Characterisation of polymeric biomacromolecules using linear dichroism and Markov chain Monte Carlo

Martyn Rittman

December 2008
Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Warwick

Acknowledgments

Undertaking a PhD thesis is a long and collaborative process. Only one name appears on the cover, but the truth is that many have influenced the work contained in the following pages. Contributions range from an encouraging word and a listening ear to wise advice on some technical aspect of the work. Inevitably some have been missed, but I hope that the names listed below go a long way towards expressing my gratitude to those who have helped me along the way.

First of all I would like to thank those who have supervised me. Alison Rodger's vision for MOAC, personal supervision and prompt proof-reading have been invaluable. Matthew Hicks' wealth of knowledge and suggestions on all sorts of subjects (relevant and otherwise) has also shaped and guided the work presented here. Bärbel Finkenstadt's guidance through the statistical elements of this work was an immense help.

Many others have offered support on scientific content. In particular, Alex Connor for help on mutagenesis of DNA; Søren Vrønning Hoffmann, the beamline scientist for the UV1 beamline in Aarhus, Denmark; Luca Sbano, for initiating the modelling of restriction enzymes and FtsZ and a member of my advisory committee; Tim Dafforn for being ubiquitously present at Warwick, Birmingham and Aarhus; Matthew Turner and David Roper, the other members of my advisory committee who offered excellent advice throughout; Florian Theil for a short conversation in which he introduced the possibility of sharp bends in DNA during Couette flow.

I would like to thank current and former members of Alison Rodger's group in chemistry, for support and friendship throughout. Especial thanks go to Nigel Dyer for developing the injection cell, Emma Gilroy for DNA experimentation, Lahari de Alwis for sharing data on a coarse-grained model, Adair Richards for many hours grappling with the concepts of persistence length, Raul Pacheco-Gomez for the data and inspiration behind the work on FtsZ and by no means least, Angeliki Damianoglou for coping with me for endless hours and bringing a very Greek sense of fun to the lab!

The writing process benefited immensely from the course run by Martin

Pumphrey whose ability to read the incomprehensible is admirable. My gratitude also goes to Ben Smith — the meetings through my last few months at Warwick helped keep me focused on the task in hand. Many thanks for those who have proof-read portions of this work.

On a personal level, I would like to thank my friends for constant encouragement and wondering where I was in the latter half of 2008. Adam Hardy earns himself a pint for knowing that I would do a PhD before I did. My eternal gratitude goes to my family: Tim, Sarah and the APs for simply being the best support anyone could wish for. Also, my thanks and praise to God - a constant source of hope. Mes remerciements à Céline qui, pour les derniers mois, a partagé avec moi le stress, les moments de succès et de déception et toujours avec encouragement et un sourire.

This work was kindly funded by the EPSRC through the MOAC doctoral training centre.

Contents

Chapter 1

Introduction

1.1 DNA

Deoxyribonucleic acid, DNA, is a molecule of great biological importance and physical and chemical interest. Sometimes termed the 'molecule of life' the DNA contained in a cell encodes all biological aspects of that cell. In this section DNA is described from an atomic level to its secondary structure and interaction with proteins and other molecules.

1.1.1 The molecular structure of DNA

[106, 20] Figure 1.1 shows the four bases of DNA. They consist of carbon, hydrogen, nitrogen, oxygen and phosphorous atoms arranged as a combination of a nitrogenous base, a ribose sugar and a phosphate group, known collectively as a nucleotide. The base in adenine and guanine is purine, whilst in thymine and cytosine it is pyrimidine. The phosphate groups can polymerise to form a chain of nucleotides bases.

Pairs of nucleotide chains can hydrogen bond between bases. In the vast majority of cases adenine and thymine are paired, as are guanine and cytosine. The former can form two hydrogen bonds and the latter three, as in figure 1.2. In polymer form the lowest energy state of DNA is a helix. Generally this is the B helix, as shown, but two other helices (known as Z and A) can also form. A and B are right-handed helices and Z is left-handed. In all of these forms the two strands run in opposite directions. Note that DNA is not symmetrical; by convention polymer sequences are listed in the 5'→3' direction. The B helix has a diameter of 20 Å and the structure repeats every 10 nucleotides, 34 Å.

Ribonucleic acid, RNA, is closely related to DNA, but contains an additional 2'-hydroxyl group in the sugar group. It is less stable than DNA, but used by the cell primarily for the translation of genes. The bases are the same,

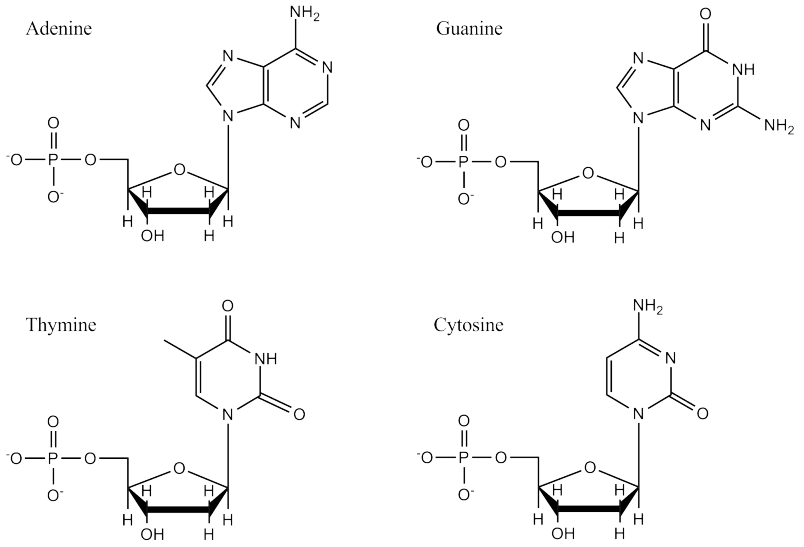


Figure 1.1: Structures of the four DNA bases

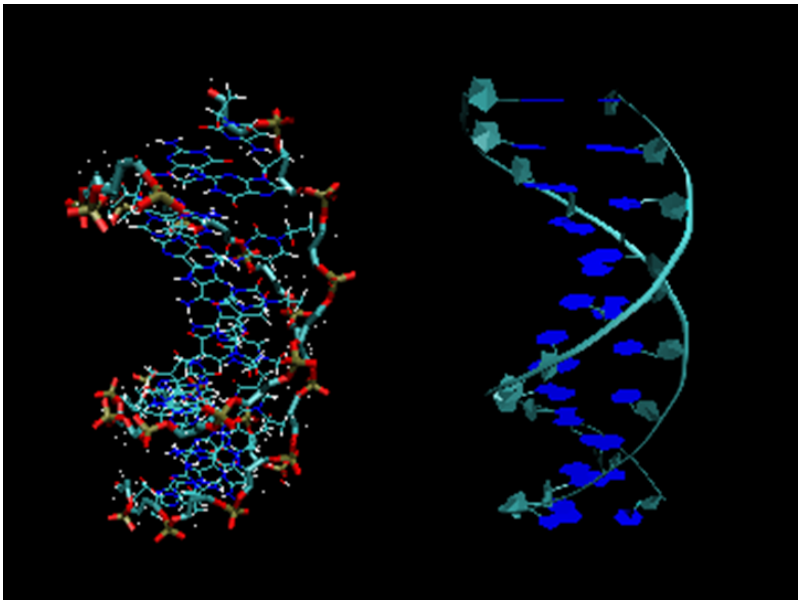


Figure 1.2: The DNA B helix. The molecular structure on the left is tilted forward compared to the cartoon on the right.

with the exception of thymine which is replaced by uracil.

1.1.2 Biological significance of DNA

[20] DNA is one of life's most important molecules. It lies at the core of the so-called central dogma of molecular biology. This is the process describing how DNA is a code for making proteins. A sequence of 3 DNA bases, termed a codon, encodes for one amino acid, with the exception of 'stop' codons that halt the interpretation process.

To express a protein, DNA is transcribed into the complementary RNA sequence, called messenger RNA (mRNA). The mRNA is exported from the nucleus (in eukaryotic cells) and binds to a ribosome protein, containing series of transfer RNA 3-mers (tRNA) matching the original DNA codons. The tRNA is bound to an amino acid and successive amino acids are joined together to create the desired protein, a process called translation. This is a highly complex and regulated process and the description here is brief. For example, transcription in eukaryotes takes place in the context of the histone, described in the next section. The process requires an efficient search method for finding the DNA sequence of any given protein. There are also processes for proof-reading and repairing DNA and copying the entire DNA chromosome (some 3 billion base pairs in humans) during cell replication.

1.1.3 DNA *in vivo* structures

In vivo there are a number of different structures in which DNA is found. Among the most interesting are the chromosomes of eukaryotic nuclei [107], packages found in viral capsids [31] and circular super-coiled plasmids in prokaryotic cell cytoplasm.

In a chromosome DNA is wrapped around octamers of histone proteins with short linking sequences between. The histones are positively charged, in contrast to DNA. Groups of histones assemble into a structure with a packing ratio of the order of 10^4 (this is the ratio of the length of the fully extended DNA to the length of the packed DNA). Many questions surround the packing of DNA in nuclei, ranging from the precise structure of packing with histones to how the correct DNA sequence is found for replication.

Virions, virus particles, consist of a protein coat which forms the capsid. This contains DNA or RNA which, like the histone, is very efficiently packed.

Plasmids are circular pieces of DNA, typically consisting of several thousand base pairs. They are widely used in laboratories for recombinant DNA work, expressing custom-made proteins. *In vivo* they exist in the cytoplasm

of bacteria where they can express proteins that give the cell properties such as antibiotic resistance. Prokaryotes are able to secrete and take up plasmids from their environment and this is one method by which they achieve genetic diversity.

Plasmids are often depicted in circles, but this is not the structure they maintain. Normally they are super-coiled, that is, the ends are twisted in opposite directions before being joined together. In topological terms the linking number of the two DNA strands is usually reduced compared with that of simply joining the ends together from the helix form (by a purely bending deformation). In this case the plasmids are said to be negatively super-coiled (a right-handed helix has positive linking number). Notable exceptions are thermophilic bacteria, some of which contain the protein reverse gyrase that introduces positive super-coiling above 55°C.

1.1.4 DNA persistence length

Of great interest for *in vivo* behaviour of DNA is its bending behaviour. This is discussed in much more detail in chapter 3. The quantity persistence length incorporates both flexibility and curvature. Flexibility is a dynamic property relating to how a material moves in solution due to hydrodynamic fluctuations, steric hindrance and other similar effects that define motion. In this context curvature is the static property that reflects the equilibrium, or lowest energy state, of the material. For example, certain DNA base motifs such as AAAATTTT introduce bends that increase curvature.

1.2 Absorbance spectroscopy

The main analytical technique used in the work presented here is linear dichroism (LD), a form of absorbance spectroscopy which can be used to gather bulk information about the structure and conformation of biomolecules. In this section the relevant theoretical and practical aspects of absorption spectroscopy are given with a particular focus on LD, followed by several examples of the use of LD.

LD is able to determine secondary structure of proteins and differentiate structural forms of DNA. In particular, it provides information on the orientation of structural motifs within molecules. For example, when DNA is in a super-coiled plasmid form it orients weakly in shear flow. This is easily differentiated from the linear form of DNA which more readily orients under shear flow conditions. Further examples are given below.

Circular dichroism (CD) is a closely related technique to LD, and more widely used. It detects chirality in molecules and is primarily used to probe secondary structure motifs, such as alpha helix, beta sheet and random coil in proteins. It is also useful for monitoring changes of molecules in solution under different conditions (e.g. a change in temperature or pH) and can detect DNA helical structure and melting transitions.

1.2.1 Molecular absorption spectroscopy

Many atoms absorb light. In other words, electromagnetic radiation is capable of exciting electrons within atoms and molecules to higher energy levels. This phenomenon is the focus of absorption spectroscopy. Measurement of the incident and transmitted radiation for a sample allows calculation of the radiation absorbed and gives information on its electronic make-up, reflecting the overall molecular structure. In this section the term 'light' refers to any electromagnetic radiation, most commonly ultra-violet.

For absorbance the main law governing measurement is the Beer-Lambert law. This is derived from the Beer law, which states that the change in power dP of light absorbed is proportional to the number of molecules present and the power of the light [37]

$$dP = -\epsilon' C P dx \tag{1.1}$$

where ϵ' is a constant of proportionality and dx is an infinitesimal distance through the sample, the total distance being x. C is molecule concentration. Integrating this and recognising that power and intensity I are equivalent quantities gives the more widely used Beer-Lambert law

$$A = \log_{10} \frac{I}{I_0} = \epsilon C x \tag{1.2}$$

where I_0 is intensity before the sample.

In the above it has been assumed that the molecules are randomly oriented, or equivalently (but erroneously) that they absorb equally in all directions. To take a more accurate view it is necessary to look more closely at individual transitions. We introduce the concept of a chromophore, defined as a region (e.g. of a molecule) that can be considered to absorb photons of a specific energy and can thus be represented by a single electric dipole. This is not a totally rigorous definition but in practice what constitutes a chromophore

is generally clear and the definition is sufficient. Note that the energy of absorbance bands due to a chromophore are experimentally not at a single wavelength, they are broadened due to environmental factors in solution.

1.2.2 Linear dichroism

Linear dichroism (LD) exploits the directionality of absorbance to provide structural and orientation information about the molecule. If one can orient a chromophore, with corresponding electric dipole $\boldsymbol{\mu}$, then absorption will be different in each of the three spatial dimensions. If $\hat{\mathbf{x}}$ is a unit vector, the absorption in that direction is given by

$$A = (\boldsymbol{\mu}.\hat{\mathbf{x}})^2 \tag{1.3}$$

Linear dichroism is defined as the difference between absorbance with polarisations parallel $A_{||}$ and perpendicular $A_{\perp}$ to the direction of molecular orientation,

$$LD = A_{||} - A_{\perp}. \tag{1.4}$$

A closely related quantity termed reduced linear dichroism LD^r is given by

$$LD^r = \frac{A_{||} - A_{\perp}}{A} \tag{1.5}$$

where A is isotropic absorbance.

Using equation (1.3) parallel to the orientation axis the absorption, $A_{||}$ is the square of the dot product of the chromophore with the unit vector in that direction. For the perpendicular direction the absorption $A_{\perp}$ is given by a rotational average around the short axis of the molecule, see figure 1.3. The rotational average arises since molecules are free to rotate about the axis of orientation. Thus

$$A_{||} = |\boldsymbol{\mu}.\hat{\mathbf{p}}_{||}|^2 \tag{1.6}$$

$$A_{\perp} = \frac{1}{2\pi}\int_0^{2\pi} |\hat{\mathbf{p}}_{\perp}.rot(\boldsymbol{\mu};\theta,\mathbf{p}_{\perp})|^2 d\theta \tag{1.7}$$

where $rot(\mathbf{v};\gamma,\mathbf{V})$ is rotation of the vector $\mathbf{v}$ by an angle γ about the axis $\mathbf{V}$.

$\hat{\mathbf{p}}_{||}$ and $\hat{\mathbf{p}}_{\perp}$ refer to unit vectors in the two directions of polarisation.

The above equations can be rewritten using the angle between the chromophore axis and axis of orientation α

$$A_{||} = \mu^2 \cos^2 \alpha \tag{1.8}$$

$$A_{\perp} = \frac{1}{2\pi} \int_0^{2\pi} |\mu \sin \alpha \cos \theta|^2 d\theta \tag{1.9}$$

and completing the integration

$$A_{\perp} = \frac{1}{2\pi} 2\mu \sin \alpha \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2}(1 + \cos 2\theta) d\theta \tag{1.10}$$

$$= \frac{\mu \sin^2 \alpha}{2}. \tag{1.11}$$

Upon substituting into equation (1.5) we obtain the reduced LD

$$LD^r = \frac{1}{\mu^2}(\mu^2 \cos^2 \alpha - \frac{\mu^2 \sin^2 \alpha}{2}) \tag{1.12}$$

$$= \frac{1}{2}(3 \cos^2 \alpha - 1). \tag{1.13}$$

Whichever method of orientation is used (as discussed below) the fraction of molecules fully orientated S is very rarely unity. S is termed the orientation parameter and is influenced by a number of considerations in a typical experimental setup as well as being dependent on the orientation method used. This is further discussed in section 1.2.4. LD^r is thus

$$LD^r = \frac{3S}{2}(3 \cos^2 \langle \alpha \rangle - 1) \tag{1.14}$$

Note that LD^r is independent of the chromophore absorption strength and concentration. The angled brackets on α indicate an averaging over possible states and along the length of the molecule, as defined in equation (3.5).

1.2.3 Experimental setup

The configuration of a linear dichroism spectrometer is much the same as for circular dichroism (figure 1.4). Light is generated from a source, which in the lab is typically xenon or mercury-xenon. For low wavelength and high

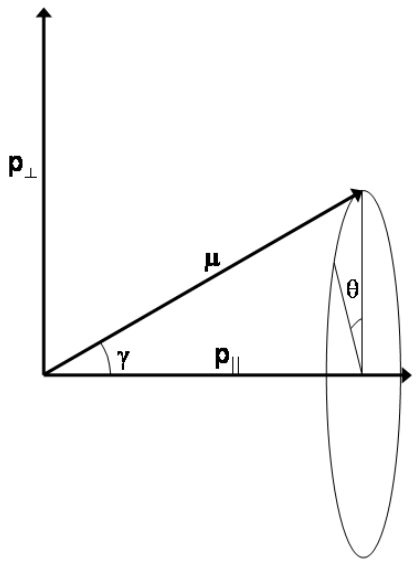


Figure 1.3: The relative positions of vectors for LD. This shows a chromophore with electric dipole $\boldsymbol{\mu}$ at an angle γ to the polarisation vector $\mathbf{p}_{||}$. To find the absorption in the perpendicular direction $\mathbf{p}_{\perp}$ the absorption is averaged around the circle described by the angle θ.

flux a synchrotron source can be used (see chapter 2). Radiation from the source is guided through a monochromator to ensure a single wavelength of light. The monochromator is often an interference grating: a series of close-set lines that have the same effect as a prism in that they separate light, emitting different wavelengths at varying angles. The light then goes to a polarising element and a photoelastic modulator (PEM). The PEM is a crystal, such as magnesium fluoride, that alters polarisation when subject to a potential difference. In practice an alternating current is used which produces phase shifts between the two perpendicular light directions, causing rapid switching between circular and linear polarisations with a frequency of the order of 100 kHz for linear polarisations. A device called a lock-in amplifier controls when the measurements are taken. After the PEM the light (now polarised) exits the spectrometer and passes through the sample chamber. A photomultiplier tube collects the light and measures its intensity. For LD this is sufficient, since a subtraction of absorbances corresponds to log (base 10) of a ratio of intensities. If the absolute absorbance is required a reference spectrum can be subtracted. Some experimental setups vary slightly, for example the monochromator can be placed after the sample chamber.

Two spectrometers were predominantly used for the work carried out here. At the University of Warwick a Biologic (Claix, France) MOS-450 spectrometer fitted with a Crystal Precision Optics (Rugby, UK) prototype thermostat-

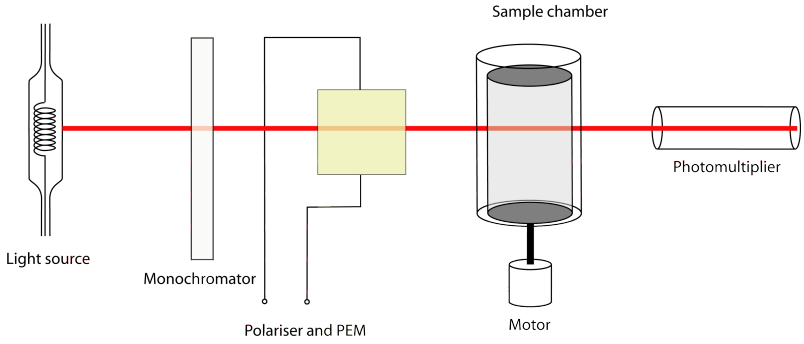


Figure 1.4: Schematic of a linear dichroism spectrometer

ted linear dichroism cell (figure 1.5) was used. At the University of Aarhus (Aarhus, Denmark) the UV1 beamline from the ASTRID synchrotron was used in conjunction with a Crystal Precision Optics linear dichroism cell. The Aarhus operational setup has been developed by Søren Hoffmann. The linear dichroism cells used have been developed jointly between Crystal Precision Optics and Alison Rodger, with the current versions based on those outlined by Marrington *et al* [70, 71]. Some of the properties of temperature control for the current prototype cells are described in chapter 5.

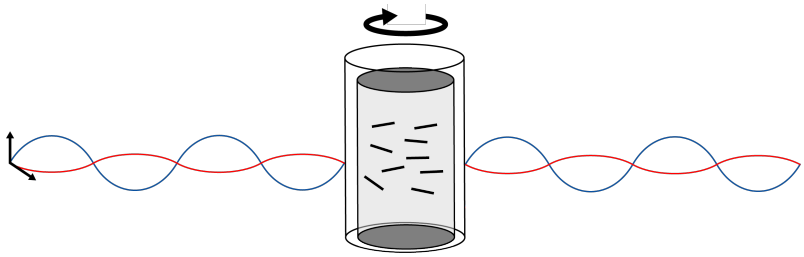


Figure 1.5: Couette flow cell. The arrows on the left indicate the directions of light polarisation. Either cylinder can be rotated, causing flow around the cylinder axes. The sample (represented by black blocks) is held in the annular gap and aligns roughly parallel to the flow direction.

Couette flow

To orient the sample, unless otherwise stated, the work and analysis here uses Couette flow, a way of generating predictable, laminar flow in fluids. The fluid is placed between two concentric cylinders and one of the cylinders is rotated (see figure 1.5). Polymers and other molecules with a high aspect ratio have been shown to align well in such a system whilst in solution. Questions remain about the precise geometric distribution of the molecules and hence the nature of the chromophore orientation distribution function which determines the orientation factor S (equation (1.14)).

To achieve Couette flow the LD cells used throughout the work here consists of a removable quartz capillary and central cylindrical quartz rod, figure 1.6. The capillary rests in a rubber holder at its base and a plastic collar at the top. The rubber base is attached to a motor that rotates. The quartz rod fits tightly into the top of the cell so as to leave the same annular gap at each point on its diameter and is stationary during measurements. The capillary has length 35 mm, internal diameter 3 mm and external diameter 5 mm. The sample is held at the bottom by an expoy resin plug. The rod is adjustable in length and is set so that there is a 1 mm gap between its end and the plug (which is variable in thickness).

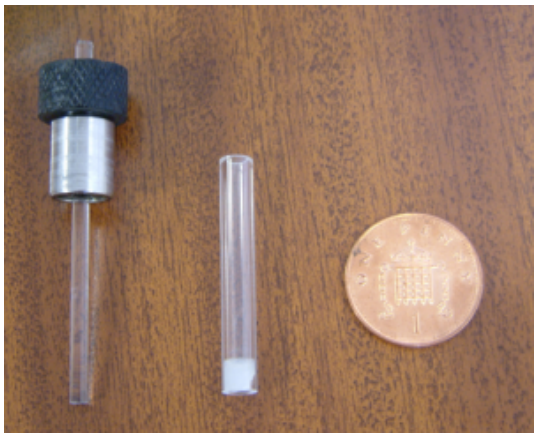


Figure 1.6: The quartz rod and capillary used for Couette flow with a 1p coin to give an idea of scale.

1.2.4 Methods of orientation

One of the key and unique considerations for linear dichroism is that the sample must be oriented. There are several ways that this can be achieved other than Couette flow, described above.

Linear flow

The equations for fluid dynamics are often complex, but two simple solutions exist. One of these is Couette flow and the other is laminar shear flow between two surfaces, both with non-slip boundary conditions. Exploiting linear shear flow has often been used to orient samples for linear dichroism [73, 10]. The setting up of such a system is relatively straightforward, although a large sample size is required, especially if there is no circular flow system (where the sample is fed back through the measurement chamber).

Intrinsic orientation

Some molecules exhibit natural orientation which can be exploited. Examples of this include crystals [54] (this example is for myglobin crystals using infrared LD), spider silk [56] and purple membrane [17, 48].

Stretched film orientation

Some molecules can be oriented by dissolving into a polymer solution (e.g. 10% poly(vinyl alcohol)) while it is cooling into a film. Once set the film is mechanically stretched which creates orientaiton, with the long axis of the molecule preferentially aligned along the axis of stretching [74, 16]. The degree of orientation is dependent on the molecule geometry.

1.2.5 Development of instrumentation

The first LD instrumentation was reported by Wada and Kozawa in 1964. Like most that have followed, it uses a circular dichroism spectrometer adapted to produce linearly polarised light along with a unit for orientation — termed a cell. Several others followed along similar principles, as summarised in table 1.1. Flow gradient, as featured in the table, is defined as

$$G = \frac{v_{out} - v_{in}}{r_{out} - r_{in}} \tag{1.15}$$

where v_{out} is linear speed of the internal surface of the capillary measured in ms^{-1}, v_{in} is the linear speed of the outer surface of the rod (equal to 0 in our case), r_{out} is the internal radius of the capillary and r_{in} is the radius of the rod.

First author	Sample size /ml	Radius of central cylinder /mm	Annular gap /mm	Maximum gradient/s^{-1}
Wada [113]	2.6	29.3	0.7	4500
Lee [60]	60 (max)	9.53	5.08	670
Oriel [84]	5.7	3.66	2.46	2000
Simonson [102]	1.8	7.25	0.5	3140
Marrington [71, 70]	0.060	1.5	0.5	3400

Table 1.1: Specifications of Couette flow LD cells. The Marrington cell is essentially the same as the cell used here. It offers a much smaller sample size compared with those previous to it, mostly due to the smaller central cylinder radius. The equation for gradient is given in equation (1.15).

1.2.6 Applications of LD

In this section several important examples are discussed.

LD is useful for studying any system with an appreciable aspect ratio (the ratio between the long and short molecular axes). Amyloid fibrils of proteins are the underlying causes of conditions such as Alzheimer's and prion diseases. Linear flow LD has been employed to study the rheology of β_2-microglobulin [1], which can be induced to form amyloid fibrils with a sufficient aspect ratio to orient in linear flow. The system used was a custom-made continuous flow system attached to a CD spectrometer adapted for LD measurements. By obtaining LD signals they were able to determine that structure exists in the fibrils and that alignment confirms previous predictions of the orientation of peptides within the fibre. The orientation factor for the fibrils was estimated to be 0.09–0.1.

A feature of LD that is often exploited is that small molecules, such as potential ligands, do not give an LD signal in Couette or linear flow since they do not possess a high aspect ratio. They will only be detected if they bind to an oriented molecule in an anisotropic manner. DNA ligands, for example, have been studied in this way [21]. One example is ethidium bromide which intercalates between DNA base-pairs. When bound it is detected in two ways, firstly as a broad peak which appears at around 520 nm and secondly by altering the magnitude of the DNA peak at 260 nm. The latter effect is due to more rigidity in the DNA which increases the orientation factor.

LD can also be used as a complementary techniques for some particularly difficult problems. Membrane proteins are one such problem [50]. Recently LD, along with circular dichroism and fluorescence spectroscopy have been used to study the kinetics of the membrance protein gramicidin inserting into liposomes (unilamellar lipid spheres). LD is sensitive to orientation, CD to overall secondary structure and fluorescence to chemical environment. Using these three together five discrete steps of gramicidin insertion were identified. Liposomes become ellipsoidal in Couette flow and thus orient. Gramicidin is too small to be detected by LD unless it is associated with a liposome. This is a great advantage as it means that LD is insensitive to protein aggregation in solution. Previous experiments with gramicidin have focused on fluorescence only, for which the process appears to have finished within 10 s. LD reveals a change in orientation over the period of a further hour. The authors are optimistic of successfully applying the same analysis to other protein–membrane systems.

1.3 Data comparison methods

[25] For most experimental data it can be surmised that it is possible to find a tractable model that produces solutions that approximate the data well. There are a number of systematic approaches that can be taken to test the suitability of models and draw inference about their parameters. Broadly, parameter fitting algorithms fall into two categories: deterministic methods and stochastic methods. The former are methods that make use of calculus to calculate parameter values whilst the latter make use of statistical convergence properties to produce an estimate of the optimum parameters and a probability distribution for each parameter.

1.3.1 Notation

For this chapter the following notation is used. Experimental data is termed $Y(t)$ and a model solution is $X(\theta, t)$ where t indexes data points (e.g. time) and θ is the vector of parameters for the model. The goal of modelling is to find a representation of the physical system with a set of parameters θ that minimises the difference between X and Y. For a more precise and descriptive explanation of how this is achieved using the likelihood function, refer to section 6.2.1. The explanations in this section are broad-brushed.

1.3.2 Deterministic methods

Deterministic methods include Newton-Raphson, steepest gradient and conjugate gradient algorithms. In general they calculate the gradient of the function for difference between $X(\theta, t)$ and $Y(t)$ with respect to θ in order to minimise the difference. For linear differential equations they are extremely efficient and can find exact solutions within two or three steps.

The advantage of these methods is that convergence to an optimum set of parameters is guaranteed (provided the function $X(\theta, t)$ is differentiable with respect to θ and either convex or concave). However, there are two major downfalls. Firstly, they do not differentiate between local and global minima. This problem can be overcome by choosing a range of starting values, although this increases the computational cost n-fold for n different starting values. Moles *et al* [78] provide a comprehensive comparison of deterministic and stochastic global optimisation methods that attempt to overcome this problem. Additionally, if there are a number of local minima it can be difficult to determine which is the global minimum simply from running the algorithm. The second major disadvantage, particularly for a complex model, is the com-

putational cost of calculating gradients. Calculation of the first differential of a function of X and Y with respect to θ increases the cost with the size of θ. Furthermore gradients (even when calculated numerically) are generally more costly than computing the solution of the undifferentiated function.

Deterministic methods have been used extensively, particularly for energy mininisation within molecular modelling problems, see for example [2], [65], [89].

1.3.3 Stochastic methods

Since the advent of the computer age stochastic algorithms have become increasingly popular. They exploit Bayesian inference, which answer the question of what can be determined about the distribution of parameters that reproduce data $X(\theta, t)$ to within a certain error. The suitability of a number of randomly sampled points in the parameter space is assessed, a method known as Monte Carlo. From these points a solution can be found. Unlike the deterministic methods discussed above, local solutions (minima or maxima) pose less of a problem, as the algorithm is able to overcome small barriers. To use the geometric analogy, if looking for a minimum a statistical method may go 'uphill' a little way and find a deeper valley on the other side, whereas a deterministic method will only ever head 'downhill'. Stochastic methods are also generally more computationally efficient than deterministic ones for large numbers of parameters, since in general only the function value at each point need be found.

There are some problems with the stochastic approach which have, as yet, no clear solution. An important one concerns convergence of the algorithm for a general function. It can be shown that for a non-linear system of differential equations, convergence cannot be absolutely guaranteed in a finite time [19]. In practice convergence can often be visually assessed from inspection of successive estimated values and values of the likelihood function, although a number of systematic convergence measures exist.

More practical problems will be discussed in the later chapter outlining the method (chapter 6). Broadly these issues are associated with pre-defined numbers that are needed to be inputted into the algorithm. For example, a number of iterations are required before the algorithm settles to a region around the solution. This is known as the burn-in and is difficult to predetermine. Other concerns are how far to move between each step and how many iterations after burn-in provide a statistically significant solution.

There are many recent examples of Monte Carlo-based methods in practice.

To choose just a few examples, in physics it has been used to look at the cosmic microwave background radiation [62], in finance for the predicting the stock market [28] and in biology it has been applied to try to understand protein folding [120].

1.3.4 Markov chain Monte Carlo

The method employed in the model fitting presented here is a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm combined with a Gibbs sampler. The reason for choosing a Monte Carlo system is that we wish to study models with a large number of parameters. Calculations of gradients for exact solutions would take a prohibitively long time. The shape of the parameter space is also unknown. Although, for the systems considered, they are expected to be mostly 'well-behaved' it is possible that some local minima exist and we wish to avoid this problem.

The Metropolis-Hastings algorithm was first developed in 1950 by Metropolis [76], and generalised in 1970 by Hastings [44]. The details of the algorithm are given in chapter 6. The Gibbs sampler is treated in the same chapter.

There is no precedent for application of MCMC to the problem of FtsZ polymerisation (see chapter 8), but it has been applied to similar systems, such as the hes1 system [47]. In this example the system was transcriptional gene regulation, which has a comparable number of estimated parameters and uses delay functions similar to those introduced in chapter 6.

1.4 FtsZ

1.4.1 Properties and behaviour

FtsZ is a vital protein for prokaryotic cell division. Much is known about cell division in eukaryotes and an ever-increasing amount for prokaryotes.

In prokaryotes, FtsZ is the first protein that localizes to the region of cell division, known as the septum or divisome. It forms a polymer ring (called the Z-ring) around the centre of the cell, dividing it into two almost equal halves. The Z ring constricts, causing the cell to invaginate and subsequently divide. FtsZ is far from the only protein involved in the process and is highly regulated. Romberg [96] has summarised current knowledge of the FtsZ mechanism and the roles of related proteins.

Much of the precise mechanism of FtsZ and its cofactors remains unknown, although some elements have been determined. A minimum concentration of

Table 1.2: Some of the proteins thought to affect formation of FtsZ polymers *in vivo*[96].

Protein name	Effect on Z ring	Summary of role
FtsA	Stabilising	Recruits other cell division proteins
ZipA	Stabilising	Recruits other cell division proteins
MinC, MinD	Destabilising	Prevent Z ring formation, particularly at cell poles
EzrA	Destabilising	Prevents FtsZ fibrilation at the early stages
SulA	Inhibition	Inhibits fibrilation and hydrolysis
Zap A	Stabilising	Promotes bundling and stabilises polymers

FtsZ is required, which has been confirmed both *in vitro* ([13], Raul Pacheco-Gomez, personal communication) and *in vivo* [61]. Super-critical concentrations are a necessary condition for *in vivo* fibrilisation but not sufficient, as regulatory proteins play a major role in determining timing and location of Z-ring formation. FtsZ requires guanine triphosphate (GTP) to polymerise. Over time GTP is hydrolysed (i.e. a phosphate is removed) to form guanine diphosphate (GDP). A summary of some of the proteins known to interact with FtsZ is given in table 1.2. This is intended to be no more than indicative, showing that a number of proteins can affect the *in vivo* formation and stability of FtsZ polymers. It should act as a caveat to any *in vitro* experiments as observations found in non-living conditions cannot be assumed to occur *in vivo*. Despite this many of the *in vivo* properties of FtsZ have been observed and measured *in vitro*, which is a much easier environment in which to work.

Monomer turnover

A noteworthy property of FtsZ is apparent turnover of monomers within the polymers. Fluorescence resonance energy transfer (FRET) [13] has revealed that the make-up of any given protofilament changes with a half-time of around 7 s. Since the average polymer length is constant, it was proposed that this occurs by replacement of intra-fibre FtsZ with free GTP-bound monomers.

FtsZ as a tubulin homologue

FtsZ is often described as a homologue of tubulin. This is true both in terms of function and structure. The sequence and x-ray structure of the FtsZ monomer and dimer has been determined for a number of bacteria. It is found to be highly conserved and the structure is analogous to that of tubulin, despite

a low sequence homology [81]. Residues at both ends of the molecule are involved in GTP hydrolysis, indicating that it can only occur in the polymer form. The make-up of the fibres is a topic of discussion. It is clear that some polymers contain GDP, since addition of GDP to monomeric FtsZ causes polymer formation. The relative rates between hydrolysis of GTP and turnover will determine what percentage of GDP is present in the polymeric form at any given time.

1.4.2 A putative *in vitro* model

For FtsZ *in vitro* it is thought that the regime outlined in figure 1.7 is accurate. The initial step of FtsZ binding to GTP is not shown. GTP bound monomers form protofilaments of lengths from 10-200 subunits dependent on concentration [12]. 6 or 7 of these collate into a Z ring. Hydrolysis is thought to be prevented by the collation of proteins such as ZapA and YgfE. When hydrolysis does occur it is though to take place at any point along the polymer. Monomer turnover occurs, as discussed above. Electron microscopy [85] has suggested a conformational change induced in polymers with large amounts of GDP present: enough to cause it to destabilise and release monomers. While GTP is available in the system the process can be cycled round. Once all of the GTP has been hydrolysed the protofilaments break down to leave only FtsZ monomers. Under certain conditions (particularly high pH), light scattering measurements have indicated that residual GDP-containing protofilaments remain.

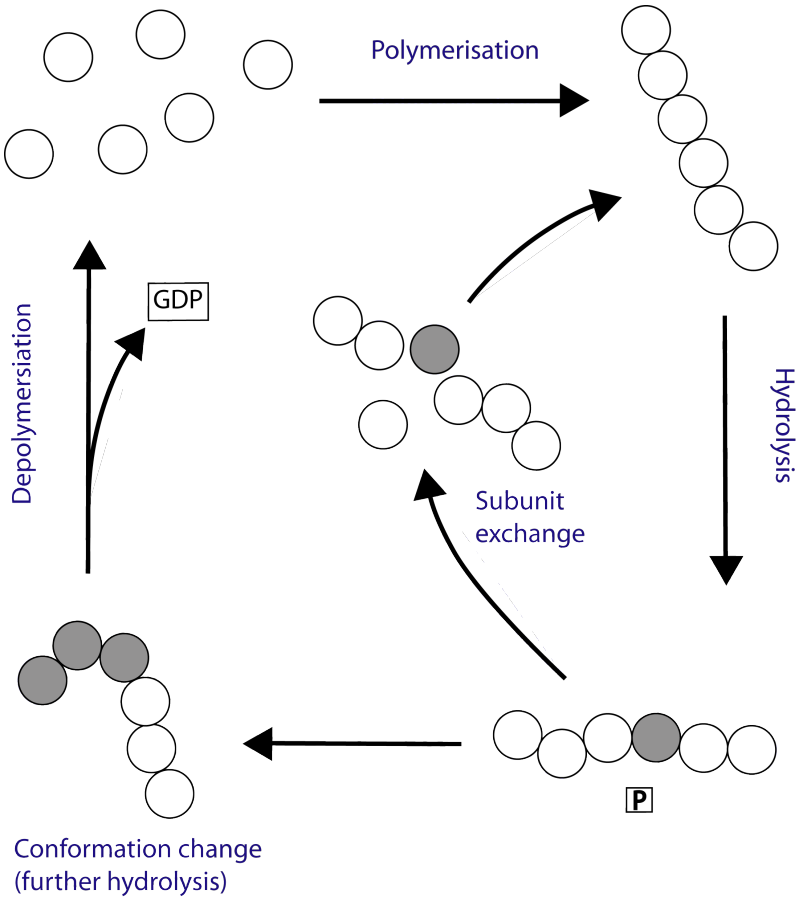


Figure 1.7: A summary of the processes thought to be involved in the action of the FtsZ protein

Chapter 2

DNA structure revealed by synchrotron radiation linear dichroism

k

2.1 Chapter summary

DNA spectra were measured using synchrotron radiation linear dichroism. For comparison, all data was measured using the UV1 beamline at ASTRID (Aarhus University, Denmark) and a Biologic MOS-450 spectrometer in the laboratory. For comparison, some data were measured using a Jasco 815 spectrometer.

Four different types of DNA were used. Contrary to expected results, DNA with G-C content of 42% and 72% exhibited a period of increasing LD signal with salt. Current theories of salt interaction with DNA do not currently predict such an effect.

For DNA with 42% G-C content and less a novel shoulder in the LD spectra was found at around 200nm. The origin of this is uncertain. In some cases the shoulder was not observed above a critical salt concentration.

The possibility of single-stranded regions within DNA at low salt concentrations was explored. Use of the DNA binding agent acridine orange was unable to confirm the presence of such regions.

2.2 Introduction

DNA is a biological molecule of great importance, as discussed in the introduction chapter. Many methods have been used to characterise its structure, but some questions still remain. Linear dichroism (LD) is a useful tool for studying polymer systems such as DNA and can provide data complementary to the other analytical methods. Recent advances in LD instrumentation allow for measurement of biologically relevant sample sizes [71]. In this chapter novel features of DNA structure are found using LD, including changes upon addition of low salt concentrations.

Methods used for previous studies of DNA include x-ray crystallography and nuclear magnetic resonance (NMR) to study short, static lengths of DNA at high resolution; circular dichroism for bulk studies of DNA determining which helical form (A,B or Z) is present and other features on a similar scale; electron microscopy for static structures and single molecule techniques such as magnetic tweezers to look at the effects of stretching DNA. Each provides different information depending on the DNA length, constraints on physical conditions (salt concentration, temperature, pressure etc.), whether the DNA is in solution, static or dynamic and so on. LD is a good complementary technique to those given above since it is amenable to a wide range of conditions and can be used with long lengths of DNA in solution. It provides data on the DNA helix conformation, orientation of conformational features and is useful for ligand binding experiments, since small ligands are only detected when bound to the DNA. The recent adaptation of a synchrotron circular dichroism apparatus for linear dichroism [22] allows investigation of DNA LD at lower wavelength than previously. In this chapter results are presented that show novel features of DNA LD at low salt concentrations.

2.2.1 Synchrotron radiation

A synchrotron is, in practice, a very intense light source — a photon generator. An electron or proton moving in a magnetic field experiences an electromagnetic force and hence emits photons. This is exploited in synchrotrons by accelerating electrons to close to the speed of light, in a circuit directed by large magnets oriented so as to bend the path of the electrons without altering their speed. At each bending magnet high energy photons are collected and passed through various optical components depending on the application. Other methods of photon generation from a beam of electrons in a synchrotron are mostly wigglers and undulators, both of which put a series of magnets together so as to rapidly change the direction of the magnetic field experienced

by the electrons. This causes the electrons to oscillate up and down during the course of their circuit and emit photons at small angles relative to their overall direction of motion.

The Astrid storage ring (Aarhus, Denmark) is a synchrotron capable of accelerating electrons to 0.58 GeV and producing photons of energies up to 700 eV, corresponding to a wavelength of 1.77 nm. There are a number of synchrotrons that can produce photons with significantly higher energy than this. In the electromagnetic spectrum, synchrotrons are capable of producing all wavelengths up to hard x-rays and gamma rays for which there is no alternative for a source of sustained intense beam-lines.

There are many uses of synchrotron radiation, including the study of surface and high temperature effects [46, 51, 43]. Until recent years it has mainly been the domain of physicists, but increasingly other disciplines are making use of the facilities. For example, small angle x-ray scattering is being used to determine structure of proteins both in solution and crystal form [97]. In the work presented here we show that synchrotron radiation with LD is another valuable application that has the potential to probe structures of biological matter.

2.2.2 DNA interaction with cations

Interest in DNA-cation interactions has been revived in the last decade with steps forward in experimental and modelling techniques. A combination of crystallography and molecular dynamics has been used to investigate the possible locations of cations binding to *in vitro* DNA, and a number of reviews pulling together the results of these experiments have been published [53, 26].

Crystallography

It has been found that monovalent (singly charged) ions are more difficult to identify than multivalent (multiply charged) ions since in crystallography studies they resemble water, although one study [52] used thallium(I) instead of sodium and predicted a number of sites that it occupies in a coordinated manner around DNA. As the authors recognise, thallium(I) is not a biological ion but it is indicative of the kind of structure that other more relevant ions may take. Occupancy of the sites was found to be roughly 10% for the minor groove of DNA and 20%–35% for the major groove, which is comparable to occupancy for studies using multivalent ions.

Studies in this area use very short sequences of DNA, most commonly the Dickerson-Drew dodecamer: CGCGAATTCGCG. This approach has been

useful for finding sequence specific interactions, but Egli [26] has questioned the applicability of these studies to other sequences and the possible importance of delocalised ions.

Modelling

Odijk, Skolman and Fixman (OSF) have proposed a charged polymer model of DNA [103, 82] which has been used over a number of years, however more recently Manning [67, 66] has questioned its validity and proposed an alternative. The OSF model does not allow for changes in structure due to electrostatic forces, unlike Manning's model which also considers the effect of transient changes in counterion concentration. Others have also introduced models able to differentiate effects of monovalent and divalent ions [108]. This is not an area where a single model has been tried, tested and proven, and more development is required.

We hope that LD will provide a complement to existing data and methods, especially since DNA can be measured in solution. Magnitude and location of peaks in LD spectra provide information on both the DNA structure and associated cations.

2.3 Methods

2.3.1 DNA

The DNA used was, with corresponding percentage G-C content, poly(A) poly(T) (AT), 0% (Amersham biosciences, Buckinghamshire, UK); clostridium perfringens (CP), 28.5% (Sigma-Aldrich, Dorset, UK); calf thymus (CT), 42% (Sigma-Aldrich); micrococcus luteus (ML), 71.0% (Sigma-Aldrich); poly(G) poly(C) (GC) 100% (Sigma-Aldrich) and single-stranded polyA (ssA), 0% (Sigma-Aldrich). Emma Gilroy prepared all DNA stock solutions and measured their concentrations using the absorbance at 260 nm.

The salt, DNA and extra water as required was mixed to give, in most cases 200 µM DNA in salt at the requisite concentration. An exception was ML DNA for which the volume of stock concentration was low, so 80 µl volumes were made using 50 µM DNA. Samples were mixed by pippetting and manual shaking. A total of 60 µl were used for each LD spectrum.

The DNAs used here are of heterogeneous lengths. Because of this little can be inferred from the relative LD magnitudes between DNA samples. For example, the ML DNA gives a much larger signal than the others.

2.3.2 Acridine orange

Acridine orange is a DNA dye that differentially stains single and double stranded DNA. It was purchased from Sigma-Aldrich and used at a concentration of 1 µM.

2.3.3 Linear dichroism spectrometers

The primary light source for LD was the UV-1 beamline at the Astrid storage ring (Aarhus University, Denmark). A Crystal Precision Optics Couette flow LD cell was used. The setup is not significantly different from a bench-top device, except that the light source occupies a rather larger area. The capillary and rod size are identical to those typically used with a bench-top spectrometer. The main difference is that the spectrometer uses a step between wavelength values rather than a continuous change, hence one of the parameters is 'number of measurements per value' rather than 'scan speed' and 'response time'. The parameters used with the synchrotron source were

- Rotation: 3 V
- Wavelength: 350–170 nm
- Number of measurements per value: 15
- Wavelength step: 1 nm
- Accumulation: 1
- Temperature: approximately 22°C

Accumulation refers to the number of times the wavelength range was measured over to give one output spectrum. The synchrotron software gives LD values in units of mV, which can be converted into absorbance units using a scaling factor although, as explained above, magnitude is not important here and this was not routinely done.

For measurements referred to as from a home source spectrometer a Biologic MOS-450 was used with a similar Couette flow LD cell to the synchrotron measurements and the following parameters

- Rotation: 3 V
- Wavelength: 350-170 nm
- Bandwidth slits: 4 nm

- Scan speed: 0.5 s nm^{-1}
- Accumulation: 3
- Temperature: 25 °C

Some measurements were taken using a J-815 spectrometer (Jasco, Tokyo, Japan). The settings used were as follows

- Rotation: 3 V
- Wavelength: 350-170 nm
- Sensitivity: standard
- Scan mode: continuous
- Bandwidth: 1 nm
- Response time: 1 s
- Scan speed: 100 nm/min
- Accumulation: 3

Temperature control was not used with these measurements.

2.3.4 Linear dichroism experimental method

The process used to take a linear dichroism measurement is described in the following.

For all three spectrometers, nitrogen gas is flowed through the parts containing the optical components. For the MOS-450 and J-715 spectrometers the flow rate is 8–10 l/min and for the synchrotron 5 l/min. The purging prevents the build-up of ozone which can cause oxidation damage to the optical components, is toxic and also reduces signal to noise ratio below 200 nm.

The quartz rod and capillary were cleaned between each sample using three washes of deionized water from a Direct-Q 5 water purification system (Millipore, Molsheim, France) followed by three washes of ethanol. If necessary the outside was wiped with a soft tissue. Both were dried using a pressurised air line. For samples particularly difficult to remove (e.g. if they have been left overnight in the capillary) a twisted lens tissue was inserted into the capillary to provide an abrasive clean (without scratching), or an acetone wash can be applied. Acetone is corrosive to the araldite plug at the base of the capillary

so these washes were relatively brief and followed immediately by the normal washing procedure.

The length of the central quartz rod was adjusted to provide a gap of 1 mm between its end and the plug at the capillary base.

The sample was loaded into the capillary and, if necessary, vigorously flicked to ensure that it reached the bottom of the capillary. The most effective method of loading was found to be when holding the capillary at an angle of 30-45° and pipetting the sample onto the highest part of the internal surface. The capillary was placed into the LD Couette flow cell followed by the rod. The rotation motor was started and allowed to settle to within less than 0.02 of the set value of potential difference, which takes 10–20 s. The measurement of LD was started from the computer software. If temperature control was used the sample was allowed to equilibrate to the set temperature (see chapter 5).

After the rotating spectrum is taken (or before) a non-rotating spectrum is taken. For only a very few highly unusual samples is residual orientation found if the non-rotating spectrum is measured after the rotating one.

2.4 Results

poly(A) poly(T) (AT) and single-stranded poly(A) DNA and DNA from clostridium perfringens (CP) DNA, calf thymus (CT) and micrococcus luteus (ML) were measured with a variety of sodium fluoride (NaF) concentrations. Since chloride ions strongly absorb at the lower end of the wavelength range used fluoride ions were used.

2.4.1 AT DNA

Figure 2.1 shows that for all salt concentrations, the LD of AT DNA gives the same spectral shape, with the familiar DNA peak at around 260 nm and a further peak at 185 nm which has not been previously reported. The 260 nm peak is due to $\pi \rightarrow \pi^*$ transitions in the pyrimidine and purine bases. The most intriguing feature of these spectra is the shoulder between 195 nm and 210 nm which has not been reported previously. Figure 2.2, where the data is scaled to have a relative LD of 1 at 256 nm, indicates that the spectral shape remains the same at all salt concentrations and that only the magnitude changes.

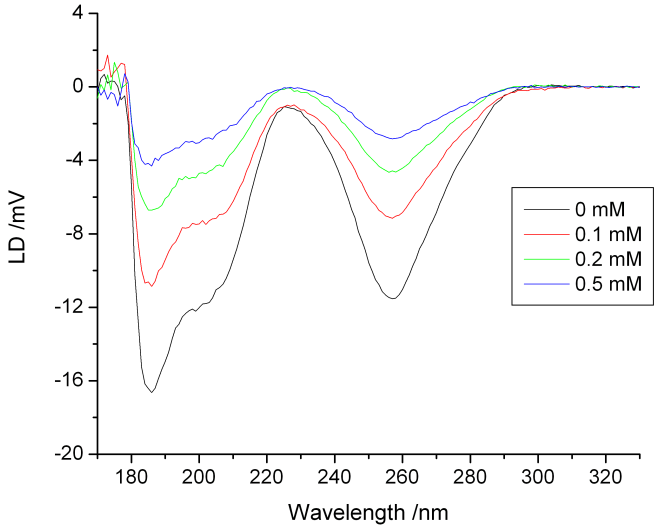


Figure 2.1: LD of AT DNA with various concentrations of NaF

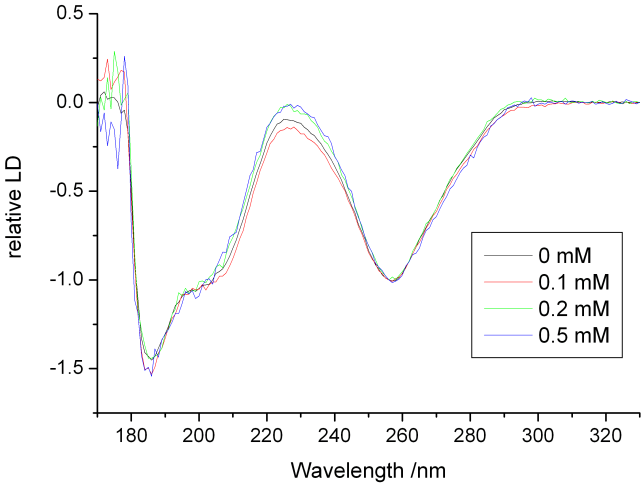


Figure 2.2: Data shown in figure 2.1 scaled so that all spectra are 1 at 256. It shows that they are similar in shape, particularly between 200 and 215 nm.

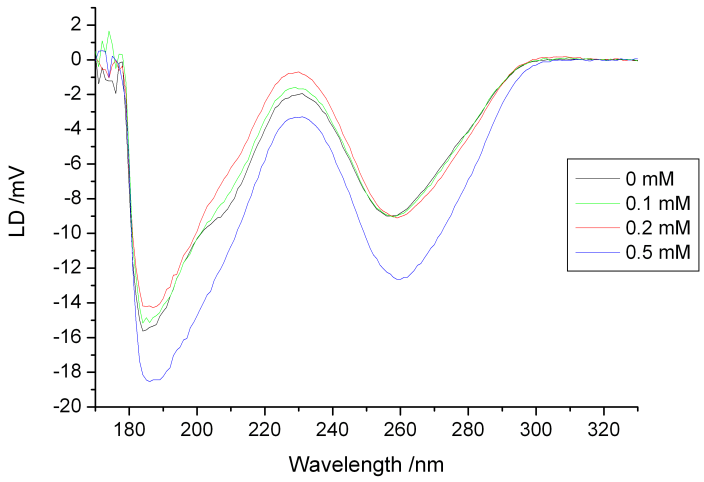


Figure 2.3: LD of CP DNA with concentrations of NaF and indicated

2.4.2 CP DNA

CP DNA has a higher G-C content than AT DNA. In figure 2.3 it is noticeable that the spectral shape changes as a function of salt concentration both in magnitude and geometry. In magnitude at 260 nm it decreases up to 0.2 M salt concentration, then vastly increases between 0.2 mM and 0.5 mM salt concentration. The peak at around 260 nm moves to progressively higher wavelength as the salt concentration increases and the lowest wavelength peak decreases in magnitude relative to that at 260 nm. This last feature is seen most clearly in figure 2.4 where the data has been scaled to 1 at 256 nm. We also observe the same shoulder present in AT DNA between 205 and 215 nm, but only when no salt is present and very slightly in 0.1 M salt. It is not present at higher salt concentrations.

2.4.3 CT DNA

The widest range of salt concentrations were used for CT DNA which is often used as a generic type of DNA because of its availability at low cost. The LD data are shown in figure 2.5. The shoulder between 200 and 210 nm observed in the LD of AT and CP is seen again at low salt concentrations but is not present above values higher than 200 mM. Also, a clear red-shift of the 260 nm peak can be seen at the highest salt concentrations.

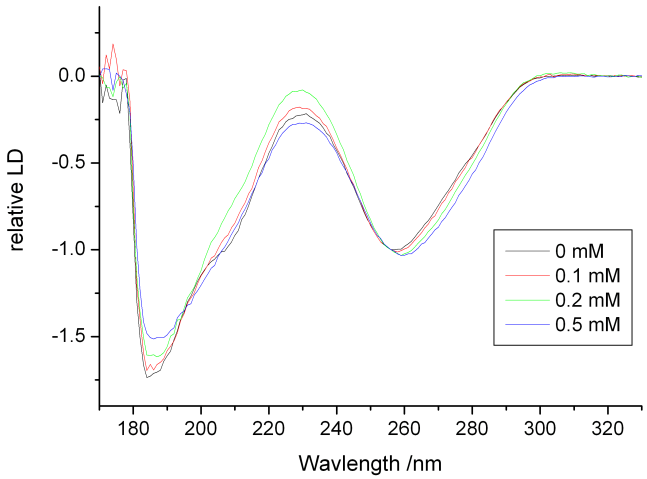


Figure 2.4: LD data for CP DNA from figure 2.3 scaled to 1 at 256 nm

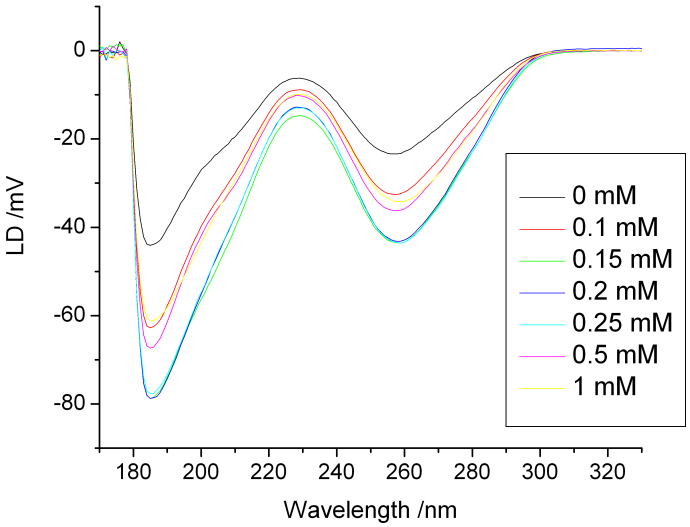


Figure 2.5: LD of CT DNA with various concentrations of NaF

2.4.4 ML DNA

Figure 2.6 shows data for ML DNA. Unlike those with lower G-C content there is no shoulder present around 200 nm, even at the lowest salt concentrations. The peaks are at the same wavelength for all spectra: they are simply multiples of each other, reflecting the effect of salt concentration.

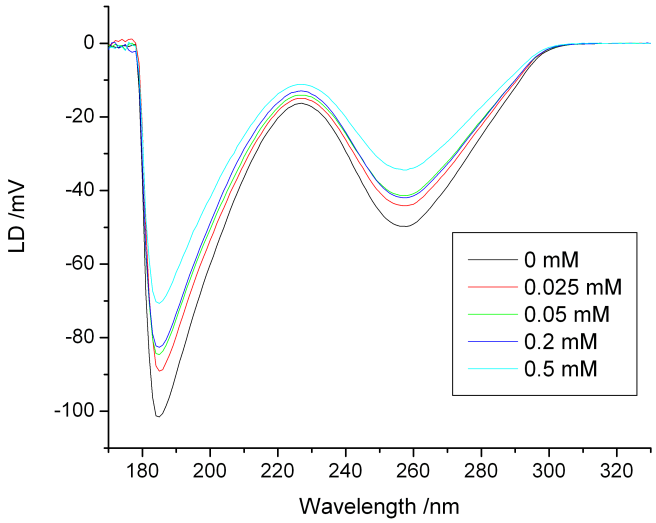


Figure 2.6: LD of ML DNA with indicated concentrations of NaF

2.4.5 Poly G poly C

Poly G poly C DNA was tested, but did not exhibit any significant LD. This is likely to be because the average length of the DNA was not long enough to give sufficient orientation.

2.4.6 ssA DNA

Data from ssA DNA are shown in figure 2.7. A shoulder at 210 nm is apparent but LD signal quickly disappears with increasing salt concentration, reflecting the flexibility of its structure compared with double stranded DNA. The peak at 257 nm is at the same wavelength for all spectra.

2.4.7 LD at 256 nm

The peak most identified with DNA spectroscopy occurs at around 260 nm. In fact, for the majority of cases the peak maximum is at 256 nm but it is common practice to quote extinction coefficients, for example, at 260 nm. Figure 2.8

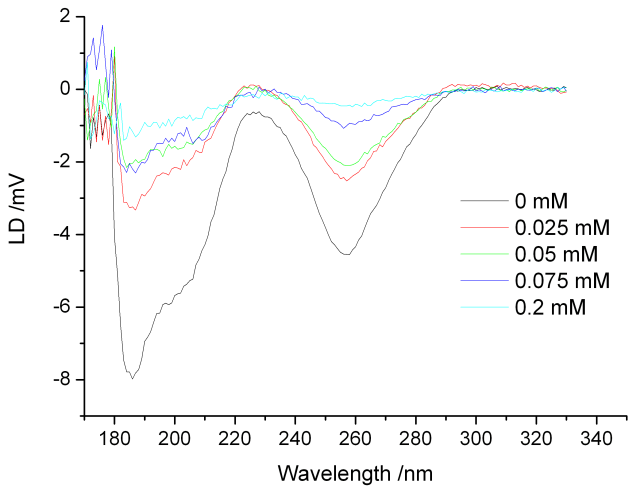


Figure 2.7: LD of ssA DNA with various concentrations of NaF.

shows the behaviour of LD for the different types of DNA at 256 nm with values scaled so that for each, the first point, at 0 nM salt, has a relative LD of -1. Three of the DNA types, ML, AT and ssA, exhibited no significant change in spectral shape in the LD spectra shown above. They also show similar behaviour here, with a roughly exponential decrease in LD magnitude as a function of salt concentration. The other two DNA types, CP and CT, exhibit different behaviour. CP DNA increases in LD signal with increasing salt. CT DNA initially exhibits increasing LD with salt, but then shows a decreasing trend above the equimolar salt-DNA concentration of 200 mM.

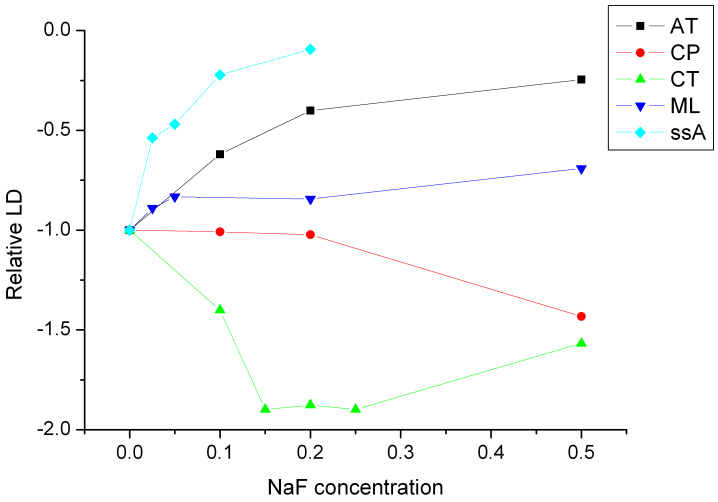


Figure 2.8: LD at 256 nm plotted against NaF concentration for different types of DNA. The data is scaled to be equal to -1 with no NaF.

2.4.8 Acridine Orange

Acridine orange is a small molecule used as a fluorescent DNA dye, especially for whole cells. It has the advantage that it fluoresces differently for single and double-stranded DNA so we speculated that it may produce differential fluorescence for the different conditions used here.

LD was measured for CT and ML DNA with no salt and 1 µM acridine orange present using the MOS-450 spectrometer (figure 2.9). If acridine orange was sensitive to single stranded regions in double stranded DNA we would expect to see different spectral shapes for the CT DNA (which may have single-stranded regions) and ML DNA (which appears not to). The same spectral shape was observed for both types of DNA. Thus, either this method is not sensitive to regions of single stranding, or single-stranded regions are not present. A peak at 500 nm was found, which suggest that the acridine orange was bound to the DNA. The apparent DNA peak was shifted to 264 nm, probably by absorbance of the acridine orange in this region.

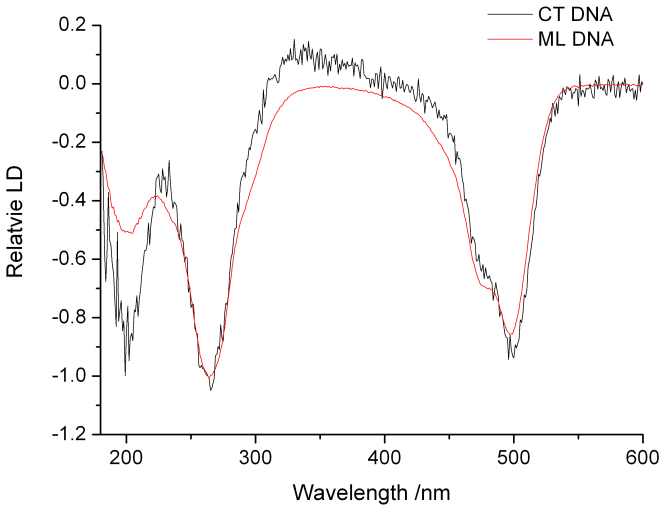


Figure 2.9: LD of CT and ML DNA measured in the presence of 1 µM acridine orange scaled to be 1 at 253 nm. The plots show no significant difference in spectral shape. The disparity at 200 nm is an artefact in the CT data since the HT is large at this point.

2.5 Discussion

2.5.1 Comparison of spectrometers

Two spectrometers were used to measure the LD of various types of DNA. One used a synchrotron source and the other a xenon lamp (home source). Data

for the home source are given in appendix 2.

There are two major advantages to using the synchrotron source: reduced noise and a greater wavelength range.

Figures 2.3 and C.2 show that at lower wavelength novel features are revealed by synchrotron LD. Below 220 nm the latter are noisy and the shoulder seen in the former at 200 nm is not visible.

A key indicator of the measurement accuracy is the high tension voltage (HT) which is the potential difference across the photomultiplier tube. The fewer the number of photons that reach the detector the higher the potential difference. Above a certain threshold, which varies for different detectors, the signal is saturated and measurements cannot be considered accurate.

Figure 2.10 shows HT for two CP plots, those at 0 mM NaF for the two spectrometers. The two plots look superficially similar with a steep gradient as the HT goes to lower wavelength, however the causes are very different in the two cases. For the home source, the Xenon lamp emits fewer photons at low wavelength causing the HT to increase and the LD to become noisier, making it more difficult to distinguish spectral features. The synchrotron emits a high flux of photons at a much wider range of wavelengths than those used here, so the reason for HT drop-off cannot be the same. Instead, it is caused by absorbance of the quartz capillary. This is currently the limiting factor for synchrotron LD. More homogeneous types of quartz, such as various grades of suprasil, are available which absorb at lower wavelengths. These are being investigated to improve the lower limit of the data.

The MOS-450 has lower HT at lower than 180 nm. It would be errant to interpret this as meaning that it can produce better data. Much effort has gone into removing stray, scattered light from the synchrotron system. This is not the case with the MOS-450 and the photomultiplier detects some of this and hence the HT is lowered.

As well as producing photons at higher energies than bench-top spectrometers the synchrotron also has much higher flux, i.e. more photons per unit area. This vastly reduces the noise of signals. This can be seen by comparison of any of the data sets, but is particularly well shown by the ssA DNA in figures 2.7 and C.5. The level of noise on the home source means the samples are barely distinguishable as DNA from an accumulation of three spectra. The synchrotron data is for a single accumulation and the peaks are clearly above the noise limit, although the 200 nm shoulder is difficult to prescribe for higher salt concentrations.

The MOS-450 spectrometer, used as a home source, is typical of circular and linear dichroism spectrometers currently used in laboratories. Not a great

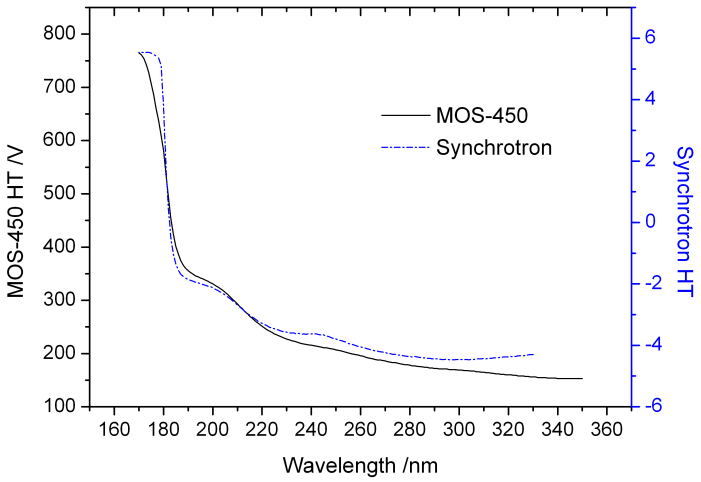


Figure 2.10: HT for CP DNA with no NaF. The same plot for other salt concentrations are very similar.

deal of optimisation of parameters was carried out for measuring the data shown in this chapter, but typically used settings were taken. More accumulations, for example, would reduce the noise. The synchrotron, on the other hand, produces data with little noise and is still good at low wavelength. Similar work has been carried out for circular dichroism, a related technique using circular instead of linearly polarised light to identify chiral molecules. Principal component analysis shows much better identifiability of structures at low wavelength, with only 2 or 3 independent structural types for data above 200 nm whereas more than 8 are at below 160 nm [77].

As a comparison, some of the samples were measured using a J-815 spectrometer, which is a newer model than the MOS-450 and represents the best benchtop spectrometers currently available. Data for ML and CT DNA from all three spectrometers is shown in figures 2.11 and 2.12. The J-815 spectrometer shows almost comparable performance to the synchrotron. It requires multiple scans for each spectrum, but the data shown uses just three and exhibits minimal noise. In addition, the HT (not shown) is similar to that for the synchrotron suggesting that stray light is not a problem. The improvements are mainly a result of better optical components and detection, rather than a different light source, hence the signal will be curtailed shortly below the observed limit. How far below is not precisely known at present and is dependent on the sample. Some light scattering by the sample is observed, in that the peak at 185 nm is slightly smaller in magnitude with the J-815. This relatively small effect for DNA is more exaggerated for protein samples.

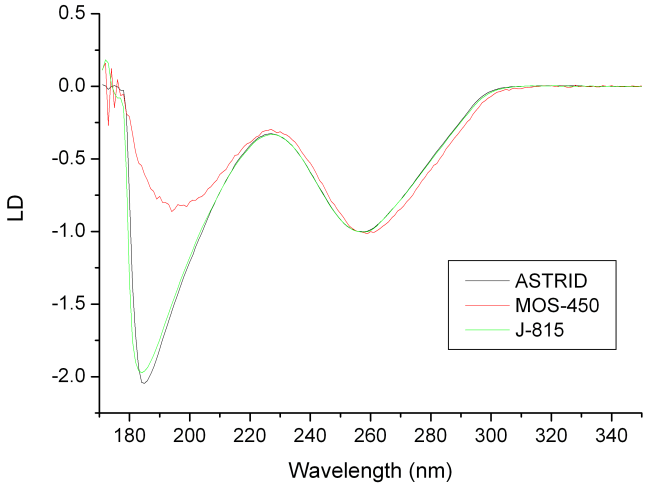


Figure 2.11: LD spectra for ML DNA with no salt from three spectrometers. ASTRID refers to the synchrotron.

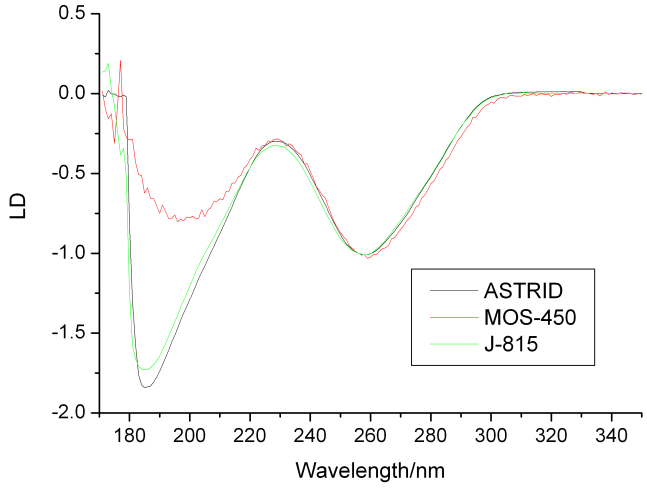


Figure 2.12: LD spectra for CT DNA with 0.2 mM NaF from three spectrometers. ASTRID refers to the synchrotron.

2.5.2 Changes in LD spectra

Before discussing the results further the general changes in LD spectra are discussed, with reference to equation (1.14). To change LD, the chromophore must undergo a change in strength, orientation relative to the molecule, or there needs to be a change in overall orientation of the molecule. Chromophore oscillator strength (equivalent to dipole magnitude) is dependent on the local chemical environment. A change in chromophore angle (realtive to the molecule) is likely to be due to a local structural change. The overall orientation is dependent on both chemical environment and molecule structure and is manifest as a scaling of the whole spectrum. Other effects that should be borne in mind during LD are light scattering or high absorbance, both of which create artefacts, although these are seldom a problem with DNA.

2.5.3 The effect of salt on DNA

It is well known that positive ions condense onto DNA although, as discussed in the introduction (section 2.2.2), the precise mechanistic details and locations are still unclear. Salt causes neutralisation of charge-charge interactions between bases.

Broadly, the main effect of salt on LD of DNA can be summarised as follows. The salt cations neutralise the negative charge-charge repulsion interactions between DNA bases. Steric hindrance present within DNA is thus reduced, which creates an overall increase in DNA flexibility. The measured LD magnitude with an increased amount of salt is thus less since the orientation parameter is smaller. LD data for ssA, AT and ML DNA (figure 2.8) are consistent with this explanation. Note that these are the DNA types with highest and lowest G-C contents. However, for CT and CP DNA other effects are more important.

The level of explanation of charge-charge interactions given is enough to explain the general trend. What is less certain is precisely how the salt arranges itself at the DNA surface. X-ray crystallography and molecular modelling have been used to address this problem in recent years, as discussed in section 2.2.2. It is thought that the true nature of salt interaction with DNA lies somewhere between weak, non-specific interactions of fully hydrated ions and tightly bound cations located at specific sites [53]. Low occupancy of specific sites limits the usefulness of x-ray crystallography, although the high resolution is valuable.

2.5.4 Differences for CP and CT DNA

The LD measurements of CP and CT DNA contain several common features distinct from the other DNA types. Firstly, the DNA peak at around 260 nm changes position with salt concentration. For CP with increasing salt the shift is from 258 to 260 nm (figure 2.3) and for CT from 257 to 259 nm (figure 2.5). Secondly, the shoulder between 195 and 210 nm disappears with increasing salt. Thirdly, as shown by the plot of LD at 256 nm in figure 2.8 there is a period of increasing LD magnitude with salt concentration for both of these types of DNA.

Changes in the chemical environment

The change at 260 nm in LD of CP and CT DNA indicates a change in chemical environment of the DNA bases. The region of increase in LD magnitude means there is an effect that stabilises the overall orientation of the DNA.

The effect causing the changes in LD spectra is clearly distinct from the charge-charge neutralisation observed for ML, AT and ssA DNA types, which causes a decrease in LD magnitude with salt. For CT and CP DNA LD magnitude increases for a period at 256 nm, indicative of an increase in rigidity. Figures 2.5 and 2.3 show that the effect is present over the whole spectral range. Assignment of the effect that causes this increase is speculation at the moment. The possibility of single-stranded regions is discussed further below. If an increase in salt stabilised the hydrogen bonds between bases, removing single-stranded regions, this would be consistent with the data.

Suppression of the shoulder

The disappearance of the shoulder around 205 nm in LD of CP and CT DNA is caused by suppression of the chromophore (or combination of chromophores) that causes it. If the peak wavelength was shifted it would result in a change of shape in another part of the spectrum, which is not observed. It is a feature associated with low G-C content, being more marked in AT DNA, for example, and not present at all in ML DNA. However, the suppression only occurs when G-C base pairs are present, suggesting that they play a role.

A question that poses itself is which transition causes the shoulder. There is no straight-forward answer to this. Recent *ab initio* calculations show general agreement with experiment and suggest that $\pi \rightarrow \pi^*$ transitions are sufficient to account for the peaks in this region with some contribution from $n \rightarrow \pi^*$ transitions [86, 87, 33, 34]. The results, however, are tentative and only apply to single bases, not base-pairs. In particular it has been a problem to estimate

polarisation directions since the calculations are made in the gas phase and solvent effects can alter the magnitude, position and direction of peaks. The position alteration estimated by Fülscher and Roos to be of the order of a few electron volts, i.e. less than 5 nm. To our knowledge there has been no such calculation carried out for base-pairs in a solvent environment.

2.5.5 Single-stranding effects

Current knowledge does not seem adequate to explain the observations presented here. The only suggestions of salt stabilising DNA structure are for concentrations much higher than those used here.

We offer the following tentative explanation: that within DNA at low salt concentrations the weaker nature of the A-T pair (having only two hydrogen bonds compared to three for G-C) allows the DNA to separate. Tracts of this result in single-stranded regions. With salt present, the DNA helix is stabilised to become more double-stranded, causing an increase in rigidity. This may occur by co-ordination of salt at specific sites, as observed in theoretical models for DNA. Since, as noted above, the G-C base pairs are involved in this effect it is likely that co-ordination sites are in G-C base-pairs or at the interface between G-C and A-T pairs.

This explanation with the data we have presented is highly speculative. It requires validation by other methods. We suggest that the effect has not been reported before because most of the techniques used by those looking at DNA structure in detail are not for molecules free in solution (x-ray crystallography, electron microscopy, magnetic tweezers etc.). Also, coarse-graining may have excluded relevant higher order terms that would result in the theoretical prediction of such an effect.

There is a possibility that the effect is an artefact of Couette flow, but the shear forces involved are small and we suggest that even if this is the case it is still of importance.

2.5.6 Future work

The work here can be considered as one piece of a jigsaw. The results show that addition of salt changes some characteristic of DNA with medium to low G-C content. For crystallography, molecular dynamics and *ab initio* techniques there is always a question as to whether solvents can have an effect on translating *in vitro* or *in silico* to those found in solution and thence to *in vivo* behaviour. Our results show that the solution state of DNA differs from crystalline or gas phases. A full explanation requires complementary data from

a number of methods and careful consideration of how they relate to *in vitro* and *in vivo* DNA.

Chapter 3

Persistence length of DNA

3.1 Chapter summary

Four definitions of persistence length commonly used in the literature are presented and shown to be equivalent under certain constraints. The sole model published relating linear dichroism and persistence length is presented and discussed. In the model is an assumption of small perturbations to the DNA molecule. Experimental data from linear dichroism, fluid dynamics studies with DNA and coarse-grained modelling suggests that in Couette flow large bends are likely to be present and the hydrodynamics does not follow the Zimm-Rouse model.

3.2 Introduction

A key consideration in the interpretation of linear dichroism (LD) of DNA is molecular orientation. In the LD measured for the data presented in this thesis, Couette shear flow was used to orient the samples. The behaviour of DNA in such an environment is clearly in some way dependent on the properties of intrinsic curvature and flexibility. Intrinsic curvature is the static measure of curvature of a polymer, i.e. at equilibrium. For DNA it is sequence dependent [111]. Flexibility is dynamic bending dependent on intrinsic polymer properties that results from the polymer environment. The bending in this case is due to factors such as thermal perturbation and hydrodynamics. These two properties are combined in persistence length, which is the topic of this chapter. The following chapter concerns experimental determination of persistence length.

A number of definitions for persistence length can be found in the literature. They are often thought to be interchangable, but this is not necessarily the case. A concept such as persistence length should be firmly grounded in measurable properties. As will be shown in the next two chapters, LD of DNA is clearly dependent on persistence length, but the state of the DNA precludes the use of persistence length definitions interchangably. Also, it is far from clear what the relation between LD at a given wavelength and persistence length should be.

The chapter starts with four definitions of persistence length, followed by a number of polymer models used to describe DNA. Following this, the equivalence of the definitions is discussed. The second part of the chapter is a presentation of the sole published model relating LD of DNA to persistence length and a discussion of some of its shortcomings.

3.3 Definitions of persistence length

Although widely used, the literature does not always contain a uniform and coherent presentation of the concepts of persistence length. Several definitions are in common use and the vast majority of references refer to one or at most two of these. In fact, to our knowledge one of the definitions (definition 4) has never been formally shown to be equivalent to the others.

We present the definitions before introducing polymer models to emphasise that the definitions are model independent. Proofs of equivalence of the definitions are dependent on certain assumptions about the nature of the polymer in question.

Definition 1: Projection definition

Given a curve $\mathbf{r}(s) : \mathbb{R} \to \mathbb{R}^3$ with unit tangent vectors $\mathbf{t}(s)$, where s is a scalar measure of length along the curve, persistence length l_p can be defined as the average projection (over the length of the curve and possible conformations) along the initial tangent as the total length L tends to infinity [30, 5]. This can be expressed as

$$l_p = \lim_{L\to\infty} \int_0^L \langle \mathbf{t}(0).\mathbf{t}(s)\rangle ds. \tag{3.1}$$

Although this has been expressed in terms of a projection, note that it has the same functional form as correlation.

This definition is usually expressed in the discrete form, but we use the continuous form here to demonstrate its independence from any polymer model. In this form it is valid for any continuous curve provided that the flexibility and curvature are the same over the region to which the definition is applied.

Note that for a rigid rod, persistence length is undefined since correlation exists along the length of the rod. Below, contour length and end-to-end distance are related to persistence length for the case of a worm-like chain.

Definition 2: Kuhn length

A curve $\mathbf{r}(s)$ as above can be approximated by a series of straight line segments of length b, such as in figure 3.1; this is called the equivalent random chain. If the segment lengths are short then they are not independent of each other (in the language of the previous definition, correlation still persists between them). If it is possible to choose a large enough segment such that consecutive segments are independent, then the approximated chain is a random walk. The minimum such length is termed the Kuhn length, l_k [58]. We show below that it is equal to twice the persistence length under certain constraints. A continuous definition is also given below.

Definition 3: Bending constant

Physical properties of a semi-flexible polymer give rise to another definition [68, 38]. For a polymer obeying Hooke's law for its bending, with bending constant B

$$l_p = \frac{2B}{k_B T} \tag{3.2}$$

where k_B is Boltzmann's constant and T is absolute temperature. Assuming Hooke's law may be a little surprising, given definitions 1 and 2, since it is explicitly dependent on an intrinsic physical property of the polymer. This is a stronger condition, but can be thought to emphasise the material from which the polymer is made, with which Hooke's law is associated, as opposed to the arising geometry, which is the emphasis of the previous two definitions. Hooke's law is applicable to short sections of semi-flexible polymers (i.e. flexibility somewhere between that of a rigid rod and a polymer with no tangential correlation); it does not necessarily describe microscopic properties well but is assumed to give the correct macroscopic behaviour.

This definition perhaps shows more clearly than the previous two that an infinitely long polymer is not required to define persistence length. In practice the minimum length is that over which a bend in the material can be measured [59].

Definition 4: Curvature definition

A definition found in several literature references arises from polymer curvature. Persistence length is defined to be the length along the polymer after which the average deflection of the tangent is 1 radian [92, 3]. That is, $\langle|\theta|\rangle = 1$ where θ is the angle between two tangents a persistence length apart in contour length.

3.4 Models of DNA

Bridging the gaps between measurements and definitions of persistence length requires a reliable polymer model. We discuss possible models with an emphasis on those that have been used for DNA.

A suitable model balances detail on the one hand and mathematical tractability and computational expense on the other. Many molecular models of DNA have been conceived. The most detailed of these require a large amount of computational power and yield results that last for time scales in the region of nanoseconds or less [95]. At the other extreme it is perfectly possible to model DNA in solution as the motion of a sphere or a rigid rod in flow [99], which is especially useful for tracing the progress of molecules through complex flows where the computational power goes into finding solutions for the Navier-Stokes equation. The most successful models are a medium between these two approaches [88, 98]. The most accurate model of DNA in solution is generally accepted to be the worm-like chain. This is a limit of several other

models, outlined below: the freely jointed chain, the random flight model and the Gaussian chain.

3.4.1 The freely jointed chain

Perhaps conceptually the simplest model of a polymer is to regard it as a series of straight segments joined end to end, with directions given by the vectors $\{\mathbf{r}_i\}_{i=1}^N \subset \mathbb{R}^3$, each of length b and allowing the joints to be completely free. This is the freely jointed chain (FJC) as shown in figure 3.1. Statistically the model is defined by the probability

$$\psi(\{\mathbf{r}_n\}) = \prod_{n=1}^{N} \psi(\mathbf{r}_n) \tag{3.3}$$

where

$$\psi(r_n) = \frac{1}{4\pi b^2}\delta(|r| - b). \tag{3.4}$$

Here δ is the Dirac delta function. This allows each segment vector to occupy any point on a sphere of radius b with equal probability.

End-to-end distance is denoted R [1] hence

$$\begin{aligned}
\langle R^2 \rangle =& \langle (r_1 + r_2 + ...r_n).(r_1 + r_2 + ...r_n) \rangle & (3.6)\\
=& \sum_{n=1}^{N}\sum_{m=1}^{N} \langle r_n.r_m \rangle = \sum_{n=1}^{N} \langle r_n.r_n \rangle & (3.7)\\
=& Nb^2. & (3.8)
\end{aligned}$$

Equation (3.7) follows because for $m \neq n$, averaging over all possible states of $r_n.r_m$ leads to 0.

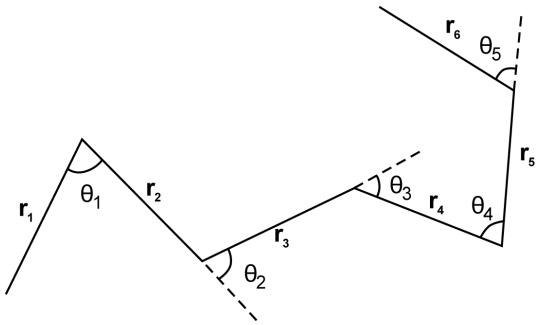


Figure 3.1: A freely jointed chain. All of the segments have the same length, b. For the random flight model $\theta_1 = \theta_2 = ... = \theta_N = \theta$

3.4.2 Random flight model

The random flight model is a special case of the FJC in which the bond angles are fixed at an angle θ, whilst the torsion angles remain free to rotate (figure 3.1). The model has the property

$$\mathbf{r}_n.\mathbf{r}_{n-1} = b^2 \cos\theta \tag{3.9}$$

and when the mean squared end-to-end length is evaluated for large N

$$\langle R^2 \rangle = Nb^2 \frac{1+\cos\theta}{1-\cos\theta}. \tag{3.10}$$

The proof of this is given in appendix B.1. Kuhn length (definition 2) of the FJC is

$$l_k = b\sqrt{\frac{1+\cos\theta}{1-\cos\theta}} \tag{3.11}$$

so that equation (3.10) is

$$\langle R^2 \rangle = Nl_k^2 = Ll_k \tag{3.12}$$

[1]Angled brackets denote averaging over all possible states, that is for a quantity $X(\phi)$ where $\psi(\phi)$ is the probability of the state variable ϕ

$$\langle X(\phi) \rangle = \frac{\int_\phi X(\phi)\psi(\phi)}{\int_\phi \psi(\phi)} \tag{3.5}$$

which is similar to equation (3.8) and hence the system imitates the FJC, which is equivalent to a random walk with step size b. This equation leads to the continuous definition for Kuhn length [23]

$$l_k = \frac{\langle R^2 \rangle}{L} \tag{3.13}$$

Another interesting quantity is the distribution of end-to-end length. This is calculated by Doi and Edwards [23] and represented by the following equation for the probability of the end-to-end vector having length R

$$\Psi(R) = \int_{\mathbb{R}^3} d\mathbf{r}_1 ... \int_{\mathbb{R}^3} d\mathbf{r}_N \; \delta \left(R - \left| \sum_{n=1}^{N} \mathbf{r}_n \right| \right) \psi(\{\mathbf{r}_n\}). \tag{3.14}$$

Using

$$\delta(r) = \frac{1}{(2\pi)^3} \int_{\mathbb{R}} dk \; e^{ikr} \tag{3.15}$$

$\langle R \rangle$ is found to have the property

$$\Psi(R, N) = \left(\frac{3}{2\pi N b^2} \right)^{\frac{3}{2}} \exp \left(-\frac{3R^2}{2Nb^2} \right) \tag{3.16}$$

that is, the end-to-end vector is Gaussian distributed.

3.4.3 Gaussian spring chain

As a progression from the random flight model we consider the model consisting of a series of spheres connected by Gaussian springs. That is, the distribution of bond lengths is given by

$$\psi(r) = \left(\frac{3}{2\pi b^2} \right)^{\frac{3}{2}} \exp \left(-\frac{3r^2}{2b^2} \right) \tag{3.17}$$

where

$$\langle r^2 \rangle = b^2. \tag{3.18}$$

This can be thought of as mimicking a series of random flight chains placed end-to-end. It is widely used since it is mathematically relatively easy to manipulate. Note that b is fixed and defines many of the properties of the polymer.

This model also results from a polymer where bond energy is

$$U(\{r_n\}) = \frac{3}{2b^2} k_B T \sum_{n=1}^{N} r_n^2 \tag{3.19}$$

where each spring is treated as a Hooke spring with spring constant $\frac{3}{2b^2} k_B T$. Equivalence is via the Boltzmann distribution, where occupancy (equivalent to probability) is given by

$$\frac{N_i}{N} = \frac{\exp\left(\frac{U_i}{k_B T}\right)}{\int_0^\infty \exp\left(\frac{U_j}{k_B T} dj\right)} \tag{3.20}$$

where N_i is the number of segments with energy U_i and N is the total number of segments. On evaluating the denominator of equation (3.20), equation (3.17) is recovered.

3.4.4 The worm-like chain

For both the random flight model and the Gaussian chain the worm-like chain (WLC) is a limiting case. The approximation is such that segment length b tends to 0 whilunst preserving contour length L. Note that this requires N to tend to infinity in the limit. The approximation is usually employed after the incorporation of terms governing the polymer environment, such as hydrodynamic terms, see for example [9, 29].

First introduced by Kratky and Porod in 1949 [57] the WLC is widely considered the most effective model for DNA, notably contributing to the most widely used model used for single-molecule stretching techniques [9] nearly half a century after its introduction. It has not been applied, however, to theoretical models of LD.

3.5 Equivalence of different definitions

The equivalence of the definitions of l_p is proven here by showing that they are each equivalent to definition 1.

3.5.1 Definition 2: Kuhn length

To start proving equivalence of the definitions given above necessitates a less general mathematical treatment of the polymer in question. Equation (3.12) gives the end-to-end distance for a polymer with Kuhn length l_k.

The end-to-end distance for a chain of finite length from definition 1 can also be calculated. Flory [30] mentions this in appendix G. Using the random flight model and projecting subsequent segments onto the first yields

$$l_p = \lim_{N\to\infty} b\sum_{k=1}^{N}(\cos\theta)^k \tag{3.21}$$

$$= \frac{b}{1-\cos\theta} \tag{3.22}$$

$$\approx \frac{b}{\log(-\cos\theta)} \tag{3.23}$$

in the limit of small b. For a finite chain the same projection, which we term X is

$$X = \sum_{k=1}^{N}(\cos\theta)^k \tag{3.24}$$

$$= l_p\left(1-\exp\left[-\frac{L}{l_p}\right]\right). \tag{3.25}$$

When one considers that $d\langle\mathbf{R}^2\rangle = 2\langle\mathbf{R}.d\mathbf{R}\rangle = 2XdL$ then

$$\frac{\langle R^2\rangle}{L} = 2l_p\left(1-\frac{l_p}{L}\left[1-exp\left(-\frac{L}{l_p}\right)\right]\right). \tag{3.26}$$

So in the limit as $L\to\infty$

$$\mathbf{R}^2 = 2Ll_p = 2Nl_p^2 \tag{3.27}$$

By comparing equations 3.12 and 3.27, for an infinitely long polymer well approximated by the random flight model, Kuhn length is twice persistence length.

3.5.2 Definition 3: bending constant

The following argument is as reported by Hagerman [38]. Firstly we retrace our steps slightly and re-express equation (3.21) in terms of the freely jointed chain. Instead of a fixed angle, the average bond angle is used:

$$l_p = \frac{b}{1 - \langle \cos\theta \rangle}. \tag{3.28}$$

If we take an alternative approximation to the one used previously, for sufficiently small segment length (and hence bond angle) $cos\theta \approx 1 - \frac{1}{2}\theta^2$

$$l_p = \frac{2b}{\langle \theta^2 \rangle} \tag{3.29}$$

If we assume that Hooke's law applies to bending energy, that is the restorative force is linear with angular deflection, then by standard results the ensemble average over squared angular deflection (i.e. bond angle) is

$$\langle \theta^2 \rangle = \frac{bk_BT}{B}. \tag{3.30}$$

Comparison with 3.29 gives that

$$l_p = \frac{2B}{k_BT} \tag{3.31}$$

Hagerman [38] discusses the limitations of this approximation, recognising that this result is valid only for isotropic, linear bending.

3.5.3 Definition 4: curvature

Previous users of definition 4 have verified its equivalence to other definitions numerically [105].

3.5.4 Discussion of proofs

We have thus proved equivalence of the definitions used for persistence length. Note, however, that the latter two proofs are valid only under the assumption of Hooke's law. In fact, the definition is not valid if this condition is not satisfied.

There is disparity between the properties of finite length polymers if the projection and Kuhn length definitions are used. Kuhn length depends on the polymer being well approximated by a FJC, since this model is identical to a random walk. The projection definition, however, does not require this. In the limit of long chains the end-to-end distance function has the same form for both definitions, but this is not the case for a polymer of finite length, as shown by comparison of equations 3.12 and 3.26. Experimental measurements are, necessarily, made on finite-length chains and so the polymers used should either be sufficiently long for the two functions to be approximately the same, or the functional form used should be clearly stated. In the overwhelming majority of cases for which end-to-end measurements are directly observed, the projection definition is used.

The development of polymer statistics happened before the discovery of DNA structure and analysis. This seems to have caused some problems with approaching theoretical descriptions of DNA. The proofs and discussion above should make it clear that the definitions given for persistence length are only equivalent under certain assumptions, thus those working with the concept should clearly state what definitions and assumptions are being employed along with a clear reference to the definition source. This is not always the case for otherwise excellent publications [83, 114, 41].

Persistence length remains only a curious and rather abstract concept if it cannot be related to experiment. Since it cannot be measured directly, a model is required to link the theory with experimental data. This is dependent on the measurement being taken and a wider discussion of this is undertaken in the following chapter. In what follows we present the sole model developed to determine l_p from LD.

3.6 Wilson and Schellman model

Thus far only one model has been published to interpret LD of DNA as a function of persistence length [119]. We summarise the model by Wilson and Schellman before a critique of its limitations.

The general scheme of the model derivation is as follows:

- find the absorbance of a single short segment of DNA in solution
- sum over all segments
- introduce the hydrodynamic matrix
- apply the Peterlin distribution to the DNA molecules

3.7 Finding segment orientation

DNA is modelled using the freely jointed chain (FJC, see 3.4.1). A single segment can be represented either in the segment axes (x, y, z), or the fixed laboratory axes (X, Y, Z). The former is defined such that z lies along the segment and x and y are any two vectors that complete an orthonormal set. The latter are defined so that X is in the direction of flow, Y is along the flow gradient and Z is along the cylinder axis. This is shown in figure 3.2.

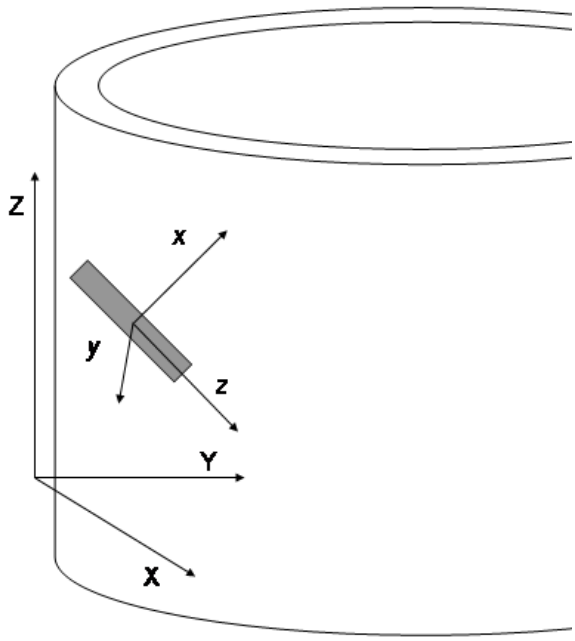


Figure 3.2: A chromophore, shown as a grey box, can be represented either by laboratory axes (capital letters) or its internal axes (lower case, italic letters). The laboratory axes are relative to the direction of Couette flow cell, indicated by the cylinder.

3.7.1 Absorbance

A chromophore absorbs along its internal z axis with extinction coefficient $\epsilon_{||}$ and perpendicular to it with extinction coefficient $\epsilon_{\perp}$, thus the matrix of absorbance in the molecular axes is

$$E = \begin{pmatrix} \epsilon_{\perp} & 0 & 0 \\ 0 & \epsilon_{\perp} & 0 \\ 0 & 0 & \epsilon_{||} \end{pmatrix}$$

Recall that by the Beer-Lambert law (equation (1.1)) extinction coefficient is proportional to absorbance. To express this in the laboratory axes note that a unit vector in the X direction of the laboratory axes can be transformed into the segment axes as follows

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} x.X \\ y.X \\ z.X \end{pmatrix}$$

and hence the change of basis matrix from the segment axes to the laboratory axes is

$$M = \begin{pmatrix} x.X & y.X & z.X \\ x.Y & y.Y & z.Y \\ x.Z & y.Z & z.Z \end{pmatrix}.$$

To obtain absorbance in the laboratory axes we change basis to the segment axes, apply the absorbance matrix and then change back again so

$$MEM^{-1}. \tag{3.32}$$

Note that the inverse of M is equal to its transpose. Writing this in full

$$\begin{pmatrix} x.X & y.X & z.X \\ x.Y & y.Y & z.Y \\ x.Z & y.Z & z.Z \end{pmatrix} \begin{pmatrix} \epsilon_\perp & 0 & 0 \\ 0 & \epsilon_\perp & 0 \\ 0 & 0 & \epsilon_{||} \end{pmatrix} \begin{pmatrix} x.X & x.Y & x.Z \\ y.X & y.Y & y.Z \\ z.X & z.Y & z.Z \end{pmatrix} \tag{3.33}$$

$$\begin{aligned} = \begin{pmatrix} x.X & y.X & z.X \\ x.Y & y.Y & z.Y \\ x.Z & y.Z & z.Z \end{pmatrix} \Bigg\{ \epsilon_\perp \begin{pmatrix} x.X & x.Y & x.Z \\ y.X & y.Y & y.Z \\ 0 & 0 & 0 \end{pmatrix} \\ + \epsilon_{||} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ z.X & z.Y & z.Z \end{pmatrix} \Bigg\} \end{aligned} \tag{3.34}$$

$$= M^{-1}\epsilon_{\perp}M + \begin{pmatrix} x.X & y.X & z.X \\ x.Y & y.Y & z.Y \\ x.Z & y.Z & z.Z \end{pmatrix}(\epsilon_{||} - \epsilon_{\perp}) \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ z.X & z.Y & z.Z \end{pmatrix} \tag{3.35}$$

$$= \epsilon_{\perp}I + (\epsilon_{||} - \epsilon_{\perp}) \begin{pmatrix} (z.X)(z.X) & (z.X)(z.Y) & (z.X)(z.Z) \\ (z.Y)(z.X) & (z.Y)(z.Y) & (z.Y)(z.Z) \\ (z.Z)(z.X) & (z.Z)(z.Y) & (z.Z)(z.Z) \end{pmatrix} \tag{3.36}$$

$$= \epsilon_{\perp}I + (\epsilon_{||} - \epsilon_{\perp})J \tag{3.37}$$

where equations 3.36 and 3.37 define the matrix J. For two perpendicular polarisations $\mathbf{P}$ and $\mathbf{P'}$ the difference in absorption (i.e. LD) is given by

$$\Delta\epsilon_{\mathbf{P}} = (\mathbf{P}^T[\epsilon_{\perp}I + (\epsilon_{||} - \epsilon_{\perp})J]\mathbf{P} - (\mathbf{P'}^T[\epsilon_{\perp}I + (\epsilon_{||} - \epsilon_{\perp})J]\mathbf{P'} \tag{3.38}$$

$$= \epsilon_{||}(\mathbf{P}^T\mathbf{P} - \mathbf{P'}^T\mathbf{P'} + (\epsilon_{||} - \epsilon_{\perp})(\mathbf{P}^T J\mathbf{P} - \mathbf{P'}^T J\mathbf{P'}) \tag{3.39}$$

$$= (\epsilon_{||} - \epsilon_{\perp})(\mathbf{P}^T J\mathbf{P} - \mathbf{P'}^T J\mathbf{P'}). \tag{3.40}$$

The last step follows since $\mathbf{P}$ and $\mathbf{P'}$ have the same magnitude, hence $\mathbf{P'P'}^T = \mathbf{PP}^T$. Here, a superscript $\mathbf{T}$ denotes the vector transpose.

3.7.2 Sum over all segments

At this point the following assumption is made

$$\Delta\epsilon_r = \langle\Delta\epsilon\rangle_0 \frac{\langle r^2\rangle}{\langle r^2\rangle_0} \tag{3.41}$$

This is based on previous work. The subscript 0 here refers to the polymer when only short-term interactions are considered, that is, excluded volume effects and other long-range interactions such as charge-charge effects are not considered. This is a simplified case that can, in principle, be calculated. This was done by the same authors in an earlier paper [118] and the relationship above established under the assumption that the polymer follows the Gaussian spring model and also that $\mathbf{r}^2$ is not much different from $\langle\mathbf{r}^2\rangle_0$, which turns out to be a strong condition. Assuming equation (3.41), the LD of a population of segments results. Hence the sum for a complete chain of N segments is

$$\Delta\epsilon_{\mathbf{P}} = \frac{\langle\Delta_{\epsilon}\rangle_0}{\langle\mathbf{r}^2\rangle_0} \sum_{i=1}^{N} \langle\mathbf{r}_i^2\rangle(\mathbf{P}^T J_i\mathbf{P} - \mathbf{P'}^T J_i\mathbf{P'}) \tag{3.42}$$

To find the LD, evaluation of the sum is required. The subject of the next section is average properties of the FJC in a Couette flow hydrodynamic environment.

3.7.3 Hydrodynamics

The inclusion of hydrodynamic effects follows Zimm, widely reported in texts on the subject [23]. First order hydrodynamic effects are given by a matrix H which expresses local deformation due to the hydrodynamic forces.

The first step for inclusion of hydrodynamic effects is a change of notation. It transpires that the following simplifies the calculations: let the vector of co-ordinates of the segment ends be X,

$$\mathbf{X} = (x_1, x_2, x_3, ..., x_N)^T$$

then

$$J' = \sum_{i=1}^{N} \langle r_i^2 \rangle J_i \tag{3.43}$$

$$= \begin{pmatrix} \langle \mathbf{X}^T A \mathbf{X} \rangle & \langle \mathbf{X}^T A \mathbf{Y} \rangle & \langle \mathbf{X}^T A \mathbf{Z} \rangle \\ \langle \mathbf{Y}^T A \mathbf{X} \rangle & \langle \mathbf{Y}^T A \mathbf{Y} \rangle & \langle \mathbf{Y}^T A \mathbf{Z} \rangle \\ \langle \mathbf{Z}^T A \mathbf{X} \rangle & \langle \mathbf{Z}^T A \mathbf{Y} \rangle & \langle \mathbf{Z}^T A \mathbf{Z} \rangle \end{pmatrix} \tag{3.44}$$

where the matrix $A_{i,j}$ is given by

$$A = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & 0 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix}. \tag{3.45}$$

To include the hydrodynamic term a matrix Q is introduced. Q diagonalises HA and A, with eigenvectors $\{\lambda_i\}_{i=1}^N$ and $\{\mu_i\}_{i=1}^N$ respectively, i.e.

$$Q^{-1} H A Q = \begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \mu_N \end{pmatrix}.$$

Now introduce the axes

$$\xi = Q^{-1}\mathbf{X} \tag{3.46}$$

$$\eta = Q^{-1}\mathbf{Y} \tag{3.47}$$

$$\zeta = Q^{-1}\mathbf{Z}. \tag{3.48}$$

These correspond to the principal directions in which the hydrodynamics stretch (or compress) the DNA molecule.

3.7.4 The Peterlin distribution

Peterlin has derived the distribution of freely jointed chain segments in shear flow, reported in a footnote by Zimm [121]. The equations of motion include a Brownian motion term, forces exerted by neighbouring atoms and an Oseen-Bergers form of interaction with the fluid given the presence of the rest of the chain. Using the notation above the distribution density of ξ,η and ζ for each molecule (where we suppress the subscripts, including for μ and η) is given by

$$\psi = \psi'(\xi,\eta)\psi''(\zeta) \tag{3.49}$$

where

$$\psi'(\xi,\eta) = \frac{\exp\left\{-\frac{1}{2(1-\rho)^2}\left[\left(\frac{\xi}{\sigma_\xi}\right)^2 - 2\rho\left(\frac{\xi\eta}{\sigma_\xi\sigma_\eta}\right) + \left(\frac{\eta}{\sigma_\eta}\right)\right]\right\}}{2\pi\sigma_\xi\sigma_\eta(1-\rho^2)^{\frac{1}{2}}} \tag{3.50}$$

$$\psi''(\zeta) = \frac{1}{\sigma_\zeta\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\zeta}{\sigma_\zeta}\right)^2\right\} \tag{3.51}$$

with the mean values

$$\sigma_\xi^2 = \frac{1}{3}\langle\mathbf{r}^2\rangle_0\left(\frac{1+2\frac{p^2}{\lambda^2}}{\mu}\right) \tag{3.52}$$

$$\sigma_\eta^2 = \frac{1}{3}\frac{\langle\mathbf{r}^2\rangle_0}{\mu} \tag{3.53}$$

$$\sigma_\zeta^2 = \frac{1}{3}\frac{\langle\mathbf{r}^2\rangle_0}{\mu} \tag{3.54}$$

for which p is a non-dimensional gradient parameter

$$p = \eta_0[\eta]\frac{M}{RT}G \tag{3.55}$$

where η is viscosity, $[\eta]$ is intrinsic viscosity[2], M is molecular weight, R is the universal gas constant and T is absolute temperature. G is shear gradient defined in equation (1.15). The constant ρ is given by

$$\rho = p\frac{\left(1+2\frac{p^2}{\lambda^2}\right)^{-\frac{1}{2}}}{\lambda} \tag{3.56}$$

To find the distribution function for a chain, the distribution of all the constituent segments is multiplied. Using this, the terms in equation (3.43) can be evaluated. For example,

$$\begin{aligned}
\langle \mathbf{X}^T A \mathbf{X}\rangle =& (Q\xi)^T A (Q\xi) && (3.57)\\
=& \xi^T Q^T A Q \xi && (3.58)\\
=& \xi^T \begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \mu_N \end{pmatrix} \xi && (3.59)\\
=& \sum_{i=1}^{N} \mu_i \langle \xi_i \rangle^2 && (3.60)\\
=& \sum_{i=1}^{N} \mu_i \sigma_\xi^2 && (3.61)\\
=& \sum_{i=1}^{N} \frac{1}{3}\langle \mathbf{r}^2\rangle_0 \left(1+2\frac{p^2}{\lambda_i^2}\right) && (3.62)\\
=& \frac{1}{3}\langle \mathbf{r}^2\rangle_0 \left(N + 2p\sum_{i=1}^{N}\frac{1}{\lambda_i^2}\right) && (3.63)
\end{aligned}$$

In the original reference β and K are defined as

$$\begin{aligned}
\beta =& \quad p\textstyle\sum_{i=1}^{N}\frac{1}{\lambda_i} && (3.64)\\
K =& \quad \frac{\sum_{i=1}^{N}\frac{1}{\lambda_i^2}}{\left(\sum_{i=1}^{N}\frac{1}{\lambda_i}\right)^2} && (3.65)
\end{aligned}$$

[2]Intrinsic viscosity is defined as the extrapolation to zero concentration when viscosity is plotted measured against DNA concentration.

so equation (3.63) may be written as

$$\frac{1}{3}\langle \mathbf{r}^2 \rangle_0 (N + 2\beta^2 K)$$

Working out the other terms of J' (equation (3.43)) similarly leads to

$$J' = \frac{1}{3}\langle \mathbf{r}^2 \rangle_0 \begin{pmatrix} N + 2\beta^2 K & \beta & 0 \\ \beta & N & 0 \\ 0 & 0 & N \end{pmatrix} \tag{3.66}$$

Combining equations 3.42, 3.43 and 3.66 the dichroism of a sample can be calculated. We rewrite equation (3.42) as

$$\Delta\epsilon_{\mathbf{P}} = \frac{\langle \Delta_\epsilon \rangle_0}{\langle \mathbf{r}^2 \rangle_0} (\mathbf{P}^T \langle \mathbf{r}_i^2 \rangle J_i \mathbf{P} - \mathbf{P'}^T \langle \mathbf{r}_i^2 \rangle J_i \mathbf{P'}) \tag{3.67}$$

$$= \frac{\langle \Delta_\epsilon \rangle_0}{\langle \mathbf{r}^2 \rangle_0} (\mathbf{P}^T J_i' \mathbf{P} - \mathbf{P'}^T J_i' \mathbf{P'}). \tag{3.68}$$

For the standard linear dichroism setup the vectors $\mathbf{P}$ and $\mathbf{P'}$ are $(1,0,0)^T$ and $(0,0,1)^T$ respectively. Hence

$$\Delta\epsilon_{\mathbf{P}} = \frac{\langle \Delta_\epsilon \rangle_0}{\langle \mathbf{r}^2 \rangle_0} \langle \mathbf{r}^2 \rangle \left\{ (1,0,0) \begin{pmatrix} N + 2\beta^2 K & 0 & 0 \\ \beta & N & 0 \\ 0 & 0 & N \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \right. \tag{3.69}$$

$$\left. (0,0,1) \begin{pmatrix} N + 2\beta^2 K & 0 & 0 \\ \beta & N & 0 \\ 0 & 0 & N \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} b \right\} \tag{3.70}$$

$$= \langle \Delta_\epsilon \rangle_0 \frac{\langle \mathbf{r}^2 \rangle}{\langle \mathbf{r}^2 \rangle_0} 2\beta^2 K \tag{3.71}$$

For stiff wormlike chains, Wilson and Schellman [118] show that

$$\langle \Delta\epsilon \rangle_0 = \langle \Delta_\epsilon \rangle_r \frac{l_p}{L} \tag{3.72}$$

Although this result is rather simple, the derivation is involved and we refer the reader to the original text, suffice to say that the derivation assumes a freely jointed chain model and lp is derived implicitly from definition 1. Thus,

$$\Delta\epsilon_{\mathbf{P}} = \frac{2}{3}\langle\Delta\epsilon\rangle_r \frac{l_p}{L}(\eta\left([\eta]\,\frac{M}{RT}G\right)^2 K \tag{3.73}$$

where β has been quoted as

$$\beta = [\eta]\,\frac{M}{RT}G \tag{3.74}$$

If both sides are divided by isotropic extinction then substitute the geometrically derived equation for reduced LD of a single segment (recalling that α is the angle between the axis of orientation and the electric dipole):

$$\frac{\langle\Delta\epsilon\rangle_r}{\Delta\epsilon}_r = \frac{3}{2}(3\cos^2\alpha - 1) \tag{3.75}$$

then

$$LD^r = (3\cos^2\alpha - 1)\frac{l_p}{L}(\eta\left([\eta]\,\frac{M}{RT}G\right)^2 K \tag{3.76}$$

All of these quantities are known, estimated or measurable and thus this is an equation that can be used to evaluate persistence length from experimental LD data.

Models of transient electric birefringence

A closely related technique to LD, in the context of analysing data for l_p, is transient electric birefringence (TEB). It will be discussed further in the following chapter. Here we wish to discuss its modelling briefly since it is closely related to modelling for LD.

Several years after Wilson and Schellman published the model described above, Hagerman and Zimm [39] criticised methods that use preaveraging of hydrodynamic interactions. We wish to emphasise that this is not a problem with the approach taken by Wilson and Schellman. The criticism was particularly directed at the work of Hearst who avoided the Rouse-Zimm approach [45], although Hagerman and Zimm calculated adjustments to the Hearst result which have been shown to give more accurate results than other models.

We describe the method in brief. Firstly, Gaussian chains are generated using a Monte Carlo method. The force field on each chain element is calculated

from motion of the chain, friction of the buffer and friction due to other chain elements (using an Oseen-Burgers hydrodynamic interaction tensor). The hydrodynamic equations are solved to provide the rotational diffusion coefficient. The results of this step were considered exact and used to provide a polynomial correction for the method of Broersma [8]. Recently this method was confirmed as providing good results when calculating persistence length [63].

3.7.5 Discussion of the model

As mentioned above this is the only model available for calculation of persistence length from linear dichroism measurements. In the following chapter experiments comparing the model to LD data are presented and show fundamental differences to the model of Wilson and Schellman. We suggest that the reasons for this lie in the hydrodynamic approximations of the model and the presence of large bends in DNA during LD measurement.

The construction of the theory above relies on several assumptions. One is that DNA perturbations are small. The majority of measurements of persistence length have used this approximation and deviation has seldom been found. However, in the presence of shear flow recent modelling has shown evidence of the DNA being able to sharply bend back on itself (Lahari de Alwis, personal communication). This phenomenon is not surprising since parts of the DNA closer to the non-rotating cylinder of the cell experience less force than those nearer the rotating cylinder and will hence decelerate. This introduces bends much larger than those for which the approximation is valid. In addition, the potential flexibility of DNA is shown by its packing into small spaces within chromosomes and virions.

Furthermore, the DNA we have used is an order of magnitude longer than that used for the majority of persistence length experiments. This presents the possibility of significant effects from higher order hydrodynamic effects.

Recently, to assess the importance of higher order hydrodynamic effects, principal component analysis (PCA) has been used to classify the modes of motion of DNA [18]. PCA was used to determine which conformations of DNA are found most commonly in videos of fluorescently labelled DNA taken at a rate of 4.5 ms per frame. Lambda DNA, with a length of 48,502 base pairs was placed in a flow cell with a variable potential difference across to keep the DNA centre of mass in the centre of the frame. The potential difference was adjusted each frame to counteract Brownian motion without altering the overall kinetics of the DNA. The principal components showed evidence of deviation from the Markovian linearity of eigenstates predicted by Zimm Rouse theory, i.e. the

DNA mode expressed in the basis of principal components (the eigenstate) **a** should evolve over time as follows

$$\mathbf{a}(k+1) = M\mathbf{a}(k) + \xi(k) \tag{3.77}$$

where k indexes the frames, M is a transition matrix and ξ is a vector of white noise. Deviations from this indicate higher order hydrodynamic effects not predicted by the original theory. The conclusion was that there is strong evidence that the motion of a given portion of a DNA helix is dependent on the motion of the parts of DNA in close proximity to it. These motions caused deviation from equation (3.77) and thus higher order terms of the hydrodynamics are important.

Our experimental disagreement with theory (chapter 4), De Alwis' observations from modelling and the PCA analysis suggest that non-linear hydrodynamic effects are important in the motion of long DNA in Couette flow and that Zimm-Rouse hydrodynamic theory is not enough to encapsulate the behaviour of DNA.

Relation to transient electric birefringence

The force field employed by Peterlin and Zimm, as used by Wilson and Schellman, is almost identical to that assumed to be exact by Hagerman and Zimm for TEB. It may, therefore, be prone to the same drawbacks as Wilson and Schellman's approach, although this has not been discussed in literature concerning TEB.

There may be useful lessons in the TEB analysis, however. The method of calculating corrections to a current equation using a more exact, time-consuming, simulated result may be a way of introducing the necessary higher order terms into the equations for LD. A coarse-grained model with known persistence length, such as that employed by De Alwis, can generate a number of DNA chains. The Wilson and Schellman equations can be applied to these chains and a term polynomial in some parameter of the system could be derived as a correction. Such a solution is beyond the scope of the work here, but highly commended to those continuing the analysis of DNA using LD.

Chapter 4

Measurement of persistence length with linear dichroism

4.1 Chapter summary

The experimental measurement of persistence length of DNA with linear dichroism (LD) is investigated. Previous methods of persistence length measurement are summarised in three categories: single molecule techniques, microscopy and solution methods.

LD is used to measure three different DNA plasmids (made linear with a restriction enzyme). Results show that LD is linear with shear gradient, at odds with the theory presented in the previous chapter. This result precludes the calculation of persistence length, but the effect of salt concentration, temperature, DNA length and sequence are assessed. Response to salt concentration appears sequence dependent. Temperature has a minimal effect in the range measured. Longer DNA did not necessarily lead to larger LD.

4.2 Introduction

Interpretation of linear dichroism (LD) is heavily dependent on knowledge of sample orientation. In the case of DNA in solution this is inextricably linked to the property of persistence length, denoted l_p, as introduced in the previous chapter. This chapter is a review of experimental determination of persistence length and reports LD experiments carried out with the aim of probing the variation of persistence length with salt concentration, sequence length and temperature.

4.3 Previous measurements of persistence length

A number of techniques have been used previously to determine persistence length. These fall into three categories: microscopy, single molecule techniques and solution methods.

4.3.1 Microscopy

Electron microscopy (EM) [3, 32] and atomic force microscopy (AFM, also called scanning force microscopy) [42, 93, 69] have been used to measure persistence length. The former is an apparatus analogous to a conventional microscope, but using electrons in a vacuum as opposed to photons to image samples, whilst the latter uses a fine cantilever to scan across the sample and detects changes in height. To obtain a value of persistence length the DNA is dried onto a surface (such as mica) or frozen (for cryo-EM), an image is taken and pixellated, and measurements of the polymer properties made.

It has been shown that freshly cleaved, untreated mica is the best surface for deposition [93]. A charged surface does not allow the DNA to deform and it becomes a projection of the 3-dimensional geometry onto a 2-dimensional surface. A desposition like this is distorted since the process preserves total contour length. An uncharged surface allows structural transitions that give a 2-dimensional polymer with the same properties as in 3 dimensions, meaning that the persistence length measured from the 2 dimensional picture is the same as that in solution.

For cryo-EM the sample is usually confined to a solution layer 40–50 nm thick, which alters the persistence length of the DNA in a non-trivial way [3]. Simulations have shown that if, however, the DNA is not significantly longer than the layer thickness (i.e. $<$150 base pairs (bp)) the alteration is less than 5%.

For microscopy, the properties from which persistence length is calculated are usually the ratio of end-to-end distance to contour length, or the average cosine of angular deflection. These methods are generally recognised as the most direct and reliable route to obtaining a value for persistence length since once the image has been measured the analysis is relatively simple. The disadvantages are that it cannot be carried out in solution and the range of conditions for measuring the DNA are not as varied as, for example, solution techniques.

4.3.2 Single molecule techniques

The single molecule class of methods for measuring persistence length covers several procedures. These include optical tweezers [93], magnetic beads [104] and tethered particle motion [80]. The general approach is the same for all of them: DNA is tethered to a surface at one end and to a movable particle at the other. The force on the particle is measured as the polymer is stretched and the information required is contained in the force-extension graph.

The main difficulty with these methods is the interpretation of the data. A successful theory, though, has been proposed by Bustamante *et al* [9] and is generally accepted as giving accurate results. Theory and experiment for DNA of pre-determined persistence length agree in the limits of long and short extension, but was shown to differ by up to 10% in between these limits. Improvements have been suggested using corrections based on numerical solutions to the worm-like chain (e.g. [7]).

There are a number of advantages of using these methods. One is a high level of precision: forces can be detected down to 0.1 pN [116]. They can also be carried out in solution and over-stretching allows observations of a novel DNA conformation, termed S-DNA. The length of DNA is, however, limited, dependent on the specific instrument used. Typically the DNA is shorter than 2,000 bp. Since only one molecule is used per measurement, an average should be taken over a number of molecules for a reliable result. The tethering of both ends of the DNA negate the observation of long-distance interactions, which can be considered as either a positive or a negative since the measurements are easier to interpret, but less similar to the normal conditions in which DNA is found.

4.3.3 Solution methods

We use the term solution methods for those that make bulk observations of DNA in solution. Of these, transient electric birefringence (TEB), cyclisation

and linear dichroism have been used to measure persistence length.

Transient Electric Birefringence

TEB is similar to LD in that it exploits anisotropic properties of molecules to produce a signal. Optically anisotropic molecules can be aligned using a square electric pulse. Typical values for the field are 5 $kVcm^{-1}$ for 8 µs. Light passed through the sample experiences altered refractive indices in perpendicular directions of polarisation. The difference between the two indices is defined as the birefringence. It is measured over time and the field-free decay time τ, characteristic of the time to reattain random orientation, can be found. τ has been related to diffusion coefficient and consequently persistence length. This relations require a polymer model of DNA.

A careful review of TEB measurements has been published by Lu *et al* [63] which concludes that the model producing the best value is that of Broersma with corrections terms added by Hagerman and Zimm, which is discussed in the previous chapter (section 3.7.5). The correction terms were found by comparing the solution to the Broersma equation with a Monte Carlo simulated chain of known persistence length.

Cyclisation

Cyclisation experiments determine persistence length from the closure probability of DNA fragments [111, 101]. Binding of the ends is promoted, either by leaving overhanging complementary sequences at each end or by adding a ligating protein. Ligated and unligated DNA is separated and quantified using agarose gel electrophoresis. Analysis based on the worm-like chain has been carried out [100], although it is noted that large fluctuations (caused by long chains) can lead to errors.

Linear dichroism

The modelling aspect of LD measurements of persistence length is dealt with in chapter 3. Measurements of persistence length using LD have had limited success. Wilson and Schellman used T7 DNA [118] and estimated persistence length to be 35 nm, slightly shorter than the generally accepted values. Rizzo and Schellman attempted to improve on this estimation, using the same equations, and generated a value of 47 nm with 0.2 M NaCl.

Facets of solution methods

Solution methods are useful for quickly obtaining average measurements over many molecules, which can be a long and pain-staking process for alternative methods. They are also quicker, cheaper and experimentally simpler to carry out. The range of conditions is generally flexible, as long as the solution does not contain another molecule that exhibits the property being measured. For example, chlorine ions absorb strongly in some regions of UV so can be problematic for LD. They are also especially suitable for long DNA samples.

Evidently, the properties of a single piece of DNA cannot be examined with solution methods. In addition resolution is low — it can be difficult to tell which part of the molecule causes a given effect. The biggest difficulty with solution methods, though, is to find a model that accurately reflects the behaviour of DNA in solution.

4.4 Experimental procedures

4.4.1 DNA

To measure persistence length three types of DNA with different sequence lengths were used: the plasmids pBR322, pUC19 and a pC3.1 variant (which is denoted pC3.1v). DNA plasmids have homogeneous sequence lengths. The native circular plasmid form was converted to linear using the restriction enzyme EcoRI. Each DNA sequence contained a unique EcoRI recognition site. Two of the DNAs were purchased and one made using a maxiprep amplification procedure.

The experiments probe the effect of rotation speed, temperature, salt concentration and DNA length.

Purchased DNA

pBR322 and pUC19 DNA were purchased from New England Biolabs (NEB, Massachussetts, US) at concentrations of 1 mg/ml and were supplied in 10 mM Tris-HCl (pH 8.0) and 1 mM EDTA. This was modified to 20 mM Tris-HCl and 2 mM EDTA in order to match the pC3.1v DNA described below. pBR322 and pUC19 are plasmids of length 4361 and 2686 base pairs respectively.

DNA amplified in the laboratory

The following subsections describe the process of making pC3.1v DNA, purchased initially from Invitrogen (Paisely, UK), by transforming a plasmid in

to competent bacteria (these have cell walls that are permeable to DNA). The bacteria are grown and the DNA extracted. It has a length of 6882 base pairs.

Competent cells

To make thermally competent cells the following method was followed. A culture of DH5α *Escherichia coli* (*E. coli*) cells was grown overnight and 1 ml added to 30 ml of luria broth (LB). The cells were grown to an optical density (OD) of about 0.3 and harvested by centrifugation at 10,000 rpm for 5 minutes. They were then resuspended in 15 ml of chilled, filtered 100 mM calcium chloride. After being left on ice for 20 minutes the cells were resuspended as before then 3 ml of calcium chloride solution were added. After a further hour on ice, 1 ml of 80% glyerol was added and the cells were stored in 200 µl aliquots at -80 °C.

Transformations

To put pC3.1v DNA plasmid into the competent *E. coli* cells, a heat shock transformation procedure was used. 30 µl of thermally competent cells were added to a 1.5 ml Eppendorf tube on ice. 1 µl pC3.1v DNA was added and the mixture incubated on ice for 20 minutes. This allows the DNA to align at the pores in the cell walls. The cells were placed in a water bath at 42 °C for 30 s. This causes the DNA to enter the cell cytoplasm. The transformed cells were left on ice for 2 minutes before adding 800 µl LB. They were then incubated at 37 °C for 1 hour in order to start the growth process. The cells were harvested by centrifugation at 13,000 rpm for 5 minutes. The majority of the supernatant was removed and the pellet resuspended in approximately 30 µl of the remaining LB. The resuspended mixture was spread onto an agar plate containing ampicillin and incubated overnight (roughly 16 hours).

Maxiprep

To obtain large amounts of DNA the transformed bacteria are grown and the DNA extracted using a process known as a maxiprep. A single colony from the agar plate as described in the previous subsection was picked and used to inoculate 250 µl LB containing 0.1 mg/ml ampicillin. The bacteria were grown overnight and suspended by centrifugation at 5,000g for 15 min.

To extract the plasmid DNA the Marligen maxiprep protocol was followed [75], after which Tris-HCl was added to a concentration of 20 mM and EDTA to a concentration of 2 mM.

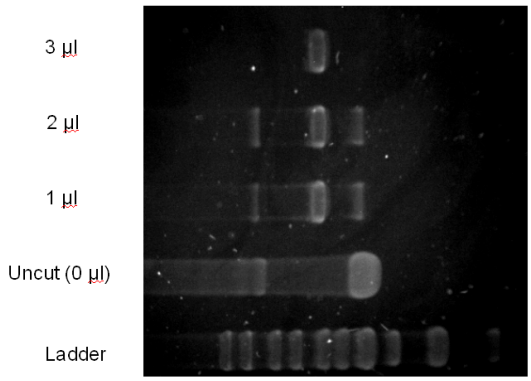


Figure 4.1: A 1% agarose gel of pBR322 DNA digested with various amounts of EcoRI. 3 µl 10X EcoRI buffer and 6 µg DNA were used in a total of 30 µl. The lanes are labelled with the volume of enzyme used. The ladder values from right to left in kbp are 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 8 and 10.

Enzymatic restriction

In order to carry out LD the plasmid DNA, which was circular, was cut once using the EcoRI type II restriction enzyme protein (NEB). Type II restriction enzymes cut DNA at specific points defined by a short, usually palindromic sequence (see chapter 7). EcoRI cuts after the first base of the motif GAATTC. Note that this leaves an overhang of 5 bases.

For all three types, 6 µg DNA were used with EcoRI buffer at 10X concentration, as supplied with enzyme (3 µl) and 3 µl EcoRI enzyme. The sample was made up to 30 µl with MilliQ water in a 1.5 ml Eppendorf tube and placed in a water bath at 37 °C for 5 hours. 2 µl of the product were run on a 1% agarose gel to check that the process had finished. A long incubation time was used to denature the protein and ensure it would not bind to the DNA non-specifically and thus affect the persistence length. Restriction enzymes do not bind only to their recognition site and such binding may affect DNA curvature. Various concentrations of EcoRI were tried and 3 µl was found to be the smallest that completely digested the DNA (Figure 4.1).

4.4.2 Spectroscopy

LD of the DNA was measured using a Biologic MOS-450 spectrometer. The settings used for spectra were as follows

- Wavelength: 180 – 350 nm
- Rotation voltages: 0, 1, 2, 3, 4 V
- Slit widths: 4 nm
- Time base: 0.5 s
- Accumulation: 3
- Temperature: 25, 30, 35, 40 °C

The three types of DNA, linearised using EcoRI as described above, were made up to 60 µl using sodium chloride solution to give final salt concentrations of 0.1 M, 0.5 M and 1 M and DNA concentration of 303 µM. The buffers for enzymatic restriction and DNA stocks gave concentrations of the following chemicals in the samples measured with LD: tri-HCl 50 mM, EDTA 0.2 mM, potassium phosphate 0.5 mM, 2-mercaptoethanol 0.5 mM, $MgCl_2$ 5 mM, triton X-100 0.08%, glycerol 2.5 %, bovine serum albumen 10 µg/ml. These were assumed to give a negligible contribution to the behaviour of DNA under the conditions measured in comparison to the relatively large concentration of sodium chloride.

Time course measurements were also taken using the following parameters

- Wavelength: 260 nm
- Rotation voltages: 0, 1, 2, 3, 4 V
- Slit widths: 4 nm
- Time base: 1 s
- Time length: 200 s
- Temperature: 25, 30, 35, 40 °C

During each scan the voltage was changed every 30 s, with the subsequent 10 s of data disregarded to allow the sample to adjust to the new rotation speed.

The method used to measure an LD spectrum was as described in section 2.3.4. Before plotting, the spectra had the average value between 330 and 350 nm subtracted from it, since the signal is known to be zero at this point.

To calculate reduced LD the spectra were divided by natural logarithm of the value of HT of the sample (HT measured divided by HT of buffer) at 260 nm. Note that this provides a value that is proportional to the true reduced LD, however it is still independent of concentration.

4.5 Results

4.5.1 Controls

A number of checks were carried out to ensure that the measurements were as accurate as possible and to eliminate unwanted effects.

LD of buffer

It is standard procedure with linear dichroism to subtract a buffer spectrum from each sample spectrum. LD of the buffer was measured with the same parameters and conditions as the DNA and was found to exhibit no LD (data not shown).

Non-specific binding of enzyme

Because a restriction enzyme is present in the sample during LD measurement, there was a concern that it could bind non-specifically to the DNA and alter the flexibility, thus altering the persistence length. The effect was expected to be relatively small since restriction enzymes are very specific and only one binding site was present for each DNA plasmid. To counter this the samples were held at elevated temperature (37 °C) for an extended period, to denature the enzyme without affecting the DNA.

Rotation voltages

Rotation of the capillary for LD is controlled by the potential difference across the motor. To verify the relationship between rotation in rpm and potential difference in volts a black line was drawn down the side of a capillary, buffer put into the capillary and the LD measured using:

- Wavelength: 260 nm
- Rotation voltage: 0, 0.2, 0.4, ... 3.0 V
- Slit widths: 4 nm
- Time base: 500 µs
- Time length: 2s
- Temperature: 30, 40 °C

Each time the black mark passes through the beam a spike appears in the LD. The number of spikes per time unit gives the rotation rate of the cell. A typical plot is given in figure 4.2.

Figure 4.3 shows the data collected at 30 °C and 40 °C. A linear fit (using Origin, Northampton, MA, USA) was used to shear gradient Ω to potential difference Λ using

$$\Omega = A + B\Lambda \tag{4.1}$$

for points above which the rate was zero ($\Sigma \geq 0.4$ V). The null values for the first two points are attributed to frictional forces in the cell stopping it from spinning at low potential difference. At 30 °C the parameters were $A = -52.7 \pm 12.5$, $B = 580 \pm 6.6$ and at 40 °C, $A = -50.5 \pm 8.8$ and $B = 591 \pm 5$. The two lines in figure 4.3 are very similar, so we conclude that there is no significant difference between the two temperatures.

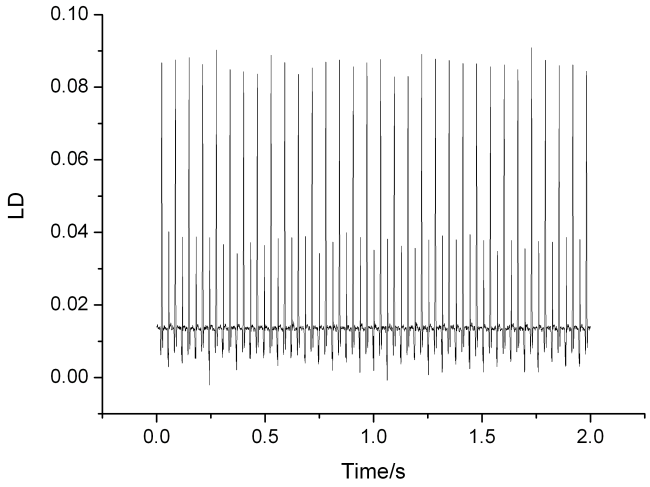


Figure 4.2: A plot to determine the rate of rotation of the LD capillary at a given voltage. These data were measured at 30 °C at a rotation of 1 V. Note that two sets of spikes appear, the taller corresponding to the black line passing through the light beam on the near side and the shorter on the opposite side.

4.6 Rotation

Equation (3.71) states that LD is proportional to the square of the rotation gradient, as defined in equation (1.15).

LD was measured at various rotation speeds for each sample. These were

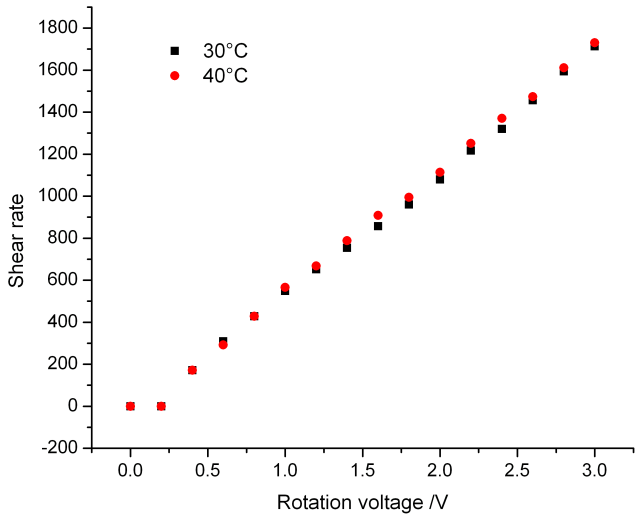


Figure 4.3: Rotation of an LD capillary versus potential difference across the rotation motor measured at 2 different temperatures.

0, 1, 2, 3 and 4 V. Figure 4.4 shows typical results for these experiments. For pC3.1v DNA with 0.1 M salt at 30 °C the LD was recorded at intervals of 0.2 V from 0 to 3.0 V, as shown in Figure 4.5. The data is linear and fits to the equation

$$LD = 0.01408 - 0.00173\Lambda. \tag{4.2}$$

For pBR322 DNA at 0.1 V, 30 °C the data fits to

$$LD = 0.01163 - 4.64 \times 10^{-4}\Lambda \tag{4.3}$$

Thus, since LD is linear with rotation voltage and rotation voltage is linear with rotation gradient, LD is linear with rotation gradient.

4.6.1 DNA length

It is well established that DNA length affects LD, although it is less clear precisely what the relationship between the two should be. The received wisdom, as shown by Simonson and Kubista [102] is that longer DNA gives a larger LD amplitude, since the aspect ratio is higher. Looking at the equation of Wilson and Schellman 3.71, the dependence on length is related to the eigenvalues of the hydrodynamic interaction matrix.

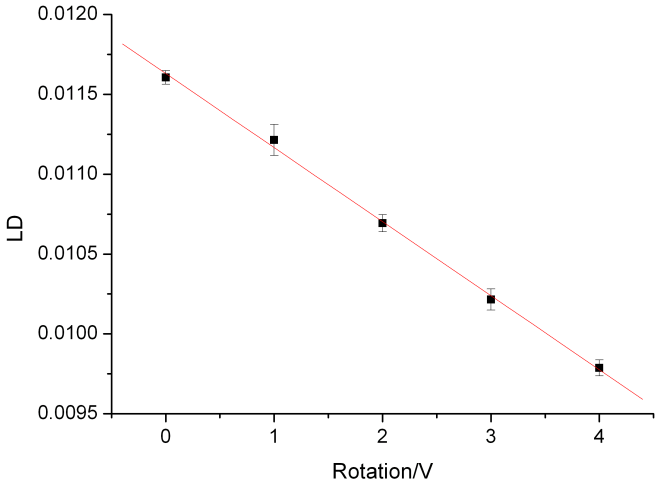


Figure 4.4: Unzeroed LD at 260 nm of pBR322 DNA with 0.1 M NaCl measured at different rotation voltages at 30 °C. A linear plot was fitted through the data. The error bars show two standard deviations calculated over the 30 s of data for each measurement.

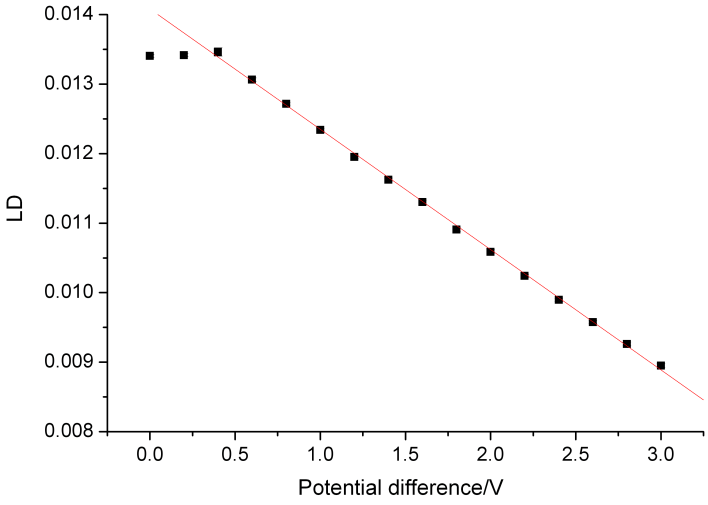


Figure 4.5: pC3.1v DNA with 0.1 M NaCl measured at 30 °C for various rotation potentials. The equation to describe the line fitted to the data is in the text.

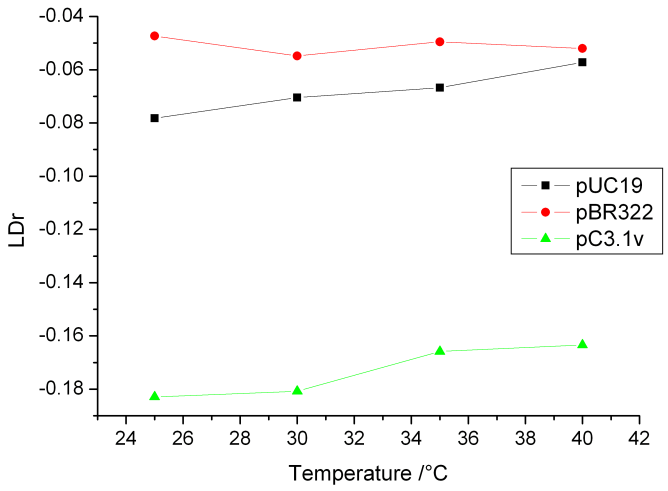


Figure 4.6: Reduced LD of the three types of DNA at different temperatures at a salt concentration on 0.1 M.

Here, three different DNA lengths were used, as detailed above. Figures 4.6, 4.7 and 4.8 show reduced LD at 260 nm for the three types of DNA at various temperatures and salt concentrations.

As expected, the pC3.1v DNA has the largest magnitude, but the two other plasmids show comparable LD with the pBR322 having slightly lower magnitudes. These results are consistent across the three different salt concentrations used. This is surprising considering the large differences between the DNA lengths. Figure 4.1 and similar ones show that the DNA has been cut and is hence linear. Figure 4.9 and other data (not shown) used gel electrophoresis to verify the DNA lengths.

4.6.2 Temperature

A relatively narrow temperature range was used because some DNAs at low salt concentrations can melt at temperatures as low as 45 °C. Generally, an increase in temperature corresponds to an increase in entropy and hence more disorder of the DNA. This process should result in a reduction of LD. However, the narrow temperature range may mean that this is not the dominant effect.

Figures 4.6, 4.7 and 4.8 show relatively little change of LD with temperature. Both pC3.1v and pUC19 show slight decreases, whereas the data for pBR322 are less clear. Measurement of calf thymus DNA in the presence of no salt have shown a complete drop-off of LD in the temperature range used (Emma Gilroy, personal communication). This has been attributed to a decrease in viscosity.

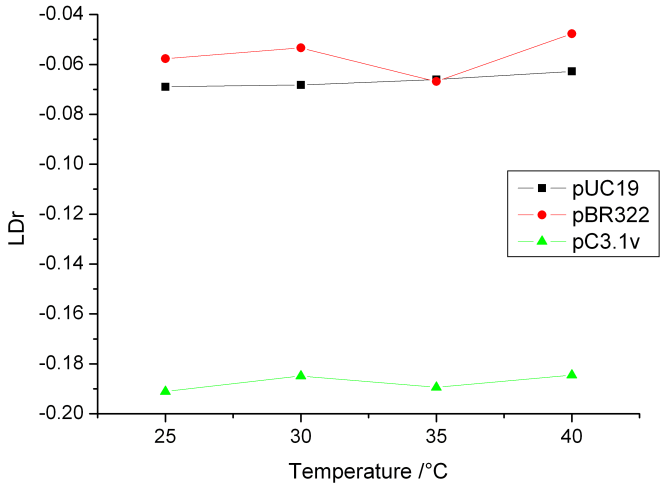


Figure 4.7: Reduced LD of the three types of DNA at different temperatures at a salt concentration on 0.5 M.

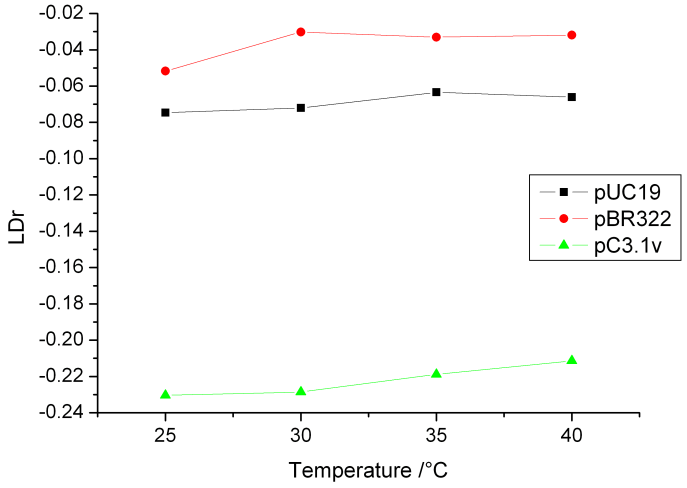


Figure 4.8: Reduced LD of the three types of DNA at different temperatures at a salt concentration on 1 M.

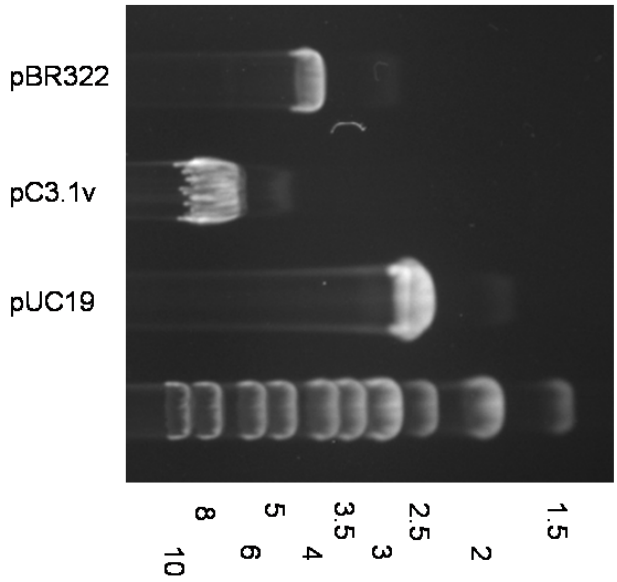


Figure 4.9: Electrophoresis of the three types of DNA after restriction. The length is indicated by the ladder, values of which are shown underneath in kbp.

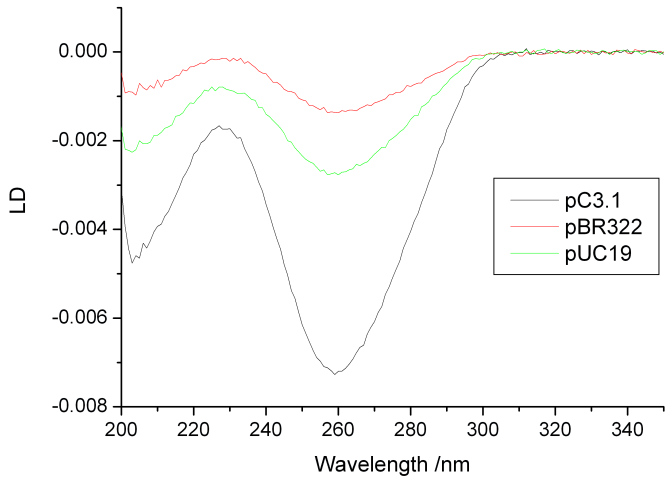


Figure 4.10: LD of the three types of DNA measured with 0.5 M sodium chloride at 4 V rotation. The data are zeroed between 330 and 350 nm.

4.6.3 Salt concentration

The effect of low salt concentrations on DNA is discussed elsewhere in this work (chapter 2).

pBR322 DNA showed negligible difference with salt concentration between 0.1 M and 0.5 M and a decrease in LD up to 1 M, as shown in figure 4.15. pUC19 DNA data is shown in figure 4.16 and exhibits no significant difference with salt concentration. The 0.1 M salt data appear to have more temperature dependence than other salt concentrations. In figure 4.14 the data for pC3.1v DNA shows a slight increase for 1 M salt.

We have used relatively high salt concentrations for two reasons: firstly, to match the concentrations used by Rizzo and Schellman [94] and also for the salt to be a significantly greater effect than the other chemicals in the buffer. The data show that in this regime the changes in LD are relatively small compared to the overall magnitude. It appears that we may have measured above the limit at which persistence length, and hence LD, varies significantly with salt concentration.

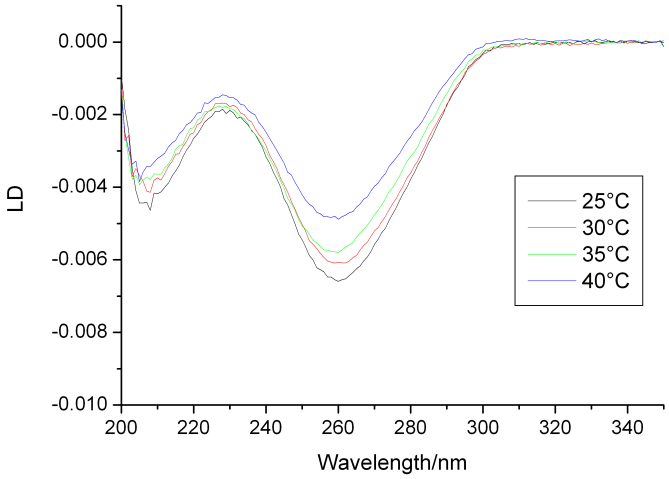


Figure 4.11: LD of pC3.1v DNA containing 0.1 M NaCl measured at various temperatures. The plots were zeroed between 330 and 350 nm.

4.7 Discussion

4.7.1 Persistence length

As discussed extensively in the previous chapter only one model has been reported for relating LD of DNA to persistence length. We have clearly shown

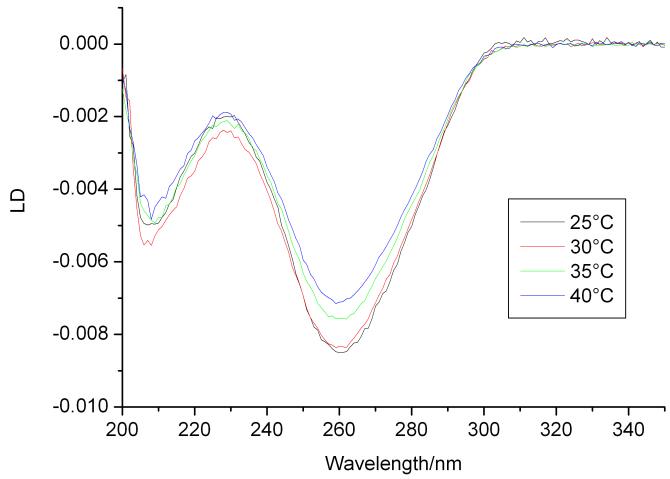


Figure 4.12: LD of pC3.1v DNA containing 0.5 M NaCl measured at various temperatures. The plots were zeroed between 330 and 350 nm.

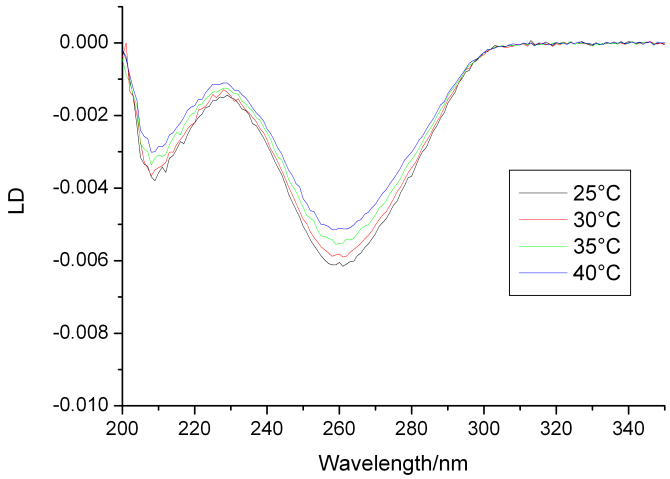


Figure 4.13: LD of pC3.1v DNA containing 1 M NaCl measured at various temperatures. The plots were zeroed between 330 and 350 nm.

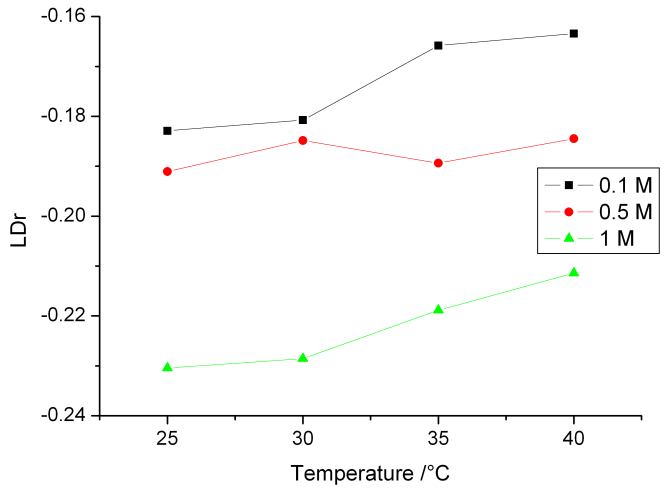


Figure 4.14: Reduced LD at 260 nm of pC3.1v DNA at different salt concentrations and temperatures.

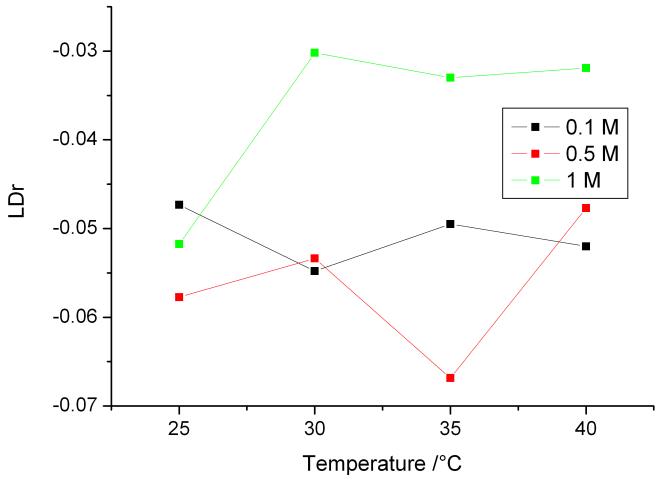


Figure 4.15: Reduced LD at 260 nm of pBR322 DNA at different salt concentrations and temperatures.

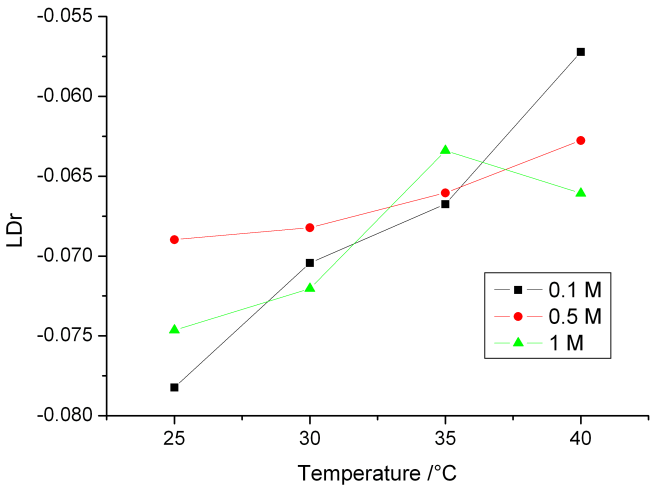


Figure 4.16: Reduced LD at 260 nm of pUC19 DNA at different salt concentrations and temperatures.

in the work presented here that LD of DNA is linear with respect to shear rate. Wilson and Schellman, however, predict it to be quadratic (see equations 3.55, 3.64 and 3.71). In the work of Simonson and Kubista [102] a quadratic form of dependence is found only for low gradients (<300 s^{-1}) and short (<3000 bp) DNA. These values are outside the range of the work here as well as being dissimilar to typical DNA measurement parameters with a capillary LD cell. The aim of the experiments described in this chapter was to calculate persistence length of DNA. However, basic facets of the model greatly differ from those measured and thus use of LD to measure persistence length is not possible with current modelling capabilities. A more modest aim is to identify the factors that most affect experimentally measured LD at 260 nm and should inform any model for calculating persistence length of DNA from LD.

4.7.2 Temperature

Temperature is a variable that appears in almost all models of persistence length, yet has been little studied as an effect. This may be partly because the temperatures of interest for DNA lie in a relatively small range. We suggest that temperature can play a significant role in measurements because it alters both thermodynamic motion of DNA and viscosity. Between 20 °C and 60 °C the dynamic viscosity of water more than halves (in a non-linear fashion). Both viscosity and temperature appear in the equation linking LD and persistence length.

The data here are in general agreement with Lu *et al* [64], who observed

a small decrease in persistence length in this temperature range. It would be beneficial to have below-ambient temperature control since the temperature range cannot be extended higher than 40 °C due to melting processes. The changes appear to be dependent on the type of DNA. In the study with calf thymus DNA mentioned above, LD disappeared completely over the same temperature range.

4.7.3 Salt concentration and DNA type

The most studied effect with persistence length is salt concentration. The salt concentrations chosen here match those used in a previous study with LD [94], although are significantly higher than many other studies, as discussed above. Our measurements show that salt has an effect on LD magnitude, although not as would be expected, with DNA type appearing to play a more significant role.

Neither the effects for salt or DNA type fitted with our expectations. For salt, a decrease of LD magnitude with increasing concentration is expected. This trend was not observed. In addition, longer DNA would be expected to produce larger LD magnitude. This is true to an extent, with pC3.1v DNA giving the largest magnitude measurements, but LD of the other two DNA types are barely distinguishable despite a factor of two between their lengths. This suggests that something else is playing a role in determining the orientation of these DNA types. Certain short motifs are known to create intrinsic bends in DNA, although precisely what rules govern the effect are unknown. Cyclisation studies with short pieces of DNA [111] have estimated the contribution of intrinsic curvature to be typically 20 times less than that due to thermal fluctuation. It is possible, although has not been reported, that pUC19 DNA could contain a large number of such motifs and hence have a much reduced persistence length. Alternatively, since LD dependence on length is not well-characterised, long-range interactions could play a much greater than expected role for pC3.1v DNA.

4.8 Conclusion

We have tried to find the persistence length of DNA using LD, where the DNA length is an order of magnitude longer than in previous studies. This has proved problematic. In the previous chapter problems with the modelling structure available, particularly hydrodynamics, were discussed. In this chapter we have tried to identify the experimental factors that have the most effect

on persistence length of the DNA by assessing their effect on the magnitude of LD at 260 nm. In doing so, we have made the assumption that LD at 260 nm decreases with an increase in persistence length and vice-versa. The ultimate goal of this work is to quantify this relationship, but that has not been achieved thus far.

Consistent with previous results, temperature plays a small part in determination of persistence length at high salt concentration. DNA length is an unresolved issue and further studies with DNAs of different type and the same length may aid clarification of the results. It appears that DNA sequence is the most predominant factor in our studies. This may explain why previous studies have produced contradictory results [63].

In conclusion, we maintain that it is important to work towards studying DNA with *in vivo* sequence length. This requires a rethink of modelling approaches and more studies of the experimental properties of DNA in such regimes using techniques such as LD.

Chapter 5

Temperature control of linear dichroism samples

5.1 Chapter summary

The temperature control ability of a current prototype thermostatted Couette flow cell is assessed. The cell controls temperature via a temperature jacket. The temperatures to which the jacket is set are compared with measurements from a thermocouple placed in the sample chamber. A sample of water takes up to 13.3 mins to equilibrate when the temperature is altered by 5 °C. With the motor running temperature can be accurately controlled to within 1 °C. This level of control is not sufficient for the requirements of many kinetic experiments.

An air flow through the cell did not help attain the set temperature. Two solutions are proposed. Firstly, an injection system where sample can be loaded directly into the annular gap of the Couette flow cell through a hollow inner capillary. Secondly, key attributes of a redesigned Couette flow cell are presented.

5.2 Introduction

As discussed in the introduction (section 1.2.5), recent advances in the design of linear dichroism (LD) cells have opened the way to much smaller sample sizes relevant to biochemical molecules such as DNA and proteins. There are few constraints on conditions in which LD can be measured, providing the sample does not contain a strongly absorbing species other than the one being measured, the viscosity is such that orientation of the molecule of interest is possible and that there is no reaction with the quartz or epoxy resin plug. This chapter reports how well temperature can be controlled with the prototype LD cell used at the University of Warwick with a Biologic MOS-450 spectrometer.

Initial motivation for this work was from kinetic studies of restriction enzyme action reported in chapter 7. It was proposed that the time for a sample to equilibrate to the temperature required in the reaction was affecting the kinetic measurements. This could alter both the reaction rate and the reported LD, since restriction enzymes are very sensitive to temperature and LD magnitude is temperature dependent (chapter 4). Controlling temperature is key to quantitative interpretations of kinetic data.

Despite the commercial availability of thermostatted capillary LD cells for several years no systematic study of this kind has yet been carried out.

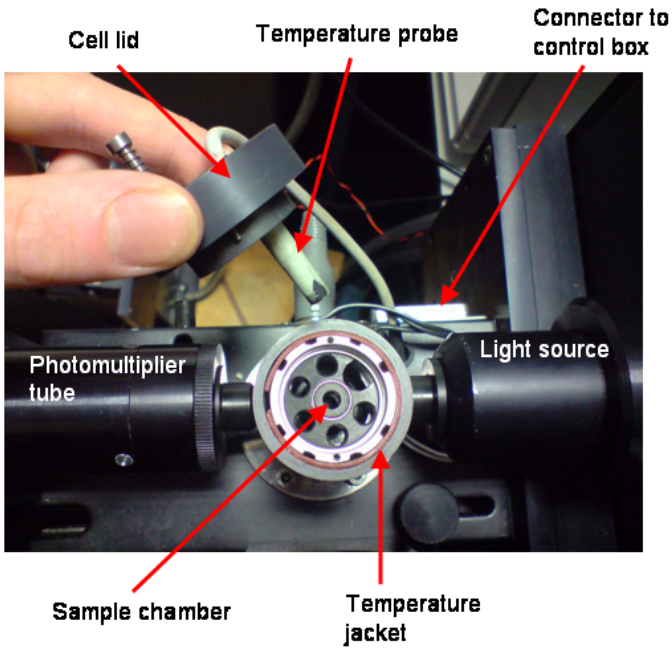


Figure 5.1: The LD cell viewed from above. The capillary slides into the sample chamber. Note the distance between the temperature jacket, which appears as a brown layer, and the sample.

5.3 Materials and Methods

For LD a Biologic MOS-450 spectrometer was used with a prototype Crystal Precision Optics thermostatted LD cell. The temperature control of the sample was carried out using a jacket around the cell attached to an external control box and a temperature probe located inside the cell, figure 5.1. The jacket heats only and cooling is attained by radiation of heat from the cell.

The temperature was measured using a Picotechnology (Eaton Socon, UK) TC-08 temperature recorder with three type J thermocouples. This device is attached to a computer which can be programmed to record the temperature at regular intervals. The settings used were:

- Collection mode: real-time continuous
- Data frequency: 10 s or 30 s
- Readings per sample: 'as many as possible'

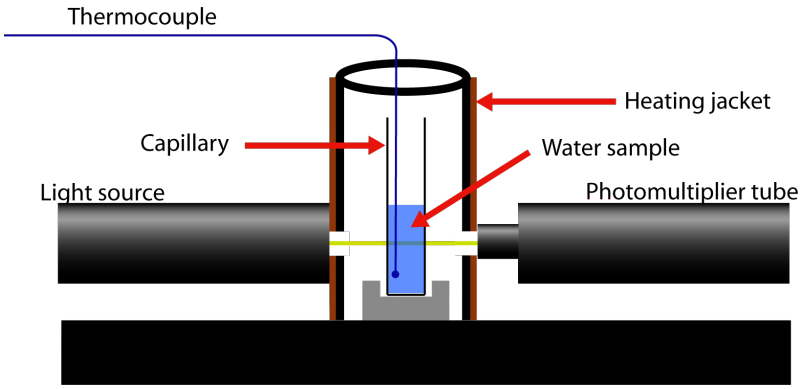


Figure 5.2: Schematic of experimental setup. When a rod was used this was placed in the centre of the capillary. When a rod was not used, the top of the cell was covered to minimise exchange of heat with the surroundings.

The main experimental setup is given in figure 5.2. The thermocouples were placed as follows: one away from the spectrometer to measure the ambient temperature; another on the cell, outside the sample chamber and the third in a quartz capillary in the sample chamber (as indicated in figure 5.2). For all of the experiments described, the sample used was 60 µl of MilliQ water placed in the capillary. The end of the thermocouple was placed in the water.

Some of the experiments also included a quartz rod and thermocouple in the capillary. Although this is closer to the normal operational setup for the spectrometer it could not be used routinely, since it would have resulted in damage to either the rod or capillary from the thermocouple when the motor was run. The data shown in figure 5.6 were recorded using a rod,

but repeats (figure 5.11) without a rod (and the same volume of water) were almost identical, indicating that the thermodynamic properties are not altered by omitting the rod.

When used, the motor that rotates the capillary was run at 4.0 V, a typical value for LD experiments. The motor used in the cell is from RS components (Kettering, UK), stock number 262-2135 and can be used up to 6 V with a maximum rotation of 6846 rpm.

The sample temperature was measured using thermocouples as described above. The temperature of the jacket, when measured, was read from the control box. Air was used to control the temperature for some measurements. This was flowed through the side of the cell lid — the inlet for the air is obscured in figure 5.1. Air flow rate was measured using a flow-meter on the incoming pipe.

5.4 Results

The influence of the following on control of temperature during LD were examined:

- the motor that drives quartz capillary rotation
- temperature control jacket
- air flow through the cell

As far as possible these effects have been considered in isolation and in combinations that mimic experimental conditions.

5.4.1 Motor

Like many electrical components, the motor that rotates the quartz capillary generates heat. It is located directly under the sample chamber a few centimetres away from the sample. With the sample at room temperature the motor was switched on and run at 4.0 V. Data measured over time are plotted in figure 5.3, showing a rapid initial increase in temperature, levelling off over the course of around half an hour. The following exponential equation was fitted through the data using the MCMC algorithm described in chapter 6, as shown in the figure

$$y = 2.9 - 2.6e^{-\frac{x}{398}}. \tag{5.1}$$

A 2.9 °C increase in temperature is observed with a half-time of around 120 s. The half-time is the time for the value to reach half of the difference between the start and end values and is calculated as $\log 2$ multiplied by the denominator in the exponential.

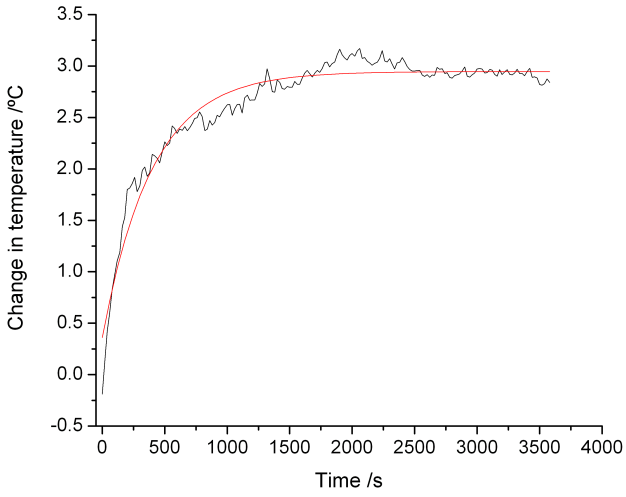


Figure 5.3: The difference between the temperature in the sample and the ambient temperature with only the motor running. Equation (5.1) shows the equation of the fitted curve.

5.4.2 Temperature control jacket

Figure 5.4 shows the effect of heating a sample from room temperature to 37 °C. The thermocouple temperature was recorded by computer and the temperature readout of the jacket control box was manually recorded at regular intervals. The graph shows that the temperature of the sample rose more than the jacket during the course of heating. In addition, the final temperatures recorded were not the same, suggesting either a calibration problem or a temperature gradient between the jacket and the sample. Surprisingly, there was no significant delay observed between the control box and the sample.

Further heating measurements were made with different start and end temperatures, shown in figures 5.6 and 5.5. All of the curves follow a similar pattern, with an initial increase in temperature and overshoot, followed by a decrease, overshoot and levelling out. Figure 5.11 shows that data collected with and without the central rod have similar discrepancies between the measured and set temperatures.

There is a period of about 1250 s (nearly 21 mins) before the plots in figure 5.6 become flat, however, within 800 s (13.3 mins) all are within 0.5 °C of the

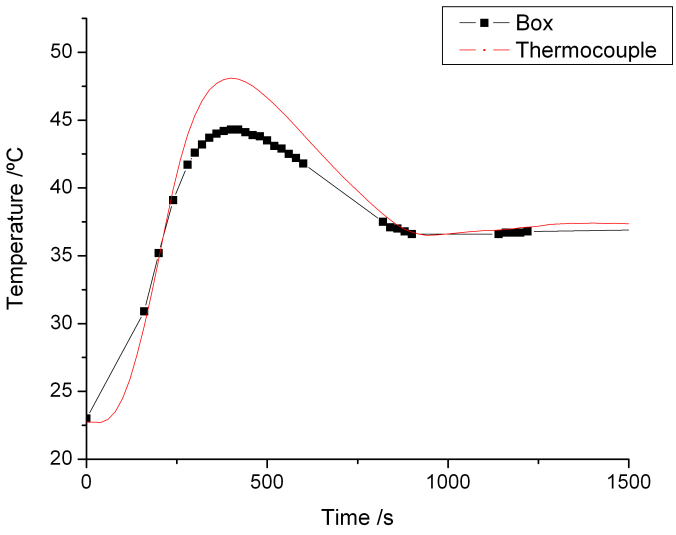


Figure 5.4: Temperatures of the thermocouple and temperature control box while heating from room temperature to 37 °C.

set temperature. As expected, the lower the set temperature the longer it takes to equilibrate since cooling is driven only by the difference between internal and ambient temperature. To quantify this effect the curves were fitted to the following equation

$$T = T_0 + e^{-\lambda t} A \cos(k_1 t + k_2) \tag{5.2}$$

where t is time divided by 1800 s (to give a reasonable magnitude for the values k_1 and k_2), T is temperature and the remaining terms are fitted variables. The term λ indicates a characteristic time for the curve. This equation corresponds to a lightly damped oscillator system. A polynomial instead of a cosine function was used, which corresponds to heavy damping, but better fits were found using a cosine. Fitting was carried out using the MCMC algorithm described in chapter 6. All values for figure 5.4 and shown in table 5.1. Values found for λ only from data in figure 5.6 are given in table 5.2 and figure 5.7. A general increase in λ is observed with temperature, confirming that the higher the set temperature the quicker it is attained.

5.4.3 Motor and temperature control jacket

Normal experimental measurement for LD requires two spectra. One is with no motor running and the other is with the motor running. Subtraction of

Parameter	Median	10% confidence	90% confidence
T_0	24.7	24.6	24.9
λ	6.60	5.58	7.59
A	15.6	11.4	21.8
k_1	6.63	5.52	7.42
k_2	4.25	4.18	4.34

Table 5.1: Fitting parameters for equation (5.2) and data changing from room temperature to 25 °C, figure 5.4

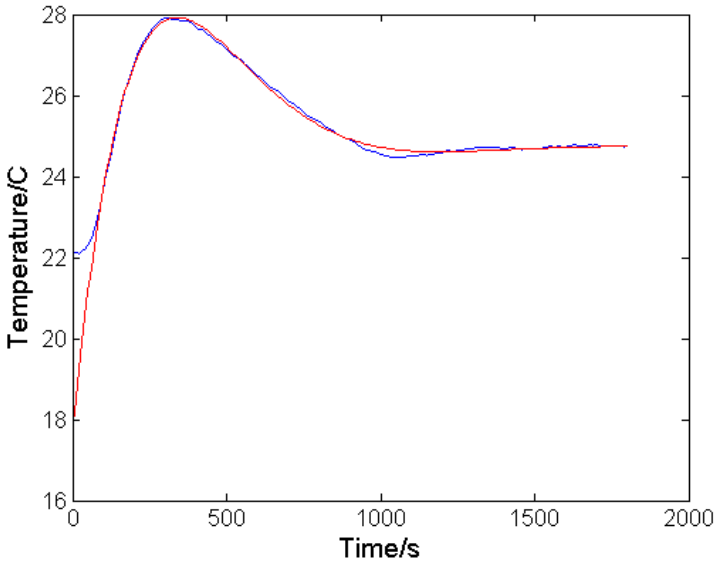


Figure 5.5: Fitting of equation (5.2) to data when changing from room temperature to 25 °C

Temperature range	λ	10% confidence value	90% confidence value
RT-25	6.60	5.58	7.59
25-30	7.96	7.04	8.76
30-35	7.63	6.50	9.10
35-40	8.78	6.74	11.55
40-45	9.68	7.82	12.52

Table 5.2: Median, 10% and 90% confidence values of λ (equation (5.2)) for fitting the data in figure 5.6.

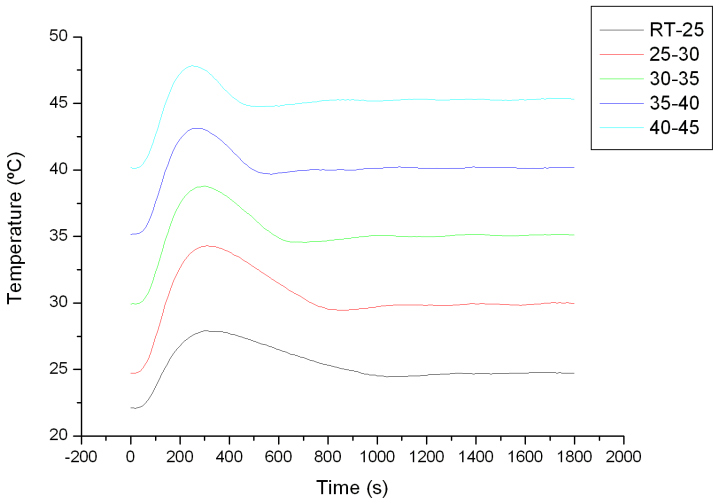


Figure 5.6: Traces of sample temperature as the set temperature changes from various starting temperatures to various finishing temperatures.

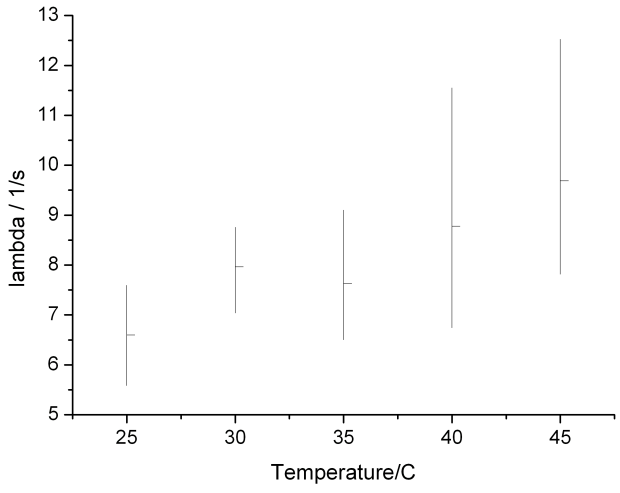


Figure 5.7: Values of lambda, as defined in equation (5.2) for fitting to data from figure 5.6

the latter from the former removes background signals from optical components, the cell etc. and leaves only LD from oriented molecules. The following imitates the latter case.

To investigate the effect of the motor combined with the temperature control jacket a sample was allowed to equilibrate to a set temperature and then the motor switched on while the sample temperature was measured. Plots of the results are shown in figure 5.8. A positive deviation from set temperature occurs for all data within 250 s. Origin 7 (Northampton, MA, USA) was used to fit exponential curves to the data and the resulting half-times along with the final temperature deviations are shown in figure 5.9. The deviation is temperature dependent but the half-time, characteristic of time taken for the temperature to stabilise, is independent of temperature. Up to a set temperature of 45°C the deviation remained below 1°C. The existence of such deviations strongly suggests the presence of a temperature gradient between the sample and the probe which causes over-heating of the sample that is not detected and adjusted for.

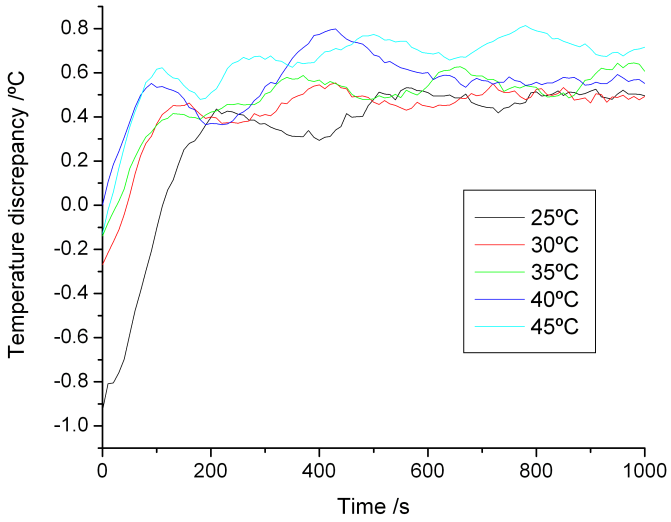


Figure 5.8: Temperature deviation from set temperature measured from various equilibrated starting temperatures whilst the motor is running.

5.4.4 Adjusting temperature using air

It was thought that a flow of air at room temperature through the instrument might adjust the temperature sufficiently to account for the discrepancy due to the motor. The idea is that as air passes through it removes some of the heat due to the motor, enabling better control by the temperature jacket. The

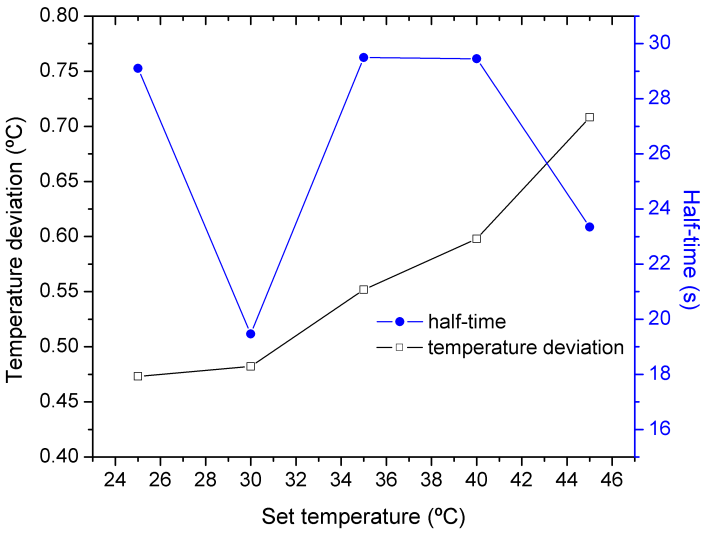


Figure 5.9: Final temperature deviations and equilibration half-times for samples with the temperature control jacket and motor, figure 5.8.

experiment proceeded as for those above with air purging the cell at different rates. Figure 5.10 shows a summary of the temperatures attained. The data show, as previously, that increasing temperature shows an increasing discrepancy between measured and set temperature. The air does not significantly lower the temperature when compared with figure 5.11. Higher air flows would be likely to act as a cooling system, counteracting the temperature jacket to such an extent that it would be difficult to predict the temperature in the sample without the presence of a thermocouple. It could also flow into the photomultiplier tubes and cause damage unless a quartz window was used in front of the tubes.

5.5 Discussion

The prototype LD cell used here does not reach the specifications required to carry out a wide variety of kinetic experiments. The sample takes minutes to attain the set temperature and the final temperature is dependent on use of the motor. Once the temperature is attained it is steady to within 1°C, but a temperature gradient across the cell introduces an unnecessary uncertainty into measurements.

We suggest a rethink in the approach to temperature control of Couette LD cells and propose two solutions to the problems given here: an injection system to allow samples to be heated before being combined and a new LD cell design.

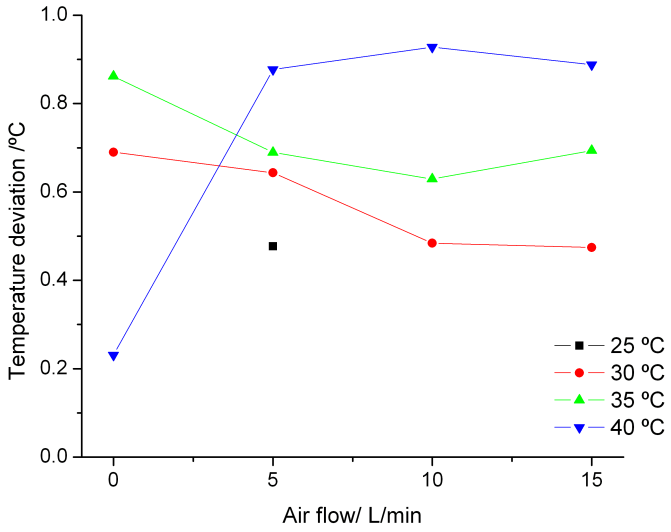


Figure 5.10: Temperature deviations found when air was flowed through the sample chamber at rates indicated. These data do not differ significantly from 5.11.

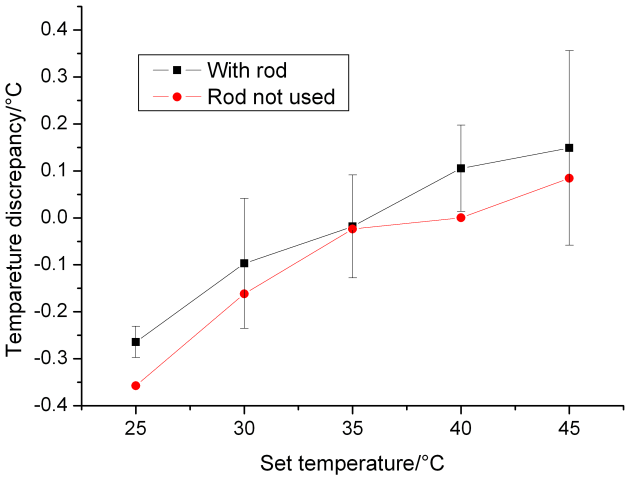


Figure 5.11: Temperature discrepancies using the temperature control box: the difference between the temperature set and measured.

5.5.1 An injection cell

An injection cell has been developed with invaluable contributions from Tim Dafforn, Matthew Hicks, Nigel Dyer, Daniel Waldron and Rhod Mortimer at Crystal Optics. It is described in detail in chapter 9. The principle is that instead of a fixed quartz rod as the inner of the two cylinders of a Couette flow cell, a hollow capillary is used through which sample can be injected into the annular gap (see figure 5.12). Injection is controlled by a syringe driver.

The advantage of such a system in this context is that a sample can be held in the capillary for an arbitrary length of time before injection, allowing the whole cell to equilibrate to a set temperature. In doing so it removes the problem of warming reactants from room temperature to reaction temperature before starting an experiment.

Additional benefits of an injection cell include the ability to measure fast kinetics with a minimal dead-time. Currently sample loading requires 20–30 s, including the installation of the capillary and rod into the sample chamber. An injection system could cut this to under a second. The limiting factors are the times taken to inject the sample and start the motor. It can, in principle, be used with any similar LD cell although the current design requires a custom lid that can be injected through.

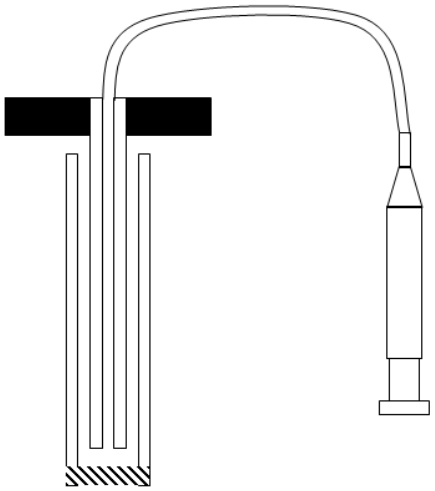


Figure 5.12: Schematic of injection cell (not to scale). A syringe driver is used to control the injection of samples and a thin tube connects this to the capillary.

5.5.2 A new cell design

A more wholesale and satisfactory solution is to propose a new design of LD cell. Based on the work presented here the most important specifications are

that:

- the sample be better insulated from the motor
- the temperature control jacket (or equivalent component) be as close to the sample as possible to remove any temperature gradient.

We also suggest a number of other specifications that would aid future work:

- employ an active cooling mechanism
- make the cell walls thinner to allow quicker temperature change
- install heat sinks to facilitate rapid cooling
- change the control of rotation from potential difference units to rpm or equivalent
- put the motor in a position from which it can be easily removed and changed periodically.

Some of these points are discussed in more detail below. A schematic of a new design of cell is shown in figure 5.13.

Sample insulation

Insulation of the sample from the motor could be achieved either with a layer of insulation between the motor and sample or by moving the motor further away from the sample. In the diagram a universal joint is shown, allowing the motor to be placed much further from the sample and for flexibility in its location. It could, for example, be placed at the side so that heat does not convect straight into the sample chamber. This would have the additional benefit of making it more accessible for replacement. The bearings in the motor last for between 6 and 9 months with frequent use.

Active cooling

The two most widely used methods for rapid cooling of small regions are Peltier control, or a combination of a heat sink and air.

Peltier control has been employed on circular dichroism cells. A thermocouple is two wires of different metallic composition, and a Peltier system is based on the principle that passing a current through a thermocouple generates a temperature gradient. This is effectively a reversal of the method by which thermocouples measure temperature. It enables precise temperature

control and rapid temperature changes, although the energy transfer is not very efficient.

Heat sinks and fans are used to cool computer components. The current linear dichroism cell is not designed for rapid cooling, meaning that the air flowed through the cell was not effective at removing temperature from the sample (in addition the air we used was at room temperature). A redesigned cell with heat sinks on the outside and a well-defined path for air to pass through is likely to be much more effective. The air could be flowed through in one of two ways: either using a pressurised air source, or using fans to pull the air through. If, in addition, the air was cooled beforehand then variation of air flow should be sufficient for control at below-ambient temperatures.

Heating

Heating the sample is more straight-forward, although the current prototype has some flaws. The distance between the heating jacket and sample can be greatly reduced (see figure 5.1). In addition, the sides of the cell are relatively thick and made of aluminium coated with nylon, although stainless steel has been used for similar designs. This conducts heat away from the sample during heating and stores it to prevent cooling. Thinner walls would not present a great technical challenge and would increase efficiency of heat transfer.

The heating element is currently located on the outside of the main sample chamber (figure 5.1). If it were on the inside of the cell with only a thin gap to the sample the temperature gradient across the cell would be effectively removed.

If a Peltier system were used it would be able to heat and cool the sample. A heat-sink-based design with air flow would require additional heating. An external bath through which the air passes can only efficiently cool samples, so we suggest that a heating pad inside the cell would be the best solution. The alternative of having a hot bath seems superfluous when heating elements are much more efficient. This raises the problem of having software that is able to control both air flow through the cell (or fan rotation) and a heating element in order to control sample temperature.

A typical commercially available Peltier unit can pump at 172 Watts over and area of 40 $\times$40 mm^2 and an air-cooled system at 194 Watts over a 135 $\times$ 135 mm^2 area. These values are comparable and both adequate for the purpose here. Since the control software is readily available, we suggest that Peltier temperature control is the better option.

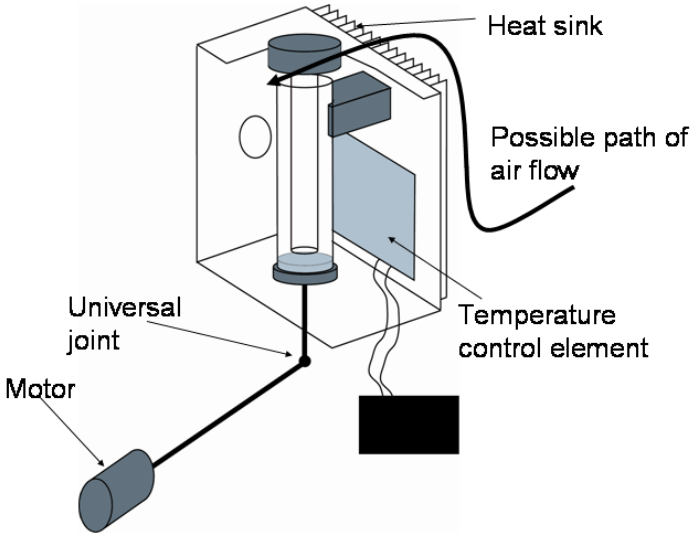


Figure 5.13: Cut away schematic of a new cell design. This shows the motor away from the sample chamber. The temperature control element (Peltier or heating pad) is connected to a control box, represented in black. The air flow could either be as shown, follow a more complex geometry or simply pass through the sample chamber. The sample chamber should be as small as possible to minimise heat loss.

Cell geometry

An additional consideration is the overall geometry of the cell. Currently the cell chamber is circular, but a short conceptual leap raises no objection to it being square, oblong or any other shape providing that the capillary is held in position and that the light beam has a straight path through the sample. This may prove an advantage as heat pads and Peltier elements are often flat. With air flow-based cooling, the air could either pass through the sample chamber or a second chamber separated by a thin, conducting wall. The advantage of the latter is that a larger volume of air could be flowed through a more complex system of heat sinks to increase efficiency of heat transfer. With the former idea the sample chamber would need to be closed, in particular quartz windows would be required at the beam entrance and exit locations.

5.6 Conclusion

After careful testing of the temperature control properties of the current thermostatted Couette LD cell, we have found that it does not meet the requisite standards for experiments that require rapid sample heating or precise temperature control. It is suitable for holding samples at a constant temperature

for long periods of time, however, a Peltier system such as those employed on modern circular dichroism spectrometer sample holders would be more suitable.

Chapter 6

Two Markov chain Monte Carlo algorithms

6.1 Chapter summary

A description of the Markov chain Monte Carlo (MCMC) algorithm used in later chapters is given. In particular the Metropolis-Hastings method and the Gibbs sampler are described. The theoretical basis and a summary of the method of implementation are given.

6.2 Introduction

When interpreting experimental data a mine of information can be inferred. Initially this is qualitative, but a quantitative answer will often be sought. Further still, an estimate of how sure we can be about these values is often extremely useful, especially for comparison with alternative methods. In short we wish to extract as much information as possible from the data without over-interpreting. Statistics offers a framework in which to accomplish much of this. In particular, Bayesian statistics and Markov Chain Monte Carlo algorithms (MCMC) are adept at quantifying the distributions of solution parameters for model fitting problems, as outlined below. The Metropolis-Hastings algorithm and the Gibbs sampler are types of MCMC explained in this chapter and applied in chapters 6 and 7.

6.2.1 Notation, Bayesian statistics and the Likelihood function

The problem to be solved is stated as follows. Given experimental data $Y(t)$ and a model with parameters θ, we wish to find the values θ^* which produce the closest fit of the model output values $X(t|\theta^*)$ to $Y(t)$. [78]

This can be considered as a Bayesian problem. Conventional probability theory determines the probability of some outcome given the experimental setup. For an outcome A and a model with parameters m we can state that we wish to know $P(A|m)$. For example if a fair die is rolled, after a while it is likely that the probability of any number will be calculated as $\frac{1}{6}$. Baysian statistics solves a different kind of problem. It may be happen that, for example, the number 1 occurs much more often than one in six and we wish to know how biased the die is. We seek the die properties from the output data, that is $P(m|A)$ [25] [35]. This is the kind of problem that Bayesian statistics solves. From Baye's theorem

$$P(m|A) \propto P(A|m)P(m). \tag{6.1}$$

$P(m|A)$ is referred to as the posterior distribution. $P(m)$ is termed the prior distribution and contains information known about the parameters before the inference. In practice the prior distribution may be very broad, even if very little is known.

$P(A|m)$ can be considered as a function of m as opposed to A. In this case we call the function the likelihood, which is thought of as the likelihood that

the value of m is correct given the values of A. The likelihood function is

$$\mathcal{L}(m|A) \equiv P(A|m). \tag{6.2}$$

Inspection of this equation instantly yields that the most suitable values for m occur at the maximum value of the likelihood function. The aim of the techniques that follow is to obtain the values of m that maximise the likelihood. Assuming that the distribution of parameters is normal, the likelihood function takes the form

$$\mathcal{L}(\theta) = \prod_{t=0}^{T} \frac{\tau}{\sqrt{2\pi}} \exp\left\{ \frac{\tau(X(t|\theta) - Y(t))^2}{2\pi} \right\} \tag{6.3}$$

Although not discussed here, the problem can also be considered as a nonlinear programming problem subject to nonlinear differential-algebraic constraints [78].

6.2.2 Markov chain Monte Carlo

The problem outlined above is solved in an iterative manner. How this is done is described by the terms 'Markov chain' and 'Monte Carlo'.

A Markov chain is a sequence of values in which any given value is dependent on only the previous term. For example, the sequence $1, 4, 7, 10, 13, ...$ is a Markov chain since the any value can be determined from the previous one by adding 3. The Fibonacci sequence is not a Markov chain, since each value is the sum of the previous two.

Monte Carlo, taking its name from the region in Monaco famed for casinos and gambling, is a term that covers a broad range of techniques. In general a Monte Carlo method is one that relies on sampling outcomes at random and accepting or rejecting these values based on some criteria.

In the work presented here two MCMCs are used — the Metropolis-Hastings algorithm and the Gibbs sampler. They differ in the knowledge about θ^*. The algorithms are introduced in general, then applied to Bayesian likelihood estimation.

6.3 The Metropolis-Hastings algorithm

The aim of the Metropolis-Hastings algorithm is to sample the parameter space of a function in such a way that the values sampled conform to some property. Consider a space $B \subset \mathbb{R}^n$, from which samples are taken, and a function termed the transition kernel $p : B \times \mathbb{B}(B) \to [0, 1]$ which is the probability that a point $x \in B$ is mapped to the set $A \in \mathbb{B}(\mathbb{R}^n)$ where $\mathbb{B}(X)$ is the Borel field of the field A (roughly equivalent to the power set).

A constraint is introduced by a distribution $\pi^*(.) : \mathbb{R}^n \to [0, 1]$ with density $\pi(.)$, describing regions in which a value is most likely to be found. The relationship between π and p can be expressed as the following

$$\pi^*(dy) = \int_{\mathbb{R}^n} p(x, y)\pi(x)dx \tag{6.4}$$

This is equivalent to saying that π is the invariant distribution of $p(., .)$.

6.3.1 A useful result

Tierney [110] proved the following:

If p is reversible, that is

$$\pi(x)p(x, y) = \pi(y)p(y, x) \tag{6.5}$$

then $\forall A \in \mathbb{B}(\mathbb{R}^n)$

$$\int_B p(x, A)\pi(x)dx = \int_A \pi(y)dy \tag{6.6}$$

which is equivalent to stating that π is the invariant distribution of $p(., .)$. This result is key to the Metropolis algorithm, as explained in section 6.4. The problem is now one of finding a reversible function.

6.3.2 The algorithm

[76, 44, 110, 15, 112] If a putative transition kernel $q(x, y)$ is proposed it is unlikely to be symmetric. Instead

$$\pi(x)q(x, y) > \pi(y)q(y, x). \tag{6.7}$$

A function $\alpha(x, y)$ can be manufactured that will correct for the difference, so that the function is reversible. The Metropolis-Hastings probability function is defined as

$$P_{MH} := q(x, y)\alpha(x, y) \tag{6.8}$$

for $y \neq x$ and

$$\alpha(x, x) = 1 - \int_A P_{MH}(x, y)dy. \tag{6.9}$$

Since, by definition of α the system is reversible

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x). \tag{6.10}$$

By equation (6.7), $\alpha(y, x)$ should be made as large as possible. Since this is a probability means that this cannot exceed 1, so

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x). \tag{6.11}$$

By rearranging

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}. \tag{6.12}$$

Hence we define

$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\} \tag{6.13}$$

A Markov chain can be built using this function. Any initial value for x_i is proposed. For any point in chain, x^i, the next value x^{i+1} in the chain a value is proposed (at random) and accepted with a probability α. If it is not accepted $x^{i+1} = x^i$. By the properties of α these values are being sampled from a symmetric distribution, hence by Tierney's result (equation (6.6)), from the invariant distribution of $p(x, y)$ as required.

6.3.3 Irreducibility and aperiodicity

Two important properties for convergence of Monte Carlo methods are irreducibility and periodicity. Here they are defined and their importance outlined. The conditions or irreducibility and periodicity are as follows. Consider

a measure χ and transition kernel $P : E \times \mathbb{B}(E) \to [0, 1]$.

Define P to be χ-irreducible if: for each $x \in E$ and $A \in \mathbb{B}(E)$ with the property that $\chi(A) > 0$, there exists $n \in \mathbb{N}$ such that $P^n(x, A) > 0$. This states that, regardless of the starting position, iterations of the transition kernel will eventually reach each point of E.

P is defined to be periodic if there exists $d \in \mathbb{N}_{\geq 2}$ and a sequence $(E_j)_{j=0}^{d-1}$ of non-empty disjoint subsets of E such that for each $i = 0, ..., d-1$ and $x \in E$, $P(x, E_j) = 1$ where $j = i + 1 \bmod d$. In other words, there is some sequence of subsets of the space that the transition kernel will cycle round in sequence. P is aperiodic if it is not periodic.

Tierney 1994 [110] has shown that under the conditions of irreducibility and aperiodicity the following is true. For a transition kernel P with invariant distribution π^* define

$$P^{(n)}(x, A) := \int_{\mathbb{R}^n} P^{(n-1)}(x, dy) P(y, A), \tag{6.14}$$

then

$$\lim_{n \to \infty} P^{n-1}(x, dy) \sim \pi^*(dy). \tag{6.15}$$

This means that iteration of a correct transition kernel will tend towards the invariant distribution. This property underpins Markov chain Monte Carlo since it allows a Markov chain to be constructed from successive iterations of the transition kernel. Using the function α from the previous section a putative transition kernel can be gradually corrected so that eventually the iterations will sample from the invariant distribution. Note that there is no general principle to determine the rate of convergence.

6.3.4 Application of the algorithm

We have applied the Metropolis-Hastings algorithm to the maximum likelihood problem. The transition kernel ($q(x, y)$ in the notation above) is the likelihood function $\mathcal{L}(\theta)$ given in equation (6.3) and the prior distribution entered as the distribution density π.

The algorithm can be summarised as follows:

1. For a parameter value x^i, propose a new value y by sampling from a normal distribution with a predetermined width (called the step-size)

2. Calculate $\alpha(x, y)$ using equation (6.13), so

$$\alpha(x, y) = \min\left\{\frac{\pi(y)\mathcal{L}(y)}{\pi(x)\mathcal{L}(x)}, 1\right\}. \tag{6.16}$$

   In practice the logarithm of this function is often used.

3. Generate u by sampling from the uniform distribution on the interval $[0, 1]$

4. If $u < \alpha(x, y)$ accept y as x^{i+1}, otherwise $x^{i+1} = x^i$ for the next iteration

Typically this is done changing each value of the vector x in turn, in which case one iteration is complete when every value of x has been altered. After a predetermined number of steps, the Markov chain $(x^1, x^2, x^3, ...)$ is returned.

6.3.5 Convergence

In many simple cases accompanying theorems guarantee convergence within a finite number of samples to a certain error [14] . However, in general this is not the case. Cowles and Carlin [19] note that neither is it possible to guarantee convergence by running diagnostics on the outcomes.

For the work presented here, note that the form of the likelihood function assumes a normal distribution of parameters. When the Markov chain appears to the eye to be sampling from normal distributions (ascertained using histogram plots) it is assumed to have converged.

6.4 The Gibbs Algorithm

The Gibbs algorithm exploits knowledge of parameter distributions to ensure an exact output. The most noticeable difference to the Metropolis-Hastings algorithm described above is that there is no decision about acceptance. The transition kernel is constructed in such a way that the probability of acceptance is 1 at each step.

To examine the basis for the algorithm, suppose the transition kernel $q(.,.)$ is the full conditional probability density function $\pi(.)$. Using the notation $x_{-i} = \{x_1, x_2, ..., x_{i-1}, x_{i+1}, ...x_n\}$ consider only changing one of the parameters of x:

$$q(y, x_i|x_{-i}) = \pi(x|x_{-i}) \tag{6.17}$$

equation (6.12) becomes

$$\alpha(x, y) = \frac{\pi(y)\pi(x)}{\pi(x)\pi(y)} \tag{6.18}$$

$$\equiv 1. \tag{6.19}$$

This is a fortuitous result, however is it possible to find such a transition kernel? The next section considers the example of precision in the likelihood function.

6.4.1 Deriving the parameters of the conjugate gamma distribution

This section considers the example of estimating the precision τ of the likelihood function. The probability distribution is given by

$$P(\tau|\theta) = \frac{P(\theta|\tau)P(\tau)}{\int P(\theta|\tau)P(\tau)d\tau} \tag{6.20}$$

by Baye's theorem, where θ is the set of parameters.

But $P(\theta|\tau)$ is the likelihood $\mathcal{L}$, where

$$\mathcal{L} = \left(\sqrt{\frac{\tau}{2\pi}}\right)^N \exp\left(-\frac{\tau\mathcal{S}}{2}\right) \tag{6.21}$$

from equation (6.3) where N is the number of time points in the data and $\mathcal{S}$ the sum of squared differences:

$$\mathcal{S} = \sum_{t=1}^{N}(X(t|\theta) - Y(t))^2 \tag{6.22}$$

for X theoretical data and Y experimental data.

$P(\tau)$ is assumed to be gamma distributed, then

$$P(\tau) \sim \Gamma(\tau|a, b) = \frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-\tau b} \tag{6.23}$$

for some parameters a and b. Note that the first Γ refers to the distribution and the second to the gamma function, given by

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt. \tag{6.24}$$

Combining equations 6.20, 6.21 and 6.23

$$P(\tau|\theta) = \frac{\left(\sqrt{\frac{\tau}{2\pi}}\right)^N \exp\left(-\frac{\mathcal{S}\tau}{2}\right) \frac{b^a \tau^{a-1} e^{-\tau b}}{\Gamma(a)}}{\int_0^\infty \left(\sqrt{\frac{\tau}{2\pi}}\right)^N \exp\left(-\frac{\mathcal{S}\tau}{2}\right) \frac{b^a \tau^{a-1} e^{-\tau b}}{\Gamma(a)} d\tau} \tag{6.25}$$

$$= \frac{\tau^{\frac{N}{2}+a-1} \exp(\tau\left\{\frac{\mathcal{S}}{2}+b\right\})}{\int_0^\infty \tau^{\frac{N}{2}+a-1} \exp(\tau\left\{\frac{\mathcal{S}}{2}+b\right\}) d\tau} \tag{6.26}$$

Consider the integral in the denominator of 6.25. Let

$$t = \tau\left\{\frac{\mathcal{S}}{2}+b\right\} \tag{6.27}$$

then

$$t|_{\tau=0} = 0 \tag{6.28}$$

$$\lim_{\tau\to\infty} t = \infty \tag{6.29}$$

$$dt = \frac{d\tau}{\left\{\frac{\mathcal{S}}{2}+b\right\}} \tag{6.30}$$

then the integral is

$$\int_0^\infty \left(\frac{t}{\left\{\frac{\mathcal{S}}{2}+b\right\}}\right)^{\frac{N}{2}+a-1} e^{-t} \frac{d\tau}{\left\{\frac{\mathcal{S}}{2}+b\right\}} \tag{6.31}$$

$$= \left(\frac{t}{\left\{\frac{\mathcal{S}}{2}+b\right\}}\right)^{\frac{N}{2}+a} \Gamma\left(\frac{N}{2}+a\right) \tag{6.32}$$

So equation (6.26) becomes

$$\frac{\left(\frac{\mathcal{S}}{2}+b\right)^{\frac{N}{2}+a}\exp\left(-\tau\left\{\frac{\mathcal{S}}{2}+b\right\}\right)}{\Gamma\left(\frac{N}{2}+a\right)} \tag{6.33}$$

$$=\frac{B^A\tau^{A-1}\exp(-\tau B)}{\Gamma(A)} \tag{6.34}$$

$$=\Gamma(\tau|A,B) \tag{6.35}$$

where

$$B = b + \frac{\mathcal{S}}{2} \tag{6.36}$$

$$A = a + \frac{N}{2}. \tag{6.37}$$

This means that the posterior distribution is gamma distributed with the parameters A and B. Successive samples of this create a Markov chain.

6.4.2 Gibbs sampler summary

The Gibbs sampling algorithm can be summarised as follows:

1. Find or choose initial values of the estimated parameter x
2. Calculate the value of x_i consistent with $P(y, x_i|x_{-i})$ using the full conditional probability (e.g. $\Gamma(x|A,B)$ in the example given above)
3. Since P is known to be exact the value for x_i is accepted

Chapter 7

Restriction enzyme kinetics

7.1 Chapter summary

The kinetic action of restriction enzyme digestion of DNA has been measured using linear dichroism, with the aim of quantifying rate constants for the *in vitro* process. Two restriction enzymes, EcoRI and BstZ17I were used. Gel electrophoresis of samples held in a water bath was also used to verify the results.

Linear dichroism is able to clearly differentiate between circular plasmid DNA and different lengths of linear, cleaved DNA. However, the rate of digestion in a water bath was different to the rate in the Couette flow cell used for linear dichroism. In addition, the time courses from linear dichroism show a rapid increase at the beginning and a gradual increase with time. These time course features are evident in a plasmid designed to contain no restriction site. The possible sources and implications of these effects are discussed.

7.2 Introduction

Restriction enzymes (REs) are proteins used by prokaryotic cells to protect themselves against rogue DNA that finds its way into the cytoplasm [90, 117]. They cut DNA in very precise ways, a property exploited by molecular biologists who use them effectively as scissors for DNA. In this chapter we investigate the ability of linear dichroism (LD) to provide qualitative data on the kinetics of REs and work towards quantisation of parameters related to a model of RE action.

In simple terms, the process followed by REs is to bind at a short recognition sequence, cleave and then release DNA. However, the action of individual REs can vary widely [36] and a number of classifications and sub-classifications based on mode of operation have been defined [117, 90]. Type I REs have four subunits and require adenosine triphosphate (ATP) and magnesium ions to operate. They also require two copies of the recognition sequence per DNA molecule and usually cleave around the midpoint between the two sequences. Type III REs consist of two subunits. Like type I they require both magnesium ions and ATP, two copies of the recognition sequence and cleave close to one of the recognition sites. Type II REs are by far the most studied and most widely used. This is because they cut at a precise point, usually within the recognition sequence. They require magnesium ions, but not ATP, for cleavage. The majority require one or two recognition sequences that are 4–8 base pairs (bp) long and palindromic. The EcoRI enzyme used in this work, for example, cleaves after the first base of the sequence GAATTC — note that this leaves a 5 bp overhang at each cleaved end.

A notable property common to all REs is that they can find the restriction site faster than is possible by diffusion. Several different mechanisms have been proposed as to how this can be achieved [40]. These are described as sliding, hopping, jumping and looping. Sliding is a 1-dimensional search achieved by moving along the DNA sequence until a recognition site is found. Hopping describes release of a piece of DNA and rebinding to a different site on the same molecule. Jumping is similar to hopping, but finishing on a different piece of DNA. Looping is a mechanism for the 75% of REs that can bind more than one site on the same piece of DNA when in their active form. Binding of both sites simultaneously creates a loop, which can be essential for the action of some enzymes [36, 27]. Each RE is different and a combination of these actions describes the reality [6].

The most commonly used template for REs is a DNA plasmid (described in section 1.1.3). Plasmids are useful for studies of REs with LD since the super-

coiled structure gives a magnitude of LD distinct from the corresponding linear sequence. The following contains a discussion of literature methods used to investigate kinetic action of REs as they cleave DNA, after which the method used for LD is introduced. The alternative methods fail either to have the time resolution of LD or do not measure cleavage free in solution.

7.2.1 Measuring kinetic action of REs

To characterise accurately the kinetic action of REs requires experimental methods for which the enzymes are in their native environment. Several methods have been used for this purpose.

The most common method to follow RE kinetics is to start the reaction *in vitro* and at a number of time points on the desired time-scale, remove an aliquot of the mixture and quench with something such as ethylenediaminetetraacetic acid (EDTA), which chelates the magnesium ions and prevents the enzyme cutting [4, 11, 27, 36]. Some of each aliquot is then analysed by electrophoresis on an agarose gel. Different lengths and conformations of DNA produce different bands on a gel, so the composition of DNA in the sample can be observed at each time point. The mass of DNA in each band is often measured using scintillation counting, which requires H^3 labelling of the DNA and has an accuracy of around 10%. Typically used time resolution of this method is a few minutes.

Others have used a combination of single molecule methods and fluorescent labelling [6, 55]. Either the enzyme or the DNA is attached to a surface and total internal reflection (TIR) or fluorescence resonance energy transfer (FRET) used to measure the reaction progress. This has the advantage of better time resolution, but since either the enzyme or substrate are tethered, the time-scales of the enzyme action are not the same as in solution.

LD has been used previously to measure RE kinetics [49]. The principle behind the measurement is that plasmid DNA and different lengths of linear DNA give differing LD magnitudes at 260 nm because the variations in conformation change the orientation parameter S (see equation (1.14)). All of the reaction components can be added without labelling or modification. LD is measured at a single wavelength over time. Time resolution can be up to milliseconds, but 1 s is usually sufficient and an order of magnitude better than can be achieved using electrophoresis methods.

7.3 Materials and methods

EcoRI and BstZ17I restriction enzymes (New England Biolabs, Ipswich, MA, USA) at concentrations of 20,000 U/ml[1] and 5,000 U/ml respectively were used. 0.167 U/µl of enzyme were used in each experiment. Buffers supplied with each enzyme were used. A pC3.1 derivative DNA plasmid (as described in section 4.4.1) with one EcoRI and two BstZ17I restriction sites was used. The plasmid has a total length of 6882 bp. Restriction was carried out at a DNA concentration of 0.0333 mg/ml (51 µM using a molecular weight of 660 Da/bp).

A variant of the plasmid with no EcoRI restriction site was produced using mutagenesis as follows. Oligomeric DNA primers were designed to act on the plasmid to alter the EcoRI restriction site from GAATTC to GATTTC. The primer sequences were CCAGAAAATTTATATGATTGA*GATTTC*CACCAC ACTGGACTAGTGG and its reverse complement, with the modified recognition sequence shown in italics. To modify the DNA, polymerase chain reaction (PCR) was carried out using 100 ng plasmid template, 0.3 pM of each primer, 40 mM Pfu DNA polymerase from Promega (Madison, WI, USA) with buffer provided and 40 mM deoxyribonucleotide triphosphates (dNTPs, a mixture of single DNA bases from which to construct the sequence). The total volume was 60 µl. 70 µl mineral oil were added on top on the reactants to prevent evaporation during PCR. A control was run with identical components, except that no primers were used. The cycle run for PCR was 95 °C for 1 min then 12 cycles of 30 s at 95 °C (denaturation), 60 s at 55 °C (annealing) and 68 °C for 14 min (elongation). In order to remove the methylated DNA template, the PCR product had 1 µl Dpn I restriction enzyme added and was incubated at 37 °C for 3 hours. Electro-competent cells of *Escherichia coli* were prepared and the digested PCR product was transformed into the cells using a standard procedure (see section 4.4.1). Cells were grown overnight. Colonies from the plate containing the cells with plasmid transformed into them were harvested, grown overnight and a mini-prep protocol (Qiagen, Crawley, UK) was carried out to purify the plasmid, following the protocol provided by the manufacturers. Confirmation of successful sequence modification was found by attempting digestion of the plasmid with EcoRI and full sequencing of the DNA (data not shown).

[1] 1 Unit (U) is the amount of enzyme required to cleave 1 µg of lambda DNA in 1 hour at the correct temperature and in the correct buffer for the enzyme in a total volume of 50 µl

7.4 Spectroscopy

A Biologic MOS-450 spectrometer adapted for linear dichroism (LD) was used with a thermostatted Crystal Precision Optics Couette flow cell. The protocol in section 2.3.3 was followed for sample loading, with the RE added immediately prior to measurement. Measurements of LD were taken every second for up to an hour at a wavelength of 260 nm and a bandwidth of 4 nm. The temperature control was set to 37 °C and DNA was aligned by rotation at a potential difference of 4 ±0.04 V. The dead-time between addition of the enzyme and commencing measurement was around 20 s. Once finished, a baseline measurement was taken for at least 60 s with the sample not rotating.

7.4.1 Gel electrophoresis

For gel electrophoresis, 600 µl of the reactants described above at the same concentration were prepared and heated in a water bath at 37 ±1°C. All other reactants were preheated toegether before the addition of the enzyme. Every 3–5 minutes for 1 hour an aliquot of 30 µl was removed and added to an equal volume of EDTA. 5 µl of each aliquot was run on a 0.7% agarose gel at 120 V for 1 hour. Before photographing in the presence of a UV lamp, the gel was soaked in a solution of the DNA stain ethidium bromide for 1 hour.

7.5 Results

7.5.1 Linear dichroism

The DNA used in this study has one restriction site for the enzyme EcoRI (GAATTC) and two for the enzyme BstZ17I (GTATAC). Following the work of Hicks *et al* [49] kinetics of these restriction enzymes cutting DNA was measured. The aim of this is to work towards a regime where the kinetics can be analysed in a similar manner to the data in chapter 8 to validate models and provide values for rate constants.

Kinetics with EcoRI

A time course showing LD at 260 nm for the DNA plasmid with the EcoRI enzyme is shown in figure 7.1. As expected, the relatively low magnitude at the beginning, corresponding to the plasmid, is gradually replaced by a signal of higher magnitude (more negative), generated by the DNA in linear form.

Kinetics with BstZ17I

For the enzyme BstZ17I, a time course with the DNA is shown in figure 7.2. The pattern here is initially the same as for EcoRI, but as the DNA is cut a second time the signal magnitude decreases. This is caused by the shorter linear DNA, which orients less well than the full length linear plasmid. An interesting feature of these data is that the second cut appears at least as rapid as the first.

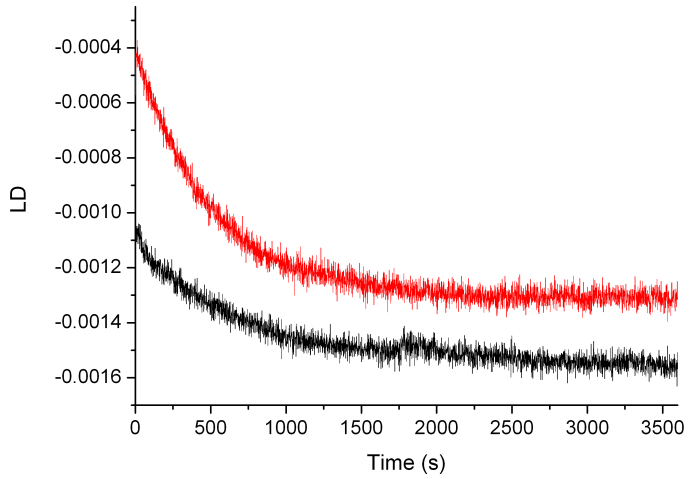


Figure 7.1: LD at 260 nm of DNA with the EcoRI enzyme measured over time. The DNA is initially in plasmid form. The trace becomes more negative as the DNA is converted into the linear conformation. Both traces (red and black) are measured under the same conditions and indicate the poor reproducibility of the data.

A mutant plasmid

A plasmid with no EcoRI restriction site was produced, as described in section 7.3. Time course measurements in the presence of EcoRI are shown in figure 7.3. Some of the time courses exhibited a slow increase with time, as was seen for BstZ17I, and almost all had a short time of increasing LD at the beginning.

DNA spectra

Quantitative modelling of the action of the restriction enzymes is aided greatly by knowledge of the relative contributions of each DNA species: the plasmid, full length linear DNA (i.e. the plasmid cut once) and the cut linear DNA (i.e. after two cutting events). Spectra of each type are shown in figure 7.4. The plasmid was measured in the presence of EcoRI buffer, but with no enzyme

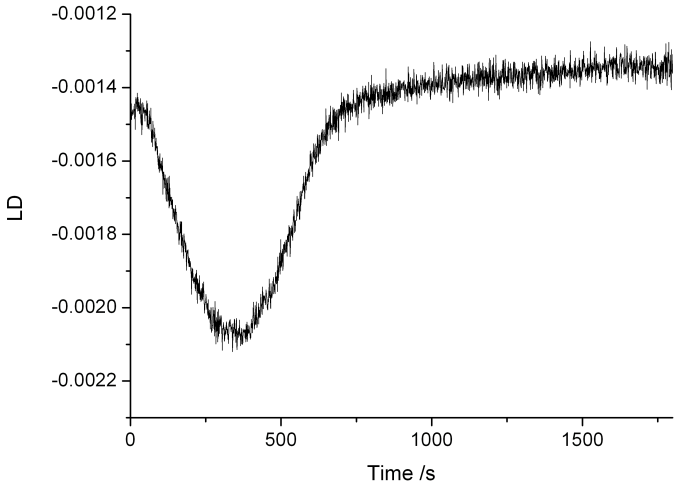


Figure 7.2: LD at 260 nm of DNA with BstZ17I enzyme measured over time. The DNA starts in plasmid form, then becomes linear and is subsequently cut again to give a shorter linear form. The three types can easily be distinguished since they give different LD magnitudes.

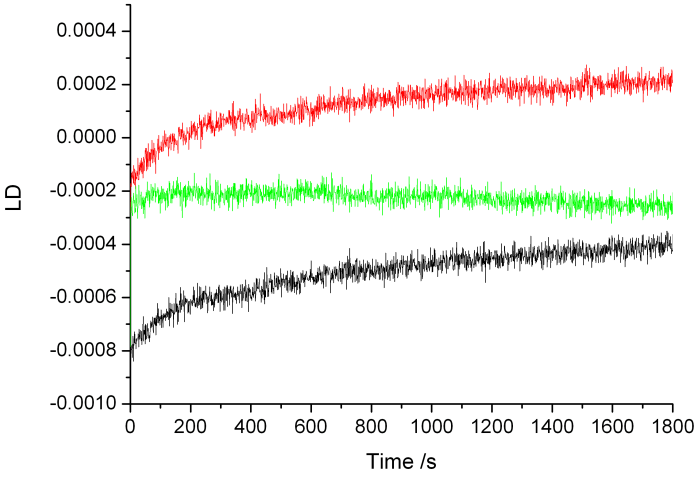


Figure 7.3: LD at 260 nm of DNA with no EcoRI restriction site. All three traces are measured with the same experimental setup.

added. The full length linear DNA was measured after a kinetic trace such as those shown in figure 7.1, where the trace had become flat and it was hence assumed that all of the DNA in the sample had been linearised. The doubly cut DNA was measured in a similar manner to the full length linear DNA, but after a BstZ17I trace instead of EcoRI. Measurements taken of BstZ17I or EcoRI with the corresponding buffer and no DNA present exhibited a small LD signal and it was assumed that the two buffers and enzymes had a similar effect on the LD of DNA at 260 nm.

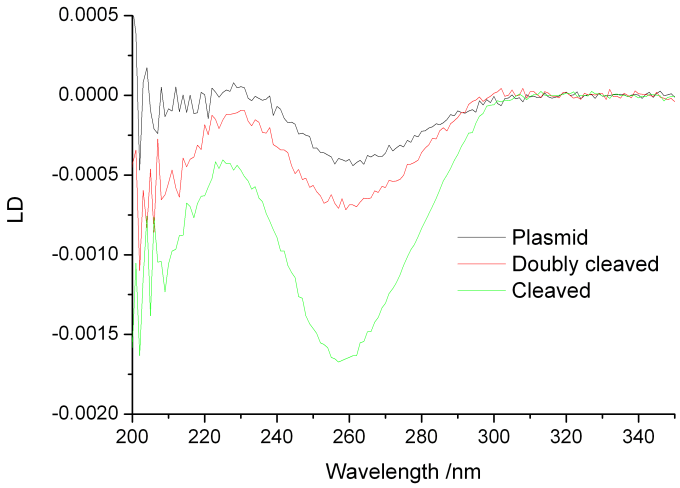


Figure 7.4: LD spectra of three conformations of DNA used: the plasmid, after cutting once with EcoRI (cleaved) and after cutting twice with BstZ17I (doubly cleaved).

7.5.2 Gel electrophoresis

Using agarose gel electrophoresis to measure aliquots of the reaction mixture quenched at various time points (as described in secion 7.4.1) has the advantage that all of the DNA conformations can be seen, since they run at different rates on a gel and produce different bands. Figure 7.5 shows samples from different time points for the BstZ17I enzyme reaction. It shows the disappearance of the band corresponding to the plasmid with time, the appearance and subsequent disappearance of the band corresponding to the full length linear plasmid and the appearance of the two bands of the product of the second cut.

To further characterise the results, an LD spectrum of each aliquot was measured. To compare with LD, the gel data was turned into a sequence of LD magnitudes as follows. The total intensity of each band was measured using UVIdoc (UVItec, Cambridge, UK) and the DNA conformation of each

band assigned as indicated in the caption for figure 7.5. Each intensity was multiplied by a factor determining its contribution to LD derived from plots such as 7.4. For each time point the LD was calculated to be the sum of the contributions of bands present. In other words, the predicted LD from gel data was

$$LD^{P}_{260} I^{P} + LD^{1}_{260} I^{1} + LD^{2}_{260} (I^{2i} + I^{2ii}) \tag{7.1}$$

where the superscript P refers to the plasmid and 1 and 2 refer to the singly and doubly cut DNA respectively, with the doubly cut DNA having two products denoted i and ii. I is normalised intensity, i.e. the measured value divided by the value of the band with the largest intensity. Nicked DNA was not considered, since the contributions are small and this method is considered to be no more than indicative.

Figure 7.6 shows broad agreement between the LD and gel data, although some individual points are not in good agreement. Between this and inspection of the gel we are confident that the broad changes seen in the LD time courses are consistent with the explanation of the changing conformation of LD. There is no indication, however, of what could cause the period of increase observed at the beginning of the time series. A slight increase at the end of the series is evident in the data. The rate of cutting is slower in the quenched series than in the LD cell, but the amount of noise associated with the quenched data makes this a tentative conclusion.

7.5.3 Temperature control

The increase observed at the beginning of many of the kinetic traces, including some in figures 7.2 and 7.3 could be caused by a temperature change. This idea is further discussed below. To see if there was a possibility of temperature change a thermocouple was inserted into the LD capillary with a sample present (the contents of the sample are unimportant), then the central rod was inserted and the cell allowed to equilibrate for 15 minutes. The rod was then removed for 20 s and reinserted. After this, a drop in temperature of around 10 °C was measured on the thermocouple. The precise drop in temperature was variable and appeared to be dependent on the amount of time for which the rod was out of the capillary. The time for which the rod was out of the sample was chosen to match that which could be used for sample loading during experiments.

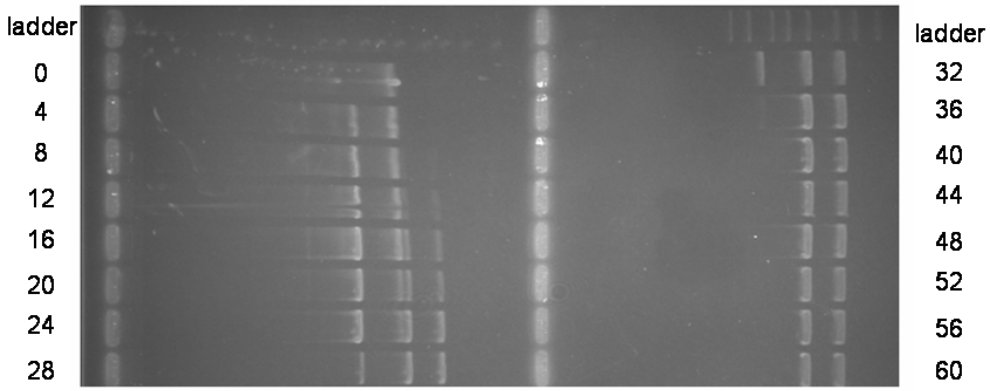


Figure 7.5: An agarose gel showing aliquots of quenched reaction mixture initially containing plasmid DNA and BstZ17I enzyme prepared as detailed in section 7.4.1. The aliquots were taken at 4 minute intervals for 1 hour and are presented in time order as indicated. The DNA runs on the gel left to right and, in general, shorter DNA runs further than longer DNA with nicked and plasmid DNA being notable exceptions. The first time point shows only plasmid DNA and a faint band corresponding to nicked DNA (DNA with a cut in one strand which disrupts the super-coiled structure). After 4 minutes a band for the full length linear DNA is present. After 28 minutes only the full length band and two bands corresponding to the two products of the second cut remain and by the end only the doubly cut DNA is present.

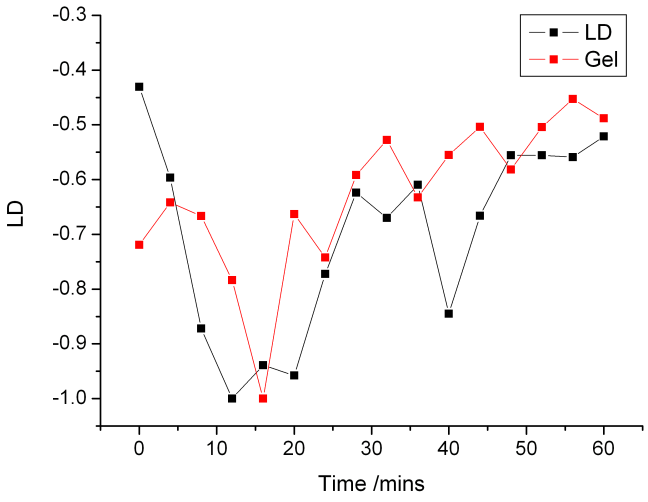


Figure 7.6: Comparison of LD at 260 nm and gel data for BstZ17I (as described in the text). The samples were quenched aliquots of the reaction mixture taken at intervals of 4 minutes.

7.6 Discussion

LD has been used to qualitatively measure the cleavage of DNA by REs. These data could be used to make quantitative predictions. The aim here is to identify key factors that will enable the data to accurately reflect the kinetic action of the RE used.

The data exhibit three main features that suggest effects that could affect the kinetics of REs. These are: different rates of restriction between cleavage measured between real-time LD and gel electrophoresis, a period of decreasing magnitude at the beginning of LD kinetic data and a slow decrease in magnitude over time. We consider each of these in turn.

7.6.1 Real-time LD compared with gel electrophoresis

The rate of action of REs on DNA is concentration dependent. Since the concentrations of all species in the real-time and quenched data were the same, the data should exhibit the same behaviour. What is found from inspection of figures 7.2 and 7.6 is that the real-time measurements show more rapid cutting than the quenched data.

The main difference between the two is that for the real-time data the reactants were in a Couette flow cell and for the quenched data they were in a water bath. Couette flow elongates the DNA and this could make it easier for the RE to find its recognition site, thus increasing the reaction rate.

Another key difference is that the mixture in the water bath was sampled from every few minutes, requiring the lid to be opened. This could have led to cooling of the sample and thus less efficient kinetics.

7.6.2 Decreasing LD magnitude at the start of measurements

At the start of many of the kinetic measurements there is a period of decreasing LD for about 300 s, as exhibited by BstZ17I with the native plasmid and EcoRI with the plasmid containing no restriction site. This occurs on the same time scale as it takes to warm a sample by 5 °C using the temperature control jacket attached to the LD Couette flow cell (chapter 5).

Since the effect is evident in the plasmid with no restriction site, it is not related to the cleaving action of the enzyme. As detailed in the results section of this chapter, removal and replacement of the rod can lead to temperature changes of 10 °C. We thus suggest that the most likely cause of the decrease in magnitude is temperature affecting the DNA.

A change in temperature also has implications for the kinetics of the RE. The effective temperature range for REs is narrow, so a long period of temperature adjustment at the beginning of measurement would affect the kinetics. The first few minutes are a crucial time in estimating rate parameters, so this casts doubt on the reliability of rates estimated from these data.

7.6.3 Gradual decrease in LD magnitude

A gradual decrease in LD magnitude was observed for a number of real-time LD measurements of DNA with a RE. It is noteworthy that this effect was not observed in all data sets, as evident in figure 7.3. After a certain time it is assumed that all of the DNA is linear and thus the time series are expected to show no change with time.

The decrease is unlikely to be caused by non-specific cleavage of the DNA. Firstly, non-specific cleavage would result in at least one extra band appearing on agarose gels after some time, and this was not observed. In addition, previous studies have found the rate of cleavage for DNA containing sequences very similar to the recognition sequence to be many orders of magnitude smaller [109].

The most likely explanation is that over time there is some evaporation of water from the sample and the remaining solution becomes more concentrated. The extent of this effect depends on ambient temperature and humidity, reflecting the non-reproducibility of the effect.

Again, this has implications for the kinetic rate of action of the RE. REs work faster at higher concentrations of enzyme and DNA.

7.7 Conclusion

A number of barriers have been identified to being able to estimate rate constants from LD data of RE digestion of DNA.

One of the most important problems is temperature control. The possible solutions to this for Couette flow LD are discussed elsewhere in this work (chapter 5) and the arguments will not be repeated here.

The difference between real-time LD and gel electrophoresis data remains a problem. Currently, it is not clear which most accurately reflects the kinetic action of REs. Experiments to determine this should be carried out with some urgency. Should it prove that Couette flow alters the action of REs, then measurements could be taken only every few minutes. This would mean the solution experiencing Couette flow for only a few seconds, which is unlikely to

significantly affect the rate of action of the RE. In this case the LD loses time resolution, but accuracy is still much better than the 10% errors for scintillation counting of agarose gels and the results are available much quicker.

We believe that the time resolution of LD combined with the ability to measure in solution mean that it has the potential to play a major role in looking at REs once the problems listed above have been solved. It is unlikely that it will be able to differentiate between the methods by which the recognition sequence is located. However, values for the cleavage rates of the different DNA conformations are likely to be achievable.

Chapter 8

FtsZ

8.1 Chapter summary

Action of the prokaryotic cell division protein FtsZ was modelled, incorporating polymerisation, hydrolysis and depolymerisation. The experimental data used was from light scattering and a phosphate release assay. A Markov chain Monte Carlo algorithm was used for fitting of a number of models.

More complexity in the modelling did not aid fitting. The models producing the best results incorporated a delay in the release of phosphate.

8.2 Background

8.2.1 FtsZ

The FtsZ protein is discussed in section 1.4. In summary, it is a prokaryotic cell division protein that polymerises *in vivo* and *in vitro* in a guanine triphosphate (GTP) dependent manner. Much of its mechanism, including reaction rates and co-ordination with other proteins, remains obscure. 90 ° light scattering and a phosphate release assay have recently been used to measure aspects of the *in vitro* action of FtsZ ([85]). Figure 1.7 summarises what is currently thought to be the mode of action of FtsZ. In this chapter we seek to assess the suitability of models based on this outline and their fit the *in vitro* data. The Markov Chain Monte Carlo (MCMC) algorithm described in chapter 6 is used for fitting.

8.2.2 Previous work

Within publications on FtsZ there is no clear precedent to the work presented here. The most significant attempt to model polymerisation kinetics of FtsZ has been made by Chen *et al* [12]. They used two assays to monitor polymer formation, one with the mutant FtsZ-L68W [1] which shows increased fluorescence upon assembly and the other using fluorescence resonance energy transfer (FRET) to monitor protofilament turnover (see below). Their modelling methods were taken from work on actin, which is also a polymerising protein. Their model did not include hydrolysis or depolymerisation and thus does not directly apply to the data presented here, which are collected where hydrolysis and depolymerisation are expected to be significant factors.

The FRET experiments mentioned above used a F268C mutant of *Escherichia coli* FtsZ. The added cysteine allowed two populations to be labelled, one with fluorescin (FRET donor) and the other with tertramethylrhodamine (TMR, FRET acceptor). When the two populations are added together with GTP, FtsZ assembles as normal with a random mixture of the two fluorophores. In some cases absorbed photons from the donor are re-emitted and absorbed by close acceptor fluorophores. In these experiments the decrease in donor fluorescence due to this effect was measured, although one can also monitor the reabsorbance of the photons from the acceptor. The kinetics of the FRET signal was shown to be proportional to FtsZ concentration. The conclusion of the fitting carried out was that a cooperative assembly process takes place, which

[1] The notation here uses the single letter representation of amino acids to describe replacement. L68W means that the leucine (L) in position 68 has been replaced by a tryptophan (W). Other amino acid labels used here are F, phenylalanine and C, cysteine.

involves a relatively slow nucleation step followed by oligomerisation. In a non-cooperative regime both the initial and subsequent steps have comparable rates.

It was noted by Chen and Erickson that the L68W mutant had a 5–10 fold lower critical concentration, below which assembly does not occur, than wild-type FtsZ. This suggests a stronger affinity and more rapid assembly kinetics than the wild-type. They note that F268C has a slightly reduced hydrolysis rate compared to the wild-type, although they assert that controls show essentially wild-type kinetics. A problem encountered with their fitting algorithms was instability in the the program, which crashed unless the input parameters gave a fairly good fit before running the algorithm. This raises the question that the parameter space may not be properly interrogated in producing the result and it is difficult to assess whether the parameters output correspond to a unique or optimal fitting.

8.3 Methods

An MCMC algorithm using a combination of the Metropolis-Hastings algorithm and Gibbs algorithm (chapter 6) was used in this work to maximise the likelihood function of model solutions described below and experimental data. In all cases the Gibbs sampler was applied to the precision in the likelihood function and the Metropolis-Hastings algorithm used for all other parameters sought.

Matlab (The Mathworks inc., Massachusetts, USA) was used to encode the algorithms and programs, which are shown in appendix A.3.

8.3.1 Experimental data

Experimental data were taken from two experiments [85]. The first of these was 90° light scattering of FtsZ [79]. Upon the addition of guanosine triphosphate (GTP) polymers form. The polymers are assumed to scatter light in a manner proportional to their length, which has shown to be true for long polymers, although some maintain that short polymers are under-represented.

The second experiment was an absorbance assay to measure the amount of phosphate released by hydrolysis [115]. The phosphate release is detected as follows: the substrate MESG (2-amino-6-mercapto-7-methylpurine riboside) is converted into ribose 1-phosphate and 2-amino-6-mercapto-7-methlpurine in the presence of phosphate ions and the enzyme PNP (purine nucleoside phosphorylase). The enzymatic conversion of MESG results in the shift of its

isotropic absorbance maximum from 330 nm to 360 nm. During the polymerisation of FtsZ, GTP is hydrolysed to guanosine diphosphate (GDP) with the release of phosphate ions. A standard assay is used to calibrate the reaction for a given pH value and the output of the assay gives the amount of GTP hydrolysed as a percentage of the initial amount of GTP, which is equivalent to the percentage release of phosphate. In the modelling it is assumed that all of the GTP present is eventually hydrolysed, although in practice the assay plateaus at below 100%. For both light scattering and the phosphate release assay, pH values ranging from 5.5–7.5 were measured. Unless otherwise stated the data used here were measured at pH 6.5, where electron microscopy shows little lateral association between FtsZ oligomers.

8.3.2 Experimental data for the model

The original number of data points (around 1000 for light scattering and 17000 for phosphate release) proved too time consuming for the fitting routines. The experimental data were taken and thinned to 100 data points, by taking equally spaced points from the original data set (figure 8.1).

Each model was programmed and saved as two files, 'diffeqns.m' and 'post.m'. The former defined the differential equations associated with the model and the latter how the data should be processed after the differential equations were solved. An example is given in appendix A.3. For some models a delay function was applied to one of the data sets after calculation as outlined below.

8.3.3 Model programming

Differential equations for each model were determined by assuming the law of mass action: that the rate of change of a species is directly proportional to the concentration of reactants that produce that species. Conservative equations were also generated from conservation of total FtsZ monomers, phosphate ions and GDP molecules. The process is described in this chapter for model 1 and in appendix A.3 for other models.

The differential equations were programmed using Matlab in a form acceptable to be solved by the inbuilt Matlab differential equation solver 'ode45'.

The non-estimated parameters used for simulations were taken from experimental quantities. They were the initial amounts of GTP and FtsZ monomer used: 200 µM and 11 µM respectively.

For each model a set of initial parameters was chosen along with initial step-sizes. According to the theory of MCMC the choice does not affect the final values, although choices closer to the final output will require fewer iterations

to converge. The algorithm was typically run several times with successively decreasing step sizes as the solution converged. The algorithm was restarted if convergence was slow or divergent.

8.3.4 Simulated solutions

To assess the ability of the algorithm to find unique solutions a simulated data set was formed. This was done by altering the parameters in the differential equation to produce a result that appeared, to the eye, a reasonable approximation to the real data and adding noise similar to that on experimental data.

The noise was generated as follows: a smoothed version of the (full) experimental data was subtracted from the data. The variance of the residuals was used as an estimate of the data variance. Normally distributed noise with this variance was added to the model solution in order to mimic the experimental data as closely as possible, but with known values for the model parameters. If the parameters are identifiable and the algorithm efficient the known values will be recovered from simulated data. The standard deviations found were 1.79 for the light scattering data and 0.507 for phosphate release (note that both of these values are unitless).

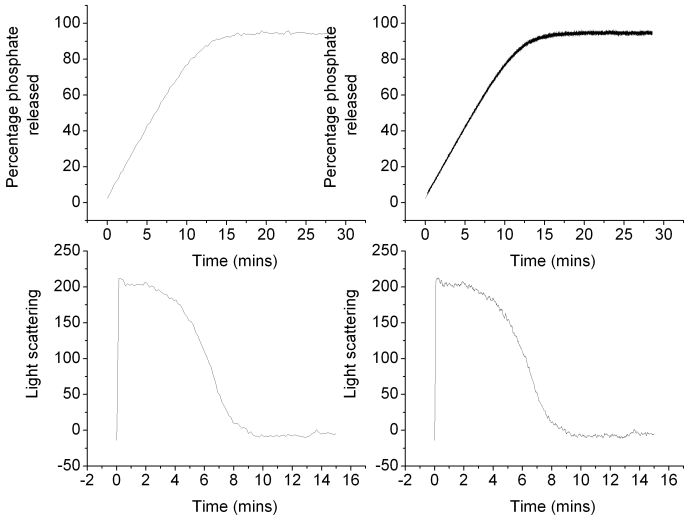


Figure 8.1: Phophate release kinetics (top) and light scattering kinetic data (bottom) showing the original experimental data (right) and thinned versions of the same data with 100 points each (left).

8.4 Results

8.4.1 Notation

Much of the notation is redefined in what follows, but here the main symbols used are summarised for reference purposes. k_i are rate constants, which are estimated by the fitting algorithm; s_X are scaling factors (as defined below) for the term X, also to be estimated. F_0 is the initial concentration of FtsZ monomers used, T_0 is the initial concentration of GTP, f is the concentration of FtsZ monomers bound to GTP in polymer form (FtsZ-GTP), g is the concentration of FtsZ monomers bound to GDP in polymer form (FtsZ-GDP) and p is the concentration of phosphate ions liberated. L_s is the measured value of light scattering and p_{obs} is the measured percentage of phosphate released by hydrolysis.

The measured observables are $L_s(t)$ and $p_{obs}(t)$ where t is time. τ_X is the precision of the data X, the inverse of the variance (which is usually denoted σ^2). As usual, $\dot{X}$ denotes differentiation of species X with respect to time.

8.5 Model 1

Model 1 is the simplest model tested. It allows FtsZ monomers to bind GTP and polymerise in a single step, then hydrolyse the GTP and depolymerise in a further step. The latter step releases FtsZ monomers, GDP and free phosphate ions. We assumed that the released phosphate ions are detected instantaneously, which is supported by evidence from controls for the phosphate release assay and thus the observed phosphate is proportional to the concentration of free phosphate ions.

The equations for model 1 are as follows

$$\text{FtsZ + GTP} \xrightarrow{k_1} \text{FtsZ-GTP} \xrightarrow{k_2} \text{FtsZ + GDP + P}$$

We alter the notation to be

$$\text{F+T} \xrightarrow{k_1} \text{f} \xrightarrow{k_2} \text{F+D+P}$$

Using the law of mass action (see section 8.3.3)

$$\dot{f} = k_1 FT - k_2 f \tag{8.1}$$

$$\dot{p} = k_2 f. \tag{8.2}$$

By conservation of FtsZ, where $F_0 = F(0)$ is the initial concentration of FtsZ monomers

$$F_0 = F + f \tag{8.3}$$

and by conservation of phosphate molecules with $T_0 = T(0)$

$$T_0 = T + f + p \tag{8.4}$$

so equations 8.1 and 8.2 become

$$\dot{f} = k_1(F_0 - f)(T_0 - f - p) - k_2 f \tag{8.5}$$
$$\dot{p} = k_2 f. \tag{8.6}$$

The observed light scattering is assumed to be proportional to the concentration of f and likewise the percentage phosphate released is assumed proportional to p so the following scaling factors are introduced

$$L_s = s_1 f \tag{8.7}$$
$$p_{obs} = s_2 p. \tag{8.8}$$

Hence

$$s_1 L_s = k_1(F_0 - s_1 L_s)(T_0 - s_1 L_s - s_2 p_{obs}) - k_2 s_1 L_s \tag{8.9}$$
$$s_2 p_{obs} = k_2 s_1 L_s \tag{8.10}$$

and finally

$$L_s = k_1 \left(\frac{F_0}{s_1} - L_s \right) (T_0 - s_1 L_s - s_2 p_{obs}) - k_2 L_s \tag{8.11}$$
$$p_{obs} = k_2 L_s \frac{s_1}{s_2}. \tag{8.12}$$

This scheme has simultaneous GTP binding and FtsZ polymersiation and also simultaneous hydrolysis and depolymerisation and is employed as a minimal regime. There are only two rate constants and one scaling factor s_1, which scales for L_s, was estimated. The scaling factor s_2 was calculated from the horizontal part of the phosphate release data where all of the phosphate is assumed to have been released. s_2 was taken to be the ratio of the average of the flat region of the graph to the initial concentration of GTP.

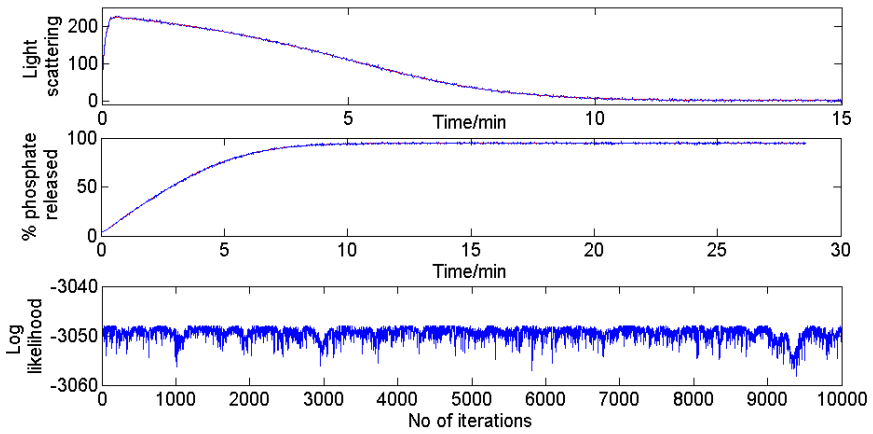


Figure 8.2: Data simulated to test model 1 and fits to it produced by the MCMC algorithm. The parameters are given in table 8.1. The upper two graphs show fitting of the algorithm output (red) to simulated data (blue) which is very close so as to obscure the majority of the former. The bottom graph shows the log likelihood of the Markov chain for the final 10,000 points, demonstrating that the algorithm has converged.

8.5.1 Simulated data for model 1

Data with known parameters, termed simulated data, were generated ,as shown in figure 8.2, by solving the equations for the model and adding noise as described above. Using the MCMC algorithm described in chapter 6, the fit shown in the same figure was determined. The obtained parameters are shown in table 8.1 which show good recovery of the values used for simulation. The corresponding Markov chains are shown in figure A.2.

Throughout the modelling convergence of solutions was determined by three subjective criteria. Firstly, inspection of the Markov chains. Generally a large amount of movement of values occurs at the beginning before the chains settle to a region around the median, after which the chain is deemed to have completed the burn-in period. If strong correlation exists between two parameters this can be seen if two chains follow roughly the same shape (one may be inverted). Secondly, histograms of the Markov chains after the burn-in period should be roughly normal in appearance. Exceptions to this include values at or near 0 since negative values were not permitted during the fitting process. Correlated values may also not give normal distributions, for which a plot of one value against another can be instructive for determining the relationship between them. Thirdly the log likelihood was plotted. When this did not have an increasing trend it was determined that convergence had occurred.

Parameter	10%	50%	90%	Simulated values
k_1	0.06997	0.07028	0.07056	0.07
k_2	4.9718	4.9837	4.9976	5
k_3	0.03501	0.03511	0.03520	0.035
τ_{LS}	0.2995	0.2997	0.3000	0.312
τ_P	4.0212	4.0235	4.0260	3.8903

Table 8.1: Table of fit parameters for model 1 simulated data. 10%, 50% and 90% refer to the percentiles of the Markov chain values; the 50% value is the median.

8.5.2 Experimental data

The best fit of model 1 to experimental data is shown in figure 8.3, the Markov chains are shown in figure A.4 and the output parameters in table 8.2. The data show obviously poor fitting, suggesting that the model is inadequate. In particular the phosphate is poorly fitted. It was possible to produce results where the phosphate fitted well, with the light scattering exhibiting poor fitting (not shown). Since the simulated data was fitted well it suggests that the model does not reflect the real chemistry of the reaction.

Parameter	10%	50%	90%
k_1	0.1901	0.2000	0.2100
k_2	3.302	3.338	3.369
k_3	0.0476	0.0481	0.0487
τ_{LS}	0.03806	0.03823	0.03841
τ_P	0.003866	0.003884	0.003901

Table 8.2: Table of fitted parameters for Model 1 using experimental data

A model (called model 2) where polymerisation and hydrolysis occurred simultaneously was considered with a similar outcome to model 1. These initial models were not expected to produce physically significant results, but were considered as the simplest possible models in order to exclude the possibility that either the hydrolysis or depolymerisation stages were fast in comparison to the rest of the reaction. They were also useful for development of the algorithm as running times are relatively short.

8.6 Model 3

To find a scheme with more physical significance than models 1 or 2, additional steps were represented explicitly. In model 3 the GTP binding and polymerisation are still considered as one step, but the hydrolysis is separated from the depolymerisation. Since, in some cases, the experimental data show residual

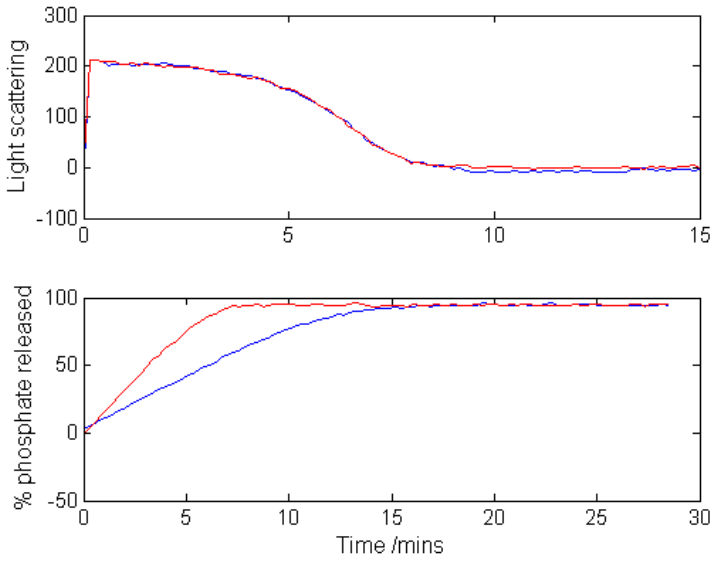


Figure 8.3: Best fit of model 1 to experimental data. Experimental data are shown in blue and the fit produced by the model is shown in red.

binding of monomers to GDP, a term to account for this is included. The chemical equations for this scheme, model 3, are as follows

$$\text{FtsZ} + \text{GTP} \xrightarrow{k_1} \text{FtsZ-GTP} \xrightarrow{k_2} \text{FtsZ-GDP} + \text{P}$$

$$\text{FtsZ-GDP} \underset{k_4}{\overset{k_3}{\rightleftharpoons}} \text{FtsZ} + \text{GDP}$$

which, as shown in appendix A.3, correspond to the differential equations

$$\dot{f} = k_1 \left(\frac{F_0}{s_1} - f - \frac{s_2 g}{s_1} \right) (T_0 - s_1 f - s_3 p) - k_2 f \tag{8.13}$$

$$\dot{g} = \frac{k_2 s_1 f}{s_2} - k_3 g + k_4 (F_0 - s_1 f - s_2 g) \left(\frac{s_3 p}{s_2} - g \right) \tag{8.14}$$

$$\dot{p} = \frac{k_2 s_1 f}{s_3} \tag{8.15}$$

along with the equations for the observables

$$p_{obs} = p \tag{8.16}$$

$$L_s = f + g. \tag{8.17}$$

The first of these is the same as for model 1, whilst the second assumes that both GTP and GDP-bound FtsZ monomers form contribute to light scattering whilst part of polymers.

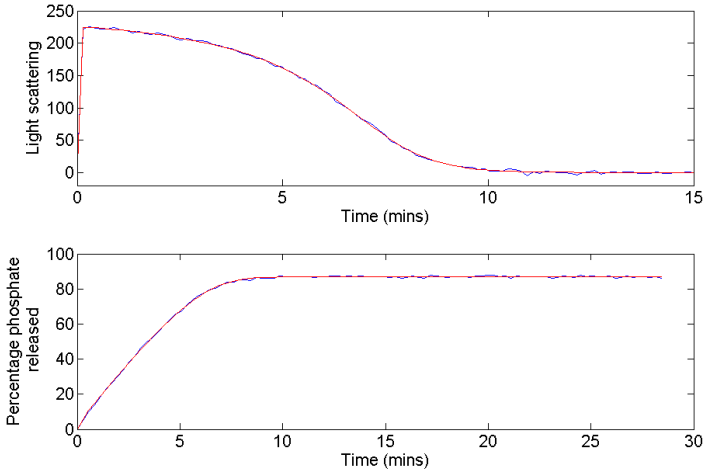


Figure 8.4: Best fit for model 3 using simulated data. Experimental data is shown in blue and the fit produced by the model is shown in red.

8.6.1 Simulated data

The data simulated for model 3 to mimic experimental data are shown in figure 8.4 along with the fit produced by the algorithm. The corresponding parameters are in table 8.3 and show good recovery of values except for k_2 and k_3. Markov chains and the corresponding histograms are found in figures A.6 and A.7. k_2 and k_3 show obvious correlation and, as shown in figure 8.6 their product is a constant. This is somewhat surprising as it is not obvious from the differential equations for the model. An explanation can be gleaned from the cyclic nature of part of the process. Before the concentration of GTP becomes 0, an FtsZ monomer cycles through being in a GTP-bound polymer, a GDP-bound polymer and as monomer in solution. There are two outputs measured in this cycle (polymer concentration and phosphate release), but three rate constants determining the forward progress. It is thus unsurprising that only two values are identifiable. In addition it seems that neither the stages before or after this central cycle are able to add to the identifiability.

Since negative values were not permitted during the fitting process the histogram for k_4 appears somewhat different from the others.

8.6.2 Experimental data and model 3

Using experimental data the fit for model 3 shown in figure 8.5 was obtained, along with the values shown in table 8.4. In a similar manner to model 1, the light scattering is found to fit well but the phosphate does not. With different starting parameters it was again possible to fit the phosphate release curve well and not the light scattering. This suggests that there are two sets

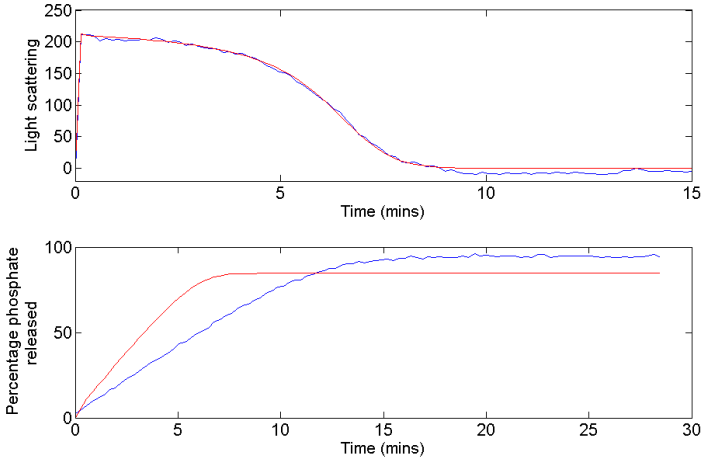


Figure 8.5: Best fit of model 3 to experimental data. Experimental data is shown in blue and the fit produced by the model is shown in red.

Parameter	10%	50%	90%	Simulated values
k_1	0.1233	0.1252	0.1271	0.122
k_2	6.62	6.91	7.26	30
k_3	6.3725	6.68	6.97	4
k_4	3.56×10^{-6}	2.23×10^{-5}	5.57×10^{-5}	0
s_1	0.0503	0.0521	0.0541	0.07
s_2	0.0352	0.0360	0.0372	0.04
s_3	2.2963	2.2983	2.3008	2.3
τ_{LS}	0.3282	0.3297	0.3314	0.3121
k_P	2.670	2.676	2.683	3.890

Table 8.3: Values found for model 3 with simulated data.

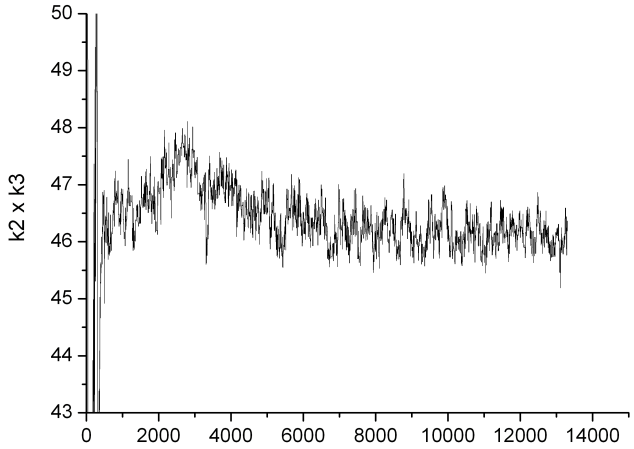


Figure 8.6: k_2 multiplied by k_3 for model 3 with simulated data. The Markov chain of the product tends to a constant mean value.

of parameters that maximised the likelihood function (see equation (6.3)) and thus the model is not sufficient to describe the data.

As with the simulated data set, a notable feature of this fitting is correlation between k_2 and k_3. Both parameters were kept in the model in the expectation that improved models would enable separate identification of these parameters.

Parameter	10%	50%	90%
k_1	0.184	0.194	0.204
k_2	5.00	5.58	6.98
k_3	6.35	8.32	9.94
k_4	1.83×10^{-6}	1.16×10^{-5}	3.80×10^{-5}
s_1	0.04329	0.04607	0.04915
s_2	0.047	0.051	0.057
s_3	2.32	2.36	2.42
τ_{LS}	0.03716	0.03736	0.03748
k_P	0.00537	0.00542	0.00546

Table 8.4: Table of fitted parameters for Model 3 using experimental data. The percentages refer to confidence interval values of the Markov chains, the 50 % value being the median.

8.6.3 Model 4

The next model was an attempt to characterise all parts of the system separately. Based on the work by Chen *et al* [12] the GTP binding to FtsZ polymer was separated, including a nucleation step after the GTP binding to FtsZ monomers. There is no explicit term included for subunit exchange, but since the formation kinetics are so fast we estimate that this will not substantially affect the parameters. In other words, we anticipate that there will be little difference between a monomer incorporated into a polymer through nucleation and one incorporated directly through subunit exchange. It is unlikely that the two mechanisms can be separated by the data and methods used. The rest of the mechanism is based on that represented in figure 1.7, with a hydrolysis phase followed by depolymerisation. Since the terms take no account of explicit molecular shape the change in conformation is implicit in the model.

We alter notation slightly, so that FtsZ-GTP and FtsZ-GDP refer to FtsZ monomers bound to GTP and GDP respectively, but not in polymer form. Polymers are represented in round brackets with a subscript n. The scheme used is shown in figure 8.7.

After several uses of the algorithm it was realised that the rate constants for the reverse processes (i.e. k_3, k_5, k_8 and k_{10}) were estimated to be close to zero for the data being used so a modified version of the system was used. The

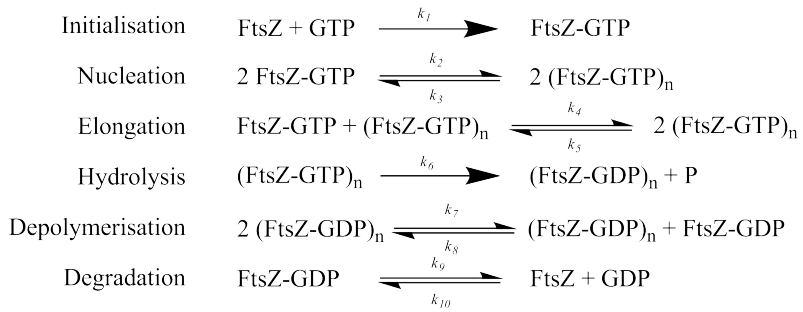
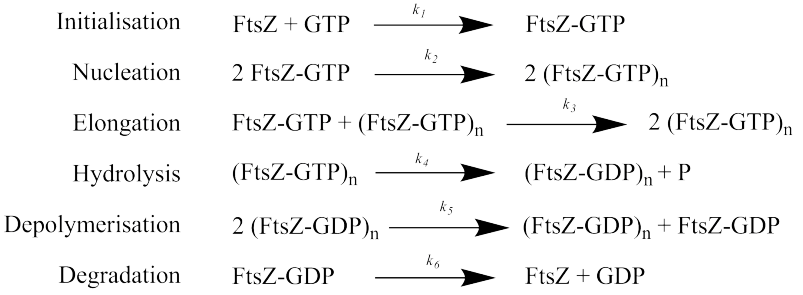


Figure 8.7: The chemical equations used for model 4

modified chemical equations were



The differential equations for this system are derived in A.3 and are

$$\dot{F_T} = k_1(F_0 - f + F_T + F_D + g)(T_0 - f + F_T + p) - 2k_2F_T^2 - k_3F_Tf \quad (8.18)$$

$$\dot{f} = 2k_2F_T^2 + k_3F_Tf - k_4f \quad (8.19)$$

$$\dot{g} = k_4f - k_5g^2 \quad (8.20)$$

$$\dot{F_D} = k_5g^2 - k_6F_D \quad (8.21)$$

$$\dot{p} = k_4f \quad (8.22)$$

where F_T is the concentration of GTP bound FtsZ monomer and F_D is GDP bound monomer.

Results for model 4

Values resulting from fitting of simulated and experimental data for model 4 are shown in tables 8.5 and 8.6 and figures A.11, 8.9, 8.8 and A.10. They exhibit many of the same features as the previous models. Once again either the light scattering or the phosphate release could be well fitted, but not both at the same time. Identification of parameters is very poor and the simulated data shows little or no recovery of the parameters used. This suggested that a different approach was needed. It was concluded that instead of further complicating the model key elements were missing from our models.

Parameter	10%	50%	90%
k_1	0.408	0.433	0.463
k_2	255	274	300
k_3	2.2	12.5	23.6
k_4	28.0	32.3	39.0
k_5	116	160	171
k_6	4.02	4.08	4.15
s_1	125	136	145
s_2	0.412	0.421	0.431
τ_{LS}	0.02641	0.02668	0.02688
τ_P	0.004970	0.004981	0.004993

Table 8.5: Values found for model 4 fitting to experimental data

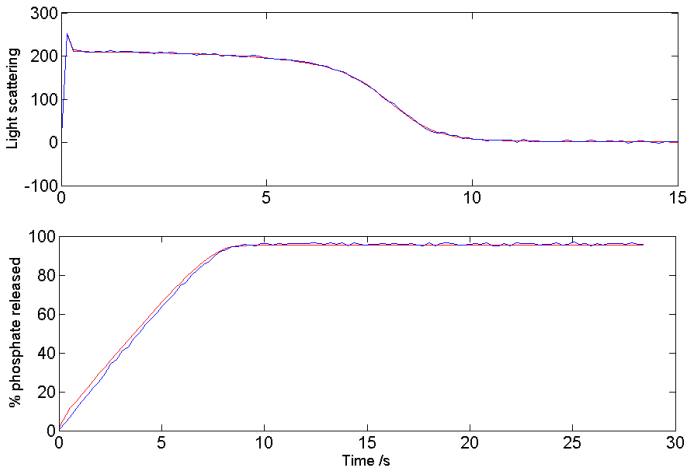


Figure 8.8: Best fit of model 4 to simulated data. Experimental data is shown in blue and the fit produced by the model is shown in red. This shows good recovery of the data.

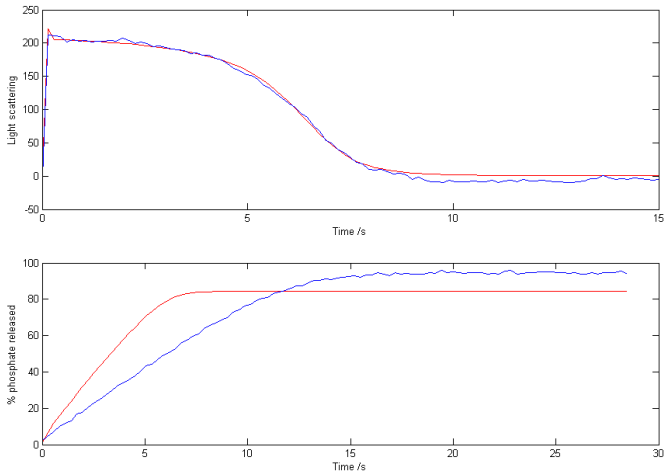


Figure 8.9: Best fit of model 4 to experimental data. Experimental data is shown in blue and the fit produced by the model is shown in red.

Parameter	10%	50%	90%	Simulated values
k_1	0.547	0.578	0.611	0.4
k_2	115	171	222	10
k_3	6.9	13.7	22.6	5
k_4	18.7	19.3	20.1	5
k_5	15.0	16.9	18.2	10
k_6	3.457	3.500	3.531	10
s_1	77.3	80.6	83.0	30
s_2	0.4760	0.4771	0.4782	4.8
τ_{LS}	0.256	0.261	0.267	0.312
τ_P	0.3959	0.4022	0.4059	3.8903

Table 8.6: Values found for model 4 fitting to simulated data. Recovery of values for this model is very poor, suggesting over-complication of the model.

8.7 Model 5: inclusion of delayed phosphate terms

In the models given above one or the other of the data sets can be fitted at once, but it seems that something is missing for a satisfactory fit. It was observed that phosphate is still being released once the light scattering has reached its minimum value (figure 8.1). This prompted the investigation of the inclusion of a delay function. Before proceeding, the possible physical basis for such a delay was considered. Bundling of FtsZ fibres has been observed using electron microscopy [85], so it was postulated that the phosphate produced during hydrolysis may become trapped in the bundles and therefore not be measurable for a time after hydrolysis. Also, FtsZ has positively charged regions on its surface which may impede release of the negatively charged phosphate. The following delay term was therefore introduced

$$p_{obs} = \int_0^t g(\tau)p(t-\tau)d\tau \tag{8.23}$$

where $g(\tau)$ is a normalised distribution function. Thus the observed phosphate is a function of the phosphate released at previous times. Note that this is applied as a post-process to the solution of the differential equations. Two functional forms of g were considered: an exponential distribution and a gamma distribution. The exponential distribution produced better fits than models 1–4, but is omitted here since the gamma distribution fit was even better.

8.7.1 A gamma distributed delay term

A gamma distribution can be considered as a generalisation of an exponential distribution. It contains two parameters, a and b (or sometimes $c = 1/b$). The mean of the distribution is ab and the variance is ab^2.

The form for the gamma distribution used as the function $g(u)$ in equation (8.23) is

$$\Gamma(u|a,b) = \frac{1}{b^a \Gamma(a)} u^{a-1} e^{-\frac{u}{b}}. \tag{8.24}$$

Model 5 is a simplified version of model 3 with a gamma distributed delay applied to the phosphate observable, as follows

$$\dot{f} = k_1(F_0 - f - g)(T_0 - f - p) - k_2 f \tag{8.25}$$

$$\dot{g} = k_2 f - k_3 g \tag{8.26}$$

$$\dot{p} = k_2 f \tag{8.27}$$

with

$$L_s = s_1(f + g) \tag{8.28}$$

$$p_{obs} = s_2 \int_0^t \Gamma(\tau; a, b) p(t - \tau) d\tau \tag{8.29}$$

This is very similar to model 3 given above, although k_4 is set to 0. Since values found for s_1 and s_2 in the notation of model 3 were very similar they are combined as one variable. This means that there is no differentiation between the contributions of GTP bound polymer and GDP bound polymer to the light scattering signal, which is likely since the cross-sections of the two forms are approximately the same.

8.7.2 Simulated data and model 5

Table 8.7 shows good recovery of parameters for model 5. As for model 3, k_2 and k_3 do not appear identifiable from a simulated fit trial. There also appears to be correlation between s_2 and b, although this does not affect identifiability of the parameters.

Parameter	10%	50%	90%	Simualted values
k_1	0.293	0.300	0.308	0.3
k_2	6.6	7.1	8.9	5
k_3	4.70	5.41	5.79	8
s_1	18.922	18.997	19.072	19
s_2	0.4500	0.4505	0.4509	0.45
a	2.863	2.961	3.042	3
b	0.91	0.94	0.97	0.9
τ_{LS}	0.2413	0.2417	0.2421	0.3121
τ_P	4.613	4.626	4.642	3.890

Table 8.7: Parameters for model 5 fitted to simulated data

8.7.3 Experimental data

Since the simulated data showed good identifiability of most values it was expected that the same would be true for experimental data. Figure 8.10 shows the correlation between k_2 and k_3 for these data. The relation between these variables was not as simple as without the delay function where the multiple tended to a constant.

Parameter	10%	50%	90%
k_1	0.182	0.191	0.201
k_2	10.2	14.5	21.6
k_3	4.06	4.45	5.14
s_1	20.650	20.880	21.129
s_2	0.392	0.404	0.412
a	0.890	0.901	0.907
b	5.17	5.39	5.71
τ_{LS}	0.038319	0.038376	0.038434
τ_P	0.15410	0.15451	0.15480

Table 8.8: Parameters for model 5 fitting to experimental data

8.8 Model 6: slow release phosphate

As an alternative to a delay function we considered the possibility that the phosphate can occupy one of two states: reacted or unreacted. Reacted refers to having been detected by the phosphate release assay, whilst unreacted has not yet been detected. In terms of the modelling, this corresponds to addition of a term, as shown in figure 8.12.

There is the possibility that this step does not measure an aspect of the FtsZ system, but is an artefact of the phosphate release assay. We wish to emphasise that controls for the assay, however, show that the reaction takes place almost instantaneously.

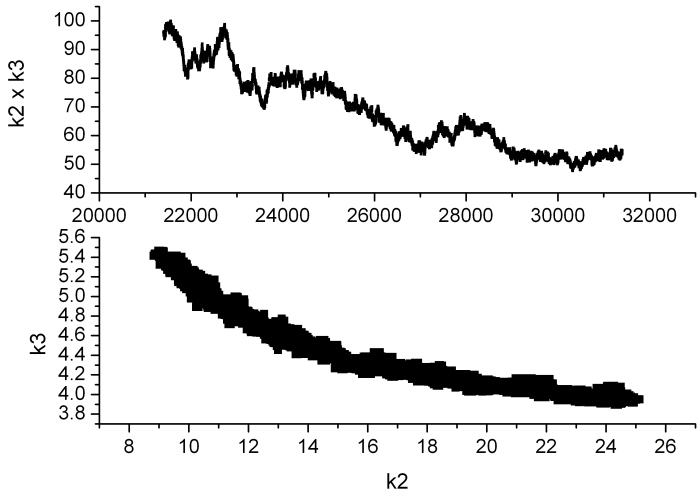


Figure 8.10: k_2 and k_3 for last 10,000 points of the Markov chain for model 5. The top graph shows their multiple as the Markov chain progresses and the bottom the corresponding values in the Markov chain plotted against each other. This shows clear correlation, but is not as simple a relationship as for model 3.

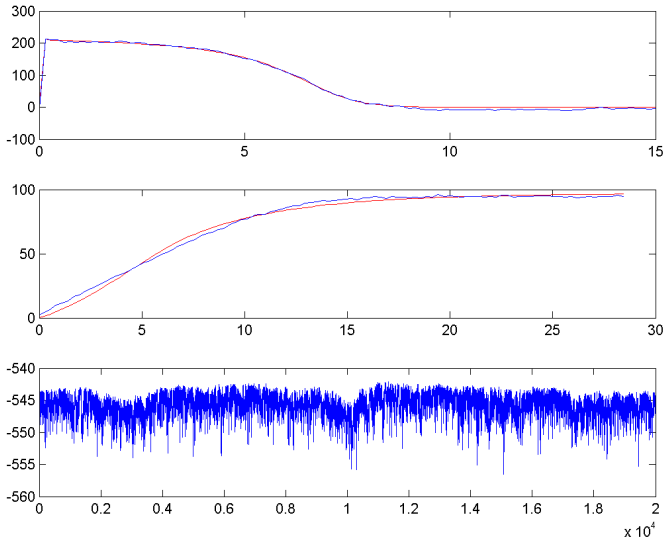


Figure 8.11: The best fit found for model 5 fitted to experimental data.

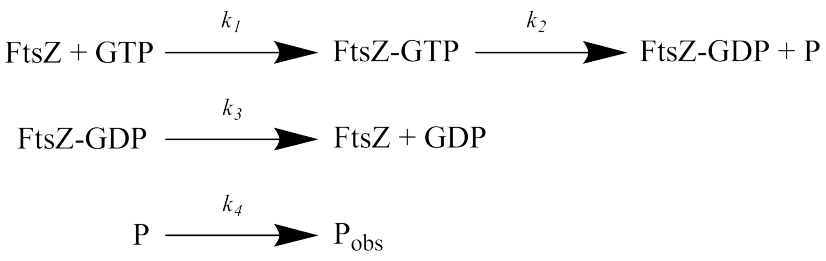


Figure 8.12: The chemical equations used for model 6.

Parameter	10%	50%	90%	Value from simulation
k_1	0.204	0.207	0.211	0.2
k_2	5.101	5.381	6.029	5
k_3	6.220	7.082	7.636	8
k_4	0.295	0.297	0.300	0.3
s_1	19.795	19.862	19.923	20
s_2	0.470	0.470	0.471	0.47
τ_{LS}	0.359	0.359	0.360	0.312
k_P	3.875	3.883	3.890	3.890

Table 8.9: Results for fitting of model 3 with slowly released phosphate to simulated data. The recovery of values is relatively good.

Parameter	10%	50%	90%
k_1	0.183	0.192	0.202
k_2	14.4	18.5	22.3
k_3	3.97	4.11	4.39
k_4	0.264	0.274	0.283
s_1	20.662	20.898	21.135
s_2	0.4727	0.4757	0.4787
τ_{LS}	0.03816	0.03825	0.03833
τ_P	0.1113	0.1125	0.1139

Table 8.10: Table of values for model 6 fitted to experimental data

8.8.1 Simulated data

As previously, to determine identifiability of parameters a simulated data set was derived and fitted to. The results are shown in figures A.15 and A.16 and table 8.9. These confirm good recovery of parameters, although the Markov chains again indicate poor recovery of k_2 and k_3.

8.8.2 Experimental data

Values found for this system are shown in table 8.10. The Markov chains (figure A.17) show that all values apart from τ_P have converged, although this is changing only a small amount. As for model 5, the fitting is much improved from models without a delay function.

8.9 Discussion

The modelling process described above was designed as a careful build up from simple models to more complex. This means that good knowledge about recovery of parameters is garnered. Conventional thinking about the mechanism of FtsZ produced poor fitting of the data (models 1 — 4). Increasing complex-

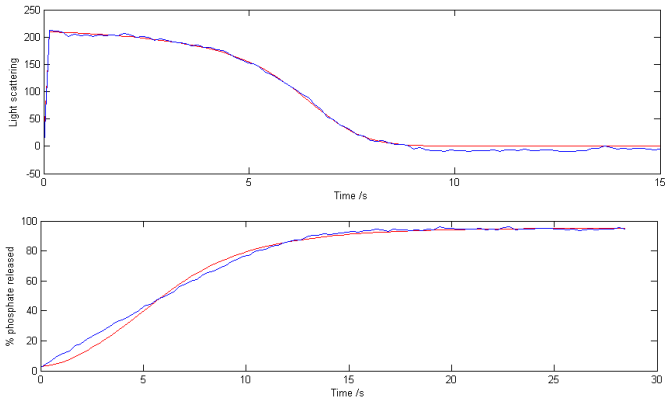


Figure 8.13: Best fit of model 6 to experimental data

ity only led to slower algorithm runs, poor identifiability of parameters and no significant improvement of the fitting. Introduction of a delay, either as a gamma distributed delay function or a slow reaction of the phosphate upon release, greatly improves fitting. This is the key conclusion from the work presented here.

8.9.1 Akaike information criteria

Akaike information criteria (AIC) is a value based on statistical entropy and assesses the goodness of fit of a model. The lower the value the better the fit is considered to be. High likelihood values are good and there is also a term that penalises large numbers of parameters. The equation from which to calculate AIC is very simple:

$$AIC = 2N - 2\log(\mathcal{L}(\theta|X)) \tag{8.30}$$

where N is the number of parameters in the model and $\mathcal{L}(\theta|X)$ is the likelihood function of the model (equation (6.3)) given the parameter set θ and experimental data X. For models 5 and 6 the AIC values are 1109 and 1126 respectively. These are remarkably similar, and on this basis it is difficult to state which provides a better fit — the values can contrasted with 1486 for model 4. This confirms what can be observed by comparison of the final fitting of both models (figures 8.11 and 8.13).

The initial motivation for inclusion of delay terms was from empirical observation of the fitting, so further thought should be given to the physical significance of such a term. At this stage suggestions are speculative, but the most likely explanations are that the phosphate is sticking to the FtsZ or its release is inhibited by the polymer structure.

8.9.2 Parameter identifiability

The problem of identifiability of two of the parameters remains an issue. k_2 and k_3 proved to be strongly correlated for all schemes based on model 3, although the precise nature of the correlation varies with the model. The data shown for model 6 provide reasonable numbers; as can be seen from the Markov chains, however, this is fortuitous and not repeatable.

Several methods can be used to remove the identifiability problem. Firstly, the nature of the correlation for each model can be determined and one of the parameters replaced with a function of the other. If this could be inferred from the differential equations that would be ideal. Here no obvious relationship exists. The relationship may be dependent on the kind of solution found, for example one relationship may be apparent when the term f is small and another when it is large.

Secondly, since the regime employed is Bayesian, knowledge about the parameter values can be input as a prior distribution. This method assumes nothing about the nature of the solution, so would be preferable, although no such information currently exists. No experiment designed so far is able to directly produce values for the rates sought, however. In addition, no other equivalent models exist, except for the very beginning of the system and as described previously there are some doubts as to the validity of these results.

Thirdly, new experiments could be designed to probe additional parts of the system. An assay that could differentiate the GTP and GDP bound FtsZ polymer would be of great use. The results for model 3 give very similar scaling values for these two. The later models based on model 3 used the same scaling constant for both as it was deemed unlikely that they could be separated based on both the model 3 results and the nature of light scattering experiments. There is some indication that linear dichroism measurements can provide this differentiation but there are currently barriers to real-time measurement. The introduction of an injection cell for LD and a new temperature control system may mean that these data are not far off.

8.9.3 Improved modelling

By looking at the resulting fits for both models 5 and 6 it appears that, although good, the result is not perfect. The first part of the experimental data is a straight line for the observed light scattering, whereas both fits contain an S-shaped feature. This suggests a defect in either the data or the model. At the start of measurement there is a dead-time of around 30 s while the sample is loaded into the spectrometer. This has been ignored in the modelling, but it

may be that the data would show an exponential-like increase. Alternatively the model could be adapted. An approach along the lines of Chen and Erickson modelling the building up of individual polymers could be of use, although their approach seemed contradictory by recognising that the average polymer length in solution is around 30 monomers but limiting the lengths in their analysis to much less than this.

8.9.4 Linear dichroism

It may come as no surprise to the reader that linear dichroism can provide useful information about the FtsZ system. The fibrils formed by the protein align well and FtsZ detects only these and neither protein monomers nor GTP or GDP in solution.

FtsZ has been used to measure the polymerisation [85, 72], but only in discrete steps after stopping the reaction after some time, as opposed to the real-time measurements of the other two assays presented previously. The data shows some interesting features.

At 210 nm LD of FtsZ polymerisation shows a very similar pattern to light scattering data over time. In this region LD detects protein backbone. Data from the two methods are separated by a small delay, which may be a result of increased bending of GDP-containing FtsZ polymers observed with electron microscopy. Unlike LD, light scattering does not differentiate between different straight and bent polymers (unless the bends are very severe).

At 278 nm, LD measured over time resembles the data collected from the phosphate release assay. It has been shown that at this wavelength a guanine conformation change upon hydrolysis is detectable with LD [72].

Currently, LD can be used to measure the real-time kinetics of FtsZ polymerisation, but similar problems to those for measurment of restriction enzyme kinetics (chapter 7) are encountered: a long dead-time at the start of the experiment and poor temperature control. The solutions proposed in the restriction enzyme kinetics chapter also apply to FtsZ.

8.9.5 Future work

There is much that can be built on from the findings here. Confirming the source of delay is a key challenge. This may be from looking closely at the FtsZ protein structure and identifying possible areas of affinity for phosphate or seeing if release of phosphate could be structurally hindered.

Variation of conditions will also provide interesting results. Higher pH values show bundling of FtsZ in electron microscopy and this may show itself

as poor fitting or altered rate constant values.

Perhaps most importantly this is *in vitro* data. Our results show properties of FtsZ in isolation and it would be fascinating to observe changes made by its regulatory proteins and how the model requires modification to incorporate their kinetics. By doing this the *in vitro* system can be built up to mimic *in vivo* behaviour.

Chapter 9

Injection linear dichroism

9.1 Chapter summary

A Couette flow cell has been modified for use with a hollow capillary to replace the central rod. This allows sample to be injected directly into the sample chamber. Accuracy of volumes injected using a syringe driver have been checked and shown to contain consistent errors that can easily be corrected for using a fitted equation. Mixing by diffusion between samples held in the inner and outer capillaries was found to be negligible.

Linear dichroism of calf thymus DNA was measured using the injection system and showed good qualitative agreement with data from conventional microvolume Couette flow linear dichroism, although the differences in cell geometry make direct comparison difficult. DNA is shown to align in the inner capillary. These results demonstrate a potential for other applications, particularly kinetic measurements.

9.2 A novel instrument design

Conventional Couette flow LD requires the sample to be put into the capillary and the capillary subsequently placed into the cell before measurement. In some cases this is not adequate, as detailed in other chapters. One solution is to replace the central rod with a narrow capillary through which samples can be injected, which we refer to as injection LD. This has some advantages over the conventional setup. The sample can be equilibrated to a certain temperature before injection. It also greatly reduces the dead-time, the time between combining the reaction components and starting measurement.

Such an instrument is novel and requires testing before being applied to experiments. Some of this work is presented here. The work of Nigel Dyer [24] is acknowledged for determining a preliminary setup and exploring some of the possibilities of injection LD. Unlike our ultimate configuration, Dyer focused on using bubbles to separate reactants and the programming of the syringe driver used. In contrast, this chapter explores preliminary experimental measurements and accuracy of injected volumes.

9.3 The experimental setup

This section details the configuration of the injection cell we have used.

9.3.1 Preparing the rod

The main difference between the injection setup and that of a conventional micro-volume LD experiment is the rod. Instead of a solid quartz rod, a narrow capillary is used through which samples can be introduced into the larger capillary. We refer to them as the inner and outer capillaries respectively. The inner capillary used in this work had an external diameter of 1.9 mm and an internal diameter of 0.73 mm and screwed securely into a modified cell lid.

9.3.2 Connecting the syringe

Figure 9.1 shows the components of the inner capillary. The flexible tubing at the top joins to a syringe, which is placed in a syringe driver to control the amount of sample injected. The syringe driver has its own user interface, as can be seen in figure 9.2, or can be programmed from a computer.

The connector on the flexible tubing screw into the cell lid, which has a hole through it. The hole is straight at the bottom and wider than the capillary, so a plastic joint was constructed. This was a pipette tip cut so that the widest

part is wider than the hole in the cell lid and the narrowest part narrower than the top of the quartz capillary. All of the joints were sealed with PTFE tape. Sealing of the joint between the flexible tubing and the syringe proved particularly important for preventing leakage of air into the system. Also, the amount of PTFE tape used on the top of the capillary provided a means of controlling capillary length, and hence the gap left between the bottom of the internal capillary and the epoxy resin plug at the bottom of the internal capillary. As with the conventional setup the gap was set to 1 mm.

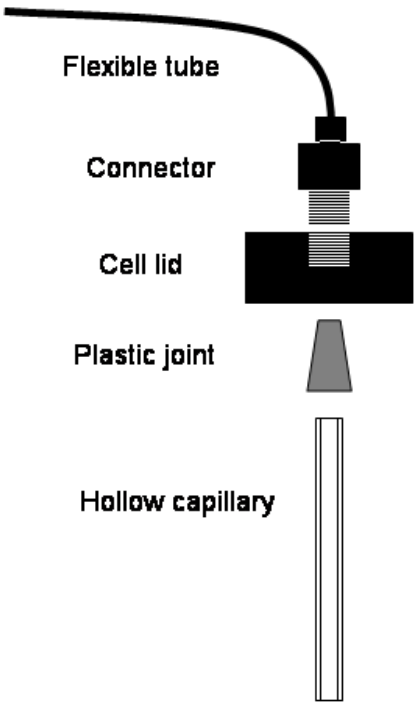


Figure 9.1: Schematic of the components of the inner capillary (not to scale). PTFE tape was used at all the joints to prevent leakage. The space inside the capillary below the joint holds about 20 µl of liquid.

9.4 Injection and withdrawal volumes

The syringe driver uses the syringe diameter to calculate the volume to inject. To verify the accuracy of this, aliquots of MilliQ water injected from the inner capillary into a small beaker were weighed. Several types of experiments were carried out, as detailed in the following.

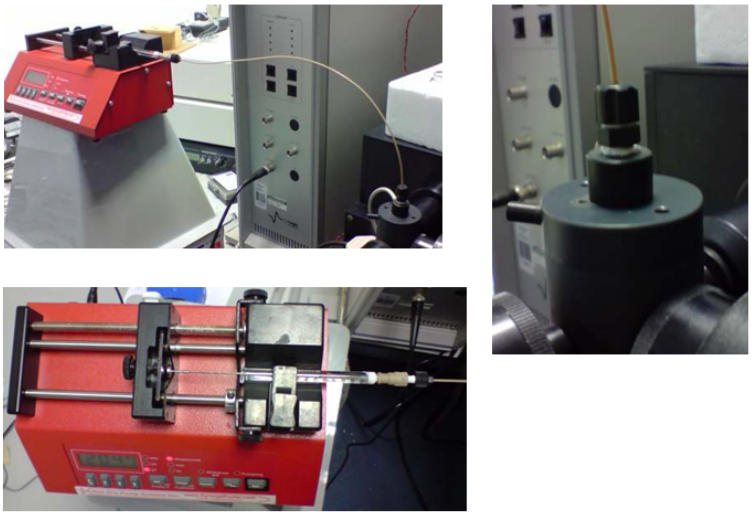


Figure 9.2: The experimental arrangement of the injection cell. The top left shows the overall setup; bottom left shows the syringe driver viewed from above, with the syringe in place; on the right is the connection showing the flexible tubing leading through the cell lid and into the Couette flow cell.

9.4.1 Immediate ejection

In the first set of measurements, different amounts of water were withdrawn through the inner capillary and the corresponding amount immediately ejected and weighed. Since the density of water is 1 g/ml, the volume in millilitres should correspond to the mass in grams. Because of a systematic hysteresis in the syringe driver action, this left a small residual amount in the capillary which was also ejected and weighed, the sum of the two being the amount originally withdrawn. Each measurement was repeated at least three times. Data are presented in figure 9.3.

A straight line was fitted through the data points (using Origin) and corresponded to the equation

$$Y = -0.0016 + 0.875X \tag{9.1}$$

for which Y is the weighed mass in grams and X is the set volume in millilitres. This means that there is a hysteresis of 1.6 µl between withdrawal and ejection. The average amount injected is 87.5% of the intended amount for large volumes.

Straight line fitting of the withdrawn volume yields

$$Y = -0.00065 + 0.901X \tag{9.2}$$

which demonstrates better accuracy than for the volumes ejected and shows smaller hysteresis.

9.4.2 Consecutive ejections

In the second set of experiments a large amount of water was withdrawn into the capillary, an initial aliquot of 10 µl ejected and discarded and subsequent aliquots of the same volume weighed. The average amount ejected was 8.9 µl with a standard deviation of 0.83 µl.

9.4.3 Consecutive withdrawals

In the third, an amount of water in a small beaker was weighed and an initial 10 µl withdrawn with the inner capillary and discarded. Successive aliquots of 10 µl were withdrawn and the remainder in the beaker weighed. The average amount withdrawn was 10.02 µl with a standard deviation of 0.73 µl.

The results from weighing ejected and withdrawn volumes of water show that the most accurate results are produced when withdrawing after a previous withdrawal action (the third set of measurements). Any other measurement carries an error of up to 13%. The errors are consistent, however, with relatively small standard deviations, so more accurate amounts can be measured simply by using fitting equations such as 9.1 and 9.2.

9.5 Mixing over time

Several methods of sample loading were considered. The initial method proceeded as follows. The syringe and flexible tube were filled with water. To withdraw a sample, firstly a small bubble was withdrawn into the inner capillary by holding it in air and withdrawing a small amount. The capillary was then placed into the sample and the sample withdrawn. Subsequent samples could be withdrawn as long as they were preceded by a small bubble. This method is similar to that used by Dyer, however, the bubbles were found to be unstable when withdrawn through the plastic joint at the top of the inner capillary.

The loading method decided upon and used for the sample loading throughout this chapter was a more simple regime. The tube and syringe were filled

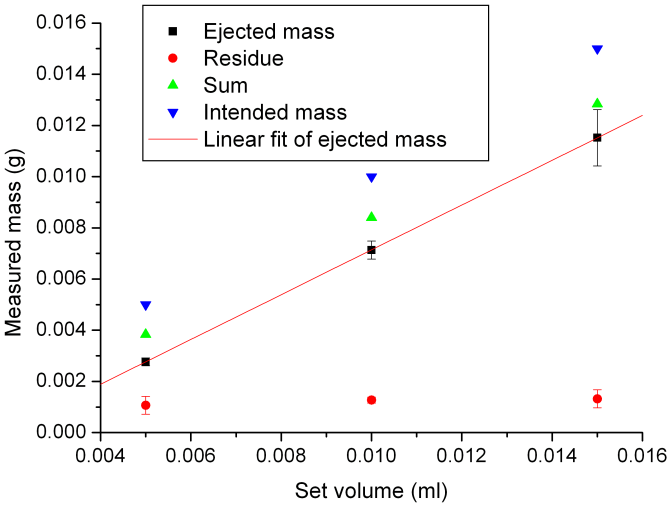


Figure 9.3: A summary of data ejected by the syringe driver as detailed in the text. Ejected mass refers to the amount ejected by the syringe driver when instructed to eject the same amount as was withdrawn. The residue is the amount left in the capillary after this step. The sum corresponds to the amount withdrawn into the capillary. Equation (9.1) gives the fitted line.

with air, the inner capillary placed into the sample solution and the sample withdrawn by the syringe driver. Providing that all of the joints were well sealed, this adequately withdrew samples and held them indefinitely in the tubing without leakage. Care was taken to wipe the outside of the inner capillary to remove excess sample before it was placed into the inner capillary.

A concern with this approach is that after insertion of the inner capillary into the outer capillary prior to measurement, the surfaces of the liquids in each capillary are in contact and diffusion between the two could cause a problem. To probe this effect a non-viscous dye was dissolved in water and withdrawn into the inner capillary. This was placed into the outer capillary outside of the cell and photographed at regular intervals for an hour. Figure 9.4 shows photographs taken immediately after insertion and after 1 hour. There is limited mixing on this time-scale. The inner capillary was found to have a capacity of around 20 µl, so the mixing observable in figure 9.2 is less than 2 µl. We recommend filling the inner capillary with chemicals for which amounts of this order will not affect spectroscopy of the components in the outer capillary.

Figure 9.4: Photographs taken immediately after insertion of the inner capillary into the outer capillary (left) and after one hour. The inner capillary contains a red dye and the outer capillary water. The inner capillary is held at the same height as in the Couette flow cell.

9.6 Linear dichroism of DNA

The injection LD system was employed to measure calf thymus (CT) DNA. CT is known to orient well in conventional Couette flow cells, so measurement in the injection cell gives a good test of its performance.

The inner capillary has a smaller diameter than the central rod normally used with the cell, giving a longer pathlength. The added number of particles in the light path should increase the measured LD, although the Couette flow will be less strong and cause a decreased alignment, implying a smaller LD. Some of the light passes through additional interfaces: solid to liquid into the centre of the rod and then back to solid at the other side. This introduces additional scattering effects. If the scattering is not equal for the two polarisations of light, this will produce a background to spectra that is not flat. The composition of the liquid in the inner capillary is also likely to affect the measurement, especially if it is absorbent.

9.6.1 Materials and methods

CT DNA was (Sigma Aldrich, Dorset, UK) used at a concentration of 200 µM in MilliQ water. LD was measured as outlined in section 2.3.3 with a few modifications. The inner capillary was cleaned only with water and wiped with a soft cloth on the outside, since Dyer [24] has suggested that use of ethanol in the cleaning process promotes mixing between the liquids in the

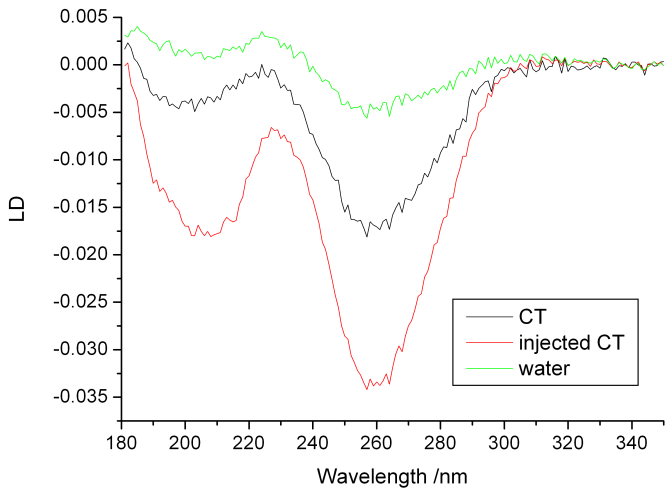


Figure 9.5: LD of CT DNA using the injection cell. A clear DNA signature can be seen in the data. The three spectra refer to CT DNA with water in the capillary (CT), CT DNA injected from the inner capillary, leaving additional CT DNA in the inner capillary (injected CT) and water with CT DNA in the capillary (water). A water spectra was subtracted and the plots were zeroed as described in the text.

inner and outer capillaries. DNA is highly water soluble, so this protocol was deemed sufficient for the samples used. The total volume used in the outer capillary after injection (if any), was 100 µl. The inner capillary was loaded as described in section 9.5. At least 20 µl of liquid were left in the inner capillary during measurement, since air in the inner capillary causes very pronounced scattering effects.

The following methods for obtaining LD spectra were measured:

1. 100 µl of water in the outer capillary and 20 µl of water in the inner capillary as a blank sample.

2. 100 µl of a 200 µM solution of CT DNA in the outer capillary and 20 µl of calf thymus DNA in the inner capillary to mimic normal microvolume LD.

3. 90 µl of water in the outer capillary and 30 µl of 2.1 mM CT DNA in the inner capillary initially. Before measurement 9.4 µl of DNA was injected, giving a total concentration of 200 µM in the outer capillary.

4. 100 µl of water was placed into the outer capillary and 20 µl of DNA in the inner capillary.

9.6.2 Results

For all data presented, a spectrum of water produced using method 1 (described in the previous section) was subtracted. Spectra are shown in figure 9.5. Each spectrum had a water spectrum subtracted and the average LD between 330 and 350 nm was subtracted to zero the data.

Calf thymus DNA with no injection

Figure 9.5 shows measurement of LD using method 2. Qualitatively, the measured LD is very similar to that obtained using a conventional microvolume setup. Direct comparison is difficult because of the difference in pathlength and scattering. It appears, however, that passing light through the inner capillary does not preclude the measurement of LD spectra. There is also no evidence of extensive scattering, which would be manifest as a superimposed increasing trend towards low wavelength.

Injection of calf thymus DNA

The spectrum in figure 9.5 using method 3 (CT DNA with injection) shows a larger LD than for method 2 (without injection). Explanations for this are given below. Again, there is no evidence of light scattering affecting the spectrum.

Spectra of water

To correctly eliminate artefacts in an LD spectrum, it is important to subtract the correct control spectrum from the data. All of the spectra in figure 9.5 have a spectrum of only water, prepared using method 1, subtracted. For injecting CT DNA by method 3, it is better to use a DNA solution in the inner capillary, since this is its contents during measurement with method 3. Method 4 was used to measure this background spectrum, and the result is presented in the figure with the other data. It shows what appears to be a small DNA LD signal, however, this is only part of the story. Figure 9.6 shows the non-rotating spectrum corresponding to methods 1 and 4. For the latter of these, there is a large DNA signal inverted compared to what would be expected. This is indicative of vertical orientation of DNA in the inner capillary.

For method 1, the blank water spectrum, figure 9.6 has an increase towards low wavelength, indicating light scattering. Note that the HT (not shown) indicates that data below 210 nm is not reliable.

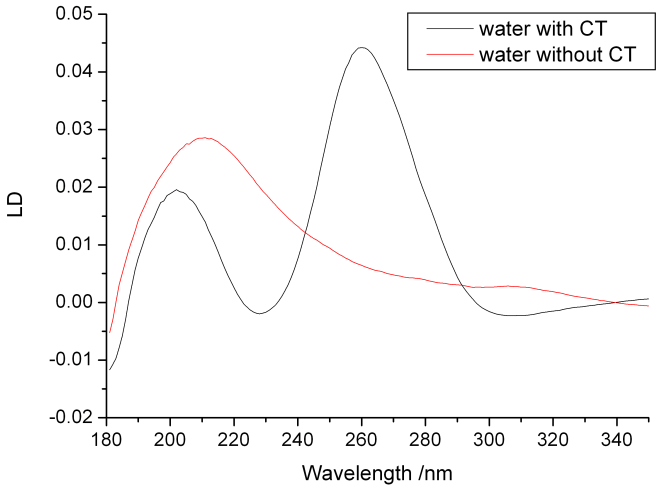


Figure 9.6: Non-rotating LD spectra corresponding to methods 1 (water without CT) and 4 (water without CT) given in the text. The difference between the two was whether CT DNA was used in the inner capillary. With DNA a strong LD signal is observed, whereas without it the increasing trend towards low wavelength is indicative of light scattering.

9.7 Discussion

We have demonstrated that an injection LD system can be constructed and that, in principle, it can be used to take LD measurements in a similar manner to conventional instruments.

9.7.1 Practical considerations

With the approach used here the equipment proved simple to operate. Assembly of the injection apparatus was also straight-forward, providing that some care is taken to seal the joints.

There are some errors measured during withdrawal and injection due to an integral hysteresis in the syringe driver, however, the errors are predictable and accurate amounts can be injection by employing equations 9.1 and 9.2. Minimal diffusion mixing between the inner and outer capillaries means that samples can be held for a period before injection, for example to adjust to a set temperature.

There remain a number of considerations and possible improvements that could be made.

- The syringe driver was simply programmed from the device itself, but software exists to automate the withdrawal and injection of samples. This allows the possibility of relatively complex injection schemes with

several species injected consecutively. Over a short time-scale, withdrawing small volumes, bubbles proved to be effective at separating different reactants in the inner capillary. One should be aware that the bubbles will end up in the cell and the effect of injecting bubbles into the outer capillary should be investigated before routine use of such a method.

- It was observed that using air in the flexible tubing sometimes resulted in erratic injection. This effect is caused by slight compression of the air to overcome friction of the solution being injected. The weighing measurements above suggest that consistent volumes were still injected, but the injection is not necessarily smooth. This could be solved by using a viscous, non-compressible fluid in the tubing. High viscosity would help to stabilise a small bubble which would be necessary to separate the fluid from the injected sample.

9.7.2 Linear dichroism of DNA

LD of DNA showed qualitatively good results, with large LD magnitudes at the appropriate points of the spectrum. This was found both with and without injection.

Water spectra

Above, it has been emphasised that the correct choice of blank spectrum to subtract from data is important. The blank spectra measured using methods 1 and 4 (given above) show some features, evident in figure 9.6.

The non-rotating spectrum using water in both the inner and outer capillary (method 1) exhibits light scattering, not observed in any other spectra. The rotating spectrum (not shown), however, is much the same and thus gives a relatively flat background spectrum.

With DNA in the inner capillary, figure 9.5 shows a small DNA LD, whereas without rotation, figure 9.6 shows a very large, inverted DNA LD. The DNA is being oriented by withdrawal into the inner capillary and the system is thus a combination of a linear flow and Couette flow LD, likely to lead to confusing data. The small LD signal in figure 9.5 is thus due to discrepancies between rotating and non-rotating data, possibly caused by vibrations from the motor disrupting alignment of the inner capillary. Since the DNA is quite concentrated, a leakage of 1 µl from the inner capillary into the outer capillary would also be sufficient to give a small LD. We recommend that experiments using an injection system are designed so as to use a species without any LD

in the inner capillary. Alternatively, the inner capillary can be used on its own to measure alignment of samples as a linear flow system.

9.8 Conclusion

We have shown that an injection system is capable of measuring LD of DNA. Direct comparison with other microvolume Couette flow systems is not possible due to different cell geometry, but the qualitative data are very encouraging. The next step should be to apply it to kinetic measurements, such as restriction enzyme digestion or FtsZ polymerisation.

Chapter 10

Conclusion

This thesis is based broadly around topics related to measuring linear dichroism (LD) of polymers. It is hoped that the reader has been given an appreciation of the possibilities and limitations of this technique. Identifications of improvements have been made at various points and these will be reviewed and discussed in this section, along with general points raised from other aspects of the work.

10.1 Linear dichroism

Couette flow LD is not currently a widely used technique, but is one that has potential for investigating properties of biophysical systems. It can measure *in vitro* polymeric systems in real time without the need for labelling and with minimal sample preparation. It accommodates a wide range of conditions and is relatively cheap and straight-forward to operate.

Many of the difficulties associated with LD lie in the interpretation of the data. Currently, accurate interpretation is limited by several factors, including a good model of semi-flexible polymers in Couette flow, characterisation of the absorbance peaks of DNA and close control of temperature. Each of these are treated in turn in the following.

Semi-flexible polymers in solution

The shape of long DNA is currently determined by models that assume linear hydrodynamics, following the methods of Zimm and Rouse. As discussed in chapter 3, recent evidence suggests that, while these approximations hold for DNA of up to a few hundred base-pairs long in solution or held within molecular tweezers (and similar apparatus), long DNA in Couette flow may contain sharp bends or long-range interactions for which the Rouse-Zimm method does

not work.

Measurements of LD of DNA of the order of a few thousand base pairs long has shown a clear difference from the predictions of the sole theoretical model relating LD and persistence length. Current models are only valid for short DNA under very mild Couette flow.

The completion of coarse-grained molecular modelling work currently being carried out should be able to inform more accurate equations for persistence length of DNA from LD, for example.

DNA absorbance

The perceived wisdom is that DNA has a strong absorbance band at 260 nm caused by a $\pi-\pi^*$ transition. Proposals of the origin of this transition, however, are much more hazy and unclear. As data from synchrotrons becomes available at lower wavelength, the nature of absorbance peaks below 210 nm is necessary for good data interpretation. Such knowledge, somewhat surprisingly, does not exist. This makes it difficult to assign the origins of a peak and shoulder at below 210 nm reported in chapter 2.

With the availability of capillaries made of higher grade quartz, lower wavelengths will become accessible at synchrotron facilities and new, previously unencountered peaks, may be observed. Currently there is no method to reliably predict where these peaks might appear or what their magnitudes might be.

In contrast to the coarse-grained approach discussed in the previous section, this requires a molecular model of DNA in a solvent environment, which is far beyond the scope of this work.

Temperature control of LD

For kinetics experiments with LD the control of temperature is vital. Both molecular orientation and reaction rate are sensitive to temperature and thus, without good control, quantifiable data are not certain to follow. This has been the conclusion of several chapters, along with the discovery that current temperature control methods are inadequate.

Two systems for getting around this problem have been proposed: the injection cell and a newly designed temperature controlled Couette flow cell (chapter 5).

The injection cell presented in chapter 9 is a relatively straight-forward alteration to the currently existing Couette flow cells used for LD. It involves a modified lid and central rod, allowing liquids to be injected directly into the sample chamber. Preliminary investigations have been made using this cell.

The results are very promising and it is hoped that it can soon be used to carry out kinetic LD experiments. Being able to allow the cell and sample to get to the right temperature before injection solves most of the problems associated with temperature control, although warming from the motor is still an issue.

A new design of Couette flow LD cell seems to be the most comprehensive solution to the temperature control problems. Such a cell will shortly be built and chapter 5 contains a possible design, in particular pointing out some of the features which would seem essential to such an instrument.

10.2 Approaches to kinetic modelling

The Markov chain Monte Carlo (MCMC) algorithm is popular in many fields. The methods given in chapter 6 describe a practical and general approach for modelling of a number of systems. The MCMC algorithm is flexible, allowing the input of prior data if it is known and automatically gives an estimate of the uncertainty of values found. The suitability of different models can be assessed by comparison of the likelihood function. A wide range of models can be used — in fact anything that produces a result purporting to represent experimental data with a finite number of variables. Speed is dependent on the complexity of solving the model and the number of parameters estimated. What was somewhat surprising is that no widely available pre-existing package appeared capable of solving the system in this manner. MCMC software is widely available, but does not generally possess the ability to solve differential equations. Equivalent systems use gradient solution methods, but the computational cost of calculating gradients for the differential equations here was deemed too much.

In this work the application of MCMC to FtsZ has revealed a fundamental delay in the way phosphate is released after hydrolysis. In the future, addition of FtsZ cofactors should lead to new models and work towards a quantitative description of prokaryotic cell division.

The future for LD

The future for LD would seem bright, mostly because of its virtual uniqueness in providing orientation information.

From the work presented in this volume, the possibilities for investigating aspects of DNA from low wavelength data gathered at synchrotron sources promises to offer new perspectives on solution structure. Also, the work with restriction enzymes offers a simple, quick method for testing the efficacy of a

given restriction enzyme. In the future it is hoped that mechanistic properties can also be probed using LD. From other work, the orientation of membrane proteins in lipid vesicles [91] by Couette flow is an exciting development.

For whatever purpose, LD offers one piece of the jigsaw. Bulk measurements are broad-brushed, but the lack of need for labelling and versatility of conditions means one can obtain real-time data imitating *in vivo* conditions of many useful classes of biomolecules. The natural place for LD would seem to be as a complement to modelling and less versatile, but more detailed experimental methods, such as microscopy or single molecule techniques. A good example of this is in the area of salt dependence of DNA structure. Results from LD presented in chapter 2 would seem to suggest that coarse-grained models and measurements carried out using x-ray crystallography are not currently providing general predictions for the interaction between DNA in solution and positively charged ions.

List of abbreviations

DNA	deoxyribonucleic acid
A	the DNA base adenine
T	the DNA base thymine
C	the DNA base cytosince
G	the DNA base guanine
RNA	ribonucleic acid
LD	linear dichroism
CD	circular dichroism
PEM	photoelastic modulator
MCMC	Markov Chain Monte Carlo
FtsZ	Filamentous temperature sensitive protein Z
FRET	fluorescence resonance energy transfer
GTP	guanine triphosphate
GDP	guanine diphosphate
NMR	nuclear magnetic resonance
OSF	the Odijk, Skolman, Fixman model of a charged polymer
AT	poly(A) poly(T)
GC	poly(G) poly(C)
CP	DNA from the bacterium *Clostridium perfringens*
CT	calf thymus DNA
ssA	single-stranded poly(A)
ML	DNA from the bacterium *Micrococcus luteus*
NaF	sodium fluoride
HT	high tension voltage
$\mathbb{R}$	the set of real numbers
FJC	freely jointed chain
$\langle . \rangle$	Averaging over all possible states. For a quantity $X(\phi)$ where $\psi(\phi)$ is the probability of the state variable ϕ, $\langle X(\phi) \rangle = \frac{\int_\phi X(\phi)\psi(\phi)}{\int_\phi \psi(\phi)}$
WLC	worm-like chain
TEB	transient electric birefringence

PCA	principal component analysis
EM	electron microscopy
AFM	atomic force microscopy
bp	base-pairs
UV	ultra-violet
E. coli	The bacterium *Escherichia coli*
LB	luria broth
rpm	revolutions per minute
g	acceleration due to gravity
10X	in reference to a buffer: the buffer should be used at a ten-fold dilution
EDTA	ethylenediaminetetraacetic acid
NaCl	sodium chloride
RT	room temperature
$[0, 1]$	the (closed) interval between 0 and 1
$\forall$	for all
$\mathbb{N}_{\geq 2}$	the set of natural numbers greater than or equal to 2
$(E_j)_{j=a}^{b}$	a sequence of numbers indexed by the integers from a to b (inclusive)
$\lim_{n\to\infty}$	the limit of a function as the value n tends to infinity
x_{-i} =	for a sequence $(x_i)_{i=1}^{n}$, $\{x_1, x_2, ..., x_{i-1}, x_{i+1}, ...x_n\}$
RE	restriction enzyme
ATP	adenosine triphosphate
H^3	the isotope of hydrogen containing 1 proton and 2 neutrons
TIR	total internal reflection
U	Unit, in reference to restriction enzymes, the amount of enzyme required to cleave 1 µg of lambda DNA in 1 hour at the correct temperature and in the correct buffer for the enzyme in a total volume of 50 µl
dNTPs	deoxyribonucleotide triphosphates
TMR	tertramethyl rhodamine
L,W,F,C	the amino acids leucine, tryptophan, phenylalanine and cysteine respectively
L68W	a protein in which the leucine at position 68 has been replaced with a tryptophan
MESG	2-amino-6-mercapto-7-methylpurine riboside

MilliQ water	Water purified to a resistivity of 18.2 Ω or greater using a MilliQ water purification unit (Millipore, Massachusetts, USA)

Bibliography

[1] R. Adachi, K. Yamaguchi, H. Yagi, K. Sakurai, H. Naiki, and Y. Goto. Flow-induced alignment of amyloid protofilaments revealed by linear dichroism. *J Biol Chem*, 282(12):8978–83, 2007.

[2] K. Balali-Mood, T. A. Harroun, and J. P. Bradshaw. Molecular dynamics simulations of a mixed dopc/dopg bilayer. *Eur Phys J E Soft Matter*, 12 Suppl 1:135–40, 2003. 1292-8941 (Print) Journal Article.

[3] J. Bednar, P. Furrer, V. Katritch, A.Z. Stasiak, J. Dubochet, and A. Stasiak. Determination of dna persistence length by cryo-electron microscopy. separation of the static and dynamic contributions to the apparent persistence length of dna. *J. Mol. Biol.*, 254:579–594, 1995.

[4] S. R. Bellamy, S. E. Milsom, Y. S. Kovacheva, R. B. Sessions, and S. E. Halford. A switch in the mechanism of communication between the two dna-binding sites in the sfii restriction endonuclease. *J Mol Biol*, 373(5):1169–83, 2007.

[5] V.A. Bloomfield, D.M. Crothers, and I. Tinoco. *Nucleic Acids*. University Science Books, Sausalito, CA, 2000.

[6] I. Bonnet, A. Biebricher, P. L. Porte, C. Loverdo, O. Benichou, R. Voituriez, C. Escude, W. Wende, A. Pingoud, and P. Desbiolles. Sliding and jumping of single ecorv restriction enzymes on non-cognate dna. *Nucleic Acids Res*, 36(12):4118–27, 2008.

[7] C. Bouchiat, M. D. Wang, J. Allemand, T. Strick, S. M. Block, and V. Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophys J*, 76(1 Pt 1):409–13, 1999.

[8] S Broersma. Rotation diffusion constant of a cylindrical particle. *J. Chem. Phys.*, 32(6):1626–1631, 1960.

[9] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith. Entropic elasticity of lambda-phage dna. *Science*, 265(5178):1599–600, 1994.

[10] P. R. Callis. Flow dichroism of dna: a new apparatus and further studies. *Biopolymers*, 7(3):335–52, 1969.

[11] L. E. Catto, S. Ganguly, S. E. Milsom, A. J. Welsh, and S. E. Halford. Protein assembly and dna looping by the foki restriction endonuclease. *Nucleic Acids Res*, 34(6):1711–20, 2006.

[12] Y. Chen, K. Bjornson, S. D. Redick, and H. P. Erickson. A rapid fluorescence assay for ftsz assembly indicates cooperative assembly with a dimer nucleus. *Biophys J*, 88(1):505–14, 2005.

[13] Y. Chen and H. P. Erickson. Rapid in vitro assembly dynamics and subunit turnover of ftsz demonstrated by fluorescence resonance energy transfer. *J Biol Chem*, 280(23):22549–54, 2005.

[14] S. Chib and E. Greenberg. Bayes inference in regression-models with arma (p, q) errors. *Journal of Econometrics*, 64(1-2):183–206, 1994.

[15] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *American Statistician*, 49(4):327–335, 1995.

[16] L. J. Childs, J. Malina, B. E. Rolfsnes, M. Pascu, M. J. Prieto, M. J. Broome, P. M. Rodger, E. Sletten, V. Moreno, A. Rodger, and M. J. Hannon. A dna-binding copper(i) metallosupramolecular cylinder that acts as an artificial nuclease. *Chemistry*, 12(18):4919–27, 2006.

[17] N. A. Clark, K. J. Rothschild, D. A. Luippold, and B. A. Simon. Surface-induced lamellar orientation of multilayer membrane arrays. theoretical analysis and a new method with application to purple membrane fragments. *Biophys J*, 31(1):65–96, 1980.

[18] A. E. Cohen and W. E. Moerner. Principal-components analysis of shape fluctuations of single dna molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12622–12627, 2007.

[19] M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[20] J.G. Voet D. Voet. *Biochemistry, 3rd Edition*. Wiley, 2004.

[21] T. R. Dafforn and A. Rodger. Linear dichroism of biomolecules: which way is up? *Curr Opin Struct Biol*, 14(5):541–6, 2004.

[22] C. Dicko, M. R. Hicks, T. R. Dafforn, F. Vollrath, A. Rodger, and S. V. Hoffman. Breaking the 200 nm limit for routine flow linear dichroism measurements using uv synchrotron radiation. *Biophys J*, 2008.

[23] M. Doi and S.F. Edwards. *The Theory of Polymer Dynamics*. The International Series of Monographs on Physics. Clarendon Press, Oxford, 1986.

[24] Nigel Dyer. Thesis in contribution to the degree of master of science, 2008.

[25] S. R. Eddy. What is bayesian statistics? *Nat Biotechnol*, 22(9):1177–8, 2004.

[26] M. Egli. Dna-cation interactions: Quo vadis? *Chemistry and Biology*, 9(3):277–286, 2002.

[27] M. L. Embleton, A. V. Vologodskii, and S. E. Halford. Dynamics of dna loop capture by the sfii restriction endonuclease on supercoiled and relaxed dna. *J Mol Biol*, 339(1):53–66, 2004.

[28] B. Eraker. Mcmc analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19(2):177–191, 2001.

[29] M Fixman and J Kovac. Polymer conformational statistics.3. modified gaussian models of stiff chains. *Journal of Chemical Physics*, 58(4):1564–1568, 1973.

[30] P.J. Flory. *Principles of polymer chemistry*. Cornell University Press, New York, 1953.

[31] C. Forrey and M. Muthukumar. Langevin dynamics simulations of genome packing in bacteriophage. *Biophys J*, 91(1):25–41, 2006.

[32] C. Frontali, E. Dore, A. Ferrauto, E. Gratton, A. Bettini, M. R. Pozzan, and E. Valdevit. An absolute method for the determination of the persistence length of native dna from electron micrographs. *Biopolymers*, 18(6):1353–73, 1979.

[33] M. P. Fulscher and B. O. Roos. Theoretical-study of the electronic-spectrum of cytosine. *Journal of the American Chemical Society*, 117(7):2089–2095, 1995.

[34] M. P. Fulscher, L. SerranoAndres, and B. O. Roos. A theoretical study of the electronic spectra of adenine and guanine. *Journal of the American Chemical Society*, 119(26):6168–6176, 1997.

[35] G.C. Tiao G.E.P. Box. *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Company, 1973.

[36] D. M. Gowers, S. R. Bellamy, and S. E. Halford. One recognition sequence, seven restriction enzymes, five reaction mechanisms. *Nucleic Acids Res*, 32(11):3469–79, 2004.

[37] Jack D. Graybeal. *Molecular Spectroscopy*. McGraw-Hill, 1988.

[38] P. J. Hagerman. Flexibility of dna. *Annu Rev Biophys Biophys Chem*, 17:265–86, 1988.

[39] PJ Hagerman and BH Zimm. Monte-carlo approach to the analysis of the rotational diffusion of wormlike chains. *Biopolymers*, 20(7):1481–1502, 1981.

[40] S. E. Halford. Hopping, jumping and looping by restriction enzymes. *Biochem Soc Trans*, 29(Pt 4):363–74, 2001.

[41] M.J. Hannon, V. Moreno, M.J. Prieto, E. Molderheim, E. Sletten, I. Meistermann, C.J. Isaac, K.J. Sanders, and A. Rodger. Intramolecular dna coiling mediated by a metallo supramolecular cylinder. *Angew. Chem. Intl. Ed.*, 40:879–884, 2001.

[42] H. G. Hansma, K. J. Kim, D. E. Laney, R. A. Garcia, M. Argaman, M. J. Allen, and S. M. Parsons. Properties of biomolecules measured from atomic force microscope images: a review. *J Struct Biol*, 119(2):99–108, 1997.

[43] M. Harbst, T. N. Hansen, C. Caleman, W. K. Fullagar, P. Jonsson, P. Sondhauss, O. Synnergren, and J. Larsson. Studies of resolidification of non-thermally molten insb using time-resolved x-ray diffraction. *Applied Physics a-Materials Science & Processing*, 81(5):893–900, 2005.

[44] W. K. Hastings. Monte-carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–, 1970.

[45] J.E. Hearst. Rotary diffusion constants of stiff-chain macromolecules. *J. Chem. Phys.*, 38(5):1062–1065, 1963.

[46] F. Hennies, A. Fohlisch, W. Wurth, P. Feulner, A. Fink, and D. Menzel. Adsorption and thermal evolution of so2 on ru (0001). *Journal of Chemical Physics*, 127(15):–, 2007.

[47] E. A. Heron, B. Finkenstadt, and D. A. Rand. Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. *Bioinformatics*, 23(19):2596–603, 2007.

[48] M. P. Heyn, R. J. Cherry, and U. Muller. Transient and linear dichroism studies on bacteriorhodopsin: determination of the orientation of the 568 nm all-trans retinal chromophore. *J Mol Biol*, 117(3):607–20, 1977.

[49] M. R. Hicks, A. Rodger, C. M. Thomas, S. M. Batt, and T. R. Dafforn. Restriction enzyme kinetics monitored by uv linear dichroism. *Biochemistry*, 45(29):8912–7, 2006.

[50] M.R. Hicks, A. Damianoglou, A. Rodger, and Dafforn T.R. Folding and membrane insertion of the pore-forming peptide gramicidin occurs as a concerted process. *Journal of Molecular Biology*, Accepted for publication, 2008.

[51] E. Holmstrom, W. Olovsson, I. A. Abrikosov, A. M. N. Niklasson, B. Johansson, M. Gorgoi, O. Karis, S. Svensson, F. Schafers, W. Braun, G. Ohrwall, G. Andersson, M. Marcellini, and W. Eberhardt. Sample preserving deep interface characterization technique. *Physical Review Letters*, 97(26):–, 2006.

[52] S. B. Howerton, C. C. Sines, D. VanDerveer, and L. D. Williams. Locating monovalent cations in the grooves of b-dna. *Biochemistry*, 40(34):10023–10031, 2001.

[53] N. V. Hud and M. Polak. Dna-cation interactions: the major and minor grooves are flexible ionophores. *Current Opinion in Structural Biology*, 11(3):293–301, 2001.

[54] D. Ivanov, J. T. Sage, M. Keim, J. R. Powell, S. A. Asher, and P. M. Champion. Determination of co orientation in myoglobin by single-crystal infrared linear dichroism. *Journal of the American Chemical Society*, 116(9):4139–4140, 1994.

[55] M. A. Karymov, A. V. Krasnoslobodtsev, and Y. L. Lyubchenko. Dynamics of synaptic sfii-dna complex: single-molecule fluorescence analysis. *Biophys J*, 92(9):3241–50, 2007.

[56] J.M. Kenney, D.P. Knight, C. Dicko, and F. Vollrath. Linear and circular dichroism can help us to understand the molecular nature of spider silk. *European Arachnology*, pages 123–126, 2000.

[57] O. Kratky and G. Porod. Rontgenuntersuchung geloster fadenmolekule. *Recl. Trav. Chim. Pays-Bas*, 68:1106–1122, 1949.

[58] W. Kuhn. Beziehungen zwischen molekulgrosse, statistischer molekulgestalt und elastischen eigenschaften hochpolymerer stoffe. *Kolloid. Z.*, 76:258, 1936.

[59] E.M. Lifshitz L.D. Landau. *Theory of Elasticity, 3rd Edition.* Butterworth-Heinemann, 1986.

[60] C. S. Lee and N. Davidson. Flow dichroism of deoxyribonucleic acid solutions. *Biopolymers*, 6(4):531–50, 1968.

[61] P. A. Levin, I. G. Kurtser, and A. D. Grossman. Identification and characterization of a negative regulator of ftsz ring formation in bacillus subtilis. *Proc Natl Acad Sci U S A*, 96(17):9642–7, 1999.

[62] A. Lewis and S. Bridle. Cosmological parameters from cmb and other data: A monte carlo approach. *Physical Review D*, 66(10):–, 2002.

[63] Y. Lu, B. Weers, and N. C. Stellwagen. Dna persistence length revisited. *Biopolymers*, 61(4):261–75, 2001.

[64] Y. Lu, B. D. Weers, and N. C. Stellwagen. Analysis of dna bending by transient electric birefringence. *Biopolymers*, 70(2):270–88, 2003.

[65] C. Machicado, M. Bueno, and J. Sancho. Predicting the structure of protein cavities created by mutation. *Protein Eng*, 15(8):669–75, 2002.

[66] G. S. Manning. The contribution of transient counterion imbalances to dna bending fluctuations. *Biophys J*, 90(9):3208–15, 2006.

[67] G. S. Manning. The persistence length of dna is reached from the persistence length of its null isomer through an internal electrostatic stretching force. *Biophys J*, 91(10):3607–16, 2006.

[68] G.S. Manning. Polymer persistence length characterized as a critical length for instability caused by a fluctuating twist. *Phys. Rev. A*, 34(1):668–670, 1986.

[69] J. Marek, E. Demjenova, Z. Tomori, J. Janacek, I. Zolotova, F. Valle, M. Favre, and G. Dietler. Interactive measurement and characterization of dna molecules by analysis of afm images. *Cytometry A*, 63(2):87–93, 2005.

[70] R. Marrington, T. R. Dafforn, D. J. Halsall, J. I. MacDonald, M. Hicks, and A. Rodger. Validation of new microvolume couette flow linear dichroism cells. *Analyst*, 130(12):1608–16, 2005.

[71] R. Marrington, T. R. Dafforn, D. J. Halsall, and A. Rodger. Microvolume couette flow sample orientation for absorbance and fluorescence linear dichroism. *Biophys J*, 87(3):2002–12, 2004.

[72] R. Marrington, E. Small, A. Rodger, T. R. Dafforn, and S. G. Addinall. Ftsz fiber bundling is triggered by a conformational change in bound gtp. *J Biol Chem*, 279(47):48821–9, 2004.

[73] M. Matsumoto. Transient electric birefringence of colloidal particles immersed in shear flow. part ii. the initial response under the action of a rectangular electric pulse and the behavior at a low alternating electric field. *Biophys Chem*, 58(1-2):173–83, 1996.

[74] Y. Matsuoka and B. Norden. Linear dichroism studies of nucleic-acid bases in stretched polyvinyl-alcohol) film - molecular-orientation and electronic-transition moment directions. *Journal of Physical Chemistry*, 86(8):1378–1386, 1982.

[75] http://www.marligen.com/inserts/powerprep/11452-018.pdf.

[76] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21(6):1987–1092, 1953.

[77] A. J. Miles and B. A. Wallace. Synchrotron radiation circular dichroism spectroscopy of proteins and applications in structural and functional genomics. *Chem Soc Rev*, 35(1):39–51, 2006.

[78] C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res*, 13(11):2467–74, 2003.

[79] A. Mukherjee and J. Lutkenhaus. Analysis of ftsz assembly by light scattering and determination of the role of divalent metal cations. *J Bacteriol*, 181(3):823–32, 1999.

[80] P. C. Nelson, C. Zurla, D. Brogioli, J. F. Beausang, L. Finzi, and D. Dunlap. Tethered particle motion as a diagnostic of dna tether length. *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys*, 110(34):17260–7, 2006.

[81] E. Nogales, K. H. Downing, L. A. Amos, and J. Lowe. Tubulin and ftsz form a distinct family of gtpases. *Nat Struct Biol*, 5(6):451–8, 1998.

[82] T. Odijk. Polyelectrolytes near the rod limit. *J. Poly. Sci. Poly. Phys. Ed.*, 15:477–483, 1977.

[83] T Odijk. Stiff chains and filaments under tension. *Macromolecules*, 28(28):7016–7018, 1995.

[84] P. J. Oriel and J. A. Schellman. Studies of the birefringence and birefringence dispersion of polypeptides and proteins. *Biopolymers*, 4(4):469–94, 1966.

[85] Raul Pacheco-Gomez. *Structural studies and assembly dynamics of the bacterial cell division protein FtsZ*. PhD thesis, University of Warwick, 2008.

[86] J. D. Petke, G. M. Maggiora, and R. E. Christoffersen. Abinitio configuration-interaction and random phase approximation calculations of the excited singlet and triplet-states of adenine and guanine. *Journal of the American Chemical Society*, 112(14):5452–5460, 1990.

[87] J. D. Petke, G. M. Maggiora, and R. E. Christoffersen. Abinitio configuration-interaction and random phase approximation calculations of the excited singlet and triplet-states of uracil and cytosine. *Journal of Physical Chemistry*, 96(17):6992–7001, 1992.

[88] A. S. Petrov and S. C. Harvey. Structural and thermodynamic principles of viral packaging. *Structure*, 15(1):21–7, 2007.

[89] M. Pillsbury, H. Orland, and A. Zee. Steepest descent calculation of rna pseudoknots. *Phys Rev E Stat Nonlin Soft Matter Phys*, 72(1 Pt 1):011911, 2005.

[90] A. Pingoud, M. Fuxreiter, V. Pingoud, and W. Wende. Type ii restriction endonucleases: structure and mechanism. *Cell Mol Life Sci*, 62(6):685–707, 2005.

[91] J. Rajendra, A. Damianoglou, M. Hicks, P. Booth, P. M. Rodger, and A. Rodger. Quantitation of protein orientation in flow-oriented unilamellar liposomes by linear dichroism. *Chemical Physics*, 326(1):210–220, 2006.

[92] A. D. Richards and A. Rodger. Synthetic metallomolecules as agents for the control of dna structure. *Chemical Society Reviews*, 36(3):471–483, 2007.

[93] C. Rivetti, M. Guthold, and C. Bustamante. Scanning force microscopy of dna deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J Mol Biol*, 264(5):919–32, 1996.

[94] V. Rizzo and J. Schellman. Flow dichroism of t7 dna as a function of salt concentration. *Biopolymers*, 20(10):2143–63, 1981.

[95] M. V. Rogacheva, A. V. Bochenkova, S. A. Kuznetsova, M. K. Saparbaev, and A. V. Nemukhin. Impact of pyrophosphate and o-ethyl-substituted pyrophosphate groups on dna structure. *J Phys Chem B*, 111(2):432–8, 2007.

[96] L. Romberg and P. A. Levin. Assembly dynamics of the bacterial cell division protein ftsz: poised at the edge of stability. *Annu Rev Microbiol*, 57:125–54, 2003.

[97] W. Ruland and B. M. Smarsly. Two-dimensional small-angle x-ray scattering of self-assembled nanocomposite films with oriented arrays of spheres: determination of lattice type, preferred orientation, deformation and imperfection. *Journal of Applied Crystallography*, 40:409–417, 2007.

[98] Jr. SantaLucia, J. and D. Hicks. The thermodynamics of dna structural motifs. *Annu Rev Biophys Biomol Struct*, 33:415–40, 2004.

[99] E. S. G. Shaqfeh and G. H. Fredrickson. The hydrodynamic stress in a suspension of rods. *Physics of Fluids a-Fluid Dynamics*, 2(1):7–24, 1990.

[100] J. Shimada and H. Yamakawa. Ring-closure probabilities for twisted wormlike chains - application to dna. *Macromolecules*, 17(4):689–698, 1984.

[101] D. Shore, J. Langowski, and R. L. Baldwin. Dna flexibility studied by covalent closure of short fragments into circles. *Proc Natl Acad Sci U S A*, 78(8):4833–7, 1981.

[102] T. Simonson and M Kubista. Dna orientation in shear flow. *Biopolymers*, 33:1225–1235, 1993.

[103] J. Skolnick and M Fixman. Electrostatic persistence length of a wormlike polyelectrolyte. *Macromolecules*, 10(5):944–948, 1977.

[104] S. B. Smith, L. Finzi, and C. Bustamante. Direct mechanical measurements of the elasticity of single dna molecules by using magnetic beads. *Science*, 258(5085):1122–6, 1992.

[105] A. Stasiack. Personal communication, 2006.

[106] L. Stryer. *Biochemistry, Fourth Edition.* W.H. Freeman and Company, 1995.

[107] L. Stryer. *Biochemistry, Fourth Edition*, pages 977–982. W.H. Freeman and Company, 1995.

[108] Z. J. Tan and S. J. Chen. Nucleic acid helix stability: Effects of salt concentration, cation valence and size, and chain length. *Biophysical Journal*, 90(4):1175–1190, 2006.

[109] V. Thielking, J. Alves, A. Fliess, G. Maass, and A. Pingoud. Accuracy of the ecori restriction endonuclease: binding and cleavage studies with oligodeoxynucleotide substrates containing degenerate recognition sequences. *Biochemistry*, 29(19):4682–91, 1990.

[110] L. Tierney. Markov-chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1701–1728, 1994.

[111] M. Vologodskaia and A. Vologodskii. Contribution of the intrinsic curvature to measured dna persistence length. *J Mol Biol*, 317(2):205–13, 2002.

[112] D. J. Spiegelhalter W. R. Gilks, S. Richardson. *Markov Chain Monte Carlo in Practice.* CRC Press, 1996.

[113] A. Wada and S. Kozawa. Instrument for the studies of differential flow dichroism of polymer solutions. *Journal of Polymer Science Part A*, 2:853–864, 1964.

[114] M. D. Wang, H. Yin, R. Landick, J. Gelles, and S. M. Block. Stretching dna with optical tweezers. *Biophys J*, 72(3):1335–46, 1997.

[115] M. R. Webb. A continuous spectrophotometric assay for inorganic phosphate and for measuring phosphate release kinetics in biological systems. *Proc Natl Acad Sci U S A*, 89(11):4884–7, 1992.

[116] M. C. Williams and I. Rouzina. Force spectroscopy of single dna and rna molecules. *Curr Opin Struct Biol*, 12(3):330–6, 2002.

[117] R. J. Williams. Restriction endonucleases: classification, properties, and applications. *Mol Biotechnol*, 23(3):225–43, 2003.

[118] R. W. Wilson and J. A. Schellman. The dichroic tensor of flexible helices. *Biopolymers*, 16(10):2143–65, 1977.

[119] R. W. Wilson and J. A. Schellman. The flow linear dichroism of dna: comparison with the bead-spring theory. *Biopolymers*, 17(5):1235–48, 1978.

[120] J. F. Zhang, S. C. Kou, and J. S. Liu. Biopolymer structure simulation and optimization via fragment regrowth monte carlo. *Journal of Chemical Physics*, 126(22):–, 2007.

[121] B. H. Zimm. Dynamics of polymer molecules in dilute solution: Viscoelasticity, flow birefringence and dielectric loss. *J. Chem. Phys.*, 24:269–278, 1956.

List of Figures

List of Tables

Appendix A

FtsZ

A.1 Derivation of differential equations

This appendix contains derivations for the differential equations of models presented in chapter 8 from chemical schemes, with the exception of model 1 which is presented in the text. The notation used is the same as in the main chapter.

A.2 Model 3

For model 3 the chemical equations are

$$\text{FtsZ} + \text{GTP} \xrightarrow{k_1} \text{FtsZ-GTP} \xrightarrow{k_2} \text{FtsZ-GDP} + \text{P}$$

$$\text{FtsZ-GDP} \underset{k_4}{\overset{k_3}{\rightleftharpoons}} \text{FtsZ} + \text{GDP}$$

where FtsZ refers to FtsZ protein in monomer form, FtsZ-GTP and FtsZ-GDP refer to the protein in oligomeric form, GTP and GDP are the concentrations of the free molecules and P is the free phosphate ion concentration. The system can be represented as

$$\text{F+T} \xrightarrow{k_1} \text{f} \xrightarrow{k_2} \text{g+P}$$

$$\text{g} \underset{k_4}{\overset{k_3}{\rightleftharpoons}} \text{F+D}$$

and using the law of mass action (see section 8.3.3)

$$\dot{f} = k_1 FT - k_2 f \tag{A.1}$$
$$\dot{g} = k_2 f - k_3 g + k_4 FD \tag{A.2}$$
$$\dot{p} = k_2 f. \tag{A.3}$$

Conservative equations can be found for FtsZ monomers, GDP molecules and phosphate ions

$$F_0 = F + f + g \tag{A.4}$$
$$T_0 = T + f + g + D \tag{A.5}$$
$$T_0 = T + f + p. \tag{A.6}$$

The last two of these imply that

$$D = p - g. \tag{A.7}$$

Upon substitution of the conservative equations into the differential equations

$$\dot{f} = k_1(F_0 - f - g)(T_0 - f - p) - k_2 f \tag{A.8}$$
$$\dot{g} = k_2 f - k_3 g + k_4(F_0 - f - g)(p - g) \tag{A.9}$$
$$\dot{p} = k_2 f \tag{A.10}$$

We assume that light scattering is a linear combination of f and g and likewise that percentage phosphate released is proportional to p, so we define f' and g' such that

$$f = s_1 f' \tag{A.11}$$
$$g = s_2 g' \tag{A.12}$$

and thus

$$L_s = f' + g' \tag{A.13}$$
$$p_{obs} = s_3 p. \tag{A.14}$$

Substituting into Equations A.8–A.10

$$s_1\dot{f'} = k_1(F_0 - s_1f' - s_2g')(T_0 - s_1f' - s_3p_{obs}) - k_2s_1f' \tag{A.15}$$

$$s_2\dot{g'} = k_2s_1f' - k_3s_2g' + k_4(F_0 - s_1f' - s_2g')(s_3p_{obs} - s_2g') \tag{A.16}$$

$$s_3\dot{p}_{obs} = k_2s_1f' \tag{A.17}$$

so the differential equations are

$$\dot{f'} = k_1\left(\frac{F_0}{s_1} - f' - \frac{s_2g'}{s_1}\right)(T_0 - s_1f' - s_3p_{obs}) - k_2f' \tag{A.18}$$

$$\dot{g'} = \frac{k_2s_1f'}{s_2} - k_3g + k_4(F_0 - s_1f' - s_2g')\left(\frac{s_3p_{obs}}{s_2} - g\right) \tag{A.19}$$

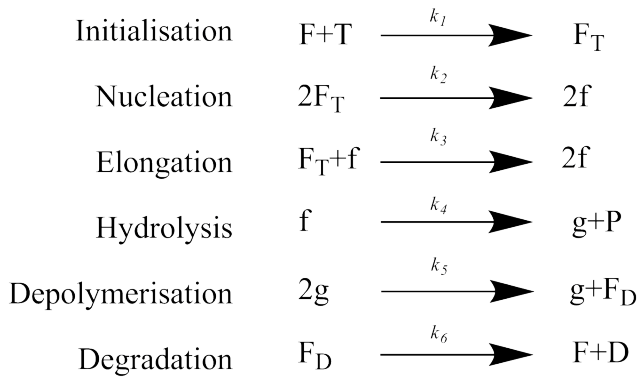
$$\dot{p_{obs}} = \frac{k_2s_1f'}{s_3} \tag{A.20}$$

A.3 Model 4

For model 4 the chemical equations are

Initialisation	FtsZ + GTP	$\xrightarrow{k_1}$	FtsZ-GTP
Nucleation	2 FtsZ-GTP	$\xrightarrow{k_2}$	2 $(\text{FtsZ-GTP})_n$
Elongation	FtsZ-GTP + $(\text{FtsZ-GTP})_n$	$\xrightarrow{k_3}$	2 $(\text{FtsZ-GTP})_n$
Hydrolysis	$(\text{FtsZ-GTP})_n$	$\xrightarrow{k_4}$	$(\text{FtsZ-GDP})_n$ + P
Depolymerisation	2 $(\text{FtsZ-GDP})_n$	$\xrightarrow{k_5}$	$(\text{FtsZ-GDP})_n$ + FtsZ-GDP
Degradation	FtsZ-GDP	$\xrightarrow{k_6}$	FtsZ + GDP

where FtsZ in monomer form is FtsZ, when bound to GTP or GDP it is represented as FtsZ-GTP and FtsZ-GDP respectively and when in polymer form $(\text{FtsZ-GTP})_n$ or $(\text{FtsZ-GDP})_n$. GTP and GDP give these molecules in monomer form and P is the free phosphate ions. Equivalently
The corresponding differential equations, via the law of mass action, are



$$\dot{F}_T = k_1 FT - 2k_2 F_T^2 - k_3 F_T f \tag{A.21}$$
$$\dot{f} = 2k_2 F_T^2 + k_3 F_T f - k_4 f \tag{A.22}$$
$$\dot{g} = k_4 f - k_5 g^2 \tag{A.23}$$
$$\dot{F}_D = k_5 g^2 - k_6 F_D \tag{A.24}$$
$$\dot{p} = k_4 f. \tag{A.25}$$

As above, conservative equations for FtsZ monomers and phosphate ions can be found

$$F_0 = F + f + F_T + F_D + g \tag{A.26}$$
$$T_0 = T + f + F_T + p. \tag{A.27}$$

Hence equation A.21 is

$$\dot{F}_T = k_1(F_0 - f + F_T + F_D + g)(T_0 - f + F_T + p) - 2k_2 F_T^2 - k_3 F_T f \tag{A.28}$$

and the other equations remain the same. For this model the observable data are given by

$$L_s = s_1(f + g) \tag{A.29}$$
$$P_{obs} = s_2 P \tag{A.30}$$

A.3.1 Introduction

This appendix contains the scripts used to carry out Markov chain Monte Carlo (MCMC) likelihood maximisation to find estimates of parameters that best fit a model, given by a set of differential equations, to a data set. The general hierarchy is given by figure A.1.

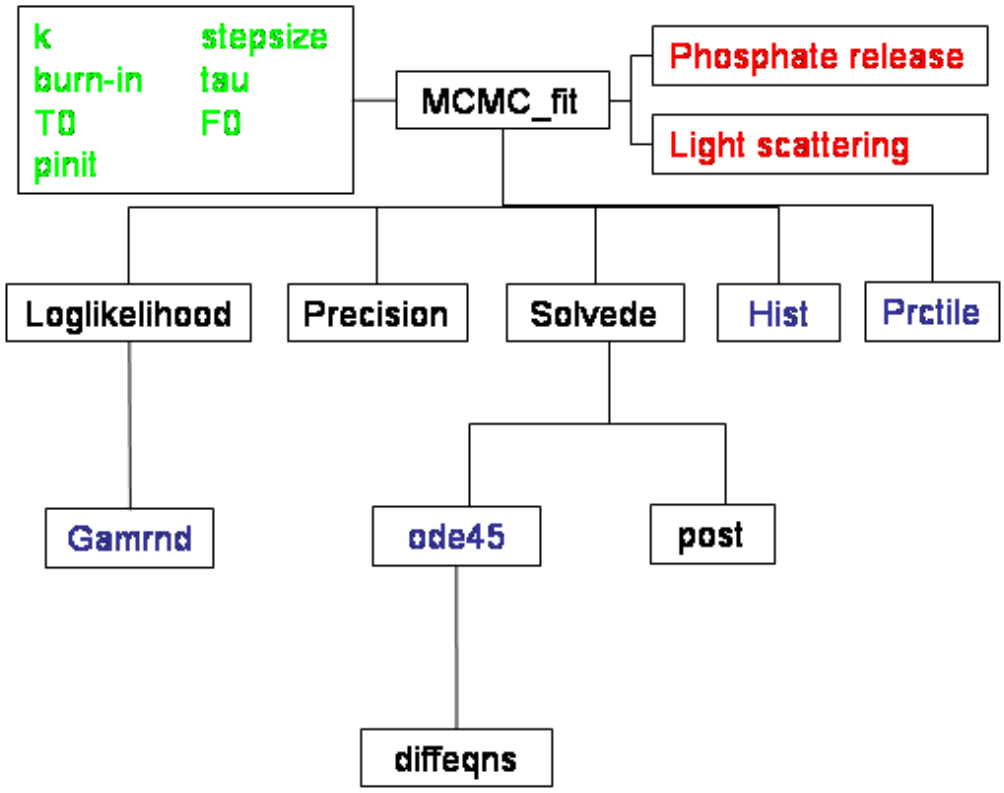


Figure A.1: The organisation of files for the MCMC algorithm. Details of the scripts are given in the main body of the text, for clarity only the titles are given here. In black are scripts written by the author. Green indicates numbers and vectors input to the main file as variables for the modelling, including initial guesses for output parameters (k, tau), variables for the algorithm (stepsize, burnin) and variables for the model (T0, F0 and pinit — the initial values for GTP concentration, FtsZ polymer concentration and released phosphate respectively). Red indicates theoretical data sets as given for FtsZ modelling. In blue are the most noteworthy Matlab functions used.

A.3.2 Summary of programs

MCMC_fit

The main script used to carry out an MCMC optimization as described in chapter 8. It is general in that it can be applied to arbitrary sets of input data provided they are given in the correct format.

Loglikelihood

Calculates the log of the likelihood function as given in equation (6.3).

A cut-off term is incorporated to ignore the first few data points during calculation. This was investigated because the models for FtsZ fitted poorly for the first few points. It was thought that a better fitting could be obtained by ignoring them. This transpired not to be the case and it did not affect the fitting. The value of the cut-off was routinely set to be 0.

Precision

Finds a value of precision by sampling from the gamma distribution with parameters given by equations 6.36 and 6.37.

Solvede

Solves the differential equations given in 'diffeqns' using the solver 'ode45' and puts the resulting into the form required using 'post'.

Post

Allows modification of the data output from the solved model, including addition of species, scalar multiplication and use of delay functions.

Diffeqns

The differential equations for each model were kept in separate folders and saved into the same folder as the other scripts listed here as appropriate.

Hist

A matlab function that plots a histogram. This was used to visually evaluate convergence of the Markov chains.

Prctile

A matlab function used to calculate percentiles of the Markov chains.

Gamrnd

A matlab function that generates a gamma distributed random number.

ode45

A matlab function that solves differential equations. Several different solvers were tried, but either produced identical results, performed slower or failed to solve the equations. This solver is recommended as a first port of call in the literature that accompanies Matlab.

A.3.3 Matlab Scripts

MCMC_fit.m

```
%%11/10/07

%% Script to fit a model to experimental data using MCMC

%% Inputs
% initial values for variables
modelno=4;    % model number
suff=suff+1;  % an index for the output files
suffix=int2str(suff);  % convert suffix to a string

% Initial guess for parameters to be estimated
k=[0.578 171 13.7 19.3 16.9 3.5 80.6 0.4771];
% step size to use in MCMC
stepsize=[0.01 1 0.1 0.05 0.05 0.05 0.1 0.01];

% no of steps and burnin for MCMC
steps=20000;
burnin=1;

% initial guesses for precision (tau)
tautau=1e-3;
tauin=[0.265/tautau 0.405/tautau; tautau tautau];

datatypes=2; % no of types of data for each data set
filenames=['FtsZ\Model 4 modii\data1_08-06-17.txt';
           'FtsZ\Model 4 modii\data2_08-06-17.txt'];
      % names of files to fit to:
      % format of data should be [time,data]; order of files
      % should be datatype1i,datatype2i,datatype1ii,datatype2ii etc

pinit=2.7582;  % initial value of ovserved phosphate
P=[200,11,40]; % (T0,F0,model) model is redundant

% Set labels for output and initial vectors for differential equations
```

```
if modelno==1
    init=[0 pinit];% Initial vector for DEs (must be correct length)
    label=['k1';'k2';'s1';'t1';'t2']; % Model 1 label;
end

if modelno==3
    init=[0 0 pinit];
    label=['k1';'k2';'k3';'k4';'s1';'s2';'s3';'t1';'t2'];
end

if modelno==31 % Exponential delay
    init=[0 0 pinit];
    label=['k1';'k2';'k3';'s1';'s2';'l ';'t1';'t2'];
end

if modelno==32 % Gamma delay
    init=[0 0 pinit];
    label=[['k1';'k2';'k3';'s1';'s2';'a ';'b ';'t1';'t2'];

if modelno==33 % Slow release phosphate
    init=[0 0 0 pinit];
    label=['k1';'k2';'k3';'k4';'s1';'s2';'t1';'t2'];
end

if modelno==4
    init=[0 0 0 0 pinit];
    label=['k1';'k2';'k3';'k4';'k5';'k6';'s1';'s2';'t1';'t2'];
    end

% Cutoff for likelihood function
cutoff=[0,0]; % ignore first X data points; no of cols = datatypes

% Allow negative parameters (1=yes,0=no)
neg=0;

%% end of inputs

% Parameters for tau's gamma distribution prior
atau=tauin(1,:);
btau=tauin(2,:);

fileind=size(filenames,1)/datatypes; % no of experimental data files

%% Load experimental data
ind=0; % initialise an index
timeind=zeros(size(datatypes)); % initialise vector of no of time points
```

```
for i=1:fileind
    for j=1:datatypes
        ind=ind+1;
        dataload=load(filenames(ind,:));

        if i==1 % Record sizes of data
            timeind(j)=size(dataload,1);

            if j==1 % Save time points
                time=dataload(:,1);
                expt=dataload(:,2);
            else % add to the end of previous time points
                time=cat(1,time,dataload(:,1));
                expt=cat(1,expt,dataload(:,2));
            end

        else

            if j==1
                expt(1:timeind(1),i)=dataload(:,2);
            else
                % add the data to the end of the file
                t=sum(timeind(1:j-1));
                expt(t+1:t+timeind(j),i)=dataload(:,2);
            end

        end

    end
end


%% 3. Find initial likelihood
% Check for negative parameters
if (min(k)<0)&&(neg==0)
    'Negative parameters'
    return % end program if any parameters are negative
end

% Solve the equations
kinit=k; % Sets output
theo=solvede(time,timeind,k,P,init,datatypes); % Solve with initial k
[tau,a,b]=precision(theo,expt,timeind,atau,btau); % find precision
likeold=loglikelihood(theo,expt,tau,timeind,cutoff); % find likelihood

% Exit if likelihood is 0 or NaN
if ¬(likeold>-1e100 && likeold<1e10)
```

```
    likeold
    'likelihood NaN'

    kout=[0,0,0];

    return
end

%% 4. Minimisation algorithm
% 4a). MCMC

% Pre-assignment of output vector
kvec=zeros(steps,size(k,2)+size(tau,2));
likevec=zeros(steps,1);

% Record initial k values
kdum=cat(2,k,tau);
kvec(1,:)=kdum;

for i=1:steps % iterate Markov steps
    i         % display step number
    for j=1:size(k,2) % iterate components of k

        % New k values and likehood
        knew=k;
        knew(j)=knew(j)+randn*stepsize(j);

        if neg==0
        % Skips rest of j-loop if proposed value is negative
            if knew(j)<0
                continue
            end
        end

        % Solve the equations
        theonew=solvede(time,timeind,knew,P,init,datatypes);

        % Skips rest of j-loop if solvede doesn't produce a result
        if size(theonew,1)≠size(time,1)
            'solvede produced no result' % error message
            continue

        end

        % Find new likelihood
        likenew=loglikelihood(theonew,expt,tau,timeind,cutoff);
```

```
        %% Accept or reject
        % Set alpha
        if ¬(likenew≥-1e100 && likenew<1e100) % reject if loglikelihood NaN
            alpha=-1e100;
        else  % otherwise use this alpha for comparison
            alpha=min(0,likenew-likeold);
        end

        % Set beta
        beta=log(rand); % randon number from uniform distribution for
                        % Metropolis-Hastings acceptance criterion

        % Compare alpha and beta
        if beta<alpha % accept
            likeold=likenew;
            k=knew;
            theo=theonew;
        % else % reject
            % k stays the same
        end

    end

    %  Find precision
    [tau,a,b]=precision(theo,expt,timeind,a,b);

    % Save outputs
    kdum=cat(2,k,tau);
    kvec(i,:)=kdum;
    likevec(i)=likeold;

end

%% 5. Outputs and save data
kout=mean(kvec(burnin:size(kvec,1),:));

% Save data
data=[steps, burnin];
data=cat(2,data,timeind,P,init,stepsize,kinit,kout,a,b)';

% file containing numbers used for fitting in the order
    % number of Markov steps
    % burn-in
    % no of time points in each data type
    % initial parameters not estimated
    % starting values for differential equations
    % step sizes
```

```
    % initial guess for k
    % estimated k
    % parameters used to estimate precision
save(['outputs\mcmc_data' suffix '.txt'],'data','-ascii','-tabs');

% file containing the Markov chain
save(['outputs\mark_chain' suffix '.txt'],'kvec','-ascii','-tabs');

% file containing the likelihood values at each point of the Markov chain
save(['outputs\likelihoods' suffix '.txt'],'likevec','-ascii','-tabs');

%% Plotting
hmark=kvec(burnin:size(kvec,1),:);

% Plot histogram and Markov chain
N=size(kvec,2);

% Settings for plots
axisfontsize=14;
labelfontsize=16;

% Plot histograms
fig1=figure;
    for i=1:N
       subplot(ceil(N/2),2,i)
       hist(hmark(:,i),30);
       set(gca,'fontsize',axisfontsize);
       set(get(gca,'YLabel'),'String',label(i,:),'fontsize',labelfontsize);
       axis tight
    end

    % Save figures in matlab, eps and png format
    saveas(gcf,['outputs\hist' suffix '.fig']);
    saveas(gcf,['outputs\hist' suffix '.eps']);
    saveas(gcf,['outputs\hist' suffix '.png']);

% Plot Markov chains
fig2=figure;
    for i=1:N
       subplot(ceil(N/2),2,i)
       plot(kvec(:,i));
       set(gca,'fontsize',axisfontsize);
       set(get(gca,'YLabel'),'String',label(i,:),'fontsize',labelfontsize);
       axis tight
    end

    % Save figures in matlab, eps and png format
```

```
    saveas(gcf,['outputs\chains' suffix '.fig']);
    saveas(gcf,['outputs\chains' suffix '.eps']);
    saveas(gcf,['outputs\chains' suffix '.png']);

% Calculate 10th,50th and 90th percentiles
percentiles=prctile(hmark,[10,50,90])';
percentiles(:,4)=std(hmark)'

% Save percentiles into a file
save(['outputs\percentiles' suffix '.txt'],'percentiles','-ascii','-tabs');

% Solve for solution: median value inputted as k
hk=percentiles(1:N-datatypes,2)'
taupramsout=[a;b]

htheo=solvede(time,timeind,hk,P,init,datatypes);

% Deconvolve data and plot best fit
fig3=figure;
    for l=1:datatypes
        % define time
        t1=sum(timeind(1:l-1))+1;
        t2=sum(timeind(1:l));
        htime=time(t1:t2);

        % define theo data
        htheodum=htheo(t1:t2);
            % plot best fit
            subplot(datatypes+1,1,l);
            plot(htime,htheodum,'r',htime,expt(t1:t2,:));
    end

% Plot likelihood
subplot(datatypes+1,1,datatypes+1)
plot(likevec);

% Save best fit and likelihood figure in matlab, eps and png formats
saveas(gcf,['outputs\plot' suffix '.fig']);
saveas(gcf,['outputs\plot' suffix '.eps']);
saveas(gcf,['outputs\plot' suffix '.png']);
```

Loglikelihood.m

```
%% Log likelihood function

function like=loglikelihood(theo,expt,tau,timeind,cutoff)

% Predefine
loglikedum=zeros(size(expt,2),1);

% Finding the mean squared error for the data
for i=1:size(expt,2) % Index data replicates
    se(:,i)=(theo-expt(:,i)).^2;
end

% Average squared difference
meanse=mean(se,2);

for j=1:size(timeind,2) % Indexes data types
    % Define time
    t1=sum(timeind(1:j-1))+1+cutoff(j); % doesn't include first 'cutoff'
    t2=t1-cutoff(j)-1+timeind(j);       % data points

    % Calculate contribution of data type to likelihood
    loglikedum(j)=(timeind(j)/2)*log(tau(j)/(2*pi))+(-tau(j)*sum(meanse(t1:t2))/2
end

like=sum(loglikedum);
```

Precision.m

```
%% Script to find the precision of the data

function [tau,aout,bout]=precision(theo,expt,tind,a,b)

% Preallocations
se=zeros(size(expt));
sse=zeros(1,size(tind,2));
tau=zeros(1,size(tind,2));

% Find squared differences
for j=1:size(expt,2) % Index data sets
    se=(expt(:,j)-theo).^2; % Find residual
end

% Average over data sets
se=mean(se,2);

% Sum over time for each data type
for i=1:size(tind,2)
    t1=sum(tind(1:i-1))+1;
    t2=t1-1+tind(i);

    sse(i)=sum(se(t1:t2));

end

% Adjust parameters
aout=a+tind/2;
bout=1./(1./b+sse/2);

% Sample from new gamma distribution
for l=1:size(tind,2)
    tau(l)=gamrnd(a(l),b(l));
end
```

Solvede.m

```
% This program solves the equations in 'diffeqns.m'
function theodata=solvede(time,timeind,k,params,init,dim)

for i=1:dim % indexes data types

    %% Define time
    t1=sum(timeind(1:i-1))+1;
    t2=sum(timeind(1:i));
    t=time(t1:t2);

    % Find solution using Matlab ode45 solver
    [tdum,theodum]=ode45(@diffeqns,t,init,[],flag,k,params);

    % Check to see if solution worked
    if size(theodum,1)~=size(t,1)
        theodum=zeros(size(time,1),size(init,2));
    end

    % post-process result - converts solution into something comparable
    % with experimental data
    theopost=post(t,theodum,k,params);

    if i==1 % output data
        theodata=theopost(:,i);
    else
        theodata=cat(1,theodata,theopost(:,i)); % add it to the end of the file
    end

end
```

A.4 Markov chain plots

This appendix contains Markov chains for model fitting referred to in chapter 8.

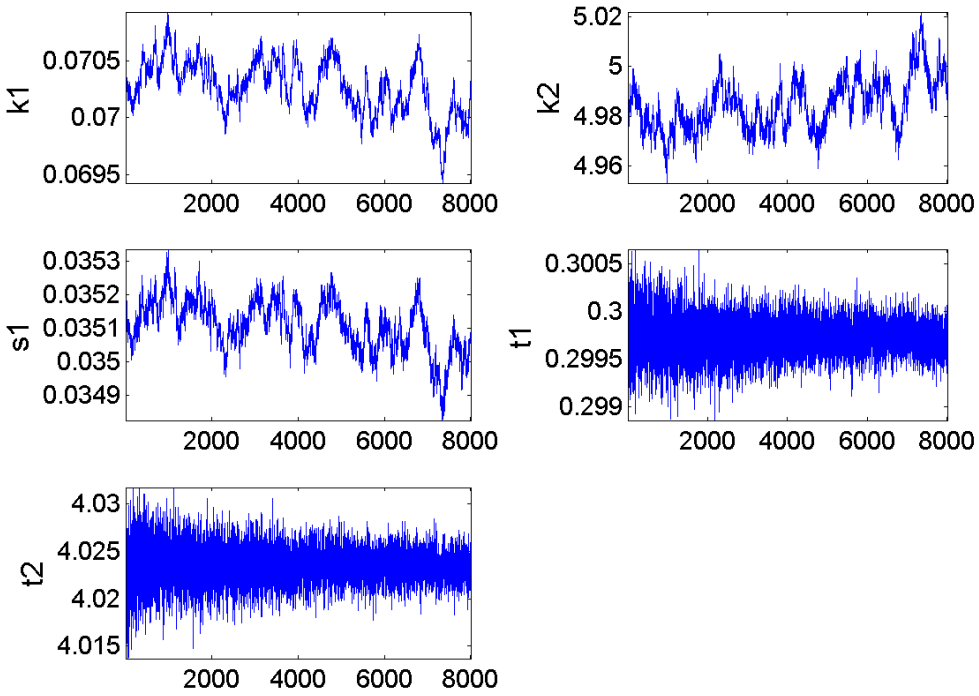


Figure A.2: Markov chains for model 1 fitting to simulated data, corresponding to the data shown in table 8.1.

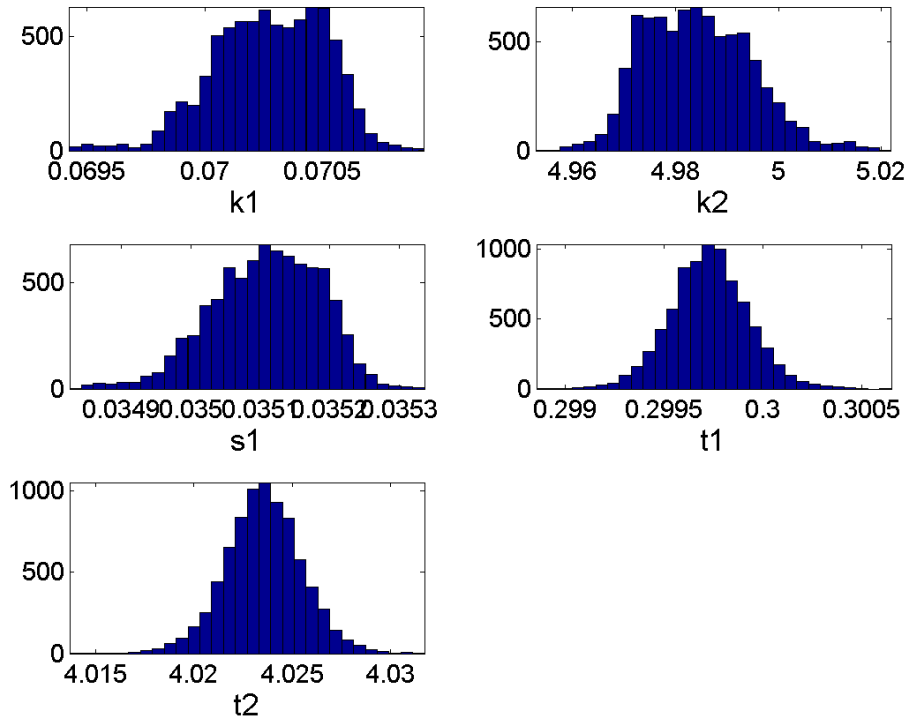


Figure A.3: Histogram of the last 8000 points of the Markov chains for fitting of model 1 to simulated data.

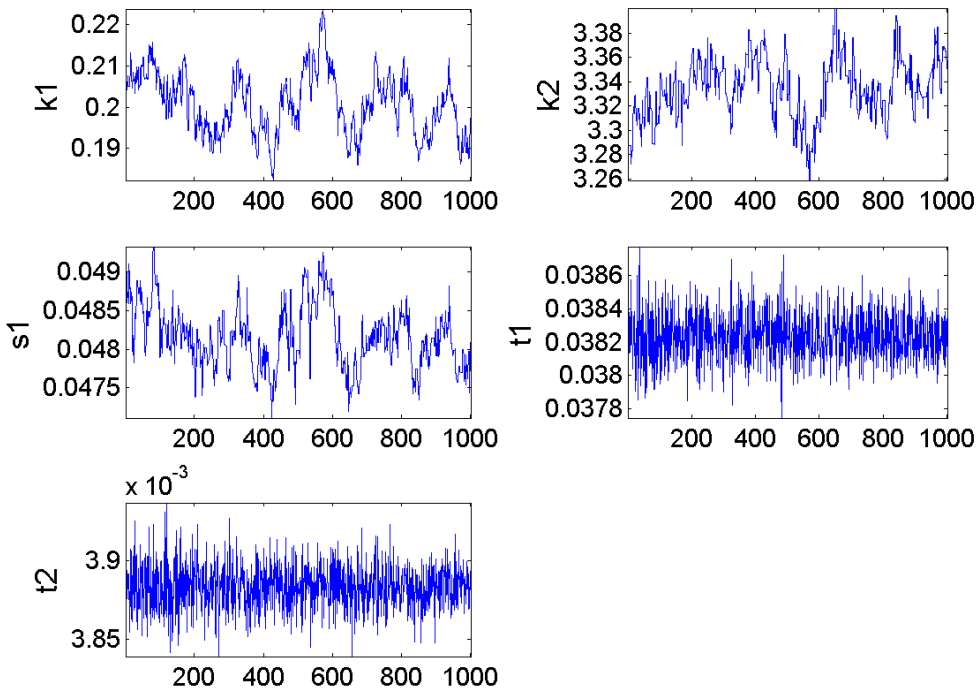


Figure A.4: The last 1000 points of the Markov chains for model 1 with experimental data.

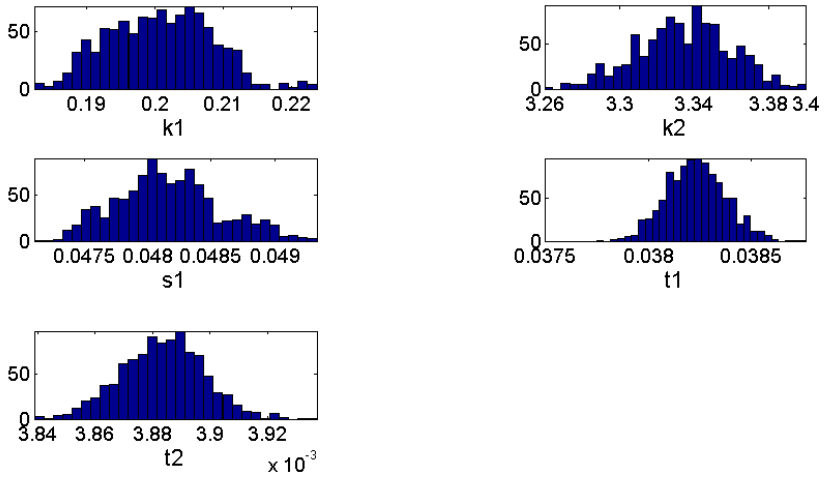


Figure A.5: Histograms of the last 1000 points of the Markov chains for Model 1 with experimental data

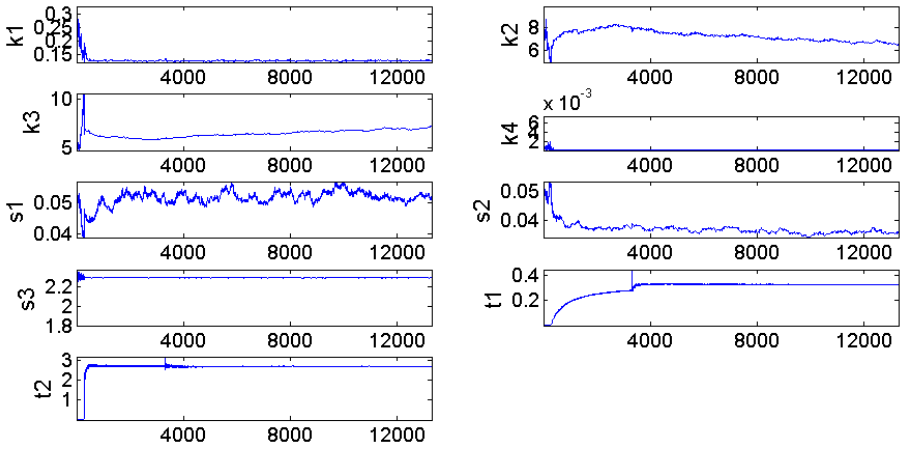


Figure A.6: Markov chains for model 3 with simulated data

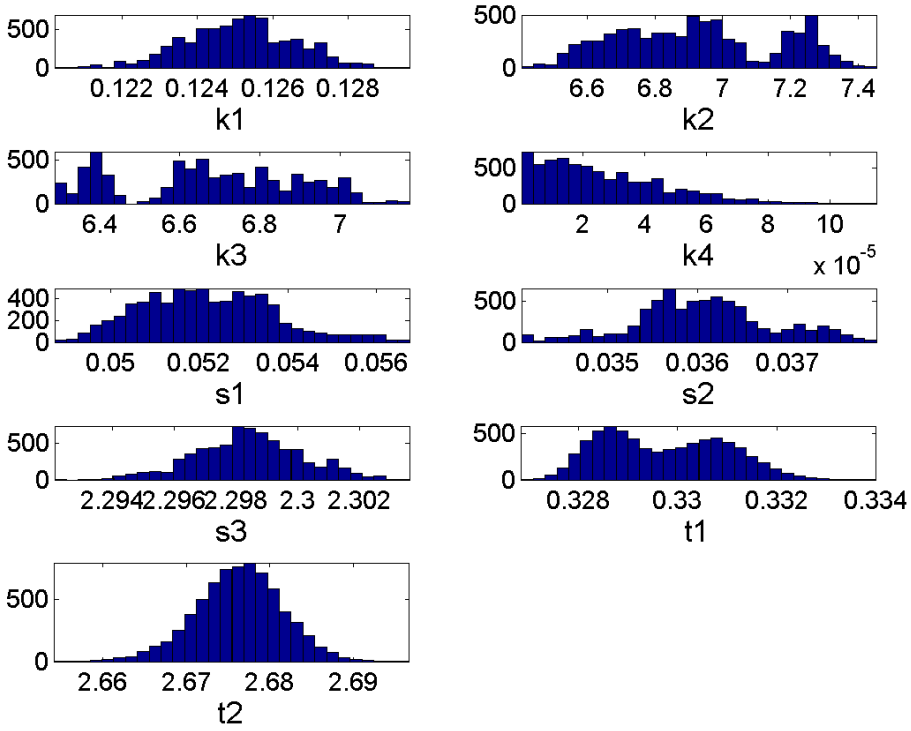


Figure A.7: Histrograms of the Markov chains for model 3 with simulated data.

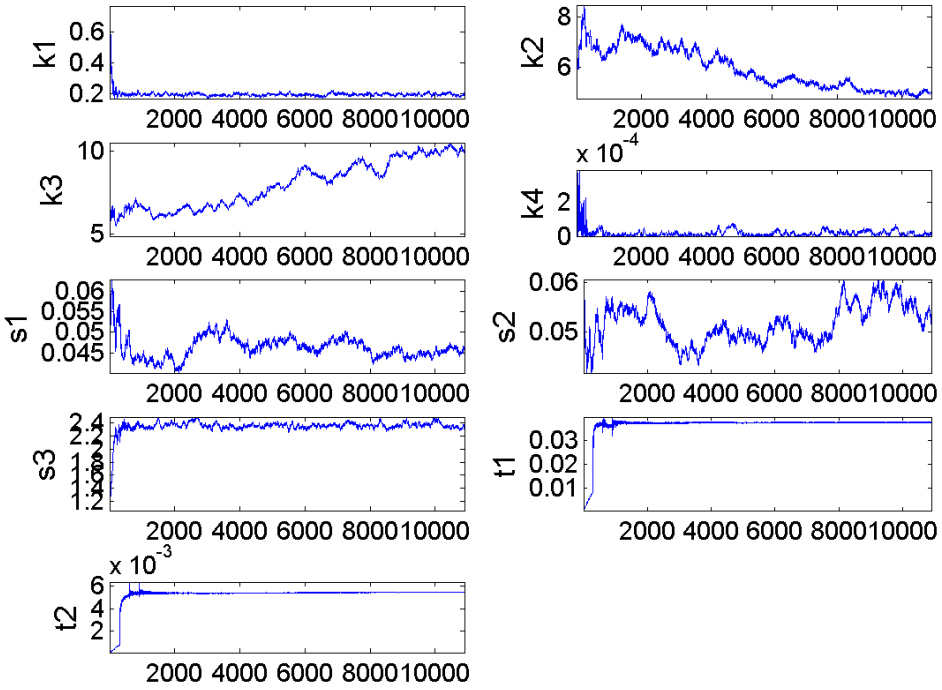


Figure A.8: Markov chains for fitting of model 3 to experimental data. The full Markov chains are shown to demonstrate how the values converge. The step sizes are altered several times during the process.

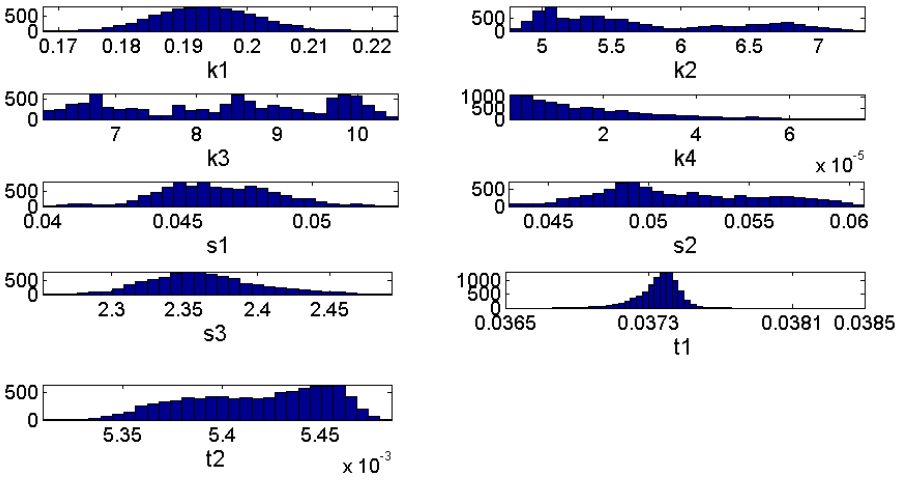


Figure A.9: Histograms of the last 19500 points of the Markov chains for model 3 with experimental data

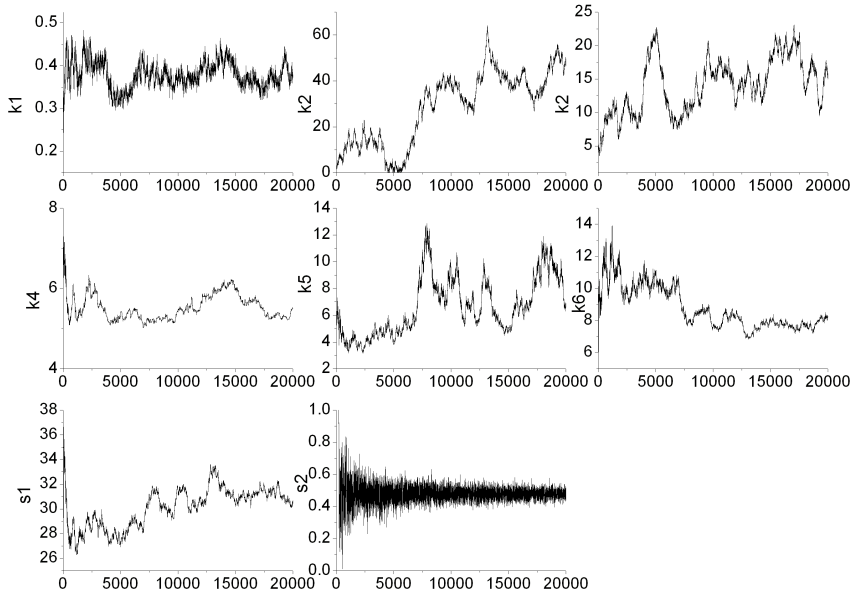


Figure A.10: Markov chains for model 4 with simulated data

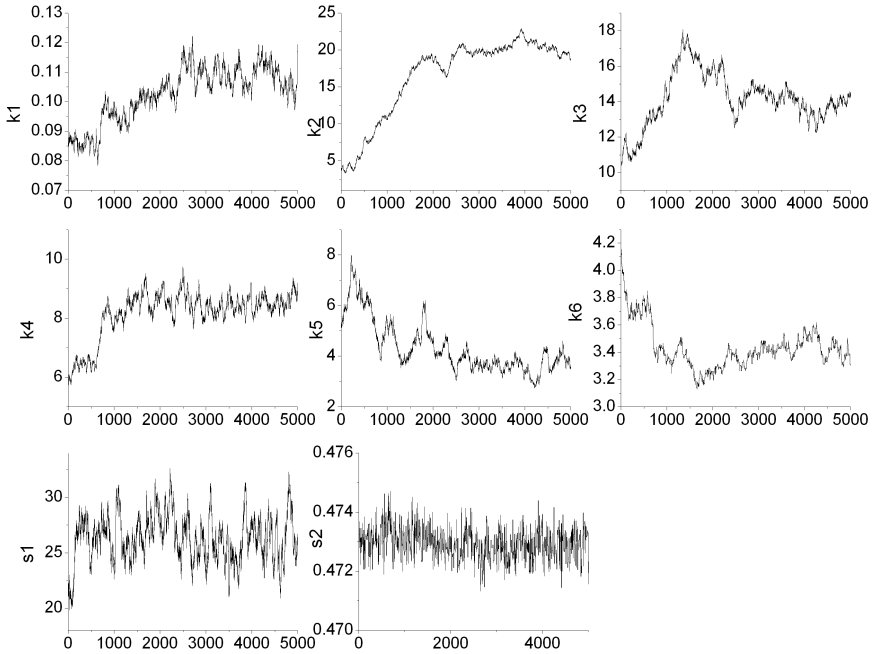


Figure A.11: Markov chains for model 4 fitting to experimental data

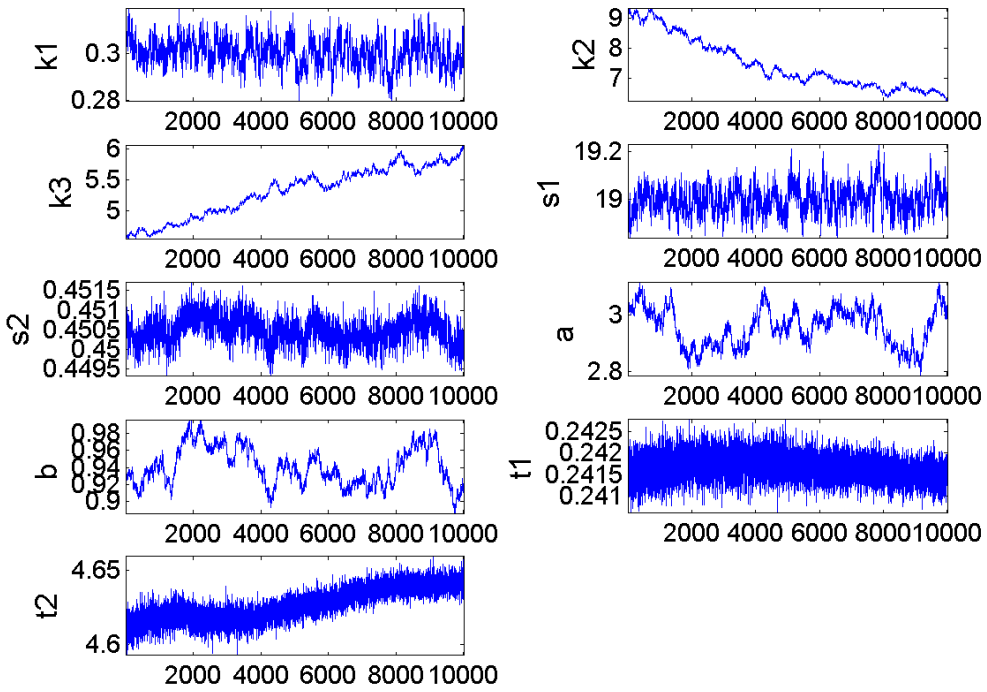


Figure A.12: Markov chains for model 3 with a gamma distributed delay for the phosphate term fitted to simulated data.

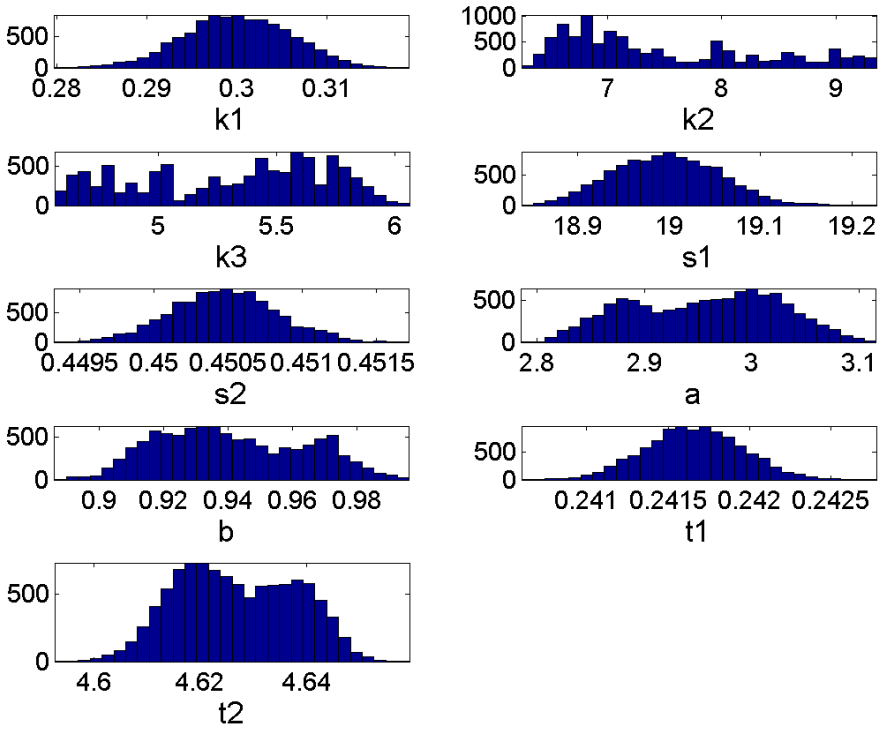


Figure A.13: Histograms for model 3 with a gamma distributed delay function fitting to simulated data.

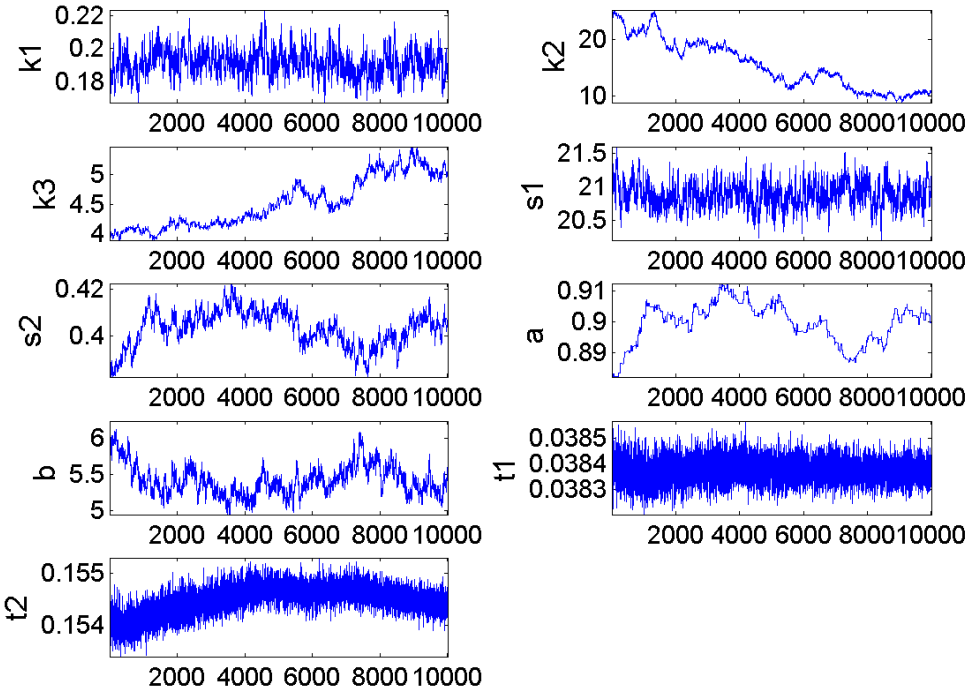


Figure A.14: Markov chains for model 3 with a gamma distributed delay for fitting to experimental data.

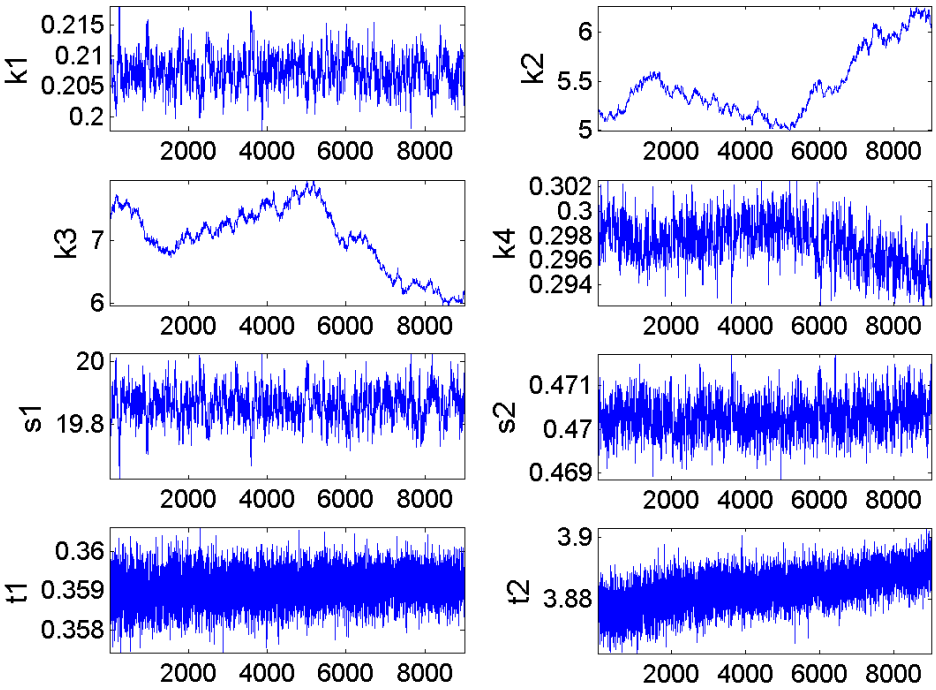


Figure A.15: Markov chains for a slowly released phosphate system fitted to simulated data.

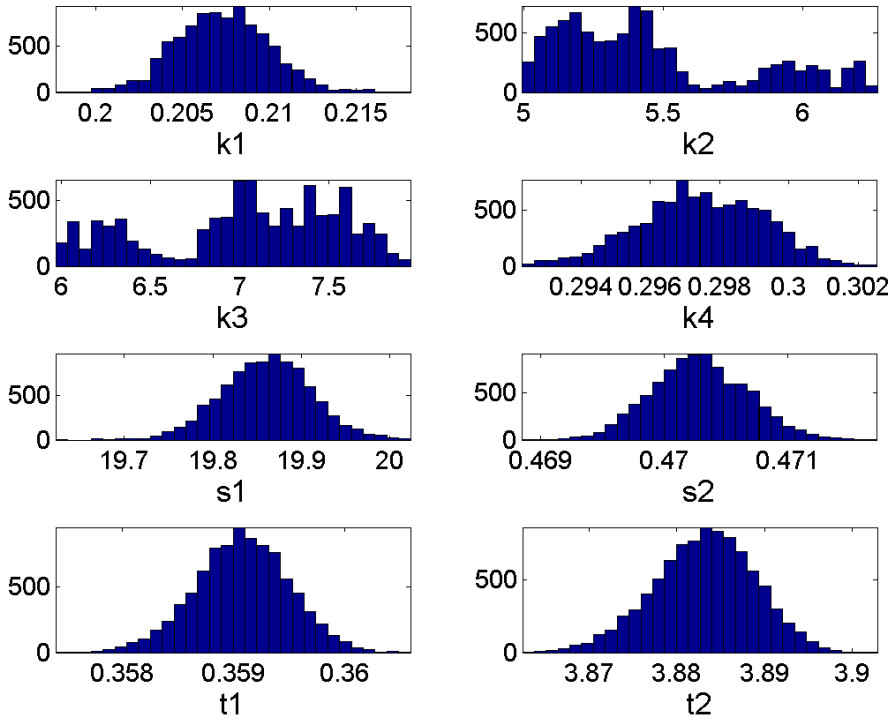


Figure A.16: Histograms for the slowly released phosphate scheme fitting to simulated data.

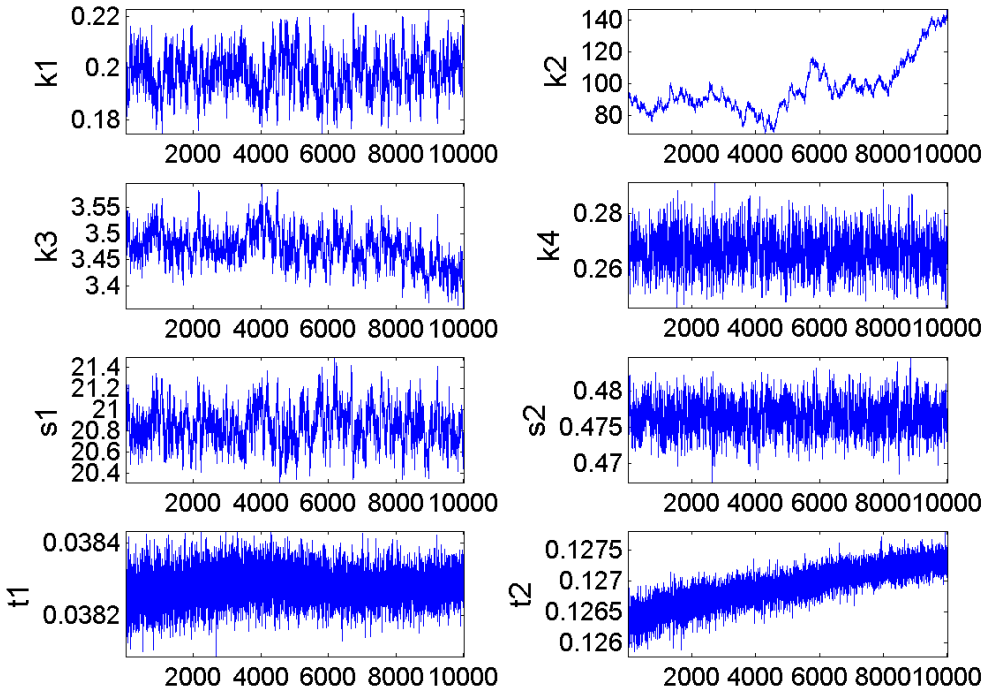


Figure A.17: Markov chains for a slowly released phosphate system fitting to experimental data.

Appendix B

Persistence length

B.1 End-to-end length of a random flight chain

As mentioned in chapter 3 the random flight model of a polymer is characterised by the condition given in equation 3.9:

$$\mathbf{r}_n.\mathbf{r}_{n-1} = b^2 \cos\theta. \tag{B.1}$$

Squared end-to-end distance R is given by equation 3.6 and is equal to the dot product of the end-to-end vector with itself

$$\begin{aligned} \langle R^2 \rangle &= \langle (\mathbf{r}_1 + \mathbf{r}_2 + ...\mathbf{r}_n).(\mathbf{r}_1 + \mathbf{r}_2 + ...\mathbf{r}_n) \rangle & \text{(B.2)} \\ &= \textstyle\sum_{n=1}^{N} \sum_{m=1}^{N} \langle \mathbf{r}_n.\mathbf{r}_m \rangle. & \text{(B.3)} \end{aligned}$$

We separate this into two different parts, one which follows the route of the FJC and the other which incorporates the cross-terms to which the random flight condition applies

$$\langle R \rangle = \sum_{n=0}^{N} < \mathbf{r}_n.\mathbf{r}_n > +2 \sum_{m \neq n} < \mathbf{r}_n.\mathbf{r}_m > \tag{B.4}$$

$$\tag{B.5}$$

Consider the second sum in the previous expression. It is equal to

$$2\sum_{n=1}^{N}\sum_{m=1}^{N-n} < \mathbf{r}_n.\mathbf{r}_m > \tag{B.6}$$

Now, $\langle \mathbf{r}_n.\mathbf{r}_{n+1} \rangle = r\cos\theta$ and hence $\langle \mathbf{r}_n.\mathbf{r}_{n+x} \rangle = b^2\cos^x\theta$. Thus equation B.6 is

$$2\sum_{n=1}^{N}\sum_{m=1}^{N-n} b^2\cos^m\theta \tag{B.7}$$

which is a geometric sum, hence the full expression in equation B.4

$$= Nb^2 + 2b^2\sum_{n=1}^{N}\cos\theta\frac{1-\cos\theta^{N-n}}{1-\cos\theta} \tag{B.8}$$

where the last step results from the sum for a finite geometric series. The expression is expanded and the finite geometric sum applied again:

$$Nb^2 + 2b^2\cos\theta\sum_{n=1}^{N}\frac{1}{1-\cos\theta} + \frac{\cos^N\theta}{1-\cos\theta}\left(\frac{1}{\cos\theta}\right)^n \tag{B.9}$$

$$= Nb^2 + \frac{2Nb^2\cos\theta}{1-\cos\theta} + \left(\frac{\cos^{N-1}\theta}{1-\cos\theta}\right)\left(\frac{1-\left(\frac{1}{\cos\theta}\right)^N}{1-\left(\frac{1}{\cos\theta}\right)}\right) \tag{B.10}$$

$$= Nb^2\left(\frac{1+\cos\theta}{1-\cos\theta}\right) + \left(\frac{\cos^{N-1}\theta}{1-\cos\theta}\right)\left(\frac{\cos^N\theta - 1}{\cos^{N-1}\theta(\cos\theta - 1}\right). \tag{B.11}$$

If the angle θ is small and N large, which is the limiting case for a smooth curve as the segment length $b \to 0$, we have that $\cos^N\theta \to 1$ and hence the last term in the previous equation tends to 0. Hence end-to-end distance is

$$\langle R^2 \rangle = Nb^2\frac{1+\cos\theta}{1-\cos\theta}. \tag{B.12}$$

Appendix C

Synchrotron radiation linear dichroism

This appendix contains spectra corresponding to those in chapter 2, but taken on a 'home source' spectrometer, i.e. a Biologic MOS-450 spectrometer. The spectra are LD of various types of DNA (indicated with the notation used in the corresponding chapter) and different salt concentrations as indicated.

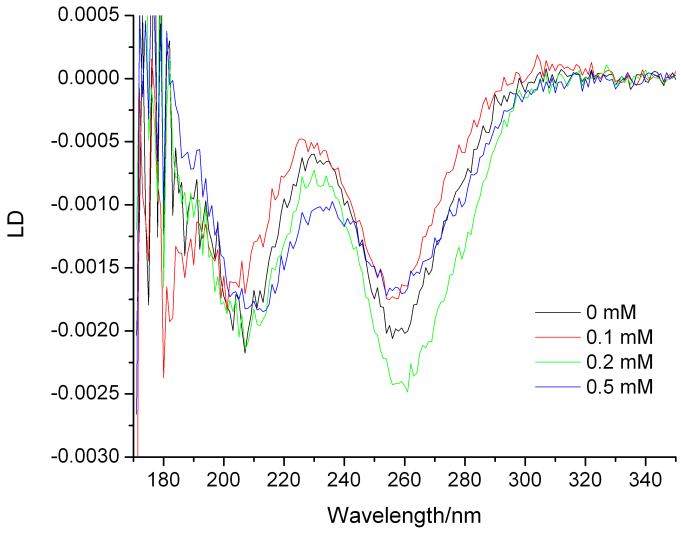


Figure C.1: LD of AT DNA measured on a home source

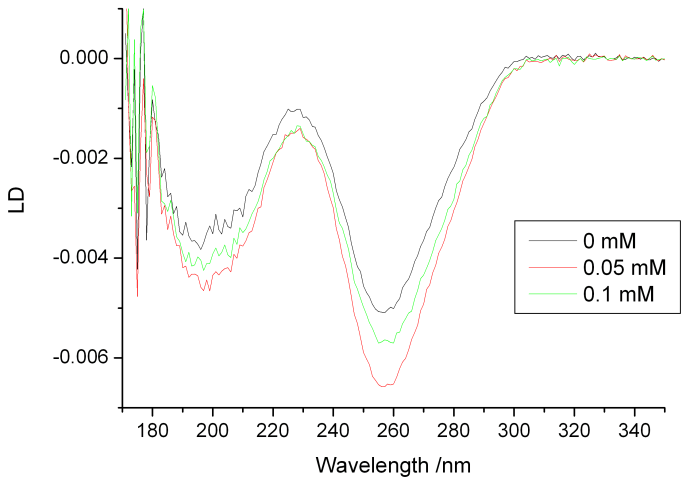


Figure C.2: CP DNA measured on a home source.

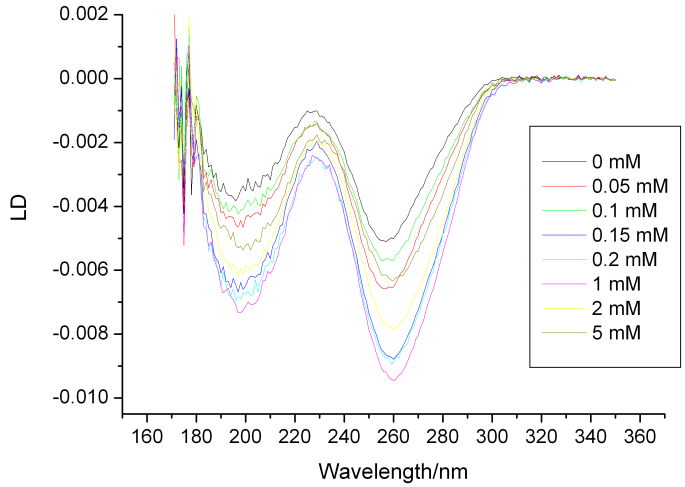


Figure C.3: Spectra of calf thymus DNA measured on a home source

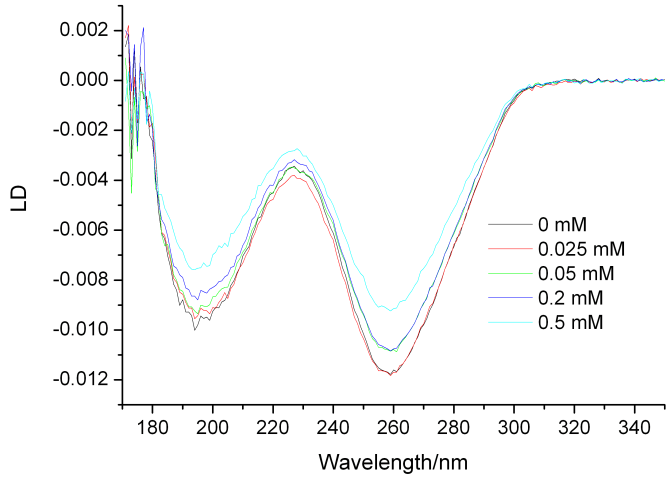


Figure C.4: Spectra of ML DNA measured on a home source

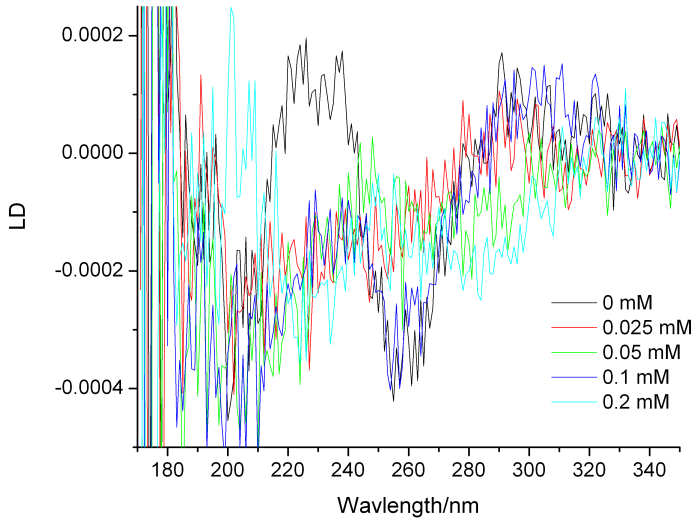


Figure C.5: Spectra of single stranded poly A ??? DNA measured on a home source