THE UNIVERSITY *of* EDINBURGH
School of Physics
and Astronomy

Senior Honours Project
DNA Unzipping and Overstretching

Ben Henderson
December 2021

Abstract

In this project the Poland-Scheraga model of DNA is used to find the transition point of non-helical DNA strands from being fully zipped to becoming denatured with respect to temperature. The flexibility of the individual hydrogen bonds are adjusted to evaluate to what extent this impacts on the temperature the transition point occurs at. I will explore this in a numerical study of results from the simulations.

Declaration

I declare that this project and report is my own work.

Signature: Date: 3/12/2021

BJ Henderson

Supervisor: Dr. D. Marenduzzo 10 Weeks

Contents

1 Introduction

DNA is one of the most important materials in a living organism and can be found everywhere, from me writing this paper to bacteria cells in a Petri dish. DNA consists of two long separate strands that are in a helix or screw-like shape around each other with opposite nucleotides being connected with hydrogen bonds of energy 1-2 k_{BT} and as such are known as a base pair. There are four 'flavours' of nucleotide within a strand of DNA. They are referred to as cytosine (C), guanine (G), adenine (A) and thymine (T)[1]. With these 4 flavours there are only 2 combinations that we find: CG pair which contains 3 H-Bonds and AT which has only 2. A more simplistic model of DNA was used in this project where we use a homogeneous model where we have only AT or CG pairs but not both. DNA also comes in a right-handed helix shape with the step between each base pair being 0.34nm [2]. The helical nature of the strands were removed and as such the two polymer strands are flat as seen in the lower image of figure (1). This removes an area of complexity when it comes to modelling the system, it also forces the H-Bonds between the base pairs to be the crucial and most important factor in holding the strands together (zipped). The final key characteristic of DNA is the persistence length of the polymer. This coincides with the stiffness of the strands and allows for how much bending can occur in the structure.

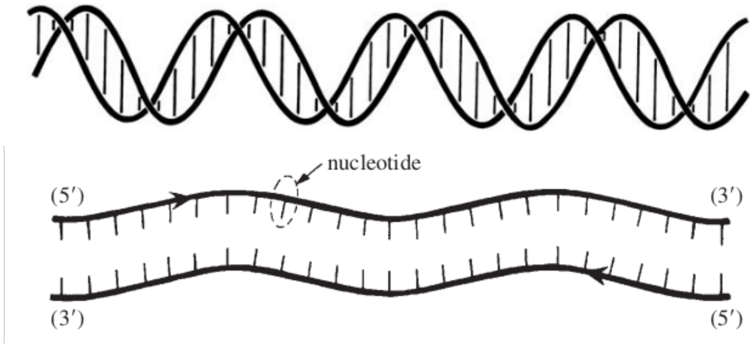

Figure 1: The top image shows how the DNA strands are coiled with hydrogen bonds between the nucleotide base pairs denoted by the straight vertical lines.
The lower image shows the simplified straight DNA strands. An individual nucleotide has been highlighted for ease of viewing[3].

In order to accurately and conveniently simulate DNA strands and their potentials, I chose to use the Poland-Scheraga model (herinafter referred to as PS) of DNA. This model defines a discrete simulation within a lattice of 3D-space. It allows for a double stranded DNA segment which is defined by a free walk path with each step being a constant size [4]. It also defines the specific angle between each of the nucleotides within the individual strands know as the stiffness. The phase transition between being fully zipped and denatured is also defined by the PS model where if the simulation contains a loop or a bubble[5] is formed by each of the strands then there is a strong chance that the section is denatured. For this project I opted for a simulation rather than an analytical solution; where an average number of pairs being open within a section is indicative of the denatured status. For a PS model we have a number of chain configurations as follows:

$$Z = \sum z_{m,n} w^m x^n \tag{1}$$

Where: z is the number of chain configurations, m is the number of contacts, n is the length, the activity x is conjugate to n and w is the Boltzmann Factor.

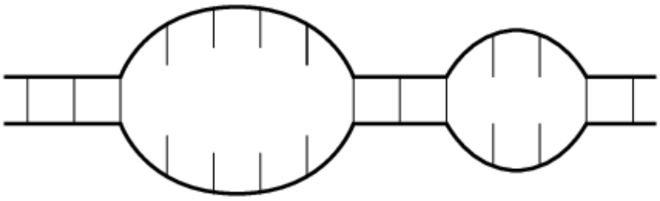

Figure 2: This image shows how the PS model describes a strand of DNA, bubbles and loops can be seen along the polymer which show the H-Bonds being broke around those areas whereas they are connected in other places[6].

The phase transition can be described by a battle; a competition between entropy and energy. The equation of entropy details the inversely proportionate relationship with temperature. At low temperature the entropy is low and as such the energy caused from H-Bonds wins and therefore the DNA is zipped. Conversely when the temperature is high then the number of configurations dominates and the DNA is unzipped, as the entropy of the 2 separated strands are larger than that of the double-stranded molecule. More quantitatively, the entropy of a DNA molecule described within the PS model can be written as:

$$S = k_b \ln\Omega \tag{2}$$

Where: S is the entropy of the system, k_b is the Boltzmann Constant and Ω is the number of configurations in the system. Then, if Ω_2, the number of configuration of two strands, is $\Omega_1^2 > \Omega_1$, if we neglect mutual avoidance between the two strands and model each as an infinitesimally thin random walk.

2 Methods

The dynamics of each monomer in each of the two strands making up the DNA molecule can be described by a Langevin equation of the following form:

$$m\frac{d^2\mathbf{x}_i}{dt^2} = -\nabla_i V - \gamma\frac{d\mathbf{x}_i}{dt} + \sqrt{2k_B T\gamma}\eta(t) \tag{3}$$

$$\langle\eta(t)\rangle = 0; \quad \langle\eta_a(t)\eta_\beta\left(t'\right)\rangle = \delta_{\alpha\beta}\delta\left(t-t'\right) \tag{4}$$

The LJ potential acts between every bead within the strands with the ith and jth potential as follows:

$$V_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{5}$$

The potential between 2 neighbouring monomers can be modelled as a 'FENE spring' in the following way:

$$\mathrm{V}_{\mathrm{FENE}}\left(r = |\mathbf{r_{i+1}} - \mathbf{r_i}|\right) = -\frac{K_{FENE}R_0^2}{2}\log\left[1 - \left(\frac{r}{R_0}\right)^2\right] \tag{6}$$

The potential in the bending of the polymers is defined by the angle between 3 neighbouring nucleotides as follows:

$$V_{bending} = K_b\left[1 + \cos(\theta)\right] \tag{7}$$

Where: $\mathbf{x_i}$ is the position of the monomers, k_B is the Boltzmann Constant, γ is the friction, T is the temperature, $\epsilon =\ k_BT$, r is the distance between monomers and K_{FENE} $=30\ k_BT/\sigma^2$ [7].

To simulate the effect of temperature on the DNA polymers we needed to use LAMMPS as a molecular dynamic simulator. We use a stochastic thermostat with the NVT ensemble where we hold N, V and T constant during the each simulation. Velocity Verlet integration [8] is used by the LAMMPS code to solve newtons equation as seen in equation (3). The raw LAMMP file was provided by my supervisor [9] and it takes multiple inputs including: a normalised random walk of two 200 atom long strands of DNA, and a file containing the epsilon value for each of these atoms. The epsilon value of the atoms is used as a substitute for the temperature, as the Lennard-Jones energy is only affected by a change in epsilon as in equation (5).The role of this energy is to ensire mutual repulsion between the 2 strands and to keep the H-Bonds intact. The Lennard-Jones energy is complemented by the FENE energy denoted in equation (6). This is the energy that keeps the polymer together, i.e the bond between neighbouring nucleotides (i and i +1).It works through 'FENE springs' which make up the polymers in each strand.

Once the LAMMP file is run the code first initialises the starting criteria. This involves normalising the random walk to make sure it is in the actual shape of a DNA polymer by avoiding any overlaps in the random walk and then making sure that during the simulation the polymers are self-avoiding. It also ensures the correct potentials are in affect and then begins to simulate all of the interactions between the atoms using the potential energy between them. This is done in timestep increments that are defined in the code - I set this to 10000 with a maximum step of 40001000 steps. A dump file is produced, indicating the x, y, and z coordinates of all 400 atoms during each timestep of the simulation. The dump file can subsequently be opened in VMD viewer to show how each simulation evolves over time. This process is completed for all epsilon values between 1.0 and 1.7 and all dump files are saved and labeled accordingly for analysis. I then changed the bending angle with the LAMMPS file from 20 degrees to 10. This change in angle is denoted by θ in equation (7). This is closely related to the persistence length which gives the amount that the bonds can bend within the polymer to detect whether the transition point is changed according to said bending angle causing the energy due to stiffness.

Once all of the simulations are complete the next stage is the post-processing analysis; this code was written by myself using python and is done in 2 stages. The first stage is to take each dump file produced during the simulations and for each timestep calculate which base pairs are 'open'. We denote open as the distance between the base pairs (1st nucleotide on strand 1 and 1st nucleotide on strand 2) being greater than half the size of the 3D space, which in this case is 50 units. This then produces a text file with 200 entries for each timestep with either a 1 or a 0 for each pair; this file can be used to create kymographs in the future. The second stage is to take an average score of how open each epsilon level is. This creates another text file with 8 entries (1.0 - 1.7) with a number between 0 and 1 indicating the average score for each simulation.

3 Results & Discussion

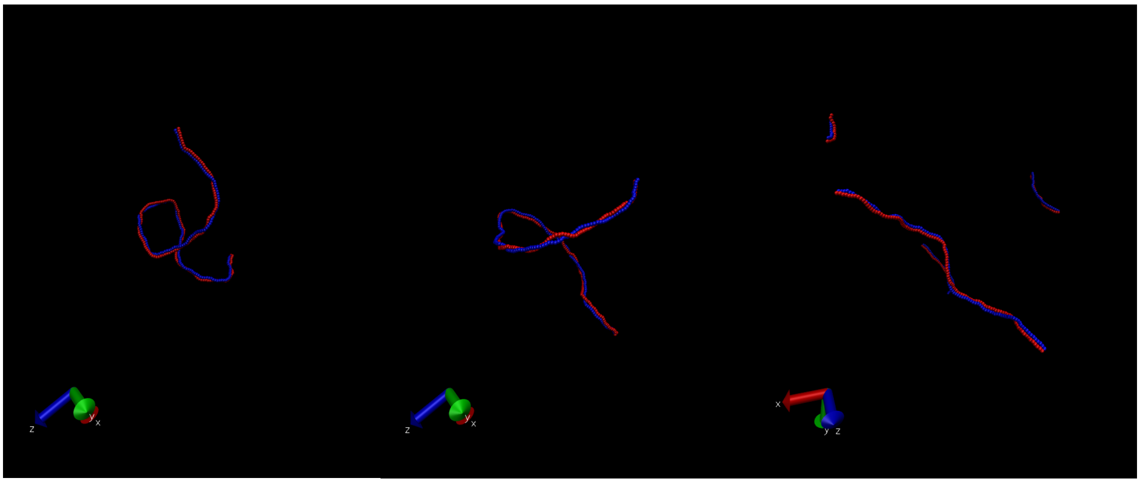

Figure 3: This image shows the start, middle and end of the simulation running with ϵ = 1.7, with $\theta = 20$. (The strands not connected are due to the periodic boundary of the 3D space.)

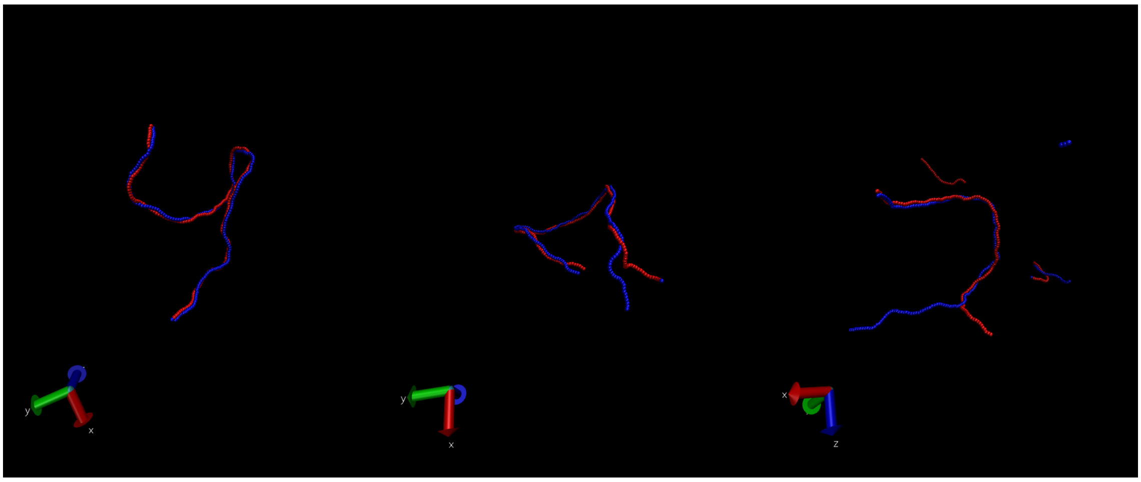

Figure 4: This image shows the start, middle and end of the simulation running with ϵ = 1.3, with $\theta = 20$. (The strands not connected are due to the periodic boundary of the 3D space.)

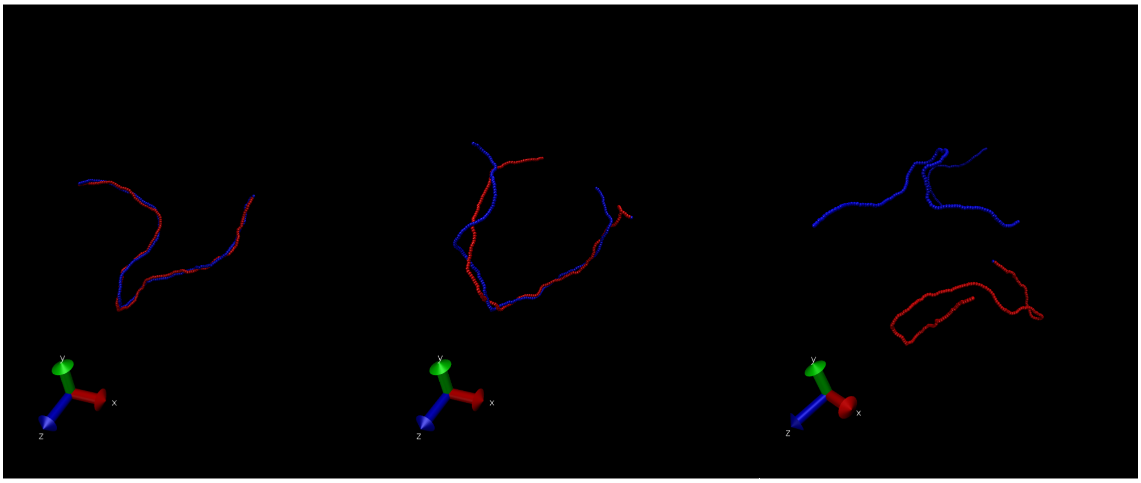

Figure 5: This image shows the start, middle and end of the simulation running with ϵ = 1.0, with $\theta = 20$. (The strands not connected are due to the periodic boundary of the 3D space.)

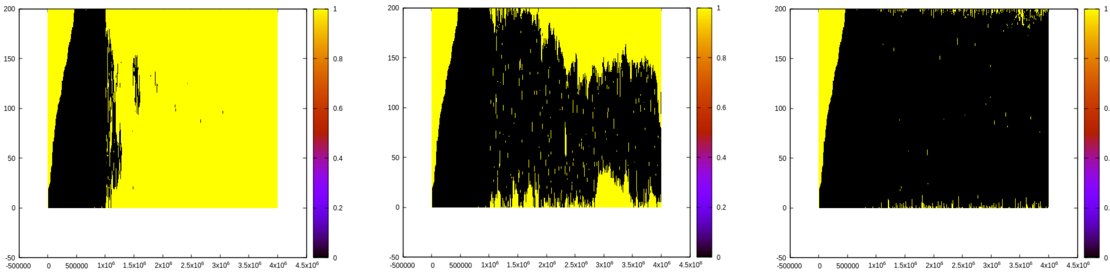

Figure 6: This image shows the kymographs from the post-simulation processing from left $\epsilon = 1.0$, 1.3, 1.7. The area before 1 x 10^6 is to be ignored as this is the initialisation stage, where the polymers initially zip up. Yellow indicates the base pair is open and black indicates closed.

From figure 3 we can see from start, middle and end the DNA strand stays completely zipped together, the polymer does move throughout the space but the internal energy is too strong to be overcome by the entropy. In figure 4, it can be seen that in the middle and the end of polymer starts to unzip from one another. From testing this is the highest ϵ to cause this behaviour. This is a strong indicator that 1.3 could be the transition point. In figure 5 it can clearly be seen that from the end it is completed unzipped and decomposed compared to the start of the simulation. This clearly shows that the transition is definitely between 1.0 and 1.7, the kymographs in figure 6 back this up by when $\epsilon = 1.0$ everywhere past the initialisation zone shows that the base pairs are defined as open. This contrasts compared to when $\epsilon = 1.7$ where everything can be seen to fully zipped up. At 1.3 you can see that the outsides of the polymer are beginning to break down but the centre remains strong.

ϵ	Average
1.0	0.99967662
1.1	0.99955976
1.2	0.78898001
1.3	0.37718905
1.4	0.05335820
1.5	0.03126865
1.6	0.02873134
1.7	0.01713930

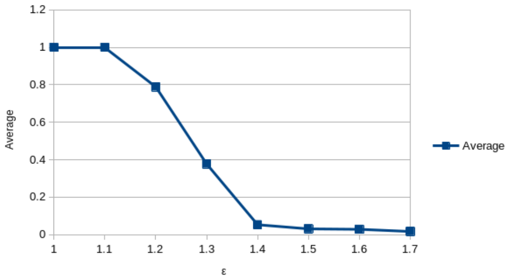

Figure 7: This graph shows the table above with ϵ on the X axis and average on the Y axis.

The table above illustrates the 2nd stage of post-simulation processing done on the data. It indicates the average of base pairs that are open; from the data the largest gap that can be seen in the data is between 1.2 and 1.4. This would show a strong agreement with the visual data seen in the kymographs and the VMD renderings. To that end, it can be concluded that by all simulation methods that when $\theta = 20$, the transition point from being zipped to unzipped occurs at $\epsilon = 1.3$.

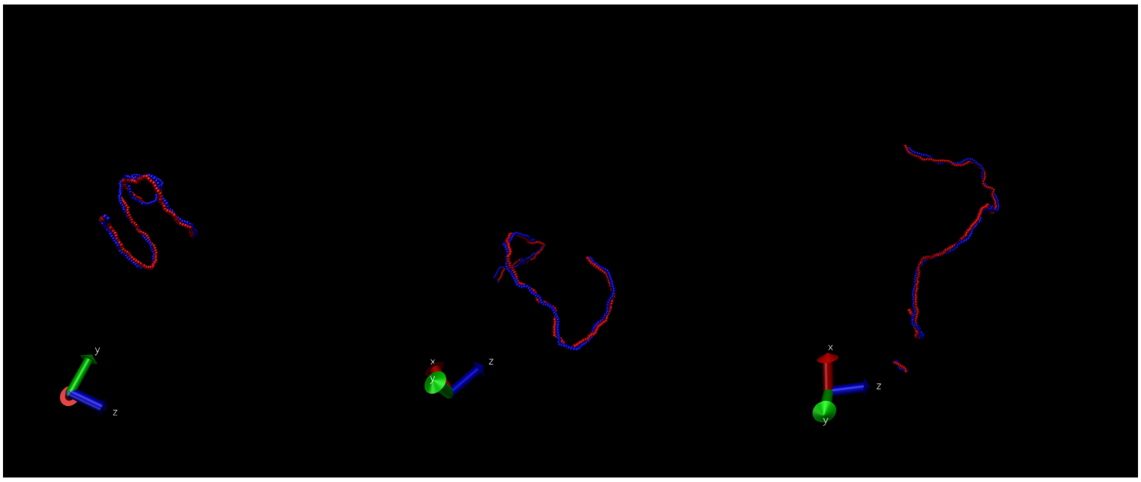

Figure 8: This image shows the start, middle and end of the simulation running with $\epsilon = 1.7$, with $\theta = 10$. (The strands not connected are due to the periodic boundary of the 3D space.)

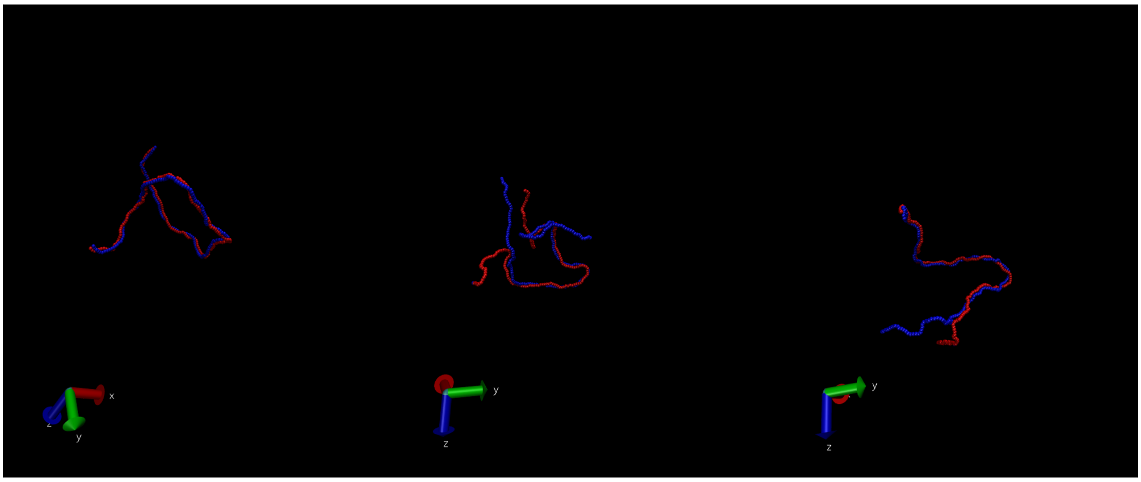

Figure 9: This image shows the start, middle and end of the simulation running with ϵ = 1.6, with $\theta = 10$. (The strands not connected are due to the periodic boundary of the 3D space.)

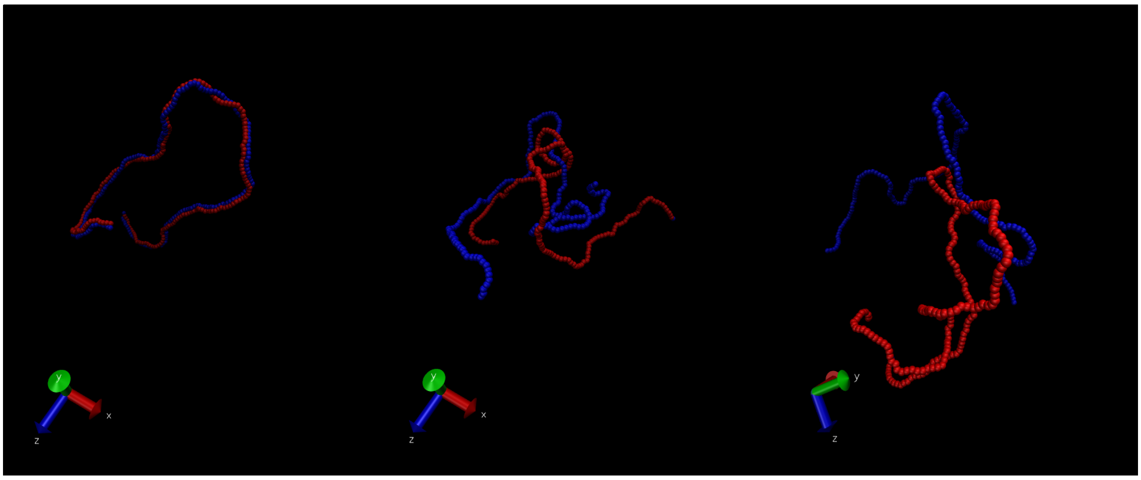

Figure 10: This image shows the start, middle and end of the simulation running with ϵ = 1.0, with $\theta = 10$. (The strands not connected are due to the periodic boundary of the 3D space.)

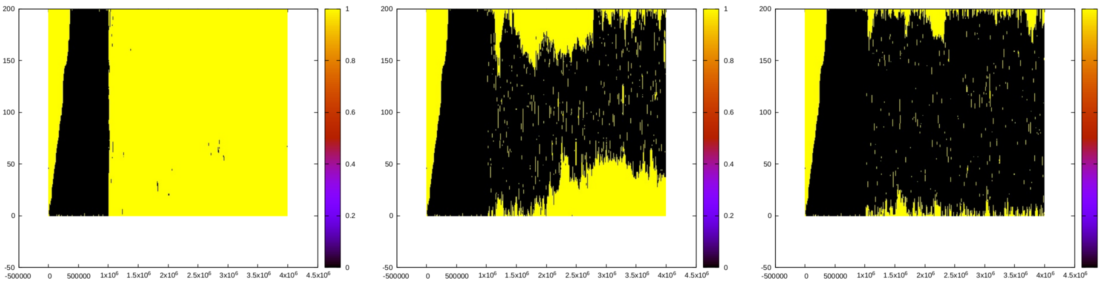

Figure 11: This image shows the kymographs from the post-simulation processing from left $\epsilon = 1.0$, 1.6, 1.7. The area before 1 x 10^6 is to be ignored as this is the initialisation stage, where the polymers initially zip up. Yellow indicates the base pair is open and black indicates closed.

Similarly as before; by looking at figure 8 it can be seen at the end of the simulation that the DNA strand is still completely zipped together. Whereas when we look at figure 10, the strand is completed unzipped and denatured. Therefore the transition point must lay within these parameters when $\theta = 10$. Figure 9 is very comparable to that of figure 4 with the difference being that for figure 9, $\epsilon = 1.6$. This is corroborated by looking at the kymographs in figure 10; this shows that at 1.6 the strand is beginning to unzip. Hinting at the idea that it could be the transition point we're searching for.

ϵ	Average
1.0	0.99920398
1.1	0.99910099
1.2	0.99900398
1.3	0.99380597
1.4	0.97241293
1.5	0.77888059
1.6	0.35181592
1.7	0.07686567

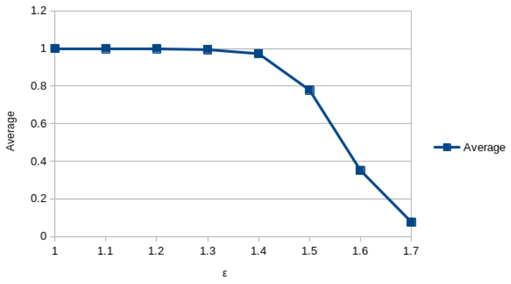

Figure 12: This graph shows the table above with ϵ on the X axis and average on the Y axis.

As before, the table above outlines the average number of open base pairs over all values of ϵ and as such we can use a similar method to conclude where the transition point occurs. From the table we can see the largest change in the average is between 1.5 and 1.7. Therefore we can say that the transition point occurs at 1.6; this is backed up by the VMD renderings as well as the kymographs. From these results we observe that there is quite a difference in ϵ, ranging from 1.3 when $\theta = 20$ compared to 1.6 when $\theta = 10$. This is a very large difference and we can solely put this down to $V_{bending}$ as described in equation (7); we can say its solely down to this because it was the only factor that was changed between the 2 simulations. This indicates that the persistence length of the polymer contains a substantial amount of energy. From equation (7) when we have a lower cosine and lower K_b, there is more entropy at a high temperature and therefore we need to have a larger ϵ to overcome the entropy.

As sound as these results may be, there are a number of things I would like to have done if the given time for the project was longer. These include decreasing the timestep between the simulations, as this would allow for the VMD rendering to play through the entire simulation much slower. As well as this, it would also give the kymographs more data points and as such, they would be smoother and more accurate to find the transition point. Given extra time, I would also have liked to increase the maximum time for the simulations, this would allow the simulations to carry on for a longer time and could find some more interesting results that way. The biggest change to this project that could be done, would be simulating the DNA strands with the correct double helix shape; however this would add far too much complexity into the project and would be very time consuming to program as well as to run. Finally, Being able to analyse more quantitatively the kymographs would allow to see where bubbles form and see how they diffuse along the strand over time.

4 Conclusion

It can be seen very easily that the transition point when $\theta = 20$ is around 1.3. Whereas when $\theta = 10$, the transition point is around 1.6. It can be deduced that because the only change made between the simulations is the potential due to bending via θ. This is due to when there is a lower cosine that a higher ϵ is required to overcome the entropy of the system and as such the transition point does change as a factor of θ when other factors are kept consistent.

5 References

[1] D. Marenduzzo. “The Physics of DNA and Chromosomes”. *IOP Publishing* **2399-2891**, (2018).

[2] M. Mandelkern, et. al. “The dimensions of DNA in solution”. *Journal of Molecular Biology* **152**, (1981).

[3] C. Calladine, et. al. “Understanding DNA The Molecule How It Works”. *Elsevier Academic Press* **0-12-155089-3**, (2004).

[4] C. Richard, et. al. “Poland–Scheraga Models and the DNA Denaturation Transition”. *Journal of Statistical Physics* **115**, (2004).

[5] D. Marenduzzo, et. al. “Dynamical Scaling of the DNA Unzipping Transition”. *Physics Review Letters* **88. 2**, (2002).

[6] Q. Berger, et. al. “Disorder and denaturation transition in the generalized Poland–Scheraga model”. *arXiv* **1807.11397**, (2018).

[7] C. Brackley, et. al. “Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization **Supporting Information**”. *Proceedings of the National Academy of Sciences* **110**, (2013).

[8] L. Verlet. “Computer ”Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”. *Physics Review* **159.1**, (1967).

[9] D. Marenduzzo. “Material for LAMMPS tutorial”. *https://www2.ph.ed.ac.uk/dmarendu/LAMMPS/tutorial.html* **Last Accessed: 19/11/2021**.