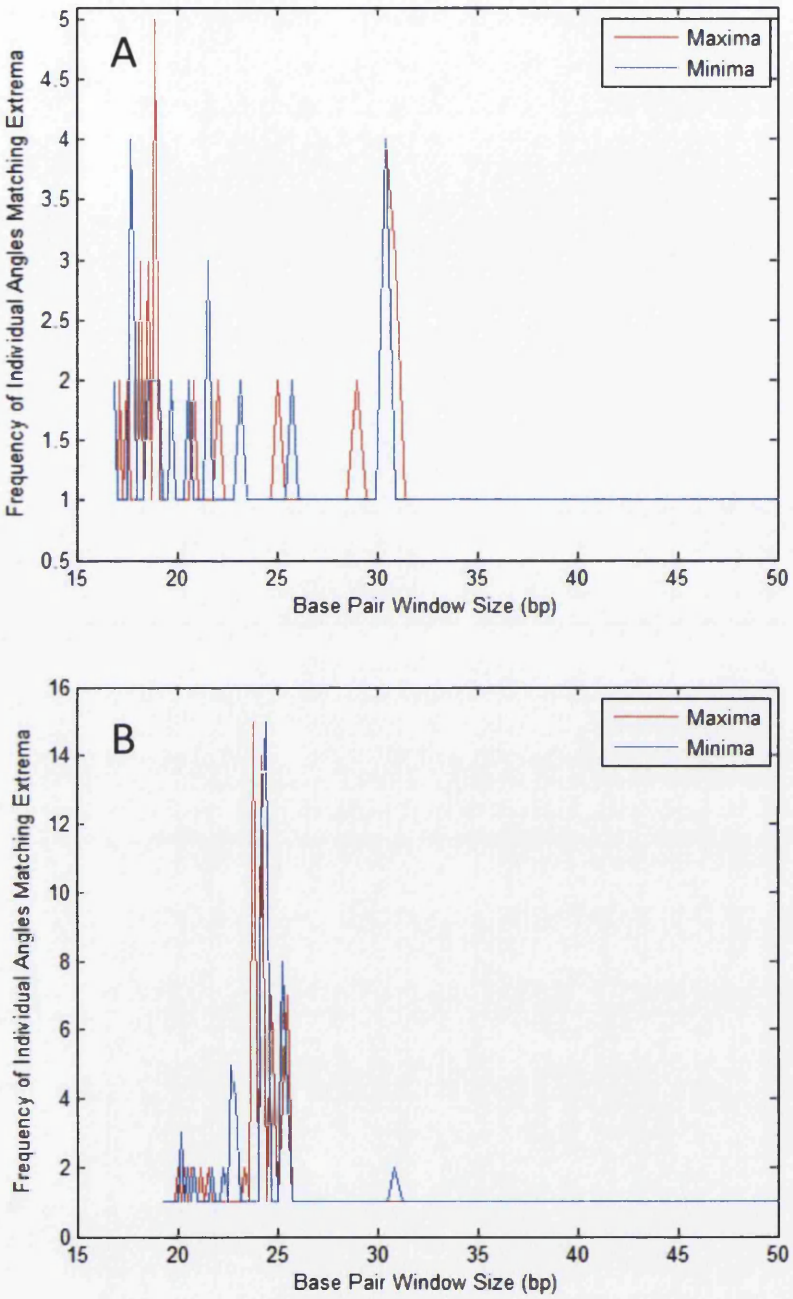

Figure 6.12. - Number of individual angles within the dataset that match the dataset extrema. A) *TP53* Exon 5-7. B) *TP53* Exon 5-9. Angles were calculated at a number of equally spaced points appropriate for the base pair window. Plots were truncated at a base pair window size of 50 bp as no values above zero occurred at larger window sizes. Matching maxima are indicated in red and matching minima in blue.

6.2.8 Unsigned Curvature for Exon 5-7

For *TP53* Exon 5-7 all base pair window sizes exhibited gradually modulating curvature across the molecule with peaks of high curvature at each end (Figure 6.13.) The highest peaks of curvature within the main body of the profiles occurred within the intron between exons 6 and 7 in the 42 bp profile and within the intronic region before exon 5 in the 63 bp profile. Exon 5 exhibited the lowest curvature of all the exon regions and the lowest curvature values of the corresponding profiles. Exon 5 was bordered by regions of moderate curvature in both profiles. Similarly, there was a slight trough in curvature for exon 6 in both profiles, bordered by regions of moderate curvature. Exon 7 occurred as a small curvature peak in all profiles.

An analysis of curvature between experimental and theoretical profiles showed no significant correlation for either window size (Spearman's Rank: 42 bp - Rho = 0.04, p = 0.71; 63 bp – rho = 0.14, p = 0.29). However, there were a number of visually identifiable similarities. A peak-trough-peak pattern in the intron between exon 6 and 7 was apparent in the experimental and theoretical profiles, but with lesser contrast in the experimental profile. A large peak at the 5′ end of the 63 bp profile could correspond with an amalgamation of the two large peaks observable within the theoretical profile. The theoretical profiles did not predict the extent of the large dip in curvature in exon 5 observed within both experimental profiles.

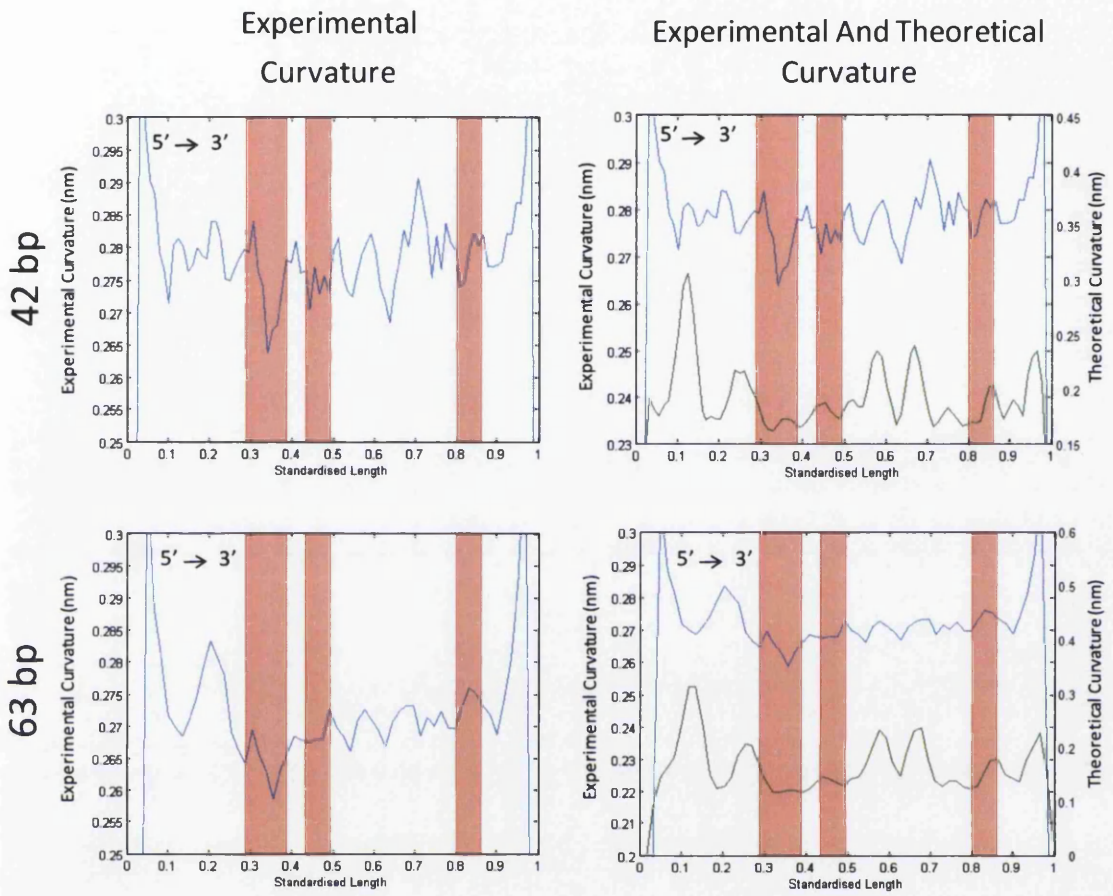

Figure 6.13. - Unsigned curvature profiles for *TP53* Exon 5-7 (n=800) calculated over a range of base pair windows and comparable theoretical profiles. The base pair window size is shown on the left hand side of the figure. Profiles were smoothed with a three point average filter. Exon positions are highlighted in red in ascending order from left to right. Experimental profiles are indicated by a blue line and theoretical profiles by a green line. Theoretical profiles were generated using the De Santis model of curvature and the Geometric Deposition method. Theoretical profiles were rescaled for easy visual comparison (z-axis shows theoretical curvature values where applicable). The length of the molecule was standardised from zero to one (5′ to 3′).

6.2.9 Unsigned Curvature for *TP53* Exon 5-9

Unsigned curvature profiles of Exon 5-9 are presented in Figure 6.14. All exons, with the exception of exon 6, exhibited a trough in local curvature at both window sizes. This was especially apparent for exon 6 in both profiles and exon 8 in the 63 bp profile. Exon 6 and exon 8 respectively bordered by or encompass the two lowest regions of curvature in both profiles. Exon 5 had a trough in local curvature towards the 3' end of the exon and the 5' end of the exon began with a region of moderate curvature. The same was observable for exon 8. Within the 42 bp profile the largest peaks of curvature, excluding the end regions, occurred before exon 5 and after exon 9. The 63 bp window size profile exhibited a large peak of curvature within the intron between exon 6 and 7.

An analysis of curvature between experimental and theoretical profiles showed no significant correlation for either window size (Spearman's Rank: 42 bp - Rho = -0.19, p = 0.04; 63 bp – Rho = <0.00, p = 0.98). The experimental and theoretical profiles have some visual similarities. The central intron, between exon 6 and exon 7 for both window sizes contained a noticeable increase in curvature in both experimental and theoretical profiles. The local region bordering exons 8 and 9 was in good agreement with theoretical profiles; two observable troughs in curvature encompassed both exons. The large peak of curvature directly preceding exon 5 corresponded with a prominent peak in the theoretical profile. Exons 6 and 7 contained a peak in experimental profiles which was not present in theoretical profiles. The large peaks of curvature predicted at either end of the sequence were notably missing from all experimental profiles. The large curvature peak may have been included in the large regions of curvature at each end of the molecule.

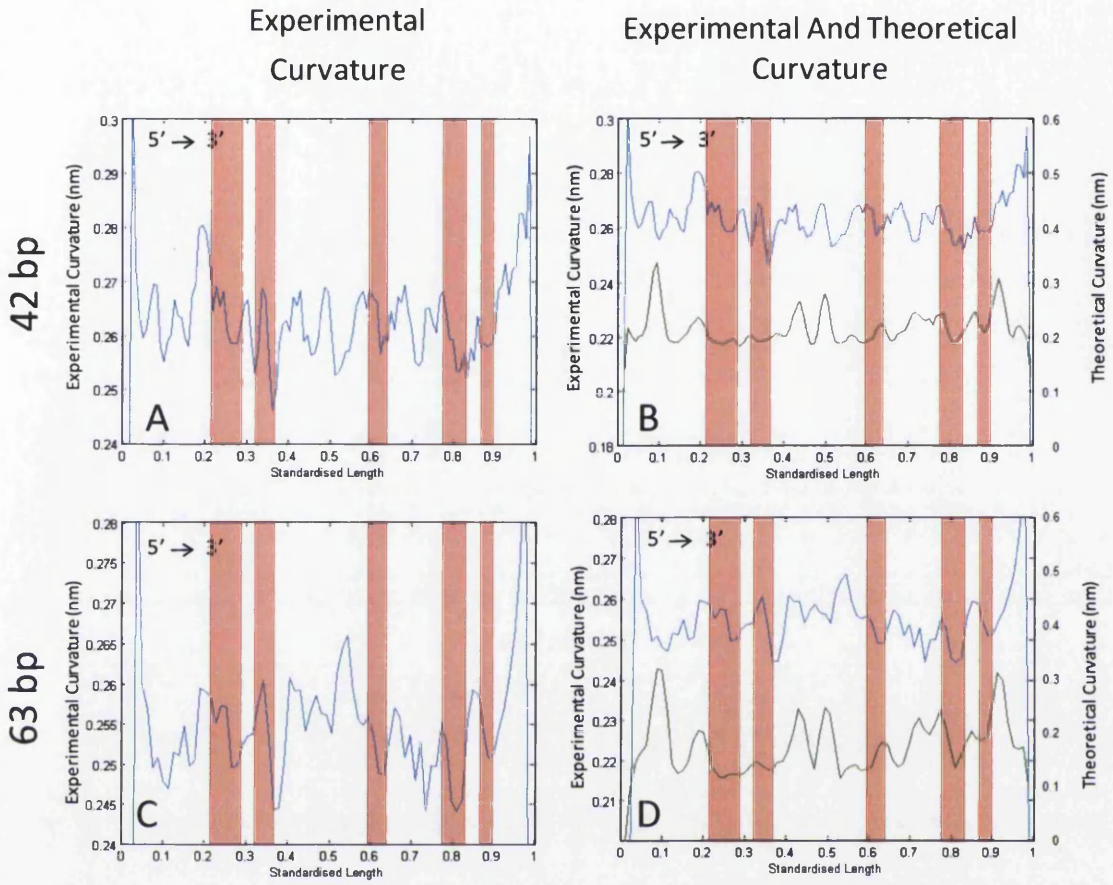

Figure 6.14. - Unsigned curvature profiles for *TP53* Exon 5-9 (n=450) calculated over a range of base pair windows and comparable theoretical profiles. The base pair window size was shown on the left hand side of the figure. Profiles were smoothed with a three point average filter. Exon positions are highlighted in red in ascending order from left to right. Experimental profiles are indicated by a blue line and theoretical profiles by a green line. Theoretical profiles were generated using the De Santis model of curvature and the Geometric Deposition method. Theoretical profiles are rescaled for easy visual comparison (z-axis shows theoretical curvature values where applicable). The length of the molecule was standardised from zero to one (5' to 3').

6.2.10 Signed Curvature for Exon 5-7

Within the signed curvature profiles of Exon 5-7 it was observed that exon positions occurred in, or were bounded by, regions of lower curvature (Figure 6.15.). Exon 5 and 6 bordered a change in the direction of curvature in all profiles. Exon 5 and Exon 7 consistently contained a slight reduction in curvature. Exon 6 occurred at a transition between two curved regions with different directions of curvature. The largest peaks of curvature were contained within intron regions at both window sizes. The largest peaks of curvature predominated at either end of the molecule. However, within the 63 bp profile one of the most extreme peaks of curvature occurred between Exon 5 and Exon 6.

An analysis of curvature between experimental and theoretical profiles showed no significant correlation for either window size (Spearman's Rank: 42 bp - Rho = 0.17, p = 0.11; 63 bp – Rho = 0.24, p = 0.07). The p-value for the 63 bp window was borderline for statistical significance and may have indicated a weak positive correlation. There were strong visual similarities between the theoretical and experimental profiles at both window sizes (42 and 63 bp). This visual similarity was apparent within the general shape of the profile, although not in the magnitude of the peaks. In addition to this, some of the peaks had opposite direction of curvature to those predicted by the theoretical profile. At both window sizes exon 5 and exon 6 contained a region of negative curvature which was also apparent within the theoretical profiles. The intron region between exon 6 and 7 contained peaks of positive curvature in experimental and theoretical profiles. The 3' end of the DNA sequence also exhibited similarities in shape, although the theoretical profile predicted almost no curvature within this region while the experimental profile exhibited a small peak in curvature. The protein labelled 5' end of the molecule showed the largest differences between experimental and theoretical profiles. The trough at around 0.1 standardised length in the theoretical profiles may have corresponded with the large peak in the experimental profiles.

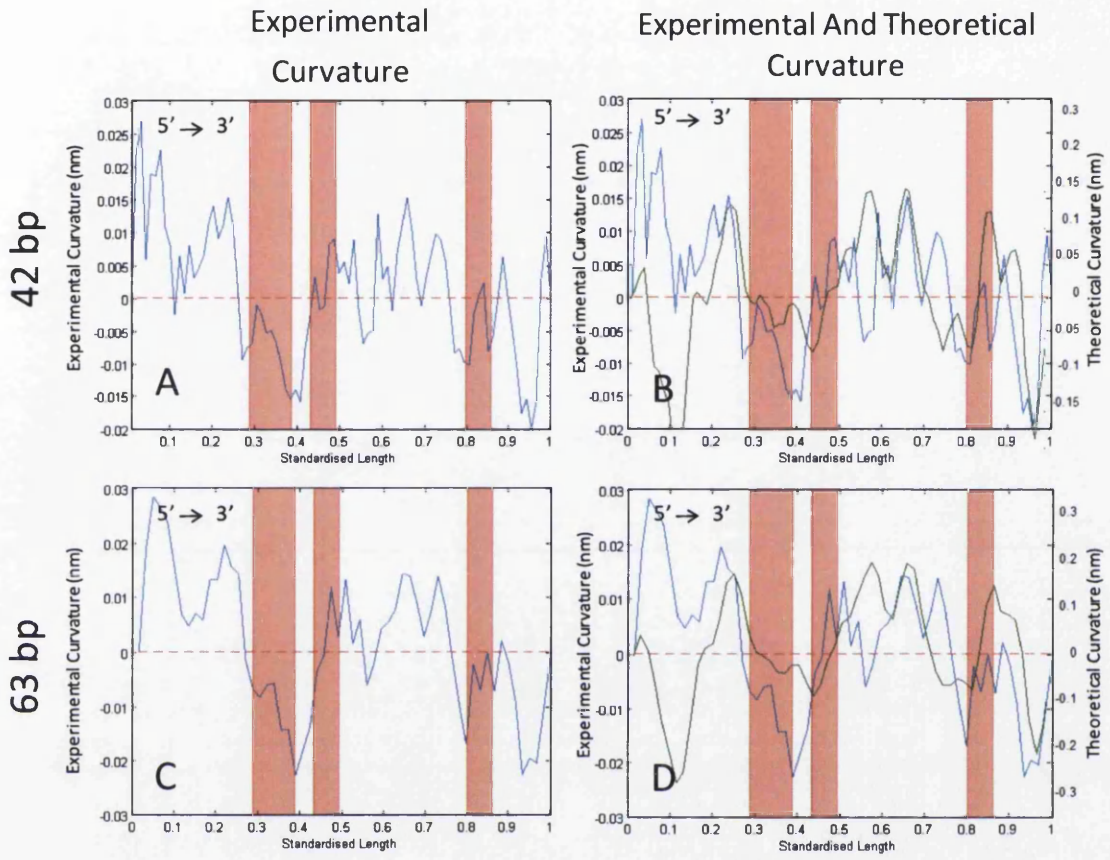

Figure 6.15. - Signed curvature profiles for *TP53* Exon 5-7 (n=800) calculated over a range of base pair windows and comparable theoretical profiles. The base pair window size is shown on the left hand side of the figure. Profiles were smoothed with a three point average filter. Exon positions are highlighted in red in ascending order from left to right. Experimental profiles are indicated by a blue line and theoretical profiles by a green line. Theoretical profiles were generated using the De Santis model of curvature and the Geometric Deposition method. Theoretical profiles are rescaled for easy visual comparison (z-axis shows theoretical curvature values where applicable). The length of the molecule was standardised from zero to one (5′ to 3′).

6.2.11 Signed Curvature for Exon 5-9

Exons 5, 7, 8 and 9 occurred within regions of low curvature relative to the rest of the profile (Figure 6.16.). This was the case at each base pair window size. Exon 6 appeared as a small peak in curvature at all window sizes. The region between exons 8 and 9 contained a large peak in curvature. Exon 9 showed a small localised curvature peak in the 42 bp profile that was absent from the 63 bp profile. The largest peak (~0.03 radians) in curvature occurred within the intron region between exons 6 and 7. Another large peak in curvature occurred between exons 8 and 9. There were two closely spaced peaks of moderate curvature at the 5' end of the sequence before exon 5.

An analysis of curvature between experimental and theoretical profiles showed no significant correlation for either window size (Spearman's Rank: 42 bp - Rho = 0.05, p=0.60; 63 bp – Rho = 0.12, p=0.30). The general shapes of the theoretical and experimental curvature profiles showed moderate visual comparability in the occurrence of peaks, if not their direction or magnitude. There was slightly less visual agreement at larger window sizes which was attributed to a larger shifting of the location of the peaks within the higher window size profile. The region containing exon 5 and exon 6 and the region containing and bordering exon 8 were comparable between experimental and theoretical profiles. The intron region between exon 6 and exon 7 showed dissimilarity to the theoretical profile.

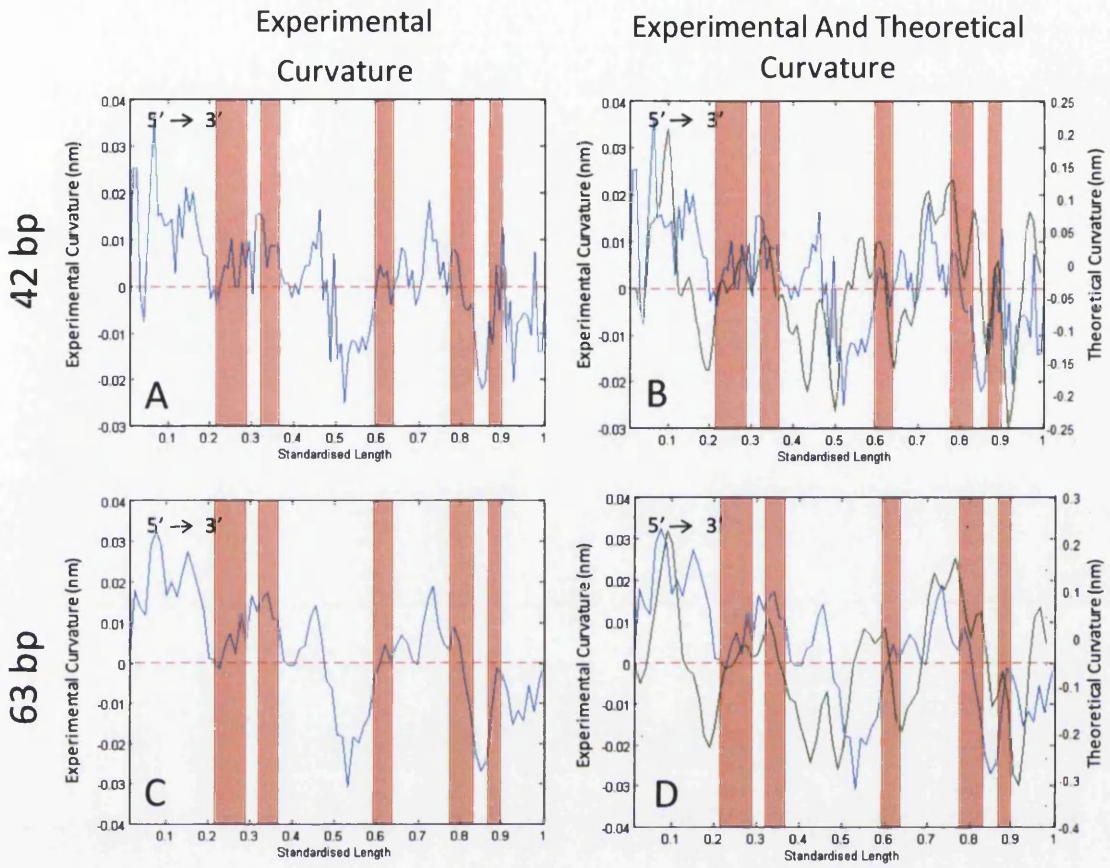

Figure 6.16. - Signed curvature profiles for *TP53* Exon 5-9 (n=800) calculated over a range of base pair windows and comparable theoretical profiles. The base pair window size is shown on the left hand side of the figure. Profiles were smoothed with a three point average filter. Exon positions are highlighted in red in ascending order from left to right. Experimental profiles are indicated by a blue line and theoretical profiles by a green line. Theoretical profiles were generated using the De Santis model of curvature and the Geometric Deposition method. Theoretical profiles are rescaled for easy visual comparison (z-axis shows theoretical curvature values where applicable). The length of the molecule was standardised from zero to one (5′ to 3′).

6.2.12 Comparison of Curvature and Flexibility Profiles between Exon 5-7 and Exon 5-9

The curvature profiles generated in Section 6.2.8 - 6.2.11. were aligned. The data points of Exon 5-9 that corresponded to the sequence in Exon 5-7 were compared using a Spearman's Rank correlation coefficient at 21bp, 42 bp and 63 bp window sizes. The results of the correlation analysis are presented in Table 6.2. The aligned profiles are presented for visual analysis in Figure 6.17. None of the aligned curvature profiles had a significant correlation using the Spearman's correlation coefficient. Visually there was little similarity between profiles. This was attributed to the shift in localisation of key peaks. Even a small amount of peak shift would have reduced the point-to-point comparability of the profiles and made the Spearman's Rank correlation coefficient unsuitable for statistical comparisons.

	Unsigned Curvature		Signed Curvature	
Window Size (bp)	**RHO**	**P-Value**	**RHO**	**P-Value**
21	0.08	0.38	<0.00	0.97
42	0.07	0.55	0.05	0.63
63	0.11	0.40	0.08	0.56

Table 6.2. - Summary of the Spearman's Rank correlation coefficient and associated P-values for two experimental *TP53* molecules. The curvature values across a range of base pair window sizes were compared for *TP53* Exon 5-7 (1855 bp) and *TP53* Exon 5-9 (2500 bp). The appropriate number of data points were selected from the exon 5 end of the *TP53* Exon 5-9 dataset and aligned with the *TP53* 5-7 data.

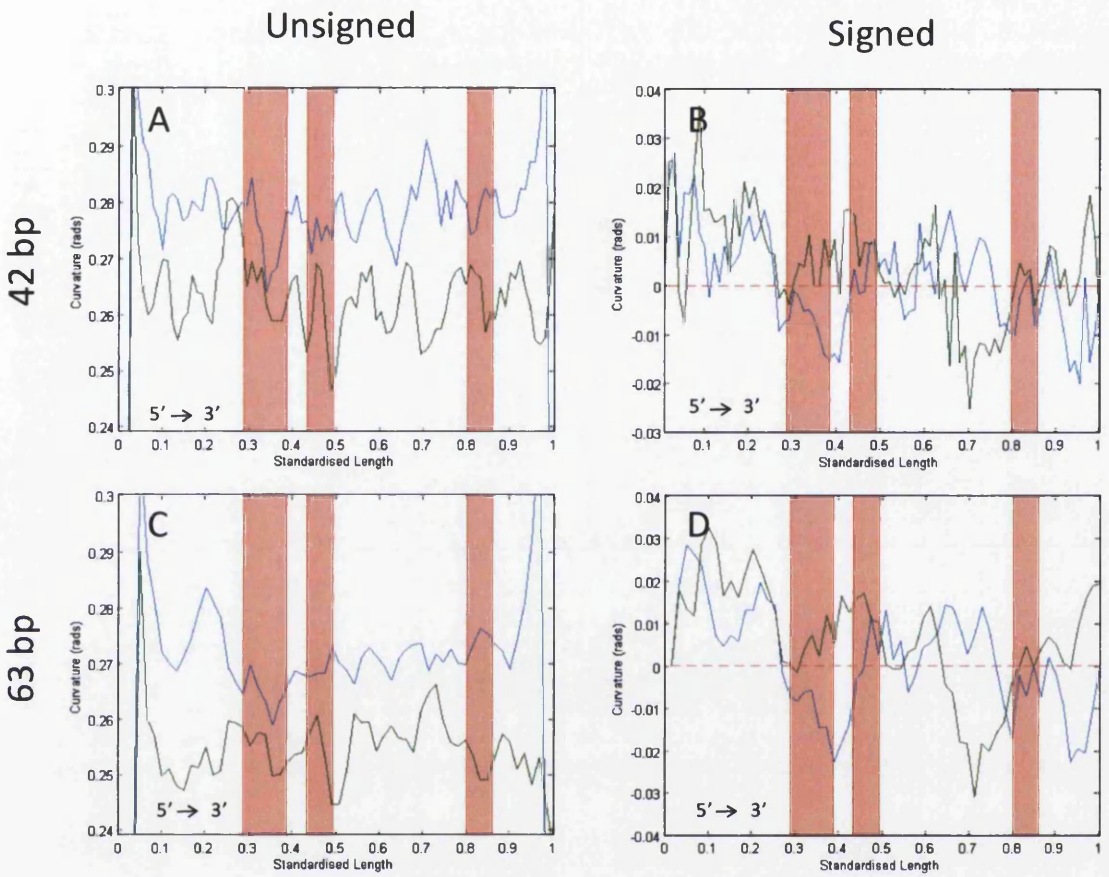

Figure 6.17. Comparison of overlapping sections of curvature profile for Exon 5-7 and Exon 5-9. The sections of the Exon 5-9 profile that correspond to the Exon 5-7 profiles were aligned. Exon 5-7 is denoted by a blue line and Exon 5-9 by a green line. The length of the molecule was standardised from zero to one (5′ to 3′). The dotted red line represents a curvature angle of zero radians. Exon positions are indicated by highlighted red regions in ascending order from left to right.

6.2.13 Analysis of Flexibility in *TP53*

Flexibility Profiles were calculated for 21, 42 and 63 bp windows sizes for both Exon 5-7 and Exon 5-9 (Figure 6.18). The sections of the Exon 5-9 profile that corresponded to the Exon 5-7 profiles were aligned. A correlation analysis showed no significant correlation (Spearman's Rank, 42 bp - Rho = 0.11, p = 0.32; 63 bp - Rho = 0.16, p = 0.22).

The 42 bp profiles indicated that the intronic region between exon 6 and 7 contained a large moderate degree of flexibility. Exons 5 and 6 both showed mixed regions of moderate to low local flexibility in both Exon 5-7 and Exon 5-9 profiles. The border of exon 6 contained the lowest flexibility value in either profile. Exon 7 showed a local trough in the Exon 5-9 profile which was in disagreement with the peak displayed in the Exon 5-7 profile. Notably within the Exon 5-9 profile all exons occurred in regions of low flexibility. Exons 6 and 8 displayed the lowest flexibility relative to the rest of the Exon 5-9 profile.

The 63 bp profile for Exon 5-7 displayed a large region of flexibility at the 5' end of the profile. This was in disagreement with the Exon 5-9 profile which displayed a reduction in flexibility. The Exon 5-7 profile displayed only one notable region of reduced flexibility in exon 5. A small local dip occurred during exon 6; however this was no larger than other dips in flexibility of surrounding regions. The Exon 5-9 profile was more heterogeneous. The central region encompassing the intronic region between exons 6 and 7, exon 7 itself and continuing up to and immediately preceding exon 8 showed the largest flexibility. The lowest regions of curvature occurred immediately before exon 5 and following exon 6. Exons 6, 8 and 9 displayed small local reductions in flexibility.

The general flexibility trend across base pair windows favoured low flexibility in exon positions. This was evident to varying degrees across the profiles. The best agreement between the profiles occurred at a window size of 42 bp. This window size indicated that exons 5 and 6 display reduced local flexibility of which exon 6 was particularly extreme. Both exons 8 and 9 also displayed lower curvature in the 42 bp and 63 bp profiles. Exon 7 displayed a less consistent pattern and there was little consensus between the numerous profiles. The intronic regions between exon 6 and 7 contained most of the flexibility peaks in all profiles.

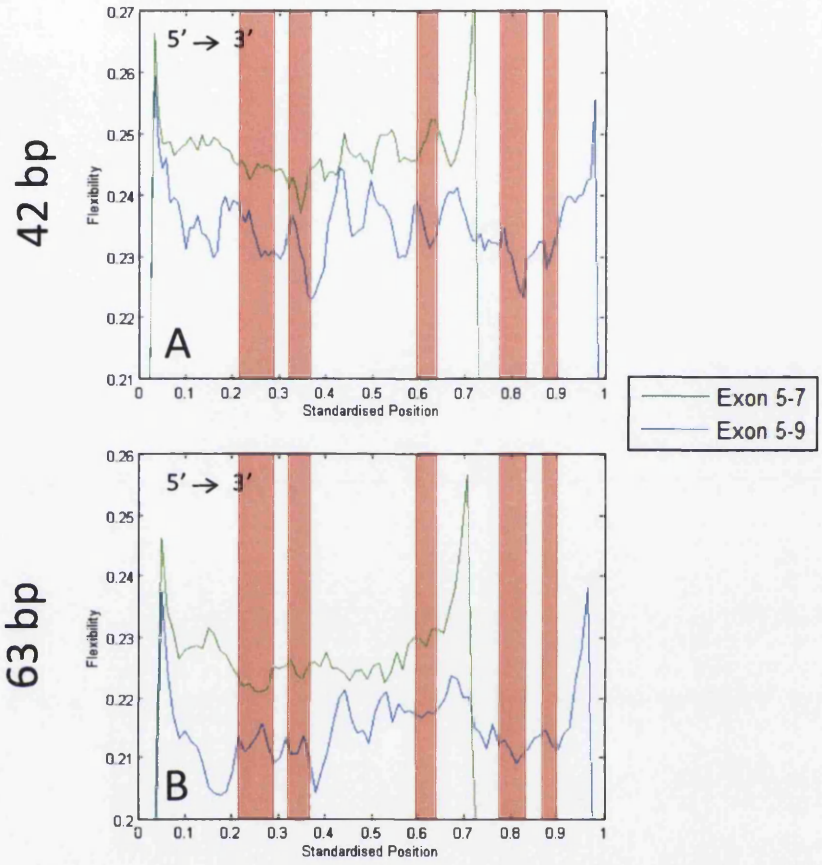

Figure 6.18. - Flexibility profiles for *TP53* calculated over a range of base pair windows. A) 42bp Window. B) 63bp Window. Flexibility values were generated for an appropriate number of linearly spaced points for the base pair window of each molecule. A flexibility profile was generated from the standard deviation of each comparable point. The Exon 5-7 sequence is indicated by a green line, the Exon 5-9 sequence in blue. The profiles were smoothed with a 3-point average filter. Exon positions are highlighted in red in ascending order from left to right.

6.2.14 Estimation of Experimental Peak Shift of Key Peaks

Ten key curvature peaks were identified in the theoretical profiles. Key peaks were defined as the peaks with maximum curvature within the theoretical profile. This was performed using the data for Exon 5-9 at a 63 bp window of curvature for signed and unsigned curvature profiles as this showed the most visually identifiable peaks (Figure 6.19.). The profiles were smoothed using a three point average filter before peak identification.

Within the signed theoretical curvature profile of Exon 5-9 nine key peaks were identified. Seven peaks within the experimental profile occurred at comparable locations. Two key peaks were missing from the experimental profile (Figure 6.19. - green circles). The mean peak shift calculated for the remaining peaks was 3.26 % or 81.37 bp. The largest individual peak shift was 8.86 % or 221.51 bp. The distribution of the magnitude of curvature of the matched peaks was significantly different between the experimental and theoretical (Wilcoxon signed-rank, $p = <0.05$).

Within the unsigned theoretical curvature profile ten key peaks were identified. The experimental profile contained nine peaks that corresponded to the location of the key peaks (Figure 6.19 - green circles). The average peak shift calculated for the remaining peaks was 3.51 % or 87.83 bp. The largest individual peak shift was 8.86 % or 221.52 bp. The distribution of the magnitude of curvature of the matched peaks was significantly different between the experimental and theoretical (Wilcoxon signed-rank test, $p= <0.05$).

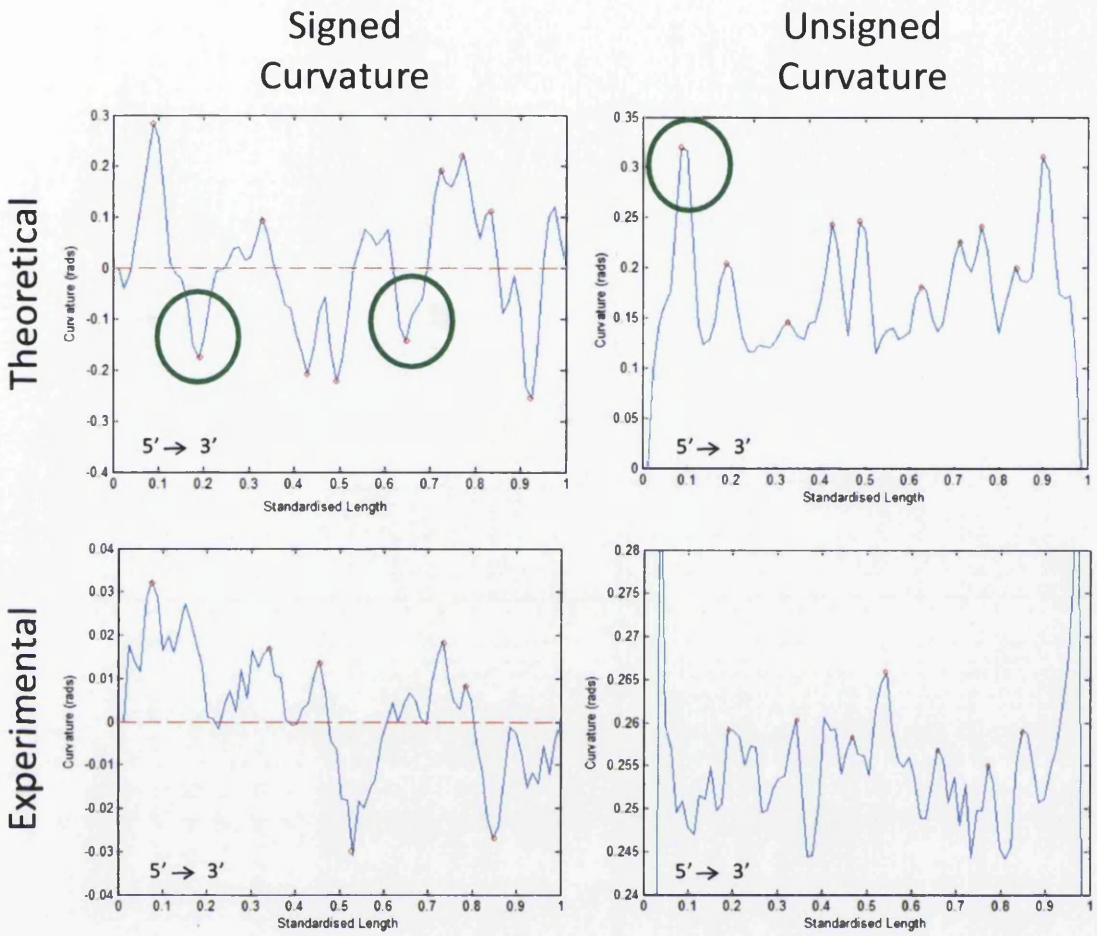

Figure 6.19. - Identification of key peaks between experimental and theoretical profiles in *TP53* Exon 5-9. A 63 bp window of curvature was used for all profiles. Key peaks were defined as those with the largest curvature values in the theoretical profiles. Key peaks are shown with small red circles. Key peaks within the theoretical profiles missing from the corresponding experimental profiles are indicated with a green circle. The profiles were smoothed using a three point average filter before peak identification.

6.2.15 Analysis of Curvature within Exon and Intron Regions

The curvature values that corresponded to exon base pair positions as designated by the IARC database were statistically compared to intronic positions (Hernandez-Boussard *et al.*, 1999). The distribution of curvature was mostly non-normal and was analysed using the non-parametric Kruskal-Wallis test. Three windows of curvature were analysed: 21 bp, 42 bp and 63 bp (Table 6.3.).

Exon 5 of *TP53* Exon 5-7 had significantly lower unsigned curvature than intronic regions at base pair window sizes of 42 and 63 bp. Exon 5 of *TP53* Exon 5-7 exhibited the lowest median curvature and exon 7 exhibited the largest. None of the signed curvature values of exon positions produced a significant result. For *TP53* Exon 5-9 none of the individual exons showed significant difference from intronic DNA in either signed or unsigned curvature profiles (Table 6.4).

The pooled curvature and flexibility of exon positions was compared to intronic curvature (Table 6.5.). Exon positions exhibited significant differences in their unsigned curvature and flexibility measurements at window sizes of 42 bp and 63 bp in Exon 5-7. Similarly, significant results were observed in unsigned curvature and flexibility profiles at a window size of 42 bp for Exon 5-9. The occurrence of significant differences in both unsigned curvature and flexibility for the same window size was not unexpected as they were related measures.

Unsigned Curvature					
Base Pair Window (bp)		**Exon 5**	**Exon 6**	**Exon 7**	**Intron**
21 bp	*Number of Sample Points*	16	10	10	-
	Median Curvature (rads)	0.29	0.29	0.29	0.29
	Kruskal-Wallis (p)	0.41	0.17	0.92	-
42 bp	*Number of Sample Points*	8	5	5	-
	Median Curvature (rads)	0.27	0.27	0.28	0.28
	Kruskal-Wallis (p)	<0.05	0.19	0.63	-
63 bp	*Number of Sample Points*	6	3	3	-
	Median Curvature (rads)	0.26	0.27	0.28	0.27
	Kruskal-Wallis (p)	<0.05	0.22	0.41	-
Signed Curvature					
Base Pair Window (bp)		**Exon 5**	**Exon 6**	**Exon 7**	**Intron**
21 bp	*Number of Sample Points*	16	10	10	-
	Median Curvature (rads)	-0.01	0.00	-0.01	-0.00
	Kruskal-Wallis (p)	0.23	0.63	0.29	-
42 bp	*Number of Sample Points*	8	5	5	-
	Median Curvature (rads)	-0.01	0.00	-0.01	0.00
	Kruskal-Wallis (p)	0.07	0.60	0.24	-
63 bp	*Number of Sample Points*	6	3	3	-
	Median Curvature (rads)	-0.01	0.01	-0.02	0.00
	Kruskal-Wallis (p)	0.11	0.40	0.41	-

Table 6.3. - Summary of measurements made in comparisons between curvature measurements of exon positions and intron positions for *TP53* Exon 5-7. The distribution of data points was non-normal and not size matched; a Kruskal-Wallis test was used to test for significant differences between median values. Significant p-values are highlighted in red.

Unsigned Curvature							
Base Pair Window (bp)		**Exon 5**	**Exon 6**	**Exon 7**	**Exon 8**	**Exon 9**	**Intron**
21 bp	*Number of Sample Points*	18	11	10	13	6	-
	Median Curvature (rads)	0.28	0.27	0.29	0.29	0.28	0.28
	Kruskal-Wallis (p)	0.47	0.09	0.16	0.05	0.44	-
42 bp	*Number of Sample Points*	8	6	6	6	3	-
	Median Curvature (rads)	0.26	0.25	0.26	0.26	0.26	0.26
	Kruskal-Wallis (p)	0.47	0.26	0.78	0.09	0.29	-
63 bp	*Number of Sample Points*	6	3	3	4	2	-
	Median Curvature (rads)	0.25	0.26	0.25	0.25	0.25	0.25
	Kruskal-Wallis (p)	0.71	1.00	0.30	0.05	0.63	-
Signed Curvature							
Base Pair Window (bp)		**Exon 5**	**Exon 6**	**Exon 7**	**Exon 8**	**Exon 9**	**Intron**
21 bp	*Number of Sample Points*	18	11	10	13	6	-
	Median Curvature (rads)	-0.00	0.01	0.00	0.00	0.01	-0.00
	Kruskal-Wallis (p)	0.74	0.53	0.90	0.92	0.34	-
42 bp	*Number of Sample Points*	8	6	6	6	3	-
	Median Curvature (rads)	-0.00	<0.00	<0.00	-0.00	0.01	0.00
	Kruskal-Wallis (p)	0.763	0.473	0.958	0.804	0.695	-
63 bp	*Number of Sample Points*	6	3	3	4	2	-
	Median Curvature (rads)	0.01	0.01	0.01	0.00	0.00	0.00
	Kruskal-Wallis (p)	0.37	0.28	0.61	0.96	0.68	-

Table 6.4. - Summary of measurements made in comparisons between curvature measurements of exon positions and intron positions for *TP53* Exon 5-9 DNA molecules. The distribution of data points was non-normal and not size matched; a Kruskal-Wallis test was used to test for significant differences between median values.

		Kruskal-Wallis (p-value)		
Exon 5-7	**Window Size**	**Unsigned Curvature**	**Signed Curvature**	**Flexibility**
	21 bp	0.25	0.30	0.50
	42 bp	<0.05	0.15	<0.05
	63 bp	<0.05	0.28	<0.05
Exon 5-9	**Window Size**	**Unsigned Curvature**	**Signed Curvature**	**Flexibility**
	21 bp	*0.35*	*0.43*	*0.44*
	42 bp	<0.05	*0.67*	<0.05
	63 bp	*0.11*	*0.35*	*0.58*

Table 6.5. - Summary of the Kruskal-Wallis test applied to the pooled curvature and flexibility of exon positions to the pooled curvature and flexibility of intron positions. The distribution of data points was non-normal and not size matched; a Kruskal-Wallis test was used to test for significant differences between median values. Significant p-values are highlighted in red.

6.3 Discussion

Experimental intrinsic DNA curvature profiles show little statistical similarity to theoretical profiles previous generated in Chapter 4. However, there is a moderate degree of visual similarity between the profiles which is more exaggerated in signed curvature profiles. This visual similarity is limited to the positional occurrence of peaks rather than the direction or magnitude of curvature. Pooled exon positions within *TP53* DNA exhibited significantly lower curvature and flexibility than intron positions. This was particularly prevalent within particular exon 5. This is in good agreement with theoretical prediction of curvature from simulated DNA images (Chapter 4). The relevance and significance of these finding is discussed in more depth in the sections below.

6.3.1 Visual Identification of Streptavidin End Labelling

The quality of the end-label biotin incorporated into the PCR primer was checked by dot blot analysis and the final product by band shift assay (Sections 6.2.1.1. and Section 6.2.1.2.). The small band shift was similar to previously reported results (Seong *et al.*, 2002). end-labelled *TP53* DNA produced a number of molecules in AFM with clearly identifiable 5' labels (Figure 6.5.). There were a number of molecules that either lacked end-labels or in which identification of the end-label was problematic. Additionally, streptavidin has four available binding sites for biotin which led to the imaging of DNA molecules bound together as dimers, trimers and tetramers (*e.g.* Figure 6.5. – red arrow). This necessitated the processing of multiple images in order to collect sufficient molecules for curvature analysis. The use of larger streptavidin fusion proteins would improve visual label identification in future studies (Rivetti *et al.*, 1996). The ideal end-label would be a monomeric avidin fusion protein (Sun *et al.*, 2001). This would present an improvement for image processing efficiency over the current method as it would provide both a one-to-one molecule binding ratio and a more easily identifiable tag. An example of this type of protein label has been used as a probe for abasic sites in DNA molecules (Sun *et al.*, 2001). Additionally, optimisation of the streptavidin to DNA ratio has previously achieved end-labelling efficiencies of greater than 90 % (Seong *et al.*, 2002).

6.3.2 Height of DNA and Streptavidin End-Labelling

The streptavidin end-label was clearly identifiable in the Z-height measurements of DNA molecules (Section 6.2.3.). Previous authors have reported variable average height values for streptavidin imaged by AFM ranging from 0.61 nm (Seong *et al.*, 2002), 1.7 nm (Woolley *et al.*, 2000) and 2.31 nm (Neish *et al.*, 2002). The average height of the streptavidin end-label investigated in this study was ~0.725 nm which is within the range of values reported by

previous studies. The height of DNA, ~2 nm, was measured from X-ray experiments (Saenger, 1984). AFM measurements of the height of DNA have been universally lower than the expected value. The heights measured vary between 0.7 (Moreno-Herrero *et al.*, 2003) and 1.28 nm (Yang *et al.*, 2007). The present study observed a DNA height of 350-475 nm. This was lower than the values reported by other authors. This was not considered a cause for concern as the variability of AFM height measurements of soft matter, such as DNA or proteins, has been well reported (Yang *et al.*, 2007). This variability in height measurements has been attributed to a number of factors including humidity (Thundat, 1992), tip size, tip loading force, salt deposition, electrostatic attraction between the molecule and the substrate (Yang *et al.*, 2007), cantilever oscillation frequency and tip adhesion to the sample (Noort van *et al.*, 1997). Another consideration after image collection was the amount of plane fitting used to produce a flat AFM image. The plane fitting used in this study was slightly more rigorous than typically used in image processing studies of AFM; two passes of flattening were applied to produce the flattest possible images for analysis. This did not negatively impact on visual end-label identification.

6.3.3 Post-Image Processing Identification of Unsuitable DNA Molecules

Any molecules that had a larger Z-height at the unlabelled 3' end in comparison to the 5' end-label were removed from the analysis (Section 6.2.3.). This ensured that only correctly labelled DNA molecules remained for further analysis. After Z-correction there remained a small Z-height increase in the unlabelled end of the Exon 5-7 dataset. This could have been due to a number of factors *e.g.* the tip may have picked up a small amount of experimental DNA causing the 'sticky' free end of the molecule to bind to the tip and appear larger than it was, there could be non-specific binding of streptavidin, unbound streptavidin or local 'bunching' of the DNA. Alternatively, it could have been due to tip-DNA interactions caused by the oscillation frequency of the cantilever (Noort van *et al.*, 1997). The potential for a small amount of incorrectly orientated molecules to have been present after Z-correction has been considered during the analysis.

6.3.4 Evaluation of Local Streptavidin-DNA Interactions

An unspecified form of DNA-streptavidin interaction has been reported by a previous study (Marilley *et al.*, 2005). Other studies have not observed any local interaction between DNA and streptavidin (Murray *et al.*, 1993; Rivetti *et al.*, 1996; Woolley *et al.*, 2000; Neish *et al.*, 2002; Seong *et al.*, 2002). In order to identify whether the streptavidin was binding local DNA or obscuring DNA by being localised on top of a DNA strand, a correlation analysis was performed between the DNA length and the streptavidin end-label height (Section 6.2.5.). A similar approach has been used by other authors to assess protein-DNA interactions (Woolley

et al., 2000). Increased streptavidin label height correlated weakly with decreased DNA contour length in the Exon 5-9 dataset. This may have impacted on both the curvature measurements of DNA at the site of the streptavidin label and on the selection of comparable points across multiple molecules. The implications of this were considered during further analysis.

6.3.5 Reconstructed Length Measurements of *TP53*

The DNA contour length was calculated for both DNA molecules, Exon 5-7 and Exon 5-9 (Table 6.1.). A number of obvious outliers were observed. These were removed before further analysis. These molecules represented broken DNA fragments, DNA molecules bound end-to-end or otherwise erroneous conformations. The removal of obvious outliers has been performed in other studies (Scipioni *et al.*, 2002a; Ficarra *et al.*, 2005b; Marek *et al.*, 2005).

The distributions of DNA contour lengths were observed to be non-normally distributed so non-parametric descriptors and statistical tests were used. This was contrary to other studies that typically use the mean and standard deviation. Parametric equivalents have been included for comparison where they are appropriate. The median contour lengths after outlier removal were lower than expectations for B-DNA. These produced underestimations of 11.17 % and 9.24 % for Exon 5-7 and Exon 5-9 respectively. The estimate of deviation from the theoretical value for the Kulpa estimator was tested to be effectively nil for simulated DNA molecules (Section 4.2.5.). However, DNA contour lengths reported in AFM studies of DNA are typically lower than the expected value for B-DNA. Examples include underestimations of 6.9 % using the Kulpa estimator in a Mg^{2+} buffer (Rivetti and Codeluppi, 2001), 8 % on polylysine coated mica (Van Noort *et al.*, 2004) and both 4.4 % and 12.13 % for DNA in a Ni^{2+} buffer (Lysetska *et al.*, 2002; Sanchez-Sevilla *et al.*, 2002). The underestimation in this study was only slightly lower than other reported contour lengths. The variation observed between streptavidin labelled and unlabelled DNA in this study maybe indicative of changes in the electrostatic potential of the mica surface, minute variations in the buffer conditions or slight interaction with the streptavidin end-label.

The contour length showed IQRs of 11.58 % and 6.35 %. In order for comparison to the work of previous authors the standard deviations of the distributions were calculated. The standard deviations were 7.21 % and 4.30 % for Exon 5-7 and Exon 5-9 respectively. The increased variability within the Exon 5-7 dataset may have partially contributed to its lower median length estimate. Similar standard deviations (6 %) have had low or negligible effects on curvature measurements by previous authors (Scipioni *et al.*, 2002a).

6.3.6 DNA Condensation or a Partial B- to A-Form DNA Transition

The reduction of contour length on adsorption to a surface has been variously attributed to a partial transformation of the DNA from B- to A-form on the mica surface (Rivetti and Codeluppi, 2001) and condensation due to interactions with the cation loaded surface (Sanchez-Sevilla *et al.*, 2002). A-form is likely to be approximately 30 % shorter than B-DNA of the same length (Ficarra *et al.*, 2005b). An underestimate of 9-11 % was observed in this study and may have been indicative of a partial transition. The expected length estimates have been included alongside the length distribution in Figure 6.7. The propensity to transition between structural forms is sequence dependent (Basham *et al.*, 1995; Ivanov and Minchenkova, 1995). During a typical experiment the length of DNA is considered to be uniform along each DNA molecule (Buzio *et al.*, 2012). Assuming a structural transition had occurred then the length of each base pair would no longer be uniform. The effect on the resulting curvature profiles would be a widening of peaks of curvature and possibly a shifting of positions of curvature peaks. There are no models currently available to simulate this effect for comparison. A length estimation method that considered both contour length and contour height when calculating the reconstructed length of a DNA molecule has been recently published and may account for some of the length variation observed in this experiment (Buzio *et al.*, 2012).

Alternatively, the reduction in length observed may have been due to local surface interactions (Sanchez-Sevilla *et al.*, 2002), the systematic underestimation by the digitised contour length estimator or a combination of both factors. It was not possible to rule out interaction with the streptavidin label as a contributing factor to the length underestimation as weak interaction was observed with Exon 5-9 (Section 6.2.5). The length reduction observed may be indicative of streptavidin-DNA interaction on or near the 5' end label. Alternatively, variations in the buffer or surface conditions could have contributed to the observed contour length variation.

6.3.7 Persistence Length of End-Labelled *TP53*

The persistence length of *TP53* for a contour length range of 0-300 was $\xi = 61$ and $\xi = 60$ for Exons 5-7 and Exons 5-9 respectively (Section 6.2.6.). This is higher than often cited ~50 nm persistence length for B-form DNA using a Mg^{2+} buffer (Rivetti *et al.*, 1996). However, considerable deviation from this consensus value has been reported by previous studies; for example, persistence lengths have been reported of 55 nm (Van Noort *et al.*, 2004), 56 nm (Podestà *et al.*, 2005), 42 nm (Marek *et al.*, 2005) and as low as 36 nm (Lysetska *et al.*, 2002). These previous studies used buffers containing the same cation, Mg^{2+}. However the previous studies had different salt concentrations, which is likely to contribute to the variations in

measured DNA persistence length (Rivetti *et al.*, 1996). In comparison to these previous studies it seems that *TP53* in this study was slightly more rigid than the expectation for B-DNA. However, when considering a contour length range of 0-200 the persistence length values were ξ = 56 and ξ = 54 for Exons 5-7 and Exons 5-9 respectively. These values are in better agreement with the consensus persistence length value of B-DNA. A central region of high curvature causing an overall decrease in the flexibility of the polymer may account for the observed persistence length, such as that predicted by theoretical curvature profiles (Section 4.2.8.). Other alternative explanations could be put forward, such as surface interactions causing modification to the typical conformation of the DNA molecules or a partial B- to A-DNA transition caused by dehydration (Charney *et al.*, 1991).

As a note of interest, some other authors have observed a systematic decrease in $<R^2>$ measurements above ~250-300 nm (Rivetti *et al.*, 1998; Moreno-Herrero *et al.*, 2006; Buzio *et al.*, 2012). This was not observed in either *TP53* sequence. The authors that observed the decrease in $<R^2>$ at larger curvilinear distances used DNA sequences either constructed with in-phase A-tracts (Rivetti *et al.*, 1998) or that contained hyperperiodicity (Moreno-Herrero *et al.*, 2006). The decrease in $<R^2>$ was considered the signature of intrinsic curvature that forced the DNA to assume a more compact structure than would be readily assumed by linear DNA of the same length. This does not seem to be the case with the experimental *TP53* molecules. Overall, the persistence length of the end-labelled DNA indicated that samples were well equilibrated on the mica. The binding observed may have been a little stronger than was expected for a weak cationic buffer (Rivetti *et al.*, 1996). This stronger binding may have been due to unexpected variation in the charge of the mica or the concentration of the binding buffer.

6.3.8 Selection of a Window of Curvature for Curvature Analysis

It has previously been shown that the effect of digitisation of the DNA contour has a measurable effect on curvature angles at low base pair windows (Section 3.2.6.3.). The Visual Threshold, developed in Section 3.2.6.5. was applied to *TP53* Exon 5-7 and Exon 5-9 (Section 6.2.7.). This analysis method suggested window sizes for both molecules were in good agreement; 38-103 bp for Exon 5-7 and 27-74 bp for Exon 5-9. The angles calculated from the suggested ranges should have been free of the effects of digitisation. As corroboration other checks were also applied. The maximum angle calculated at a number of window sizes was identified alongside the numbers of individual angles that matched the appropriate maxima (Figure 6.12.). It is observable that below a window size of 31 bp there were multiple individual angles that match the dataset extrema. Curvature profiles produced from base pair windows below this value risked introducing variation due to effects of DNA contour digitisation.

The base pair window sizes of 42 bp and 63 bp lay within the experimentally determined optimal ranges and were used for the analysis that followed. These window sizes had previously provided good peak to background contrast for *TP53* (Section 4.2.6.). These window sizes were multiples of one helical turn (10.5 bp in B-DNA) and so could be discussed in terms of a biologically relevant measure.

6.3.9 Curvature Analysis of *TP53*

Both signed and unsigned curvature profiles were produced for Exon 5-7 and Exon 5-9 and compared to theoretical profiles. Although the experiment included no method for determining the direction of curvature *i.e.* whether a molecule was up or down on the mica. However, directionality was observed in the experimental profiles. This was likely due to the thymine rich strand of DNA preferentially binding to the mica surface (Sampaolese *et al.*, 2002). The direction of the curvature was aligned with the experimental profiles to give the best visual similarity.

6.3.9.1 Unsigned Curvature Profiles of *TP53* Exon 5-7

The only exon that was expected to produce a statistically significant reduction in curvature was exon 5 (Section 4.2.15.). This proved to be the case for the Exon 5-7 sample which showed a visible trough within the curvature profile corresponding to the region containing exon 5 (Section 6.2.8. + 6.2.15.). Exon 6 was expected to show low curvature and exon 7 a small peak. This pattern was observed in the curvature profile for Exon 5-7. The curvature in exon 6 and 7 was not significantly different from intronic regions, in line with the expectation from simulated images.

There was no significant correlation between the curvature profiles and corresponding theoretical profiles (Section 6.2.8.). There were some visual similarities between the experimental and theoretical profiles, notably large regions of curvature in the intron between exons 6 and 7 and a large peak in curvature at the 5' end of the 63 bp window of curvature. This may indicate that although there was no statistically significant correlation between experimental and theoretical profiles this may be due to peak shift. A small amount of peak shift would have reduced the effectiveness of the correlation analysis by removing the assumption of point-to-point comparability between experimental and theoretical profiles.

6.3.9.2 Unsigned Curvature Profiles of *TP53* Exon 5-9

The expectation from theoretical studies was that all exons in the experimental molecules would exhibit a local reduction in curvature, with perhaps the exception of exon 7. The experimental Exon 5-9 profiles exhibited dips in curvature at all exons with the exception

of exon 6 which bordered a region of low curvature (Section 6.2.9.). The curvature values corresponding to exon positions were not significantly different from intronic curvature values at either window size (Section 6.2.15.). There was no significant correlation between the curvature profiles and corresponding theoretical profiles (Section 6.2.9.). There were some visual similarities between the experimental and theoretical profiles, such as the aforementioned dips in curvature at exon positions, peaks of curvature in the intron between exon 6 and exon 7 and the occurrence of a peak in curvature immediately before exon 5. The working hypothesis for this study was that exons would exhibit reduced curvature with respect to intron positions. This seemed to be true visually; however, it was not statistically proven.

6.3.9.3 Signed Curvature Profiles of *TP53* Exon 5-7

Exon 5-7 showed good visual agreement with signed theoretical profiles of *TP53*. The major regions of positive and negative curvature were present in both profiles and follow similar patterns (Section 6.2.10.). The exon positions showed low levels of curvature or bordered regions of low curvature. The largest curvature peaks occur in intron regions, in line with the expectation. Interestingly, there was a small peak of curvature between exon 5 and 6 that was not present in the unsigned profiles. The cause of this peak was unknown and the peak was not observed to such an obvious degree in Exon 5-9. This peak may have been produced by the inclusion of erroneously oriented molecules (Section 6.2.3.) or may be a region of curvature that was not predicted by the De Santis dinucleotide wedge model.

The 5' end region exhibited the least visual similarity. This may have been caused by the presence of the streptavidin end label, either through weak local interactions or the inability of image processing software to correctly trace a straight line through a circular streptavidin molecule. Correlation between experimental and theoretical profiles was not significant (Section 6.2.10.). However, the 63 bp window produced a p-value (Spearman's Rank: Rho = 0.24, p = 0.07) which was borderline, perhaps indicating that with a smaller amount of variation the visible similarities would also have been statistically significant.

6.3.9.4 Signed Curvature Profiles of Exon 5-9

Exon 5-9 showed good visual agreement with signed theoretical profiles of *TP53*. The major regions of positive and negative curvature were present in both profiles and followed similar patterns (Section 6.2.11.). The exon positions showed low levels of curvature or bordered regions of low curvature with the exception of exon 6 which occurred at a small peak. The largest peaks of curvature occurred in intronic regions. These results were in good agreement with theoretical predictions. The region with the least visual similarity was the central intronic region, between exons 6 and 7. Statistical correlation between experimental

and theoretical profiles was not significant (Section 6.2.11). It seems likely that there was sufficient variation between curvature profiles to limit the applicability of correlation tests, but that the profiles still retained some visually identifiable trends.

6.3.10 Comparability of Profiles between Experiments

The main objective for analysing two overlapping DNA molecules of *TP53* was to evaluate the reproducibility of AFM based curvature analysis. Experimental profiles were compared with an aim to assess this (Section 6.2.12.). The correlation analysis for comparable sections of each molecule showed no significant correlation. Visually there was little similarity between the profiles. This would indicate that there was a considerable degree of variability between the curvature profiles resulting from multiple experiments.

6.3.11 Flexibility Profiles

Exon 7 appeared to be the most flexible exon (Section 6.2.13). Exons 8 and 9 appeared to be the least flexible. Exon 5 was observed to have regions of moderate flexibility and also regions of low flexibility across multiple profiles. The same was observed for Exon 6, which was variably the least flexible exon or occurred immediately before a region of low flexibility. The region where least agreement was observed between the profiles was the site of 5' end label. This variation in the observed flexibility may have been caused by the weak protein-DNA interaction observed in the Exon 5-9 sample (Section 6.2.5.).

6.3.12 The Curvature of Exons in *TP53*

The statistical analysis of curvature showed that exon 5 exhibited significantly reduced curvature. This is the same trend predicted from simulated AFM images of *TP53*. The other profiles did not produce any statistically significant differences from intronic positions. The major trend within curvature profiles was that the majority of exons had reduced curvature in comparison to the surrounding regions. Exon 5 had greatly reduced curvature in all profiles, signed and unsigned. This trend was less clear for exon 6 which variably exhibited very low curvature in unsigned profiles and a peak in the signed profiles. Exon 6 was observed as having low to moderate flexibility. Exon 7 was a similar case, exhibiting both small peaks and reduced curvature in different profiles. The expectation for exon 7 was a small peak; however, the peak may have been masked in some profiles by moderate to high flexibility. Overall the curvature profiles support the original hypothesis of low curvature in exon positions. Additionally, the signed profiles had very good visual agreement with the theory. However, due to the lack of correlation between the profiles there was a level of doubt about agreement between the theoretical and experimental profiles.

Notably absent from the profiles were the large peaks of curvature before exon 5 and after exon 8. The end regions had less physical restraints upon them as they were only constrained by DNA at one end. This could have led to increased flexibility in the end regions of the DNA molecules. The increased flexibility and curvature observed at both ends of the molecule provided some corroboration of this explanation (Section 6.3.9. and Figure 6.18.).

The trends observed in this study could be investigated further by extension of the current method. A suitable further analysis would include end-labelled PCR products for each exon and intron region considered as separate samples. The curvature and flexibility of each exon may be investigated at higher resolutions using sharper tips such as carbon nanotubes (Woolley *et al.*, 2000). Furthermore the use of liquid imaging and time-lapse based experimentation may provide improved flexibility profiles for *TP53* and allow the application of other theoretical models for DNA dynamics (Scipioni *et al.*, 2002b; Marilley *et al.*, 2005). This type of single molecule experiment is ideally suited to end-labelled DNA molecules.

6.3.13 Differential Effect of Experimental Variation on Signed and Unsigned Profiles

The signed curvature profiles had improved visual similarity between experimental and theoretical profiles when compared to the unsigned profiles. This improvement of signed over unsigned profiles was predicted by simulated images (Section 4.2.11.). It was likely due to the differential impact of image noise between the two types of profile. The signed profiles had both magnitude and direction of curvature, while the unsigned had only magnitude. Experimental variation may reduce the magnitude of curvature in a signed profile. However, it is unlikely to change the direction of curvature. Unsigned profiles were only comparable on the magnitude of curvature which is effected by variation or noise. The signed profiles were compared using both magnitude and direction of curvature, of which direction was likely to be less effected by noise or variation.

6.3.14 Identification of Sources of Experimental Variation

Simulated AFM images of *TP53* indicated that, with little experimental noise, the base pair windows used in this study were likely to produce comparable theoretical and experimental profiles (Section 4.2.11.). However, this was clearly not the case for real AFM images as not a single curvature profile exhibited a strong significant correlation to theoretical profiles. A number of sources of experimental variation could have contributed to this. The reconstructed length measured for *TP53* indicated that the DNA may have undergone a partial transition to A-form DNA or DNA condensation; both possibilities were not accounted for by the theoretical models. The curvature and flexibility differences between A- and B-DNA were

not accounted for by the theoretical models, rendering comparison to theoretical profiles problematic.

The persistence length of the experimental DNA also indicated that it was slightly more rigid than expected. This is likely to have effected the magnitude of curvature rather than its position. The measured unsigned curvature was lower for all profiles than predicted which might be a corroboration of the reduced experimental flexibility of *TP53*. Alternatively, reduced magnitude may be attributed to sample preparation. Washing with distilled water after adsorption of the DNA to the mica decreases the cationic charge on the phosphate backbone of DNA by removing Mg^{2+} and causing increased repulsion between negatively charged DNA molecules (Moreno-Herrero *et al.*, 2006; Marilley *et al.*, 2007a). This increased repulsion between DNA molecules would have reduced the measurable curvature.

Impulsive image noise was previously shown to negatively effect the comparability of curvature profiles from simulated images (Section 4.2.11.). The sources and intensity of noise in real images was much higher than in simulated images. There were also other sources of variation that were unaccounted for such as fragmented molecules, small scale looping or sharp kinking below the resolution of the AFM image and possible false positive in the end-labelling analysis (Section 6.2.3.). All of these factors were possible sources of variation within the final curvature profile either through modification of the magnitude of curvature or through shifting of key peaks.

6.3.15 Peak Shift in Curvature Profiles

Peak shift was evaluated for the major peaks within the Exon 5-9 profile as an experimental estimate of peak shift within this study (Section 6.2.14.). The average peak shift was ~3.39 % of the standardised length of the sequences. The maximum peak shift measured was 8.86 %. For the individual peaks where this was the case this was likely to represent a different curvature peak altogether. Only one large deviation was observed per profile. The inclusion of these peaks with a large amount of peak shift may have caused the average peak shift to be an overestimation. It is difficult to compare peak shift with other studies as this measure is often not quantified. However, the peak shift reported in a previous study was lower than the average peak shift observed in the current study by ~2% (Ficarra *et al.*, 2005b). A partial B- to A-form DNA transition may provide an explanation for the observed peak shift. This was discussed in Section 6.4.5.

Changes in the magnitude of curvature were unlikely to effect the outcomes of correlation analyses. However, experimental variation leading to peak shift within the profiles would have had a negative impact on correlation analyses as point-to-point comparability is an

underlying assumption. The observed peak shift was likely to have had a large impact on the outcome of correlation analyses that have been carried out between curvature profiles.

It should be noted that different peaks were missing from the signed and unsigned curvature profiles, highlighting the need to use both types of profiles where possible. The amount of peak shift detected was very similar between the methods. As the data has not been treated differently, other than the direction of the curvature before smoothing, this is in line with the expectation.

6.3.16 A Potential Role of Curvature in Post-Transcriptional Modification

It was observed that, in signed curvature profiles, exons often occurred in regions of low curvature that bordered a change in the direction of curvature (Sections 6.2.10-11.). This could prove to be informative for modelling DNA deposition and adsorption. Regions of low curvature and flexibility may be less likely to undergo structural changes during adsorption. The impact of this would be that flexible regions bordering inflexible regions would preferentially kink in order to conform to the 2D surface.

The change in the curvature regime at the border of exons suggests a role in post-transcriptional modification of RNA, namely the removal of introns from RNA transcripts. GC content and related sequence motifs are recognised during post-transcriptional splicing of RNA (Amit *et al.*, 2012). GC content was explicitly investigated in the study, which is closely linked to intrinsic DNA curvature. The change in the structural regime of curvature on the border of exons may represent a recognition factor for the spliceozome or other associated proteins. This is an area that could be investigated in more depth in future studies.

6.4 Conclusion

The end-labelling approach to curvature investigation has produced an overarching assessment of the physical properties of the region of *TP53* that codes for the sequence-specific binding domain of the p53 protein. The most interesting and prevalent trend was a lowering of intrinsic curvature in the exon positions of *TP53*. This trend was in good agreement with theoretical predictions of *TP53* and has interesting biological implications for DNA transcription, mutagenesis and repair. Furthermore, a potential role has been identified for curvature in post-transcriptional modification that will require further investigation by future studies.

A number of methodological considerations have been identified by the present study including the importance of the window size over which to consider calculating curvature angles. The need for improved physical models of DNA deposition has been highlighted as well as a need for improved statistical analysis methodologies. It is also clear from the present study that the current methods for statistical analysis of curvature profiles that have been used by previous authors, such as visual comparison and peak comparison, are unsuitable for profiles with a small degree of peak shift or inter-experiment variation. An additional aim for future studies would be to identify or develop statistical tools that could provide more application to intrinsic DNA curvature and flexibility profiles produced by AFM imaging of DNA.

CHAPTER 7: CONCLUSION

7.1 The Investigation of Intrinsic DNA Curvature in *TP53*

TP53 is a key cancer gene. The mutation and dysfunction of *TP53* is considered a hallmark of carcinogenesis. The primary aim of the present study was to evaluate the intrinsic curvature of the region of *TP53* that codes for the DNA sequence-specific binding domain of the p53 protein. This region of *TP53* is critical for the correct functioning of the p53 protein which regulates the main cellular defences against chemical insults and tumourogenesis. *TP53* is highly conserved within evolution, occurring in a recognisable form in even the most simple of multi-cellular organisms. *TP53* has codons that exhibit slow DNA repair; these codons are also major mutation hotspots. In order to evaluate the intrinsic DNA curvature of *TP53* over a large scale the gene was investigated using both theoretical methods and AFM.

7.2 ADIPAS – A Software Suite for AFM Based Analysis of DNA Curvature

There is currently a lack of available software for the analysis of AFM images of DNA. The first objective of the present study was to create a software platform with the capability of calculating intrinsic curvature from AFM images of DNA. A complete software suite of image processing and analyses tools was developed in order to facilitate the AFM based study of DNA (Chapter 3). This analysis suite was named ADIPAS (AFM DNA Image Processing and Analysis Software). ADIPAS was developed with the primary aim of analysing intrinsic curvature of *TP53* DNA molecules. To this end ADIPAS was able to analyse AFM images of DNA and calculate curvature from the resulting coordinate data. The software incorporated analysis methods from a range of previous studies. ADIPAS allowed for a more comprehensive analysis of the structural properties of DNA molecules than any other available software pipeline. ADIPAS presented the image analysis portion of its package in a GUI that would allow even unskilled operators to process AFM images of DNA after only limited training. Other estimates of statistical and physical DNA measurements, such as DNA contour length and persistence length, were implemented into the software. The software is aimed at online distribution and publication with the hope that it will be of use to researchers within the field and also to encourage further investigation of DNA curvature by allowing research groups to overcome the large technological hurdle of in-house software development necessary for this type of investigation.

During the development of ADIPAS a number of novel considerations were identified. The most important of which was that at low base pair window sizes there was a significant influence of DNA contour digitisation of the resulting curvature angles. A novel method for identifying the effects of digitisation on angle measurements was developed and named the Visual Threshold. This consideration has not been investigated or even discussed to any great extent by previous researchers. The Visual Threshold was applied to real AFM images of *TP53*

and was invaluable in the selection of appropriate curvature window sizes. Additionally, the prediction of the Visual Threshold was shown to coincide with accuracy of the FF algorithm showing its utility for application to real AFM images of DNA (Chapter 5). These developments provide a strong foundation for researchers to build upon in future studies and represent progress towards improving accessibility to the field of DNA curvature investigation as AFM technology becomes more widespread.

7.3 The Investigation of Intrinsic DNA Curvature of *TP53* using Theoretical Curvature Models

Theoretical dinucleotide wedge models were utilised for the investigation of DNA curvature (Chapter 4). The De Santis model of curvature, previously shown to be appropriate for AFM based studies of DNA, was used as the primary dinucleotide wedge model of curvature. The De Santis model predicted significantly reduced curvature in exon 5, 6 and 7 in comparison to DNA curvature of intronic regions. Furthermore, the model indicated that both exons 8 and 9 were regions of locally reduced curvature that were not significantly different from the curvature of intron regions. The lack of statistical significance may be attributed to the curvature peaks that flanked the exons and the methodological necessity of averaging over a base pair window of at least one or two helical turns. Both exons 5 and 6 were implicated as being DNA linker regions between nucleosomes by theoretical models.

Additionally, codons within *TP53* that have been shown by previous studies to exhibit impaired DNA repair were shown in the present study to have significantly reduced intrinsic curvature in comparison to the rest of the *TP53* DNA sequence.

7.4 The Investigation of Intrinsic DNA Curvature of *TP53* using AFM

The generation of simulated AFM images of *TP53* based upon the De Santis model allowed for predictions about potential observation in real AFM images of DNA. The AFM portion of the study was approached using two separate investigative methodologies. The first methodology used was the post-image processing orientation of *TP53* molecules by the FF algorithm (Chapter 5). The results for this indicated a significant positive correlation between the theoretical predictions and the experimental curvature profiles for *TP53* and raised some methodological considerations for the FF algorithm that were overcome during the study.

The second approach was the use of streptavidin end-labelling for DNA orientation. The curvature profiles generated using this method did not show significant correlation to curvature profiles produced from simulated AFM images. This was attributed to a curvature peak shift caused by DNA condensation on the mica surface or a partial B- to A-form DNA transition. However, the expected trends in DNA curvature were still evident. This included a

significant reduction in curvature at exon regions in comparison to intron regions. Exon 5 also individually exhibited significantly lower DNA curvature than intron regions in both experiments. Both approaches provided corroboration of the predictions made using the De Santis dinucleotide wedge models of DNA curvature. In addition to this exon position experimentally exhibited moderate to low flexibility with the exception of exon 7 (Section 6.2.13.).

Of the two methodologies the FF algorithm provided curvature profiles most statistically comparable to theoretical profiles. However, it was more analytically demanding to implement and there were many methodological considerations that needed to be addressed when considering the output. Streptavidin end-labelling was analytically simpler to implement, but the output was less comparable to theoretical profiles. For future studies, the FF algorithm would be recommended for smaller, palindromic DNA molecules. Larger molecules should be approached using streptavidin end-labelling followed by application of the FF algorithm as a final corroboration of the resulting curvature profiles. This would remove any remaining uncertainly about molecule orientation and allow for a very high degree of confidence in the resulting profiles.

7.5 Exons as Regions of Low Intrinsic DNA Curvature.

TP53 is heavily conserved in evolution due to its key importance in cell regulation, maintenance and repair (Lane *et al.*, 2010). The reduced curvature of exon positions within *TP53* may indicate that the structural architecture of the coding regions has been selected for during evolution. Alternatively, low intrinsic DNA curvature could be a by-product of the accumulation of GC base pair content in coding sections of DNA throughout evolutionary time (Galtier *et al.*, 2001). If intrinsic DNA curvature has been actively selected for, then it is most likely to be due to the influence of curvature on nucleosome positioning and the maintenance of nucleosome structure (Shrader and Crothers, 1990; Virstedt *et al.*, 2004). Although curvature has been shown to influence transcription and replication, the impact of curvature is predominantly in the origins of replication and promoter regions of genes (Ohyama, 2005; Marilley *et al.*, 2007b). As the *TP53* sequences that were investigated contained no promoters or replication origins, the role of intrinsic curvature in *TP53* is likely to be structural. Low levels of DNA curvature in genes have been related to open chromatin and active transcription (Vinogradov, 2003). *TP53* is constantly transcribed at a low level within the cell, and its transcription is tightly regulated, so evolutionary selection for DNA architecture to enhance stable transcription would be beneficial to *TP53* (Hollstein and Hainaut, 2010). The theory of evolutionary selection for architectural features in genes has been proposed previously and

favours active selection for intrinsic curvature rather than selection for GC content leading to reduced curvature (Vinogradov, 2003).

7.6 Low Intrinsic DNA Curvature at Sites of Slow Repair in *TP53*

There are a number of sites in *TP53* that have shown impaired DNA repair of bulky chemical adducts (Tornaletti and Pfeifer, 1994; Denissenko *et al.*, 1998; Zhu, 2000). Codons exhibiting slow repair were found to have a significant reduction in curvature in comparison to the remaining *TP53* sequence using theoretical models (Section 4.2.14.). However, regions of slow repair were localised to exons which independently exhibited reduced curvature. Therefore, the current experiment did not discount the possibility that the low curvature in slow DNA repair codons was due to localisation within exons rather than a structural feature of the slow DNA repair codons themselves.

Reduced curvature in codons of slow repair implied a role for intrinsic curvature in the repair of DNA in *TP53*. The local DNA sequence bordering a chemical bulky adduct has been shown to have a measurable effect on the repair efficiency via the NER pathway (Cai *et al.*, 2009, 2010). Two of the key proteins, XPA and RPA, in the NER pathway specifically recognise DNA structural deformities due to chemical adduction and are also required to deform DNA in order to function (Missura *et al.*, 2001). Studies have concluded that DNA curvature may have a role as a stabilising factor in the presentation of DNA adducts for repair (Cai *et al.*, 2009, 2010). Gel electrophoretic experiments and molecular dynamics simulations indicate that rigidly bent DNA sequences present a wider minor groove leading to more efficient excision and repair of the DNA lesions. The DNA adduct used in these studies was BPDE, derived from benzo[a]pyrene, a chemical carcinogen heavily involved in the initiation and progression of lung cancer (Hecht, 2002; Kometani *et al.*, 2009). BPDE has also been implicated as a causative agent for the three lung cancer specific mutation hotspot codons that exhibited slow DNA repair (Denissenko *et al.*, 1998; Hussain *et al.*, 2001). Although the evidence for low flexibility in exon positions is not as strong as the evidence for low curvature it seem likely that flexibility also plays a role in adduct repair. The combination of low flexibility and low curvature in *TP53* exons may collaborate to reduce the presentation of chemical adducts at these sites. Therefore, it was hypothesised that the regions of slow repair in *TP53* may be due, at least in part, to the straight and rigid nature of the DNA causing a reduced presentation of chemical adducts for removal by the NER pathway.

7.7 Low DNA Curvature and Nucleosome Occupancy in *TP53*

Nucleosome affinity algorithms applied to *TP53* indicated that both exon 5 and exon 6 were unlikely to be occupied by nucleosomes (Chapter 4). The mechanism underlying reduced

sequence-specific DNA repair efficiency has also been attributed to the accessibility of the DNA due to the local chromatin structure (Bohr, 1987). As curvature has an active role in nucleosome positioning and the maintenance of nucleosome structure it may also affect DNA repair efficiency indirectly through nucleosome positioning (Shrader and Crothers, 1990; Anselmi *et al.*, 1999). The potential for exon 5 and 6 to be excluded from the nucleosome core has interesting implications for DNA damage models. For example, exon 5 is highly mutated in lung cancer (Denissenko *et al.*, 1996). One of the major carcinogens involved in lung cancer, BPDE, has been shown to preferentially bind to DNA not contained in the nucleosome core (Jack and Brookes, 1982; Kurian *et al.*, 1985). Intrinsic DNA curvature has an active role in nucleosome positioning and the maintenance of nucleosome structure (Shrader and Crothers, 1990; Anselmi *et al.*, 1999). It may therefore influence DNA damage rates and DNA repair rates indirectly via control of nucleosome architecture.

As has been observed in the present study all exons show a local reduction in intrinsic curvature. Nucleosomes have been shown to exhibit affinity for curved DNA sequences of low flexibility rather than uncurved DNA sequences (Anselmi *et al.*, 1999). This local reduction in curvature may represent large scale structural motif for exclusion from the nucleosome core. The reduced flexibility of exons within *TP53* may provide another energetic barrier to nucleosome formation. It is unlikely that this large scale structural motif would be identified by nucleosome affinity algorithms working on smaller scales (Xi *et al.*, 2010). Exon 7 shows the most deviation from the trend within *TP53* of low curvature and low flexibility, having moderate curvature and moderate to high flexibility. The high flexibility alone could provide an energetic barrier to nucleosome formation. The evolutionary benefit to *TP53* would be to promote open chromatin structure and enhanced transcription of *TP53*. The downstream effects of this selection would be increased mutation rate by environmental carcinogens due to the increased affinity of key carcinogens for open chromatin and a decrease in repair rates because of poor presentation of chemical adducts for excision.

7.8 A Potential Role of Curvature in Post-Transcriptional Modification

It was observed that, in signed curvature profiles, exons often occurred in regions of low curvature that bordered a change in the direction of curvature (Chapter 6). The change in the curvature regime at the border of exons suggests a role in post-transcriptional modification of RNA, namely the removal of introns from the RNA transcript. GC content and related sequence motifs are recognised during post-transcriptional splicing of RNA (Amit *et al.*, 2012). GC content was explicitly investigated in the study, which is closely linked to intrinsic DNA curvature. The change in the structural regime of curvature on the border of exons may

represent a recognition factor for the spliceozome or other associated proteins. This is an area that could be investigated in more depth in future studies.

7.9 Future Studies on the Intrinsic DNA Curvature of *TP53*

The present study has successfully evaluated the intrinsic DNA curvature of *TP53* and has indicated that exons have recognisable structural differences from introns within the same gene. The study has hypothesised that DNA repair efficiency of mutation hotspots may be influenced by DNA curvature via both the structural presentation of adducts for excision and the control of chromatin architecture within exons. This suggests a number of beneficial avenues for further investigation. The curvature in individual exons could be further quantified at improved resolution by the use of smaller, end-labelled PCR products of individual exons using sharper AFM tips such as carbon nanotubes (Woolley *et al.*, 2000). Alternatively, this could be achieved using small palindromic dimers and the application of the FF algorithm. Time-lapse DNA dynamics experiments on *TP53* would elucidate the relationship between intrinsic DNA curvature and flexibility in exon positions (Suzuki *et al.*, 2011).

Furthermore, AFM has been successfully applied to visualise nucleosome affinity and dynamics in both air and liquid (Van Vugt *et al.*, 2009; Filenko *et al.*, 2012). The application of these techniques to *TP53* DNA would allow for the experimental testing of the hypotheses developed as an outcome of the present study. An alternative experimental route would be to directly investigate the repair efficiency of NER repair enzymes on damaged *TP53* by AFM imaging in real time (Lysetska *et al.*, 2002).

Finally, the investigation of exonic DNA curvature could be quickly extended to other highly evolutionarily conserved genes using the theoretical framework for investigation developed in this study. The control of chromatin architecture by DNA curvature has already been established by previous authors. The investigation of large scale curvature features may provide another tool for the evaluation of nucleosome affinity and could potentially be used to identify regions of evolutionary conservation.

A.1.2. Optimisation of PCR Conditions

Genomic DNA was used as a template for PCR amplification using primer sets *TP53* Exon 5-7 and Exon 5-9. A basic PCR protocol (95 ^{0}C for 30 s, 60 ^{0}C for 30 s, 72 ^{0}C for 30 s, repeat 30 times) was adopted as a framework and modified. The extension times were extended to 60 s and 90 s for Exon 5-7 (1855 bp) and Exon 5-9 (2500 bp) respectively as they are both considered large DNA templates. An additional extension step of 72 °C for 5 min was added at the end of the protocol to ensure full primer extension. The addition of a hot start and hot stop (95 °C for 10 min) was found to be necessary to stop dimerisation of primers and products in certain amplifications. A range of annealing temperatures was used in order to find an optimal temperature (Figure A.1.1.). This was experimentally identified as 60 °C for both Exon 5-7 and Exon 5-9 (Figure A.1.2.).

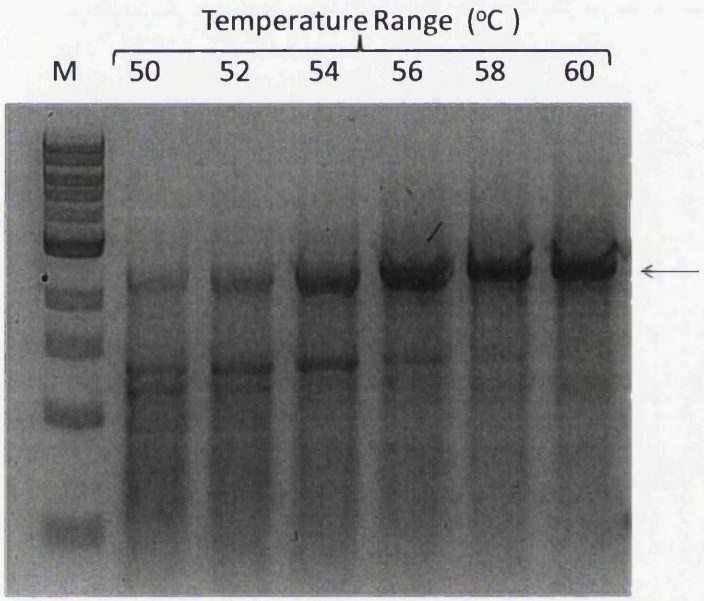

Figure A.1.1. - Comparison of PCR amplification products of *TP53* Exon 5-9 using a range of annealing temperatures (50 °C – 60 °C). Lane M contains a New England Biolabs 1 Kb DNA ladder. The expected (2500 bp) band is indicated with a black arrow. There was observable multiple banding within the 50°C- 56°C temperature range.

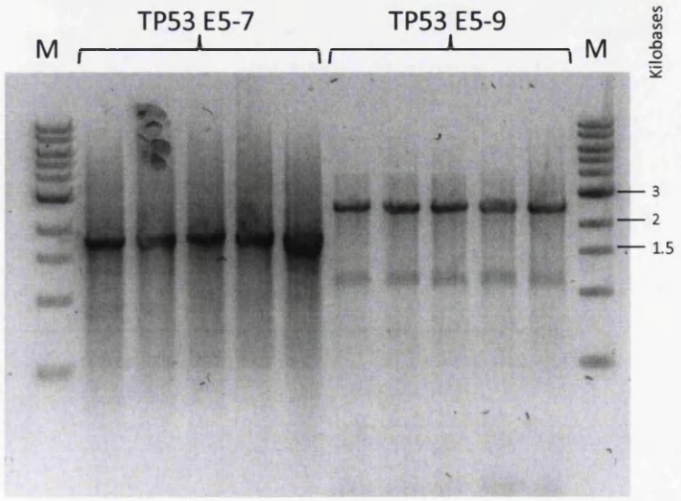

Figure A.1.2. - Replicates of *TP53* samples prepared for AFM analysis. Lanes M contain NEB 1 Kb DNA ladder. The first five sample lanes contained replicates of *TP53* Exon 5-7 (1855 bp). The second five sample lanes contained replicates of *TP53* Exon 5-9 (2500 bp). The right hand labels show the size of the pertinent marker bands in kilobases (kb).

A.1.2. Elimination of Non-Specific Bands

In some PCR amplifications of the biotinylated *TP53* Exon 5-9 PCR products there was an observable non-specific band at around 1.4 Kb. Due to the non-reproducibility of this band it was likely due to uncontrollable variations in either the thermal cycler temperature or the quality of purchased PCR reagents. In order to remove the non-specific band another series of PCR amplifications were performed on the *TP53* Exon 5-9 PCR sample. No non-specific bands were observed in the second round of amplifications (Figure A.1.3.). TP33 Exon 5-7 products are all amplified from the first round of *TP53* Exon 5-9 PCR products.

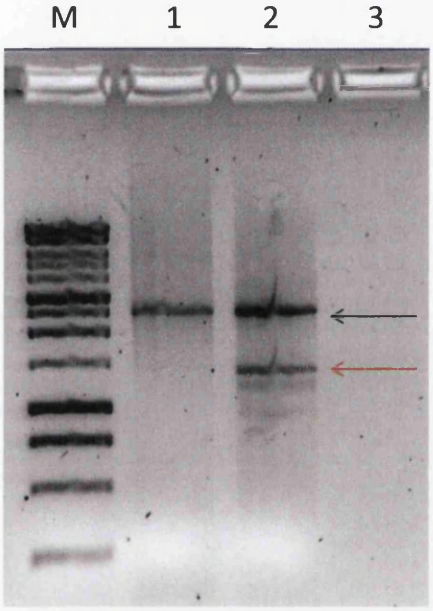

Figure A.1.3. - Comparison of initial and secondary PCR amplification of *TP53* Exon 5-9 PCR product. Lane M contains Generuler 1 Kb DNA Ladder. Lane 2 contained *TP53* Exon 5-9 reamplified PCR product. Lane 3 contained *TP53* Exon 5-9 of PCR product from human genomic DNA. The expected (2500 bp) band is indicted with a black arrow. Non-specific band at 1.4 kb is indicated with a red arrow.

A.1.3. Estimation of Error Rate of PCR for AFM Analysis.

The error rate of PCR is non-negligible (Cha and Thilly, 1993). In order to minimise potential base pair errors a High Fidelity Taq polymerase blend with a proofreading protein was used during this study for all samples prepared for AFM. The Expand High Fidelity^{PLUS} PCR System (Roche,UK) has a reported error rate of 2.4×10^{-6} base pairs per cycle. The average cycle number for amplification was 30 cycles. This gives an error rate of 7.2×10^{-5} per base pair per amplification (30 cycles multiplied by an error rate of 2.4×10^{-6}). *TP53* Exon 5-7 is a 1855 bp DNA molecule and gives an error rate of 0.134 per amplification. *TP53* Exon 5-9 was a 2500 bp DNA molecule and gave an error rate of 0.18 per amplification.

A.1.4. Sequencing Summary

DNA sequencing was carried out on parts of the amplified experimental molecule to confirm amplification fidelity and to detect any polymorphisms that may be within the human genomic DNA sample compared to a consensus sequence taken from the IARC *TP53* database (Hernandez-Boussard *et al.*, 1999). The IARC *TP53* database is a compilation of *TP53* sequences taken from human population studies. The 2500 bp molecule (from 11828 to 14328 in IARC Database notation) was sequenced from 11828-12669 and then from 13223-14328. This covered all of the major exons except for exon 6, which was only partially sequenced. Only two points of deviation from the consensus IARC sequence were detected. The SNP (validated in human populations) at 25051, commonly guanine, was detected as a thymine in the experimental DNA. Additionally, the consensus sequence of CCAGC**TTTCAAAAA**GA (14311-14327) was detected as CCAGC**TTCAAAAAA**GA. A thymine was deleted and an adenine inserted in the A-tract of the experimental molecule.

A.1.5. Estimating the Effect of DNA Polymorphisms on Theoretical Curvature

The effect of the base pair deviations from the IARC consensus sequences was estimated using theoretical models. CURVATURE was used to model both sequences (Figure A.1.4.). The C-T transition at the 5' end of the sequences had only a small net effect on the curvature profile. The insertion/deletion at the 3' end of the sequences had a more noticeable, although small, effect on the curvature profile. Neither polymorphism was likely to have a measurable effect on the experimental outcome.

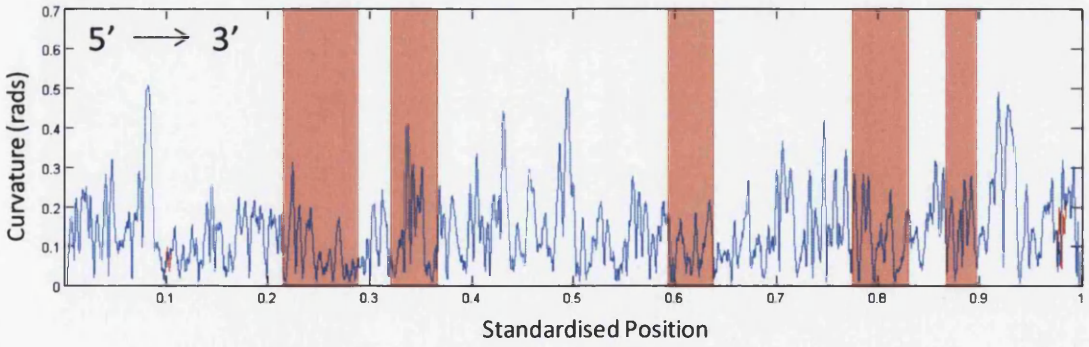

Figure A.1.4. - Effect of base pair deviations on intrinsic DNA curvature of *TP53*. The IARC consensus sequence is shown in blue and the results of sequencing of PCR product in red. The length of *TP53* was standardised using a scale of zero to one. Exon positions are highlighted in red in ascending order from left to right. Curvature profiles were generated using CURVATURE (Shpigelman *et al.*, 1993). The default settings and the De Santis model of curvature were used to produce curvature profiles.

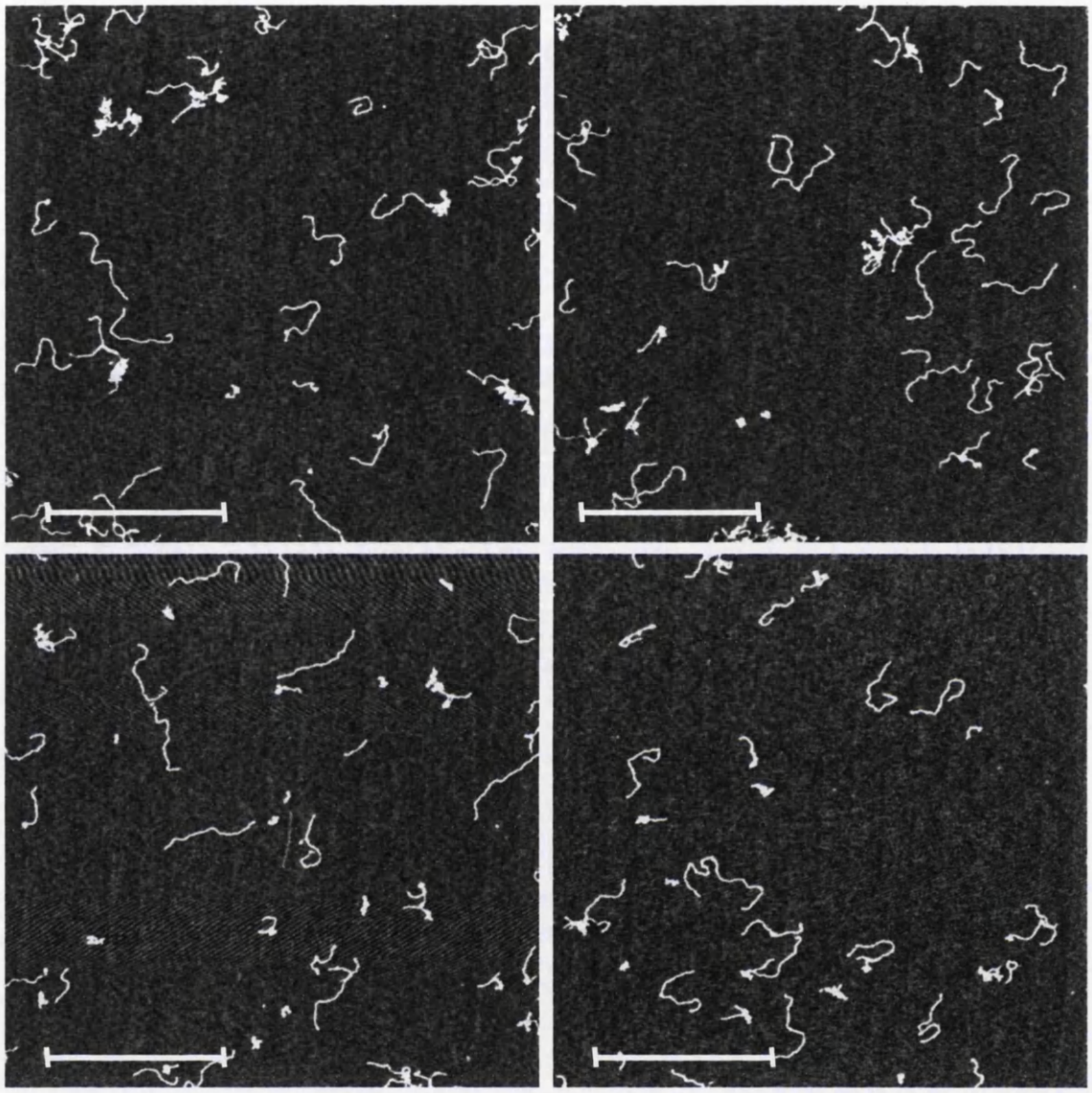

Figure A.2.1. – Example of AFM images of *TP53* Exon 5-7 1855 bp DNA molecules. Images were captured at a size of 3x3 μM, a resolution of 1024x1024 and with 6 nm ROC cantilevers. The scale bar (white) represents 1 μM.

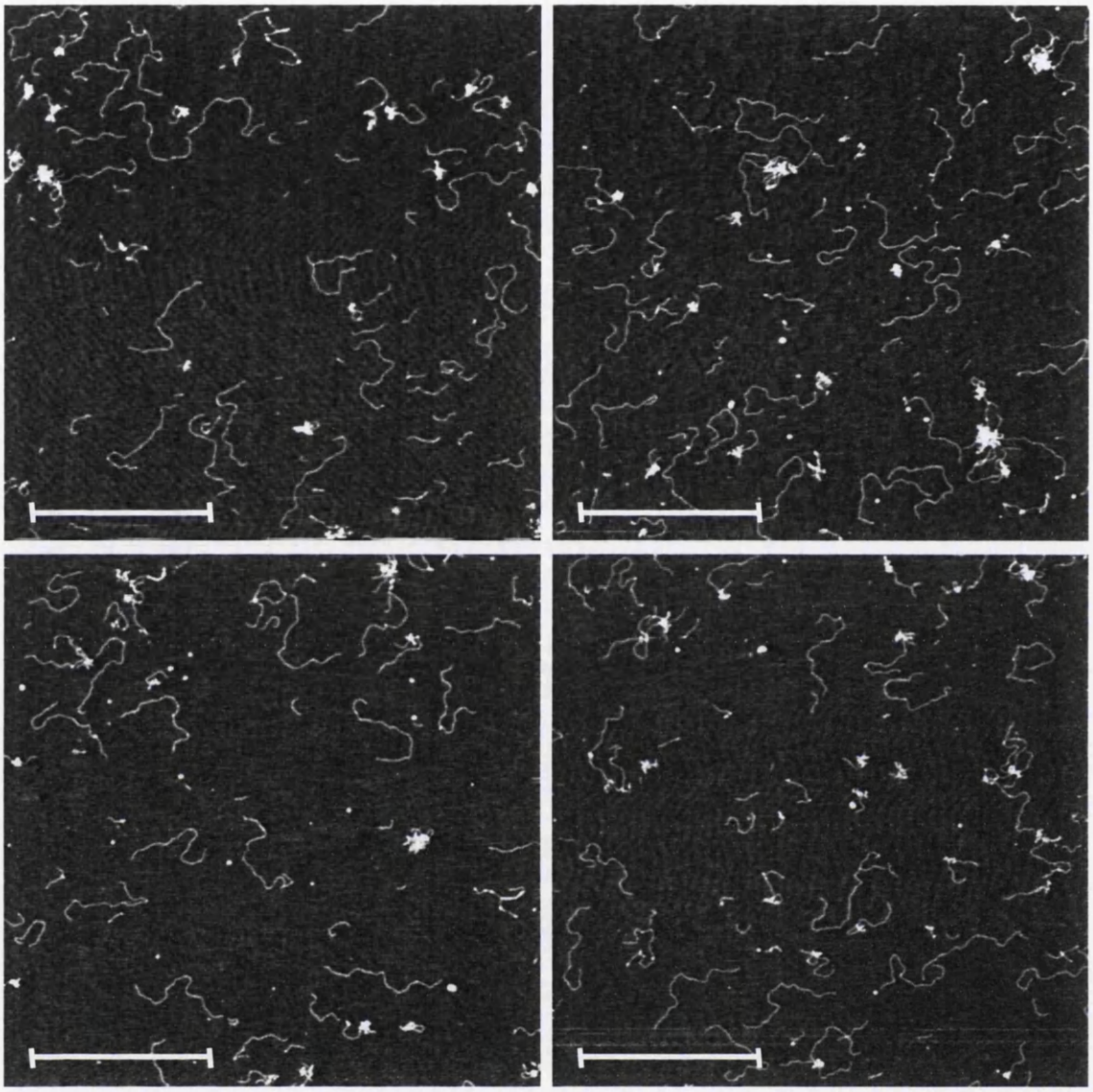

Figure A.2.2. – Example of AFM images of *TP53* Exon 5-9 2500 bp DNA molecules. Images were captured at a size of 3x3 μM, a resolution of 1024x1024 and with 6 nm ROC cantilevers. The scale bar represents 1 μM.

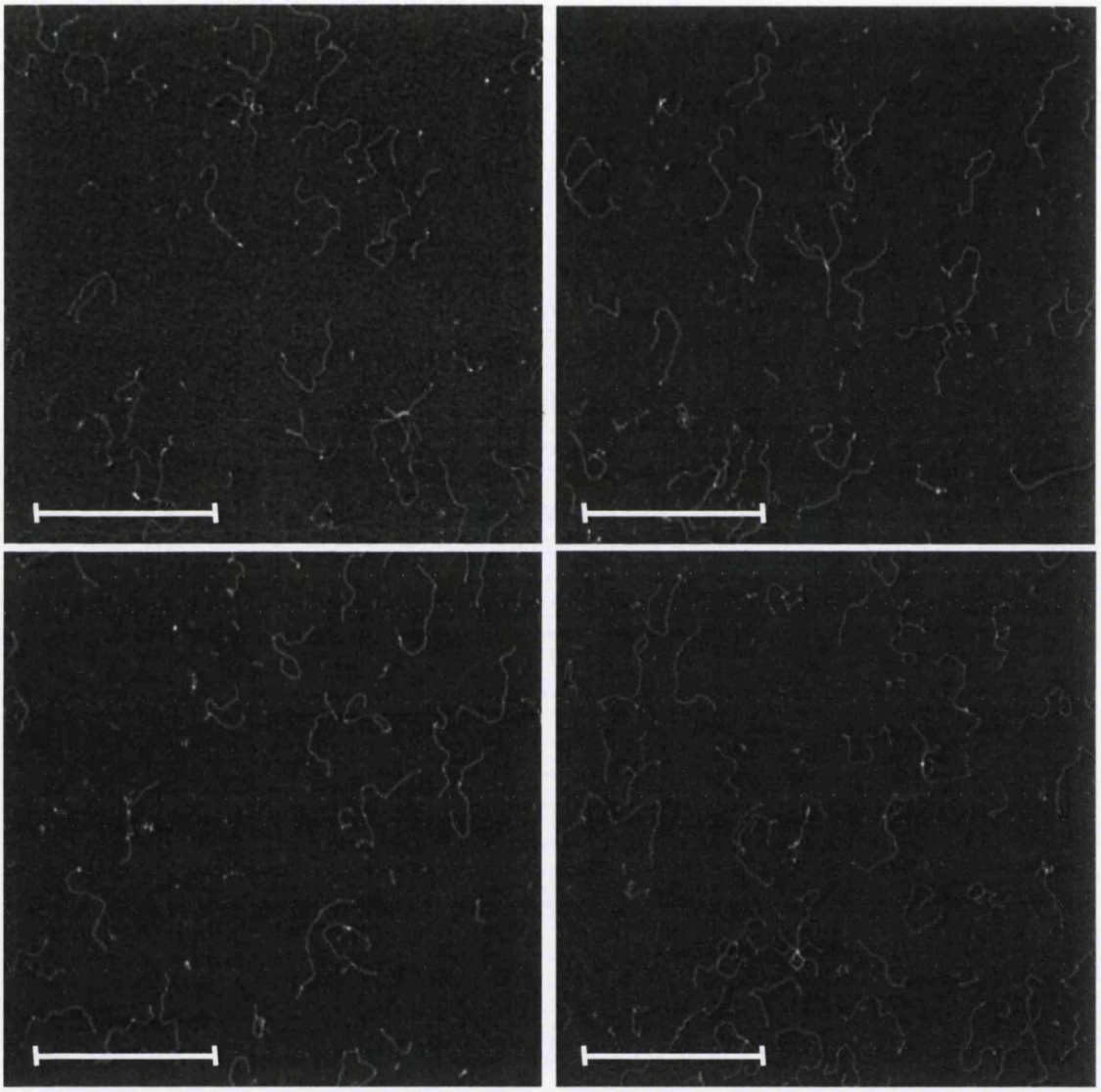

Figure A.2.3. – Example of AFM images of *TP53* DNA molecules 5' end-labelled with streptavidin. Images were captured at a size of 3x3 µM, a resolution of 1024x1024 and with 6 nm ROC cantilevers. The scale bar represents 1 µM.

Bibliography

Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., and Van de Peer, Y. (2008). Generic eukaryotic core promoter prediction using structural features of DNA. Genome Research *18*, 310–323.

Abel, J., and Mrázek, J. (2012). Differences in DNA curvature-related sequence periodicity between prokaryotic chromosomes and phages, and relationship to chromosomal prophage content. BMC Genomics *13*, 188.

Akiyama, T., and Hogan, M.E. (1997). Structural analysis of DNA bending induced by tethered triple helix forming oligonucleotides. Biochemistry *36*, 2307–2315.

Allen, M.J., Hud, N. V, Balooch, M., Tench, R.J., Siekhaus, W.J., and Balhorn, R. (1992). Tip-radius-induced artifacts in AFM images of protamine-complexed DNA fibers. Ultramicroscopy *42-44*, 1095–1100.

Allison, D.P., Bottomley, L.A., Thundat, T., Brown, G.M., Woychik, R.P., Schrick, J.J., Jacobson, K.B., and Warmack, R.J. (1992). Immobilization of DNA for scanning probe microscopy. Proceedings of the National Academy of Sciences of the United States of America *89*, 10129–10133.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., *et al.* (2012). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. Cell Reports *1*, 543–556.

An, H., Guo, Y., Zhang, X., Zhang, Y., and Hu, J. (2005). Nanodissection of single- and double-stranded DNA by atomic force microscopy. Journal of Nanoscience and Nanotechnology *5*, 1656–1659.

Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M., and Scipioni, A. (1999). Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability. Journal of Molecular Biology *286*, 1293–1301.

Anselmi, C., Bocchinfuso, G., De Santis, P., Savino, M., and Scipioni, A. (2000). A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. Biophysical Journal *79*, 601–613.

Asayama, M., and Ohyama, T. (2000). Curved DNA and Prokaryotic Promoters: A Mechanism for Activation of Transcription. In DNA Conformation and Transcription, T. Ohyama, ed. (Landes Bioscience),.

Barret, S.D. (2008). Image SXM. [online] Available at: <http://www.ImageSXM.org.uk> [Accessed: 30-06-2012].

Bartels, F.W., McIntosh, M., Fuhrmann, A., Metzendorf, C., Plattner, P., Sewald, N., Anselmetti, D., Ros, R., and Becker, A. (2007). Effector-stimulated single molecule protein-DNA interactions of a quorum-sensing system in *Sinorhizobium meliloti*. Biophysical Journal *92*, 4391–4400.

Basham, B., Schroth, G.P., and Ho, P.S. (1995). An A-DNA triplet code: thermodynamic rules for predicting A- and B-DNA. Proceedings of the National Academy of Sciences of the United States of America *92*, 6464–6468.

Baumann, C.G., Smith, S.B., Bloomfield, V.A., and Bustamante, C. (1997). Ionic effects on the elasticity of single DNA molecules. Proceedings of the National Academy of Sciences of the United States of America *94*, 6185–6190.

Bednar, J., Furrer, P., Katritch, V., Stasiak, A.Z., Dubochet, J., and Stasiak, A. (1995). Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. Journal of Molecular Biology *254*, 579–594.

Beveridge, D.L., Barreiro, G., Byun, K.S., Case, D.A., Cheatham III, T.E., Dixit, S.B., Giudice, E., Lankas, F., Lavery, R., Maddocks, J.H., *et al.* (2004). Molecular synamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(G) steps. Biophysical Journal *87*, 3799–3813.

Bewley, C.A., Gronenborn, A.M., and Clore, G.M. (1998). Minor groove-binding architectural proteins: structure, function, and DNA recognition. Annual Review of Biophysics and Biomolecular Structure *27*, 105–131.

Bezanilla, M., Drake, B., Nudler, E., Kashlev, M., Hansma, P.K., and Hansma, H.G. (1994). Motion and enzymatic degradation of DNA in the atomic force microscope. Biophysical Journal *67*, 2454–2459.

Bezanilla, M., Manne, S., Laney, D.E., Lyubchenko, Y.L., and Hansma, H.G. (1995). Adsorption of DNA to mica, silylated mica, and minerals: Characterization by atomic force microscopy. Langmuir *11*, 655–659.

Biesalski, H.K., Bueno de Mesquita, B., Chesson, A., Chytil, F., Grimble, R., Hermus, R.J., Köhrle, J., Lotan, R., Norpoth, K., Pastorino, U., *et al.* (1998). European Consensus Statement on Lung Cancer: risk factors and prevention. Lung Cancer Panel. CA: a Cancer Journal for Clinicians *48*, 167–176.

Binnig, G., Quate, C.F., and Gerber, C. (1986). Atomic Force Microscope. Physical Review Letters *56*, 930–933.

Binnig, G., and Rohrer, H. (1993). Scanning Tunneling Microscopy. Surface Science *126*, 236–244.

Bohr, V.A. (1987). Preferential DNA repair in active genes. Danish Medical Bulletin *34*, 309–320.

Bolshoy, A., McNamara, P., Harrington, R.E., and Trifonov, E.N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. Proceedings of the National Academy of Sciences of the United States of America *88*, 2312–2316.

Bram, S. (1973). Variation of Type-B DNA X-Ray Fiber Diagrams with Base Composition. Proceedings of the National Academy of Sciences of the United States of America *70*, 2167–2170.

Brugal, G., and Chassery, J.M. (1977). A new image-processing system designed for densitometry and pattern analysis of microscopic specimen. Application to the automated recognition and counting of cells in the various phases of the mitotic cycle. Histochemistry *52*, 241–258.

Brukner, I., Susic, S., Dlakic, M., Savic, A., and Pongor, S. (1994). Physiological concentration of magnesium ions induces a strong macroscopic curvature in GGGCCC-containing DNA. Journal of Molecular Biology *236*, 26–32.

Brukner, I., Sánchez, R., Suck, D., and Pongor, S. (1995a). Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. The EMBO Journal *14*, 1812–1818.

Brukner, I., Sánchez, R., Suck, D., and Pongor, S. (1995b). Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. Journal of Biomolecular Structure Dynamics *13*, 309–317.

Di Bucchianico, S., Poma, A.M., Giardi, M.F., Di Leandro, L., Valle, F., Biscarini, F., and Botti, D. (2011). Atomic Force Microscope nanolithography on chromosomes to generate single-cell genetic probes. Journal of Nanobiotechnology *9*, 27.

Burkhoff, A.M., and Tullius, T.D. (1987). The unusual conformation adopted by the adenine tracts in kinetoplast DNA. Cell *48*, 935–943.

Bustamante, C., Marko, J., Siggia, E., and Smith, S. (1994). Entropic elasticity of lambda-phage DNA. Science *265*, 1599–1600.

Bustamante, C., and Rivetti, C. (1996). Visualizing protein-nucleic acid interactions on a large scale with the scanning force microscope. Annual Review of Biophysics and Biomolecular Structure *25*, 395–429.

Bustamante, C., Smith, S.B., Liphardt, J., and Smith, D. (2000). Single-molecule studies of DNA mechanics. Current Opinion in Structural Biology *10*, 279–285.

Buzio, R., Repetto, L., Giacopelli, F., Ravazzolo, R., and Valbusa, U. (2012). Label-free, atomic force microscopy-based mapping of DNA intrinsic curvature for the nanoscale comparative analysis of bent duplexes. Nucleic Acids Research *40*, 1–14.

Cai, Y., Kropachev, K., Xu, R., Tang, Y., Kolbanovskii, M., Kolbanovskii, A., Amin, S., Patel, D.J., Broyde, S., and Geacintov, N.E. (2010). Distant neighbor base sequence context effects in human nucleotide excision repair of a benzo[a]pyrene-derived DNA lesion. Journal of Molecular Biology *399*, 397–409.

Cai, Y., Patel, D.J., Geacintov, N.E., and Broyde, S. (2009). Differential nucleotide excision repair susceptibility of bulky DNA adducts in different sequence contexts: hierarchies of recognition signals. Journal of Molecular Biology *385*, 30–44.

Cairns, B.R. (2009). The logic of chromatin architecture and remodelling at promoters. Nature *461*, 193–198.

Calladine, C.R., Drew, H.R., and McCall, M.J. (1988). The intrinsic curvature of DNA in solution. Journal of Molecular Biology *201*, 127–137.

Le Cam, E., Fack, F., Ménissier-de Murcia, J., Cognet, J.A., Barbin, A., Sarantoglou, V., Révet, B., Delain, E., and De Murcia, G. (1994). Conformational analysis of a 139 base-pair DNA fragment containing a single-stranded break and its interaction with human poly(ADP-ribose) polymerase. Journal of Molecular Biology *235*, 1062–1071.

Cam, E.L., Culard, F., Larquet, E., Delain, E., and Cognet, J.A. (1999). DNA bending induced by the archaebacterial histone-like protein MC1. Journal of Molecular Biology *285*, 1011–1021.

Carasso, A.S. (1999). Linear and Nonlinear Image Deblurring: A Documented Study. SIAM Journal on Numerical Analysis *36*, 1659–1689.

Cassina, V., Seruggia, D., Beretta, G.L., Salerno, D., Brogioli, D., Manzini, S., Zunino, F., and Mantegazza, F. (2011). Atomic force microscopy study of DNA conformation in the presence of drugs. European Biophysics Journal *40*, 59–68.

Cha, R.S., and Thilly, W.G. (1993). Specificity, efficiency, and fidelity of PCR. PCR Methods And Applications *3*, 18–29.

Champ, P.C., Maurice, S., Vargason, J.M., Camp, T., and Ho, P.S. (2004). Distributions of Z-DNA and nuclear factor I in human chromosome 22: a model for coupled transcriptional regulation. Nucleic Acids Research *32*, 6501–6510.

Charney, E., Chen, H.H., and Rau, D.C. (1991). The flexibility of A-form DNA. Journal of Biomolecular Structure & Dynamics *9*, 353–362.

Cheatham, T.E., and Young, M.A. (2000). Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. Biopolymers *56*, 232–256.

Cirrone, S. (2007). Automatic Recognition and Analysis of DNA molecules by AFM Image Processing. Master's Thesis. University of Catania.

Cocco, S., Marko, J.F., and Monasson, R. (2002). Theoretical models for single-molecule DNA and RNA experiments: from elasticity to unzipping. Comptes Rendus Physique *3*, 569–584.

Cognet, J.A., Pakleza, C., Cherny, D., Delain, E., and Cam, E.L. (1999). Static curvature and flexibility measurements of DNA with microscopy. A simple renormalization method, its assessment by experiment and simulation. Journal of Molecular Biology *285*, 997–1009.

Collins, T.J. (2007). ImageJ for microscopy. Biotechniques *43*, 25–30.

Cooper, J.P., and Hagerman, P.J. (1987). Gel electrophoretic analysis of the geometry of a DNA four-way junction. Journal of Molecular Biology *198*, 711–719.

Coury, J.E., McFail-Isom, L., Williams, L.D., and Bottomley, L.A. (1996). A novel assay for drug-DNA binding mode, affinity, and exclusion number: scanning force microscopy. Proceedings of the National Academy of Sciences of the United States of America *93*, 12283–12286.

Crothers, D.M. (1998). DNA curvature and deformation in protein-DNA complexes: a step in the right direction. Proceedings of the National Academy of Sciences of the United States of America *95*, 15163–15165.

Crothers, D.M., Drak, J., Kahn, J.D., and Levene, S.D. (1992). DNA bending, flexibility, and helical repeat by cyclization kinetics. Methods in Enzymology *212*, 3–29.

Crothers, D.M., Haran, T.E., and Nadeau, J.G. (1990). Intrinsically bent DNA. The Journal of Biological Chemistry *265*, 7093–7096.

Denissenko, M.F., Pao, A., Pfeifer, G.P., and Tang, M. (1998). Slow repair of bulky DNA adducts along the nontranscribed strand of the human p53 gene may explain the strand bias of transversion mutations in cancers. Oncogene *16*, 1241–1247.

Denissenko, M.F., Pao, A., Tang, M., and Pfeifer, G.P. (1996). Preferential Formation of Benzo[a]pyrene Adducts at Lung Cancer Mutational Hotspots in P53. Science *274*, 430–432.

Dickerson, R.E., and Chiu, T.K. (1997). Helix bending as a factor in protein/DNA recognition. Biopolymers *44*, 361–403.

Dickerson, R.E., and Drew, H.R. (1981a). Kinematic model for B-DNA. Proceedings of the National Academy of Sciences of the United States of America *78*, 7318–7322.

Dickerson, R.E., and Drew, H.R. (1981b). Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. Journal of Molecular Biology *149*, 761–786.

Dickerson, R.E., Drew, H.R., Conner, B.N., Wing, R.M., Fratini, A. V, and Kopka, M.L. (1982). The anatomy of A-, B-, and Z-DNA. Science *216*, 475–485.

Dickerson, R.E., Goodsell, D.S., and Neidle, S. (1994). "...the tyranny of the lattice...". Proceedings of the National Academy of Sciences of the United States of America *91*, 3579–3583.

Dixit, S.B., Pitici, F., and Beveridge, D.L. (2004). Structure and axis curvature in two dA6 x dT6 DNA oligonucleotides: comparison of molecular dynamics simulations with results from crystallography and NMR spectroscopy. Biopolymers *75*, 468–479.

Dlakic, M., and Harrington, R.E. (1998a). DIAMOD: display and modeling of DNA bending. Bioinformatics *14*, 326–331.

Dlakic, M., and Harrington, R.E. (1998b). Unconventional helical phasing of repetitive DNA motifs reveals their relative bending contributions. Nucleic Acids Research *26*, 4274–4279.

Dornberger, U., Flemming, J., and Fritzsche, H. (1998). Structure determination and analysis of helix parameters in the DNA decamer d(CATGGCCATG)2 comparison of results from NMR and crystallography. Journal of Molecular Biology *284*, 1453–1463.

Drouin, R., and Therrien, J.P. (1997). UVB-induced Cyclobutane Pyrimidine Dimer Frequency Correlates with Skin Cancer Mutational Hotspots in p53. Photochemistry and Photobiology *66*, 719–726.

Elias, J.G., and Eden, D. (1981). Transient electric birefringence study of the persistence length and electrical polarizability of restriction fragments of DNA. Macromolecules *14*, 410–419.

Fang, Y., Spisz, T.S., Wiltshire, T., D'Costa, N.P., Bankman, I.N., Reeves, R.H., and Hoh, J.H. (1998). Solid-state DNA sizing by atomic force microscopy. Analytical Chemistry *70*, 2123–2129.

Ficarra, E., Benini, L., Macii, E., and Zuccheri, G. (2005a). Automated DNA fragments recognition and sizing through AFM image processing. IEEE Transactions on Information Technology in Biomedicine *9*, 508–517.

Ficarra, E., Macii, E., Benini, L., and Zuccheri, G. (2004). A Robust Algorithm for Automated Analysis of DNA Molecules in AFM Images. In Proceedings of IASTED Biomedical Engineering.

Ficarra, E., Masotti, D., Macii, E., Benini, L., Zuccheri, G., and Samorì, B. (2005b). Automatic intrinsic DNA curvature computation from AFM images. IEEE Transactions on Bio-medical Engineering *52*, 2074–2086.

Filenko, N.A., Palets, D.B., and Lyubchenko, Y.L. (2012). Structure and dynamics of dinucleosomes assessed by atomic force microscopy. Journal of Amino Acids *2012*, 650840.

Franklin, R.E., and Gosling, R.G. (1953). Molecular configuration in sodium thymonucleate. Nature *171*, 400–401.

Freeman, H., Lipkin, B.S., and Rosenfeld, A. (1970). Picture Processing and Psychopictorics. In Picture Processing and Psychopictorics, pp. 241–266.

Gabrielian, A., and Pongor, S. (1996). Correlation of intrinsic DNA curvature with DNA property periodicity. FEBS Letters *393*, 65–68.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics *159*, 907–911.

Gatos, B., Pratikakis, I., and Perantonis, S.J. (2008). Efficient Binarization of Historical and Degraded Document Images. In The Eighth IAPR International Workshop on Document Analysis Systems, pp. 447–454.

Gohlke, C. (1994). CURVATURE. [online] Available at: <http://www.lfd.uci.edu/~gohlke/dnacurve/> [Accessed: 30-06-2012].

Goodsell, D.S., and Dickerson, R.E. (1994). Bending and curvature calculations in B-DNA. Nucleic Acids Research *22*, 5497–5503.

Goodsell, D.S., Kaczor-Grzeskowiak, M., and Dickerson, R.E. (1994). The crystal structure of C-C-A-T-T-A-A-T-G-G. Implications for bending of B-DNA at T-A steps. Journal of Molecular Biology *239*, 79–96.

Goñi, J.R., Fenollosa, C., Pérez, A., Torrents, D., and Orozco, M. (2008). DNAlive: a tool for the physical analysis of DNA at the genomic scale. Bioinformatics *24*, 1731–1732.

Greenblatt, M.S., Bennett, W.P., Hollstein, M., and Harris, C.C. (1994). Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Research *54*, 4855–4878.

Grove, A., Galeone, A., Mayol, L., and Geiduschek, E.P. (1996). On the connection between inherent DNA flexure and preferred binding of hydroxymethyluracil-containing DNA by the type II DNA-binding protein TF1. Journal of Molecular Biology *260*, 196–206.

Hagerman, P.J. (1988). Flexibility of DNA. Annual Review of Biophysics and Biophysical Chemistry *17*, 265–286.

Hamarneh, G. (2005). 3D live-wire-based semi-automatic segmentation of medical images. Proceedings of SPIE *5747*, 1597–1603.

Hamon, L., Pastré, D., Dupaigne, P., Le Breton, C., Le Cam, E., and Piétrement, O. (2007). High-resolution AFM imaging of single-stranded DNA-binding (SSB) protein--DNA complexes. Nucleic Acids Research *35*, 58.

Hansma, H.G., and Hoh, J.H. (1994). Biomolecular imaging with the atomic force microscope. Annual Review of Biophysics and Biomolecular Structure *23*, 115–139.

Hansma, H.G., Kim, K.J., Laney, D.E., Garcia, R.A., Argaman, M., Allen, M.J., and Parsons, S.M. (1997). Properties of biomolecules measured from atomic force microscope images: a review. Journal of Structural Biology *119*, 99–108.

Hansma, H.G., and Laney, D.E. (1996). DNA binding to mica correlates with cationic radius: assay by atomic force microscopy. Biophysical Journal *70*, 1933–1939.

Hansma, H.G., Vesenka, J., Siegerist, C., Kelderman, G., Morrett, H., Sinsheimer, R.L., Elings, V., Bustamante, C., and Hansma, P.K. (1992). Reproducible imaging and dissection of plasmid DNA under liquid with the atomic force microscope. Science *256*, 1180–1184.

Hansma, P.K., Cleveland, J.P., Radmacher, M., Walters, D.A., Hillner, P.E., Bezanilla, M., Fritz, M., Vie, D., Hansma, H.G., Prater, C.B., *et al.* (1994). Tapping mode atomic force microscopy in liquids. Applied Physics Letters *64*, 1738–1740.

Haran, T.E., Kahn, J.D., and Crothers, D.M. (1994). Sequence elements responsible for DNA curvature. Journal of Molecular Biology *244*, 135–143.

Hardwidge, P.R. (2002). Charge neutralization and DNA bending by the Escherichia coli catabolite activator protein. Nucleic Acids Research *30*, 1879–1885.

Harris, C.C., and Hollstein, M. (1993). Clinical implications of the p53 tumor-suppressor gene. The New England Journal of Medicine *329*, 1318–1327.

El Hassan, M.A., and Calladine, C.R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. Journal of Molecular Biology *259*, 95–103.

El Hassan, M.A., and Calladine, C.R. (1997). Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences *355*, 43–100.

Hecht, S.S. (2002). Cigarette smoking and lung cancer: chemical mechanisms and approaches to prevention. The Lancet Oncology *3*, 461–469.

Hernandez-Boussard, T., Rodriguez-Tome, P., Montesano, R., and Hainaut, P. (1999). IARC p53 mutation database: a relational database to compile and analyze p53 mutations in human tumors and cell lines. Human Mutation *14*, 1–8.

Hisada, M., Garber, J.E., Li, F.P., Fung, C.Y., and Fraumeni, J.F. (1998). Multiple Primary Cancers in Families With Li-Fraumeni Syndrome. JNCI Journal of the National Cancer Institute *90*, 606–611.

Hollstein, M., and Hainaut, P. (2010). Massively regulated genes: the example of *TP53*. The Journal of Pathology *220*, 164–173.

Hollstein, M., Sidransky, D., Vogelstein, B., and Harris, C. (1991). p53 mutations in human cancers. Science *253*, 49–53.

Horcas, I., Fernández, R., Gómez-Rodríguez, J.M., Colchero, J., Gómez-Herrero, J., and Baro, A.M. (2007). WSXM: a software for scanning probe microscopy and a tool for nanotechnology. Review of Scientific Instruments *78*,.

Hussain, S.P., Amstad, P., Raja, K., Sawyer, M., Hofseth, L., Shields, P.G., Hewer, A., Phillips, D.H., Ryberg, D., Haugen, A., *et al.* (2001). Mutability of p53 hotspot codons to benzo(a)pyrene diol epoxide (BPDE) and the frequency of p53 mutations in nontumorous human lung. Cancer Research *61*, 6350–6355.

Hussain, S.P., and Harris, C.C. (1999). p53 mutation spectrum and load: the generation of hypotheses linking the exposure of endogenous or exogenous carcinogens to human cancer. Mutation Research *428*, 23–32.

Ivanov, V.I., and Minchenkova, L.E. (1995). The A-form of DNA: in search of the biological role. Molecular Biology *28*, 1258–1271.

Jack, P.L., and Brookes, P. (1982). Mechanism for the loss of preferential benzo [a] pyrene binding to the linker DNA of chromatin. Carcinogenesis *3*, 341–344.

Jauregui, R. (2003). Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. Nucleic Acids Research *31*, 6770–6777.

Jiang, Y., and Marszalek, P.E. (2011). Atomic force microscopy captures MutS tetramers initiating DNA mismatch repair. The EMBO Journal *30*, 2881–2893.

Jiao, Y., Cherny, D.I., Heim, G., Jovin, T.M., and Schäffer, T.E. (2001). Dynamic interactions of p53 with DNA in solution by time-lapse atomic force microscopy. Journal of Molecular Biology *314*, 233–243.

Kaemmer, S. (2011). Introduction to Bruker's ScanAsyst and PeakForce Tapping AFM Technology. [online]. Available at: <http://nanoscaleworld.bruker-axs.com/nanoscaleworld/media/p/1548.aspx>. [Accessed on: 30-06-2012]

Kahn, J.D., Yun, E., and Crothers, D.M. (1994). Detection of localized DNA flexibility. Nature *368*, 163–166.

Kanhere, A. (2003). An assessment of three dinucleotide parameters to predict DNA curvature by quantitative comparison with experimental data. Nucleic Acids Research *31*, 2647–2658.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, A.D. (2002). The Human Genome Browser at UCSC. Genome Research *12*, 996–1006.

Klug, S.J., Wilmotte, R., Santos, C., Almonte, M., Herrero, R., Guerrero, I., Caceres, E., Peixoto-Guimaraes, D., Lenoir, G., Hainaut, P., *et al.* (2001). *TP53* polymorphism, HPV infection, and risk of cervical cancer. Cancer Epidemiology, Biomarkers & Prevention *10*, 1009–1012.

Kometani, T., Yoshino, I., Miura, N., Okazaki, H., Ohba, T., Takenaka, T., Shoji, F., Yano, T., and Maehara, Y. (2009). Benzo[a]pyrene promotes proliferation of human lung cancer cells by accelerating the epidermal growth factor receptor signaling pathway. Cancer Letters *278*, 27–33.

Kozobay-Avraham, L., Hosid, S., and Bolshoy, A. (2006). Involvement of DNA curvature in intergenic regions of prokaryotes. Nucleic Acids Research *34*, 2316–2327.

Krautbauer, R., Rief, M., and Gaub, H.E. (2003). Unzipping DNA Oligomers. Nano Letters *3*, 493–496.

Kruskal, W.H., and Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association *47*, 583–621.

Kulpa, Z. (1977). Area and perimeter measurement of blobs in discrete binary pictures. Computer Graphics and Image Processing *6*, 434–451.

Kurian, P., Jeffrey, A.M., and Milo, G.E. (1985). Preferential binding of benzo[a]pyrene diol epoxide to the linker DNA of human foreskin fibroblasts in S phase in the presence of benzamide. Proceedings of the National Academy of Sciences of the United States of America *82*, 2769–2773.

Lam, L., Lee, S.-W., and Suen, C.Y. (1992). Thinning methodologies-a comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence *14*, 869–885.

Lane, D.P., Cheok, C.F., Brown, C., Madhumalar, A., Ghadessy, F.J., and Verma, C. (2010). Mdm2 and p53 are highly conserved from placozoans to man. Cell Cycle *9*, 540–547.

Lane, D.P., and Crawford, L. V. (1979). T antigen is bound to a host protein in SY40-transformed cells. Nature *278*, 261–263.

Lankaš, F., Šponer, J., Langowski, J., and Cheatham, T.E. (2003). DNA basepair step deformability inferred from molecular dynamics simulations. Biophysical Journal *85*, 2872–2883.

Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D., and Zakrzewska, K. (2009). Conformational analysis of nucleic acids revisited: Curves. Nucleic Acids Research *37*, 5917–5929.

Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dixit, S., Jayaram, B., Lankas, F., Laughton, C., *et al.* (2010). A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. Nucleic Acids Research *38*, 299–313.

Leng, F., and McMacken, R. (2002). Potent stimulation of transcription-coupled DNA supercoiling by sequence-specific DNA-binding proteins. Proceedings of the National Academy of Sciences of the United States of America *99*, 9139–9144.

Levene, S.D., and Crothers, D.M. (1986). Ring closure probabilities for DNA fragments by Monte Carlo simulation. Journal of Molecular Biology *189*, 61–72.

Levine, A.J., Momand, J., and Finlay, C.A. (1991). The p53 tumour suppressor gene. Nature *351*, 453–456.

Li, W.J. (2007). AFM operating-drift detection and analyses based on automated sequential image processing. In 2007 7th IEEE Conference on Nanotechnology, pp. 748–753.

Llewellyn, K.J., Cary, P.D., McClellan, J.A., Guille, M.J., and Scarlett, G.P. (2009). A-form DNA structure is a determinant of transcript levels from the Xenopus gata2 promoter in embryos. Biochimica Et Biophysica Acta *1789*, 675–680.

Lonskaya, I., Potaman, V.N., Shlyakhtenko, L.S., Oussatcheva, E.A., Lyubchenko, Y.L., and Soldatenkov, V.A. (2005). Regulation of poly(ADP-ribose) polymerase-1 by DNA structure-specific binding. The Journal of Biological Chemistry *280*, 17076–17083.

Lu, X.-J., and Olson, W.K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Research *31*, 5108–5121.

Lu, X.-J., and Olson, W.K. (2008). 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nature Protocols *3*, 1213–1227.

Lu, X.J., Shakked, Z., and Olson, W.K. (2000). A-form conformational motifs in ligand-bound DNA structures. Journal of Molecular Biology *300*, 819–840.

Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome Biology *1*, 1–37.

Luykx, P., Bajić, I. V, and Khuri, S. (2006). NXSensor web tool for evaluating DNA for nucleosome exclusion sequences and accessibility to binding factors. Nucleic Acids Research *34*, 560–565.

Lysetska, M., Knoll, A., Boehringer, D., Hey, T., Krauss, G., and Krausch, G. (2002). UV light-damaged DNA and its interaction with human replication protein A: an atomic force microscopy study. Nucleic Acids Research *30*, 2686–2691.

Lyubchenko, Y. (1993). Atomic Force Microscopy of Long DNA: Imaging in Air and Under Water. Proceedings of the National Academy of Sciences *90*, 2137–2140.

Lyubchenko, Y.L., Jacobs, B.L., Lindsay, S.M., and Stasiak, A. (1995). Atomic force microscopy of nucleoprotein complexes. Scanning Microscopy *9*, 705–724.

Lyubchenko, Y.L., Shlyakhtenko, L.S., and Ando, T. (2011). Imaging of nucleic acids with atomic force microscopy. Methods *54*, 274–283.

Marek, J., Demjénová, E., Tomori, Z., Janácek, J., Zolotová, I., Valle, F., Favre, M., and Dietler, G. (2005). Interactive measurement and characterization of DNA molecules by analysis of AFM images. Cytometry. Part A: The Journal of the International Society for Analytical Cytology *63*, 87–93.

Marilley, M. (2000). Structure-function relationships in replication origins of the yeast Saccharomyces cerevisiae: higher-order structural organization of DNA in regions flanking the ARS consensus sequence. Molecular & General Genetics *263*, 854–866.

Marilley, M., Milani, P., and Rocca-Serra, J. (2007a). Gradual melting of a replication origin (*Schizosaccharomyces pombe* ars1): in situ atomic force microscopy (AFM) analysis. Biochimie *89*, 534–541.

Marilley, M., Milani, P., Thimonier, J., Rocca-Serra, J., and Baldacci, G. (2007b). Atomic force microscopy of DNA in solution and DNA modelling show that structural properties specify the eukaryotic replication initiation site. Nucleic Acids Research *35*, 6832–6845.

Marilley, M., Sanchez-Sevilla, A., and Rocca-Serra, J. (2005). Fine mapping of inherent flexibility variation along DNA molecules: validation by atomic force microscopy (AFM) in buffer. Molecular Genetics and Genomics *274*, 658–670.

Marini, J.C., Levene, S.D., Crothers, D.M., and Englund, P.T. (1982). Bent helical structure in kinetoplast DNA. Proceedings of the National Academy of Sciences of the United States of America *79*, 7664–7668.

Masotti, D., Ficarra, E., Macii, E., and Benini, L. (2004). Techniques for enhancing computation of DNA curvature molecules. In Fourth IEEE Symposium on Bioinformatics and Bioengineering,.

Merlitz, H., Rippe, K., Klenin, K. V, and Langowski, J. (1998). Looping dynamics of linear DNA molecules and the effect of DNA curvature: a study by Brownian dynamics simulation. Biophysical Journal *73*, 773–779.

Mihara, M., Erster, S., Zaika, A., Petrenko, O., Chittenden, T., Pancoska, P., and Moll, U.M. (2003). p53 has a direct apoptogenic role at the mitochondria. Molecular Cell *11*, 577–590.

Mikheikin, A.L., Lushnikov, A.Y., and Lyubchenko, Y.L. (2006). Effect of DNA supercoiling on the geometry of holliday junctions. Biochemistry *45*, 12998–13006.

Milani, P., Marilley, M., and Rocca-Serra, J. (2007). TBP binding capacity of the TATA box is associated with specific structural properties: AFM study of the IL-2R alpha gene promoter. Biochimie *89*, 528–533.

Milani, P., Marilley, M., Sanchez-Sevilla, A., Imbert, J., Vaillant, C., Argoul, F., Egly, J.-M., Rocca-Serra, J., and Arneodo, A. (2011). Mechanics of the IL2RA Gene Activation Revealed by Modeling and Atomic Force Microscopy. PLoS ONE *6*, 10.

Missura, M., Buterin, T., Hindges, R., Hübscher, U., Kaspárková, J., Brabec, V., and Naegeli, H. (2001). Double-check probing of DNA bending and unwinding by XPA-RPA: an architectural function in DNA repair. The EMBO Journal *20*, 3554–3564.

Moreno-Herrero, F., Colchero, J., and Baró, A.M. (2003). DNA height in scanning force microscopy. Ultramicroscopy *96*, 167–174.

Moreno-Herrero, F., Seidel, R., Johnson, S.M., Fire, A., and Dekker, N.H. (2006). Structural analysis of hyperperiodic DNA from *Caenorhabditis elegans*. Nucleic Acids Research *34*, 3057–3066.

Morris, P. (1993). The breakout method for escaping from local minima. In Proceedings of the Eleventh National Conference on Artificial Intelligence, (AAAI Press), pp. 40–45.

Mou, J., Czajkowsky, D.M., Zhang, Y., and Shao, Z. (1995). High-resolution atomic-force microscopy of DNA: the pitch of the double helix. FEBS Letters *371*, 279–282.

Murray, M.N., Hansma, H.G., Bezanilla, M., Sano, T., Ogletree, D.F., Kolbe, W., Smith, C.L., Cantor, C.R., Spengler, S., and Hansma, P.K. (1993). Atomic force microscopy of biochemically tagged DNA. Proceedings of the National Academy of Sciences of the United States of America *90*, 3811–3814.

Muzard, G., Théveny, B., and Révet, B. (1990). Electron microscopy mapping of pBR322 DNA curvature. Comparison with theoretical models. The EMBO Journal *9*, 1289–1298.

Nair, T.M. (2010). Sequence periodicity in nucleosomal DNA and intrinsic curvature. BMC Structural Biology *10 Suppl 1*, S8.

Neish, C.S., Martin, I.L., Henderson, R.M., and Edwardson, J.M. (2002). Direct visualization of ligand-protein interactions using atomic force microscopy. British Journal of Pharmacology *135*, 1943–1950.

Nečas, D., and Klapetek, P. (2011). Gwyddion: an open-source software for SPM data analysis. Central European Journal of Physics *10*, 181–188.

Van Noort, J., Verbrugge, S., Goosen, N., Dekker, C., and Dame, R.T. (2004). Dual architectural roles of HU: Formation of flexible hinges and rigid filaments. Proceedings of the National Academy of Sciences of the United States of America *101*, 6969–6974.

Noort van, S.J.T., Werf van der, K.O., Grooth de, B.G., Hulst van, N.F., and Greve, J. (1997). Height anomalies in tapping mode atomic force microscopy in air caused by adhesion. Ultramicroscopy *69*, 117–127.

Ohyama, T. (2005). Curved DNA and Transcription in Eukaryotes. In DNA Conformation and Transcription, T. Ohyama, ed. (Landes Bioscience).

Oliveira Brett, A.M., and Chiorcea Paquim, A.M. (2005). DNA imaged on a HOPG electrode surface by AFM with controlled potential. Bioelectrochemistry *66*, 117–124.

Olivier, M., Hollstein, M., and Hainaut, P. (2010). *TP53* mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harbor Perspectives in Biology *2*, 1–17.

Olson, W.K., Gorin, A.A., Lu, X.-J., Hock, L.M., and Zhurkin, V.B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. Proceedings of the National Academy of Sciences of the United States of America *95*, 11163–11168.

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions On Systems Man And Cybernetics *9*, 62–66.

Pastré, D., Piétrement, O., Zozime, A., and Le Cam, E. (2005). Study of the DNA/ethidium bromide interactions on mica surface by atomic force microscope: influence of the surface friction. Biopolymers *77*, 53–62.

Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S. V, Hainaut, P., and Olivier, M. (2007). Impact of mutant p53 functional properties on *TP53* mutation patterns and tumor phenotype: lessons from recent developments in the IARC *TP53* database. Human Mutation *28*, 622–629.

Pfeifer, G.P., Denissenko, M.F., Olivier, M., Tretyakova, N., Hecht, S.S., and Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene *21*, 7435–7451.

Pietrasanta, L.I., Smith, B.L., and MacLeod, M.C. (2000). A Novel Approach for Analyzing the Structure of DNA Modified by Benzo[a]pyrene Diol Epoxide at Single-Molecule Resolution. Chemical Research in Toxicology *13*, 351–355.

Podestà, A., Indrieri, M., Brogioli, D., Manning, G.S., Milani, P., Guerra, R., Finzi, L., and Dunlap, D. (2005). Positively charged surfaces increase the flexibility of DNA. Biophysical Journal *89*, 2558–2563.

Pope, L.H., Davies, M.C., Laughton, C.A., Roberts, C.J., Tendler, S.J., and Williams, P.M. (2000). Atomic force microscopy studies of intercalation-induced changes in plasmid DNA tertiary structure. Journal of Microscopy *199*, 68–78.

Raney, V.M., Harris, T.M., and Stone, M.P. (1993). DNA conformation mediates aflatoxin B1-DNA binding and the formation of guanine N7 adducts by aflatoxin B1 8,9-exo-epoxide. Chemical Research in Toxicology *6*, 64–68.

Reed, J., Mishra, B., Pittenger, B., Magonov, S., Troke, J., Teitell, M.A., and Gimzewski, J.K. (2007). Single molecule transcription profiling with AFM. Nanotechnology *18*, 1–15.

Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. Nature *423*, 145–150.

Rigel, D.S. (2008). Cutaneous ultraviolet exposure and its relationship to the development of skin cancer. Journal of the American Academy of Dermatology *58*, 129–132.

Rivetti, C., and Codeluppi, S. (2001). Accurate length determination of DNA molecules visualized by atomic force microscopy: evidence for a partial B- to A-form transition on mica. Ultramicroscopy *87*, 55–66.

Rivetti, C., Guthold, M., and Bustamante, C. (1996). Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. Journal of Molecular Biology *264*, 919–932.

Rivetti, C., Walker, C., and Bustamante, C. (1998). Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. Journal of Molecular Biology *280*, 41–59.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annual Review of Biochemistry *79*, 233–269.

Saenger, W. (1984). Principles of Nucleic Acid Structure (Springer Advanced Texts in Chemistry).

Sampaolese, B., Bergia, A., Scipioni, A., Zuccheri, G., Savino, M., Samori, B., and De Santis, P. (2002). Recognition of the DNA sequence by an inorganic crystal surface. Proceedings of the National Academy of Sciences of the United States of America *99*, 13566–13570.

Sanchez-Sevilla, A., Thimonier, J., Marilley, M., Rocca-Serra, J., and Barbet, J. (2002). Accuracy of AFM measurements of the contour length of DNA fragments adsorbed on mica in air and in aqueous buffer. Ultramicroscopy *92*, 151–158.

De Santis, P., Fuà, M., Savino, M., Anselmi, C., and Bocchinfuso, G. (1996). Sequence Dependent Circularization of DNAs: A Physical Model to Predict the DNA Sequence Dependent Propensity to Circularization and Its Changes in the Presence of Protein-Induced Bending. The Journal of Physical Chemistry *100*, 9968–9976.

De Santis, P., Palleschi, A., Savino, M., and Scipioni, A. (1988). A theoretical model of DNA curvature. Biophysical Chemistry *32*, 305–317.

Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. Journal of Molecular Biology *191*, 659–675.

Schitter, G., Astrom, K.J., DeMartini, B.E., Thurner, P.J., Turner, K.L., and Hansma, P.K. (2007). Design and Modeling of a High-Speed AFM-Scanner. IEEE Transactions on Control Systems Technology *15*, 906–915.

Scholkopf, B., Smolam, A.J., and Muller, K.R. (2005). Principle Components Analysis. In Lecture Notes in Computer Science, pp. 583–588.

Scipioni, A., Anselmi, C., Zuccheri, G., Samori, B., and De Santis, P. (2002a). Sequence-dependent DNA curvature and flexibility from scanning force microscopy images. Biophysical Journal *83*, 2408–2418.

Scipioni, A., Zuccheri, G., Anselmi, C., Bergia, A., Samorì, B., and De Santis, P. (2002b). Sequence-dependent DNA dynamics by scanning force microscopy time-resolved imaging. Chemistry & Biology *9*, 1315–1321.

Selsing, E., Wells, R.D., Alden, C.J., and Arnott, S. (1979). Bent DNA: visualization of a base-paired and stacked A-B conformational junction. The Journal of Biological Chemistry *254*, 5417–5422.

Seong, G.H., Yanagida, Y., Aizawa, M., and Kobatake, E. (2002). Atomic force microscopy identification of transcription factor NFkappaB bound to streptavidin-pin-holding DNA probe. Analytical Biochemistry *309*, 241–247.

Shaiu, W.L., Larson, D.D., Vesenka, J., and Henderson, E. (1993). Atomic force microscopy of oriented linear DNA molecules labeled with 5nm gold spheres. Nucleic Acids Research *21*, 99–103.

Shapiro, S.S., and Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika *52*, 591–611.

Shimada, J., and Yamakawa, H. (1984). Ring-closure probabilities for twisted wormlike chains. Application to DNA. Macromolecules *17*, 689–698.

Shlyakhtenko, L.S., Lushnikov, A.Y., and Lyubchenko, Y.L. (2009). Dynamics of nucleosomes revealed by time-lapse atomic force microscopy. Biochemistry *48*, 7842–7848.

Shore, D., and Baldwin, R.L. (1983). Energetics of DNA twisting. I. Relation between twist and cyclization probability. Journal of Molecular Biology *170*, 957–981.

Shpigelman, E.S., Trifonov, E.N., and Bolshoy, A. (1993). CURVATURE: software for the analysis of curved DNA. Computer Applications in the Biosciences *9*, 435–440.

Shrader, T.E., and Crothers, D.M. (1990). Effects of DNA sequence and histone-histone interactions on nucleosome placement. Journal of Molecular Biology *216*, 69–84.

Sobel, E.S., and Harpst, J.A. (1991). Effects of Na2+ on the persistence length and excluded volume of T7 bacteriophage DNA. Biopolymers *31*, 1559–1564.

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology *100*, 441–471.

Spisz, T.S., D'Costa, N., Seymour, C.K., Hoh, J.H., Reeves, R., and Bankman, I.N. (1997). Length determination of DNA fragments in atomic force microscope images. Proceedings of International Conference on Image Processing *3*, 154–157.

Spisz, T.S., Fang, Y., Reeves, R.H., Seymour, C.K., Bankman, I.N., and Hoh, J.H. (1998). Automated sizing of DNA fragments in atomic force microscope images. Medical & Biological Engineering & Computing *36*, 667–672.

Strahs, D., and Schlick, T. (2000). A-Tract bending: insights into experimental structures by computational models. Journal of Molecular Biology *301*, 643–663.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell *98*, 1–4.

Subashini, P., and Bharathi, P.T. (2011). Automatic Noise Identification in Images using Statistical Features. International Journal for Computer Science and Technology *2*, 467–471.

Suck, D. (1994). DNA recognition by DNase I. Journal of Molecular Recognition *7*, 65–70.

Sun, H.B., Qian, L., and Yokota, H. (2001). Detection of abasic sites on individual DNA molecules using atomic force microscopy. Analytical Chemistry *73*, 2229–2232.

Sundstrom, A. (2008). Measuring biomolecules: an image processing and length estimation pipeline using atomic force microscopy to measure DNA and RNA with high precision. Master's Thesis. New York University. Available at: <http://cs.nyu.edu/~aes/publish/thesis/aes.thesis.pdf>

Surrallés, J., Ramírez, M.J., Marcos, R., Natarajan, A.T., and Mullenders, L.H.F. (2002). Clusters of transcription-coupled repair in the human genome. Proceedings of the National Academy of Sciences of the United States of America *99*, 10571–10574.

Sushko, M.L., Shluger, A.L., and Rivetti, C. (2006). Simple model for DNA adsorption onto a mica surface in 1:1 and 2:1 electrolyte solutions. Langmuir: The ACS Journal Of Surfaces And Colloids *22*, 7678–7688.

Suzuki, Y., Yoshikawa, Y., Yoshimura, S.H., Yoshikawa, K., and Takeyasu, K. (2011). Unraveling DNA dynamics using atomic force microscopy. Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology.

Thomson, N.H., Kasas, S., Hansma, H.G., and Hansma, P.K. (1996). Reversible Binding of DNA to Mica for AFM Imaging. Langmuir *12*, 5905–5908.

Thundat, T. (1992). Atomic force microscopy of deoxyribonucleic acid strands adsorbed on mica: The effect of humidity on apparent width and image contrast. Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films *10*, 630.

Timoshenko, S., and Goodier, J.N. (1986). Theory of Elasticity. Journal of Elasticity *49*, 427–143.

Tiner, W.J., Potaman, V.N., Sinden, R.R., and Lyubchenko, Y.L. (2001). The structure of intramolecular triplex DNA: atomic force microscopy study. Journal of Molecular Biology *314*, 353–357.

Tolstorukov, M.Y., Choudhary, V., Olson, W.K., Zhurkin, V.B., and Park, P.J. (2008). NuScore: a web-interface for nucleosome positioning predictions. Bioinformatics *24*, 1456–1458.

Tornaletti, S., and Pfeifer, G. (1994). Slow repair of pyrimidine dimers at p53 mutation hotspots in skin cancer. Science *263*, 1436–1438.

Travers, A.A. (2004). The structural basis of DNA flexibility. Philosophical Transactions of the Royal Society - Series A: Mathematical, Physical and Engineering Sciences *362*, 1423–1438.

Trifonov, E.N., and Sussman, J.L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. Proceedings of the National Academy of Sciences of the United States of America *77*, 3816–3820.

Ulanovsky, L.E., and Trifonov, E.N. (1987). Estimation of wedge components in curved DNA. Nature *326*, 720–722.

Vesenka, J., Manne, S., Yang, G., Bustamante, C.J., and Henderson, E. (1993). Humidity effects on atomic force microscopy of gold-labeled DNA on mica. Scanning Microscopy *7*, 781–788.

Villarrubia, J.S. (1997). Algorithms for Scanned Probe Microscope Image Simulation, Surface Reconstruction and Tip Estimation. Journal Of Research Of The National Institute Of Standards And Technology *102*, 425–454.

Vinogradov, A.E. (2003). DNA helix: the importance of being GC-rich. Nucleic Acids Research *31*, 1838–1844.

Virstedt, J., Berge, T., Henderson, R.M., Waring, M.J., and Travers, A.A. (2004). The influence of DNA stiffness upon nucleosome formation. Journal of Structural Biology *148*, 66–85.

Vlahovicek, K., and Pongor, S. (2000). Model.it: building three dimensional DNA models from sequence data. Bioinformatics *16*, 1044–1045.

Vogelstein, B., and Kinzler, K.W. (1992). p53 function and dysfunction. Cell *70*, 523–526.

Van Vugt, J.J., De Jager, M., Murawska, M., Brehm, A., Van Noort, J., and Logie, C. (2009). Multiple aspects of ATP-dependent nucleosome translocation by RSC and Mi-2 are directed by the underlying DNA sequence. PLOS One *4*, 1-14.

Wang, H., Yang, Y., and Erie, D.A. (2007). Characterization of Protein–Protein Interactions Using Atomic Force Microscopy. In Protein Interactions, P. Schuck, ed. (Boston, MA: Springer US), pp. 39–77.

Wang, H., Yang, Y., Schofield, M.J., Du, C., Fridman, Y., Lee, S.D., Larson, E.D., Drummond, J.T., Alani, E., Hsieh, P., *et al.* (2003). DNA bending and unbending by MutS govern mismatch recognition and specificity. Proceedings of the National Academy of Sciences of the United States of America *100*, 14822–14827.

Wang, J.C. (1979). Helical repeat of DNA in solution. Proceedings of the National Academy of Sciences of the United States of America *76*, 200–203.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weber, P., Ohlendorf, D., Wendoloski, J., and Salemme, F. (1989). Structural origins of high-affinity biotin binding to streptavidin. Science *243*, 85–88.

Whibley, C., Pharoah, P.D.P., and Hollstein, M. (2009). p53 polymorphisms: cancer implications. Nature Reviews. Cancer *9*, 95–107.

Wiggins, P., Van der Heijden, T., Moreno-Herrero, F., Spakowitz, A., Phillips, R., Widom, J., Dekker, C., and Nelson, P.C. (2006). High flexibility of DNA on short length scales probed by atomic force microscopy. Nature Nanotechnology *1*, 137–141.

Woolley, A.T., Guillemette, C., Li Cheung, C., Housman, D.E., and Lieber, C.M. (2000). Direct haplotyping of kilobase-size DNA using carbon nanotube probes. Nature Biotechnology *18*, 760–763.

Wu, H.M., and Crothers, D.M. (1984). The locus of sequence-directed and protein-induced DNA bending. Nature *308*, 509–513.

Xi, L., Fondufe-Mittendorf, Y., Xia, L., Flatow, J., Widom, J., and Wang, J.-P. (2010). Predicting nucleosome positioning using a duration Hidden Markov Model. BMC Bioinformatics *11*, 346.

Xiong, Y., and Muttaiya, S. (2001). Protein – Nucleic Acid Interaction: Major Groove Recognition Determinants. Encyclopedia of Life Sciences 1–8.

Yaneva, M., Kowalewski, T., and Lieber, M.R. (1997). Interaction of DNA-dependent protein kinase with DNA and with Ku: biochemical and atomic-force microscopy studies. The EMBO Journal *16*, 5098–5112.

Yang, C.W., Hwang, I.S., Chen, Y.F., Chang, C.S., and Tsai, D.P. (2007). Imaging of soft matter with tapping-mode atomic force microscopy and non-contact-mode atomic force microscopy. Nanotechnology *18*, 1–8.

Yoshimura, S.H., Ohniwa, R.L., Sato, M.H., Matsunaga, F., Kobayashi, G., Uga, H., Wada, C., and Takeyasu, K. (2000). DNA Phase Transition Promoted by Replication Initiator. Biochemistry *39*, 9139–9145.

Young, M.A., Srinivasan, J., Goljer, I., Kumar, S., Beveridge, D.L., and Bolton, P.H. (1995). Structure determination and analysis of local bending in an A-tract DNA duplex: comparison of results from crystallography, nuclear magnetic resonance, and molecular dynamics simulation on d(CGCAAAAATGCG). Methods in Enzymology *261*, 121–144.

Zhang, H., Yu, H., Ren, J., and Qu, X. (2006). Reversible B/Z-DNA transition under the low salt condition and non-B-form polydApolydT selectivity by a cubane-like europium-L-aspartic acid complex. Biophysical Journal *90*, 3203–3207.

Zhang, T.Y., and Suen, C.Y. (1984). A fast parallel algorithm for thinning digital patterns. Communications of the ACM *27*, 236–239.

Zheng, G., Lu, X.-J., and Olson, W.K. (2009). Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. Nucleic Acids Research *37*, 240–246.

Zhu, Q. (2000). Decreased DNA Repair Efficiency by Loss or Disruption of p53 Function Preferentially Affects Removal of Cyclobutane Pyrimidine Dimers from Non-transcribed Strand and Slow Repair Sites in Transcribed Strand. Journal of Biological Chemistry *275*, 11492–11497.

Zinkel, S.S., and Crothers, D.M. (1990). Comparative gel electrophoresis measurement of the DNA bend angle induced by the catabolite activator protein. Biopolymers *29*, 29–38.

Zuccheri, G., Bergia, A., Gallinella, G., Musiani, M., and Samorì, B. (2001a). Scanning force microscopy study on a single-stranded DNA: the genome of parvovirus B19. European Journal Of Chemical Biology *2*, 199–204.

Zuccheri, G., Scipioni, A., Cavaliere, V., Gargiulo, G., De Santis, P., and Samorì, B. (2001b). Mapping the intrinsic curvature and flexibility along the DNA chain. Proceedings of the National Academy of Sciences of the United States of America *98*, 3074–3079.

Zuo, X., Cui, G., Merz, K.M., Zhang, L., Lewis, F.D., and Tiede, D.M. (2006). X-ray diffraction "fingerprinting" of DNA structure in solution for quantitative evaluation of molecular dynamics simulation. Proceedings of the National Academy of Sciences of the United States of America *103*, 3534–3539.