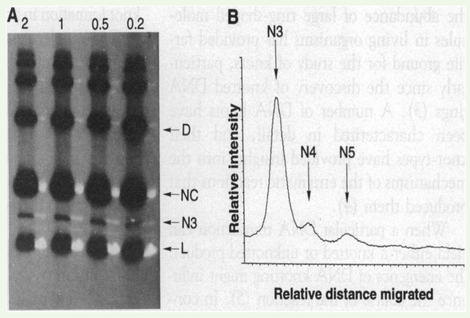
Optimizing A Lattice Polygon Model To Fit To Experimental DNA Knotting Data

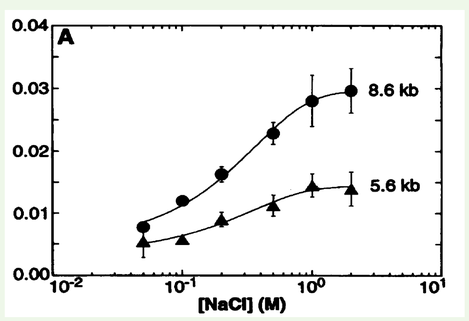
SURI - 2018

Prateek Verma with Matthew Schmirler, supervised by Dr. Chris Soteros
Department of Math & Statistics, University of Saskatchewan

Abstract

- DNA strands exist in a supercoiled state which can result in the formation of knots. During the cell replication process, DNA must be unknotted in order to carry out the division process.
- To study knotting in DNA through computer simulations, we used cubic lattice Self Avoiding Polygon model. Our goal is to optimize the model parameters such that the Simulation data closely fit with Experimental Data.
- Shaw and Wang estimated the probability of knot formation in the presence of NaCl for 5.6-kbp and 8.6-kbp DNAs. [1]





- Their experiment observed that as the concentration of NaCl increases, the probability of the DNA becoming knotted also increases.
- An unoptimized model of DNA in simple cubic lattice shows similar trend when compared with the experimental results of Shaw and Wang.

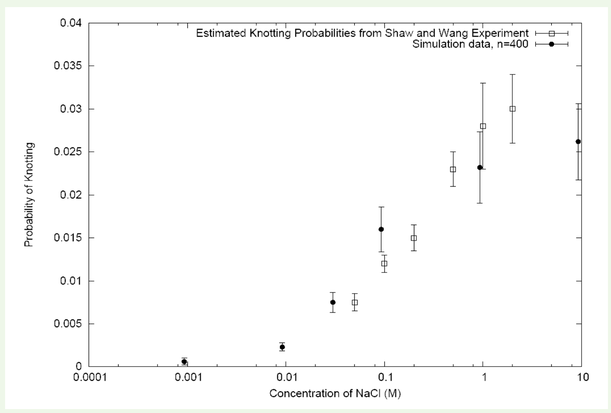
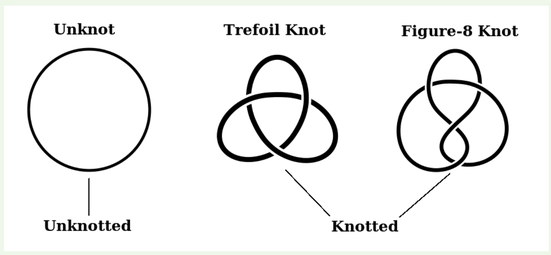


Fig. A comparison of equilibrium knotting probabilities for SAPs of length 400 and DNA molecules with 8600 base pairs as a function of NaCl concentration. [3]

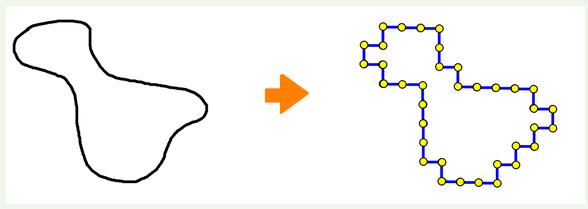
What is a Knot?

- A Mathematical knot is a closed, non-self-intersecting curve in 3 space.
- Self-avoiding polygons can be classified as knots. They fall into the category of knots known as polygonal knots (knots with edges).
- Most trivial knot type is a *unknot*. This knot type consist all the knots which can be continuously deformed into a circle.
- If the knot cant be continuously deformed into a circle, we say that it is knotted. The simplest knot type after unknot is a trefoil knot. It consist of three crossings.



The Model

- We use self-avoiding polygons (SAPs) in the simple cubic lattice as a model for randomly cyclized DNA.



- The goal is to develop a model to fit to Shaw and Wang Experimental Data, i.e. knotting probability as a function of NaCl concentration. To account for the effect of added salt, we model DNA-DNA and DNA-solvent interactions by assigning each SAP ω an energy. [2]
- The energy ($U(\omega)$) consists of DNA monomer interaction affected by parameter υ (negative in value). It is a representative of whether monomer prefer to be surrounded by solvent molecules or other monomer. Energy also consists of a long range screened coulomb potential affected by parameter A.
- For equilibrium distribution for a fixed SAP length n, the canonical distribution is given as

$$P(\omega) = \frac{e^{-\frac{U(\omega)}{K_B T}}}{\sum_{\omega`} e^{-U(\omega`)/K_B T}}$$

- The *interacting pivot algorithm* (A Markov Chain Monte Carlo program) is used to sample SAPs from this canonical distribution and KNOTPLOT packages are used to determine the knot types in the sampled polygons.

Optimizing Method

- For each choice of A and v, one can run simulations of the interacting pivot algorithm for the same salt concentrations as the Shaw/Wang data. Using data from these simulations, knotting probabilities can be estimated and compared to the experimental data. How close the simulation data is to the experiment is measured by a sum of weighted square errors (Loss function).
- In order to find the values of A and v that minimize this error, a stochastic optimization technique called Finite Difference Stochastic Approximation can be used.
- The Finite Difference Stochastic Approximation (FDSA) method is analogous to deterministic steepest descent algorithm with the gradient estimate ($\hat{g}_k(\hat{\theta}_k)$) replacing the direct gradient.

$$(A,\upsilon)_{k+1} = (A,\upsilon)_k - a_k \hat{g}_k(A,\upsilon)_k \quad ------ \quad (1)$$

- The essential part is the gradient estimation which involves measurement of the Loss function (sum of weighted square errors) in the form,

$$L\big((A,\upsilon) \pm perturbation\big)$$

- Lower the value of loss function, more optimized the parameters are.
- The gradient $\hat{g}_k$ is calculated as,

$$\hat{g}_k(A,\upsilon)_k = \begin{pmatrix} \frac{L(A+c_k) - L(A-c_k)}{2c_k} \\ \\ \frac{L(\upsilon+c_k) - L(\upsilon-c_k)}{2c_k} \end{pmatrix}$$

- Common forms for the gain sequence a_k and c_k are given as,

$$a_k = \frac{a}{(k+1+A)^\alpha}\ ,\ c_k = \frac{c}{(k+1)^\gamma}$$

coefficients a, c, α, and γ are strictly positive and $A \geq 0$

- The values of these coefficients were chosen accordingly, by using trial and error. Under appropriate conditions, the iteration (1) converges to the optimized $A^*$and υ^*.

Results & Discussions

- We performed FDSA optimization for the first three data points of Shaw and Wang Experimental Data. (ζ = 0.733, 1.034 and 1.466)
- The initial guess was taken as A=0.025, υ=-0.24. After trial and error, the values of coefficients in the sequences were chosen as follows, c=0.00065, γ=0.101, a=0.000001, A=1, α=0.602, SAP length n=400, simulation time=75 Million.
- The avg. loss function value before optimization was 36.62, after 13 iterations it went down to **0.63**. Optimized parameter values are, ***A*=0.0135 and υ=-0.2422**
- One issue with the FDSA program is it is very **time consuming** as it requires us to run multiple MCMC programs for the calculation of the loss function. A single simulation for n=400 polygon length with 50 million steps takes 5 hours (approx.) to complete. Using open API in c **we reduced the overall simulation time** by running multiple simulations across multiple CPU threads.

Future Work

- To perform optimization for all six points of Shaw and Wang Data.
- To implement Simultaneous Perturbation Stochastic Approximation (SPSA) technique for the optimization and check for improvements (if any) over FDSA.

References & Acknowledgement

- I would like to thank my supervisor Dr. Chris Soteros, for her advice and explanation of different concepts during the internship. I'd also like to thank Matthew Schmirler for providing me his pivot algorithm program, explanation of different approximation techniques and his help in writing the optimization program.

References:

1. SY Shaw and JC Wang. Knotting of a dna chain during ring closure.Science, 260(5107):533{536, 1993.
2. M. C. Tesi, E. J. Janse van Rensburg, E. Orlandini, D. W. Sumners, and S. G. Whittington. Knotting and supercoiling in circular dna: A model incorporating the effect of added salt. Phys. Rev. E, 49:868{872, Jan 1994.
3. Matthew Schmirler. Strand Passage and Knotting Probabilities in an Interacting Self-Avoiding Polygon Model. M.Sc. Thesis. University of Saskatchewan, 2012