Mathematical Modelling of DNA
Prof. John Maddocks
Session 7 A. Patelli

1 Explicit computation of apparent persistence length for a tractable probability density function (the HWLC)

During the Lecture of 3.04.2017, we introduced a simplified model of DNA in order to compute explicitly the expectation of the tangent-tangent correlation. In this exercise we will numerically check this result. A PDF version of the Lecture notes containing all the details of the simplified model and the computation can be found on the web page of the course as a supplement to the note.

1. Implement in MATLAB the Euler-Rodrigues formula (Eq. 2 session 3) for computing $Q(u)$. Check that your code produces a rotation matrix for any u.

2. Let $\langle Q(u)\rangle$ be the expectation of $Q(u)$ with respect to $\rho(u) = \frac{1}{Z}\exp\{\frac{1}{2}(u-\hat{u})\cdot K(u-\hat{u})\}$, where $\hat{u}$ is a Cayley vector with 0° of tilt and roll, and 36° of twist and $K = \mathrm{diag}(100, 100, 100)$. Compute using Monte Carlo the values of the following expectations $\langle Q(u)\rangle_{(1,3)}$ and $\langle Q(u)\rangle_{(2,3)}$, for $N = 100, 1000, 10000, 100000$ samples. What do you obtain and why?

3. Now we focus only on the entries $\langle Q(u)\rangle_{(3,3)}$ which represent the tangent-tangent correlation. Compare now the explicit result obtained in class for $\langle Q(u)\rangle_{(3,3)}$, i.e,

$$\langle Q(u)\rangle_{(3,3)} = 1 - \frac{2}{1+\hat{u}_3}\left(\frac{1}{K_1} + \frac{1}{K_2}\right), \tag{1}$$

   with the result obtained via Monte Carlo simulation, for the following cases:
   (use $N = 100, 1000, 10000, 100000$ samples)

   i) $\hat{u}: 0^\circ$ tilt, 0° roll, 36° twist , $K = \mathrm{diag}(543, 543, 543)$, what is the persistence length?
   ii) $\hat{u}: 0^\circ$ tilt, 0° roll, 0° twist , $K = \mathrm{diag}(600, 600, 600)$, what is the persistence length?
   iii) $\hat{u}: 0^\circ$ tilt, 0° roll, 72° twist , $K = \mathrm{diag}(393, 393, 393)$, what is the persistence length?

4. For each case, redo the computations with smaller values of K_3. What can you say?

5. For each case, redo the computations with smaller values of K_1. What can you say?

6. For each case, redo the computations with smaller values of K_1 and K_2 (not necessarily with same value). What can you say?

7. (optional) How small can K_1 and K_2 be taken with the analytical formula remaining as a good approximation ?

2 Monte Carlo simulation with the cgDNA model II

Download here, `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/lambda_phage.mat`, the sequences you will work on in this exercise. The five sequences, called $\texttt{lambda}_{1\ldots5}$ have 300 base-pairs and are all part of the λ phage genome (see the details in Sanger et al. 1982).

2.1 cgDNA reconstruction of the five lambda phage segments

Using the cgDNA MATLAB package reconstruct the ground state of the five sequences using cgDNAparamset2, then compare the 3D reconstruction of all the ground states.

2.2 Monte Carlo simulation of the cgDNA model with MATLAB

See Ex. 1.1 Session 9 for details about Monte Carlo simulation with MATLAB. Use only, for brevity, the cgDNAparamset2.

Sample 250 configurations for each of the five sequences, and plot each sampled configuration plus the ground state on one figure. Here you can just plot the xyz position of all the base–pairs for each configuration. Are the five plot similar? Compare them to the plot obtained with the six poly-dinucleotides (Ex. 1.1 Session 9).

2.3 Cloud of points of last base-pair positions

Use cgDNAmc to sample 10^5 configuration for each sequence and save for each sample the last base-pair position (see Ex. 1.3 Session 9). Plot using MATLAB the cloud of point, what can you say?Also compare the obtained clouds with the poly dinucleotide ones from Ex. 1.3 Session 9.

2.4 Compute the persistence length using the cgDNAmc code

The tangent–tangent correlation

For the following Monte Carlo simulations use 10^5 number of samples.

i) Plot in a single figure all the tangent–tangent correlation (ttc) values for each sequence. Explain the different behaviour and compare the obtained plot with the plot of the ttc values of the six poly-dinucleotides.

ii) Compute the persistence length for each sequence.

iii) Define by $\hat{\mathbf{t}}_1 \cdot \hat{\mathbf{t}}_i$ the intrinsic tangent-tangent for the base pair i, i.e, the tangent-tangent computed at the i–th base pair of the ground state. For all the six lambda sequence plot the following

$$\ln\langle \mathbf{t}_1 \cdot \mathbf{t}_i \rangle - \ln \hat{\mathbf{t}}_1 \cdot \hat{\mathbf{t}}_i \text{ vs base pair } i \tag{2}$$

Equation (2) is the so called factorized tangent tangent correlation and will be introduced during the next class (10.04.2017).

The Flory vector

For the following Monte Carlo simulations use 10^5 number of samples. Plot in a single figure (use `plot3` and plot only each base–pair position) the ground state and the Flory persistence vector (F), moreover on the Flory persistence vector plot a cross each 25 base–pairs. What can you say about the position of the crosses? Is the Flory persistence vector converged? (the convergence here is on the number of base–pairs and not on the number of samples).

3 Effect of the sparsity on the Monte Carlo simulation efficiency

Nowadays Monte Carlo simulation for univariate normal distribution can be done in an extremely efficient way. Hence the univariate case is also used to sample from a multivariate normal distribution, but for taking advantage of the univariate case one as to diagonalise the stiffness matrix. Here you have two methods to diagonalise the cgDNA stiffness matrix.

- Use the spectral decomposition to diagonalise the cgDNA stiffness matrix, i.e, write $K = M\Lambda M^T$.
- Use the Cholesky factorisation to decompose the cgDNA stiffness matrix, i.e, write $K = L^T L$.

1. Using the MATLAB function `mvnrnd` sample 500 configurations using both methods of diagonalisation on poly$(AT)_N$ with different N.

2. The computation of ttc and F depends only on the value of the inter variables. Thus, formally the expectation of these two quantities is with respect to the marginal distribution of the inter variables. More precisely, let $x = (y, z) \in \mathbb{R}^{12n-6}$ be a cgDNA configuration where $y \in \mathbb{R}^{6n}$ are all the intra variables and $z \in \mathbb{R}^{6(n-1)}$ are all the inter variables. Denote by $\phi(z)$ a non linear function of the inters. Thus the expected value of ϕ with respect to the cgDNA distribution $\rho(y, z)$ is

$$\langle \phi(z) \rangle = \int_{\mathbb{R}^{6n}} \int_{\mathbb{R}^{6(n-1)}} \phi(z)\rho(y, z)dzdy = \int_{\mathbb{R}^{6(n-1)}} \phi(z)\rho(z)_y dz = \langle \phi(z) \rangle_z, \tag{3}$$

where $\langle \cdot \rangle_z$ is the expectation with respect to the inter's marginal distribution $\rho(z)_y$. We want now to compare the efficiency of sampling from the whole cgDNA distribution against sampling from the marginal with respect to the intra variables, using Monte Carlo simulation. For doing that consider different repetitions of poly(AT) and, as method of decomposition, use the Cholesky factorisation. For this exercise you can fix the number of samples to 500.

Remark: The marginal distribution, of a subset of variables, of a Gaussian distribution can be obtained by selecting, in the covariance and in the mean, the corresponding blocks. For example, in the case of a cgDNA distribution the marginal distribution of the inters is obtained by extracting all the inter-inter blocks of the covariance matrix, and the inter part of the mean.