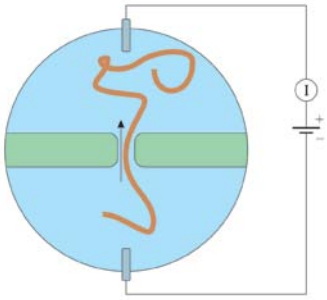
Stochastic Sampling of Stretched DNA

Tamas Szalay, AM207

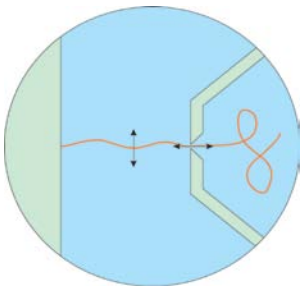
Introduction.

Reducing the cost and error rates of DNA sequencing remains one of the primary outstanding targets of biotechnology research, having the potential to revolutionize medical diagnosis and pharmaceuticals development. One of the most promising among the emerging technologies is called nanopore sequencing, and it operates by threading a single strand of DNA through a hole ('pore') not much larger than the strand itself. In doing so, an electric current passing through the pore is modulated by the presence of each base, which can in principle allow the sequence to be read extremely efficiently and without complicated cutting/reassembly procedures as with other sequencing techniques.



Nanopore sequencing is not without challenges, however. In the simplest configuration, a strand of DNA goes through the pore far too quickly for any useful sequence information to be obtained. In addition, the progress of the DNA is affected by the thermal Brownian motion acting on the strand as it moves through, causing it to travel in a jittery, haphazard fashion, even moving backwards at times.

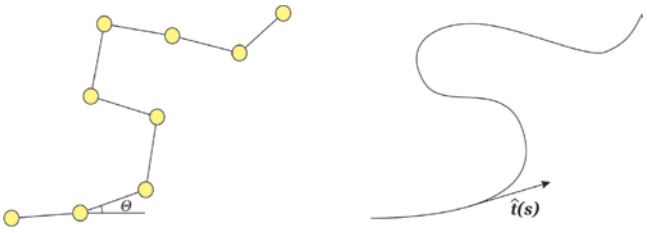
One solution to this problem is to forcibly pin the DNA to a manipulator of some sort, and then feed it in and out of the pore. The current through the pore pulls on the DNA, stretching it out, changing the manner in which it fluctuates but overall holding it stationary. The farther from the pore the DNA is anchored, the larger the fluctuations will be; at a certain point, these fluctuations will exceed the inter-base spacing and the sequence cannot be obtained.



The goal of this project is to use Markov Chain Monte Carlo sampling techniques in comparison with theory in order to determine at which point thermal fluctuations in stretched DNA become untenable. A great deal of previous work exists in realizing physically accurate models of DNA, stretched and unstretched, and single and double stranded, and here we apply it to a system where one end is pinned and along its length the rest is constrained to lie within a nanopore.

Modeling DNA.

In coming up with predictive models for linear polymers, two general approaches have been previously employed: continuous and discrete. In the continuous case, the polymer is parameterized as a flexible curve of fixed total length, with a continuous tangent vector. One can define the behavior of the polymer using various potential energies, assigning an energy penalty for bending, twisting, electrostatic interactions, etc. This potential energy then defines the equilibrium conformations of the polymer according to a Boltzmann thermal distribution.



Freely jointed chain *Wormlike chain*

The problem with continuous models is that they require complex contour integrals in order to make predictions analytically, so discrete models are often adopted. In these models, the polymer is split up into N beads and $N-1$ links of equal length; many integrals then reduce to more tractable sums. The energetic can be modeled similarly, with bending energy penalties similarly assigned. One added benefit is that a (physically realistic) elastic stretching can easily be added simply by including a spring potential term based on the inter-bead position.

The separate potential energy terms are thus

$$U_{i,s} = \frac{1}{2} k_s \left(l_{i,i+1} - l_0\right)^2$$

$$U_{i,b} = -k_b \cos(\theta_i)$$

where s and b denote stretching and bending, respectively. The bending energy is chosen as being the discrete approximation to the continuous model commonly referred to as the "wormlike chain", for which

$$U = A \int_0^L \left| \frac{\partial \hat{\boldsymbol{t}}(s)}{\partial s} \right|^2 ds$$

where $\hat{\boldsymbol{t}}$ is the unit tangent vector in the image above, and the integrand is the square of the curvature.

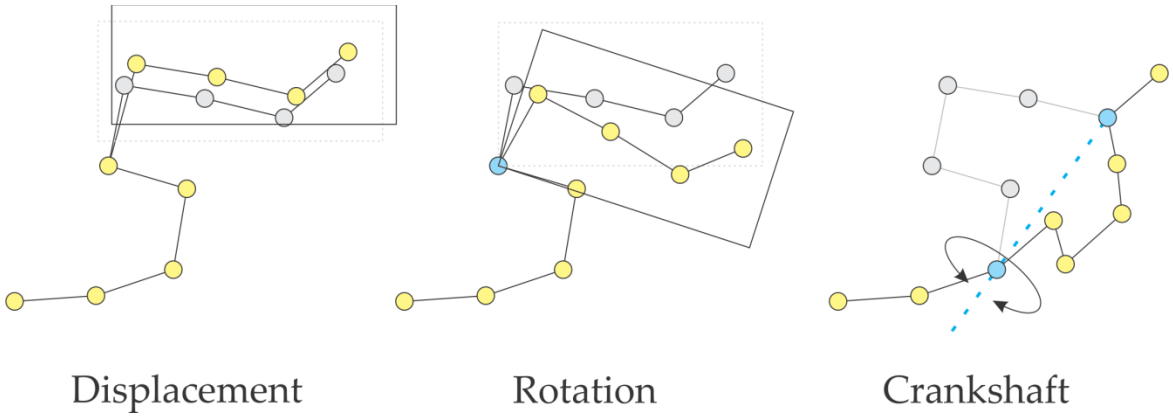
For the purposes of this project, other interactions (electrostatic and hydrodynamic) are neglected, in both cases because the effects become less important when the DNA is extended, as in our systems of interest.

Polymer Monte Carlo.

Given the models in the previous section, we ask a very important question: why Monte Carlo? One's natural reaction when posed with this problem might be to simply perform a dynamical simulation of the system and see what happens. While in many cases a dynamical simulation can be advantageous, it also has a few drawbacks. First, it can take an extremely long time to explore configuration space; a properly constructed Monte Carlo method, on the other hand, can make much larger jumps between states. Second, even though one can include many more interactions and effects, it can often be more difficult to get the model to behave in a physically accurate way: there is no guarantee that the dynamical simulation will yield the equilibrium thermal distribution, and can be affected by integration timesteps and numerical errors. A Markov Chain Monte Carlo sampler, on the other hand, is guaranteed to produce samples from the desired distribution, provided that the space can be explored quickly enough (correlations die off).

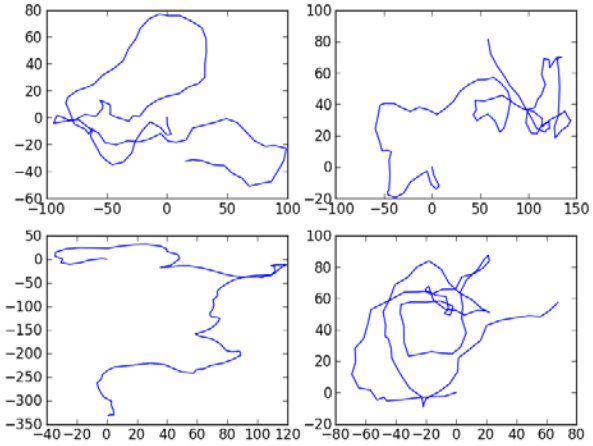
Of critical importance for the performance of MCMC is the distribution of the proposed steps. A naïve approach might be to simply pick a bead at random and perturb its position by some amount Δ, and then accept or reject the new configuration if $r \sim U(0,1) < e^{-\frac{\Delta U}{kT}}$ (standard Metropolis-Hastings). However, this means that the motion of one end requires $O(N)$ moves to affect the other end, and it will take the polymer nigh forever to explore state space.

Thus, we instead turn to so-called "global moves", of which I have implemented three types for this computation. Global moves have the property that they move multiple beads together, while changing the bond energy in only one or two locations. The types are illustrated in the image below. In a displacement, instead of just one bead in the chain being moved locally, the rest of the chain past a randomly chosen index is (randomly) displaced together. Only one length and two angles change. In a rotation, a pivot is randomly chosen and the rest of the chain is randomly rotated. No lengths and only one angle change. Finally, in a crankshaft, two distinct beads are randomly chosen, and the beads between them are randomly rotated along the axis connecting the two pivots. No lengths and two angles change. In this way, the chain can rapidly explore configuration space without changing the energy in such large jumps that no moves are accepted.

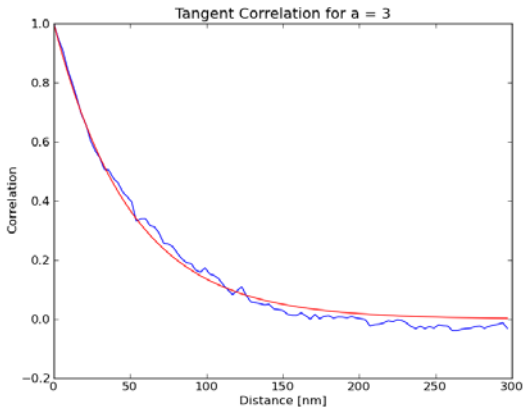


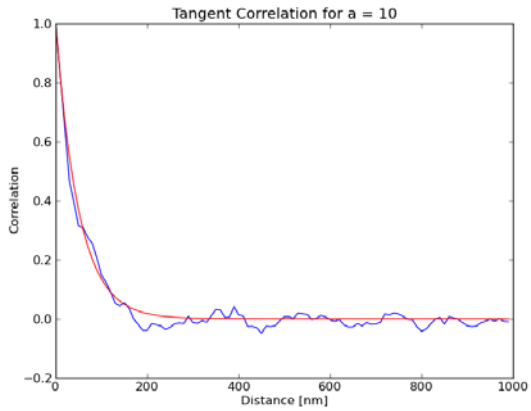
Results.

I have implemented the sampling in Python, using the Numpy library. At each Metropolis-Hastings iteration, one of the three moves above are proposed at random, the energy recalculated, and then the move accepted or rejected based on the formula given above. The sizes of the proposed steps are tuned to give accept-reject ratios between $30\% - 70\%$ or so. The proposal generation for displacement is straightforward, as well as for crankshaft, which only requires an angle θ. For the rotation move, however, what we would ideally like is a small-angle random rotation matrix, one that is uniformly distributed within a small range (which is not strictly necessary, it need only be symmetric, but uniform is simpler). To do this, I implemented the algorithm from Graphics Gems III by Arvo (1992), which does exactly that. The parameter $thinning = 1$ corresponds to a sample being taken every N global steps, where N is the number of beads, since each step updates ~ 3 effective degrees of freedom. For the calculations of the energies involved, I have chosen physically accurate values of the spring constants k_s (elastic modulus $1000\ pN$) and k_b (corresponding to correlation length of $50\ nm$). Some example projections of the 3D configurations can be seen to the right.



We can test the performance of this sampling by looking at the correlation of the tangent vector as a function of contour distance. Denoting one end as $\hat{\boldsymbol{t}}(0)$, the continuum wormlike chain model predicts that $\langle \hat{\boldsymbol{t}}(0) \cdot \hat{\boldsymbol{t}}(s) \rangle = e^{-\frac{s}{L_p}}$, where L_p is the persistence length. We can see from the plots below that this does in fact work, where the two plots correspond to different inter-bead spacings, but otherwise the same physical parameters.



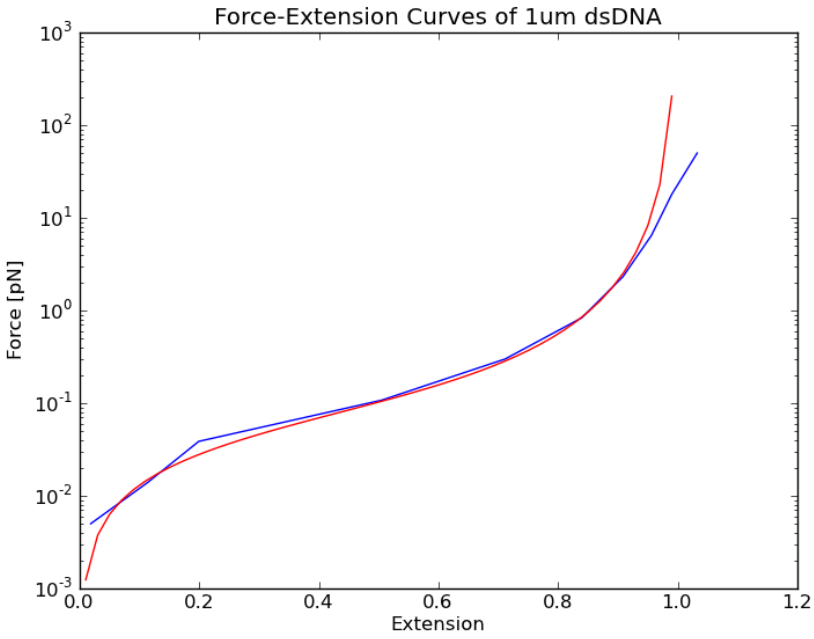


Stretching DNA.

We can make the DNA stretch by pinning one end and then applying a force F and adding a potential energy $U_F = -Fz$, where z is the coordinate of the far end of the DNA. This corresponds to the DNA being pulled by a constant force in the z direction, such as would be applied eg. by a magnetic bead. Based on the wormlike chain model, there is a theoretical approximate curve from Marko et al. 1995 that predicts the following (with L the full contour length of the DNA):

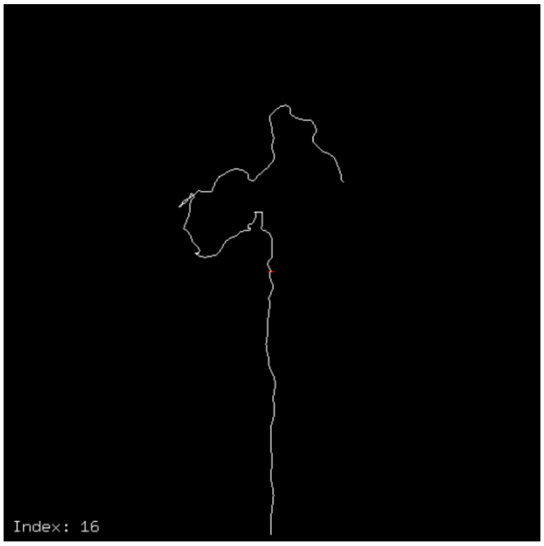
$$\frac{FL_p}{kT} = \frac{z}{L} + \frac{1}{4\left(1 - \frac{z}{L}\right)^2} - \frac{1}{4}$$

When we run MCMC sampling at a variety of applied forces and calculate the expectation of the extension $\frac{z}{L}$, we get the following plot:



The red curve is the theoretical prediction given above. Of course, the wormlike chain is an inextensible model, so it cannot be stretched past $z = L$. Ours, on the other hand, includes the natural stretchiness of DNA, which is why the force required to produce a particular extension is lower than it would otherwise be. In the previous section, the correlation length was effectively a trivial consequence of the model used; this one is a bit more subtle and it is reassuring to see that it also agrees nicely with theory.

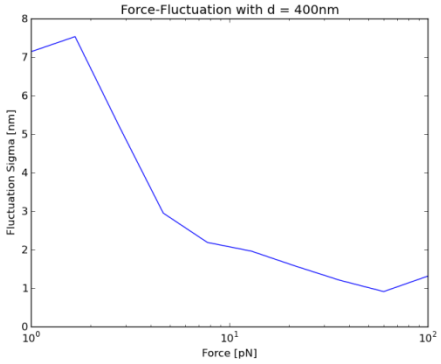
DNA in Nanopores.

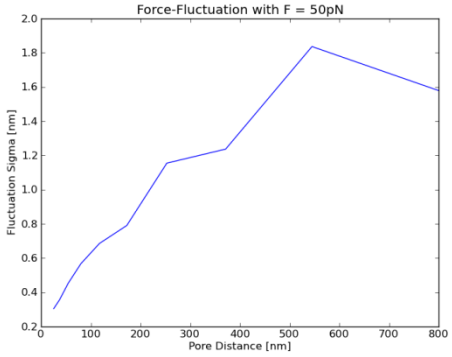


Finally, we can include the constraint of the nanopore, which sits on the z-axis a distance d away from the pinned end of the DNA. We impose a hard constraint on collision: if any part of the DNA passes through the plane $z = d$ outside of the circle $r = r_p$, we automatically set $U = \infty$. Then, if we let the contour coordinate along the DNA that is currently at $z = d$ (in the pore) be denoted $s(z = d)$, we can write the potential energy as $U_p = F_p\, s(z = d)$.

In the image to the right, taken from the 3D viewer I put together, the nanopore is the red dot directly in the center; the portion of the DNA below the nanopore is clearly stretched out, and the rest of it flops around freely, since the force acts directly at the nanopore.

Then, what we wish to measure is the fluctuations in $s(z = d)$ as the other end is held fixed, and see whether they would be large enough so as to make sequencing infeasible. The standard deviation of these fluctuations are plotted below, both as a function of the force F and the distance d:





We can see that the general behavior is as we might expect: with increasing force and decreasing distance, the magnitude of the fluctuations decreases. The inter-base spacing in stretched DNA is approximately $0.5 - 0.7$ nm; of particular note is that fairly large forces (50+ pN) are required to stretch it sufficiently.

More work is required to investigate the effects of single stranded DNA, as well as the structural transitions that can occur when double stranded DNA is sufficiently stretched, but this software provides a good framework for answering future questions.