Note: Following exercises will use notation poly$(\alpha\beta)_N$ which was introduced in Qu4 Serie 7. Download here, `https://lcvmwww.epfl.ch/teaching/modelling_dna/protected_files/codes_exercises/cgDNA+ps2`, the parameter set `cgDNA+ps2.txt` to be used with cgDNApmc code. Note that the parameter set `cgDNA+ps1.txt` is given in Serie 8, which is different than `cgDNA+ps2.txt` in a sense that different MD protocol is used.

1 Computing persistence length (using cgDNApmc) part-1

In this exercise we will compute in two different ways the persistence length of all the six poly-dinucleotide (see Qu4 Serie7 for definition of six poly-dinucleotide).

- Use the command **-e t0** with run_cgDNAmc_from_seq to compute the tangent-tangent correlation (ttc).
- Use the command **-e r** with run_cgDNAmc_from_seq to compute the Flory persistence vector.

Remark: The tangent-tangent correlation is the (3,3) entry of the quantity $\langle R_1^T R_i \rangle$, while the Flory persistence vector is the quantity $\langle R_1^T (r_i - r_1) \rangle$.

1.1 Convergence study of the Monte Carlo simulation

Use parameter set `cgDNA+ps1.txt` for the following simulations. The estimation of a mean value of an observable strongly depends on the number of samples used in the Monte Carlo simulation. In this part we will empirically study this convergence for the tangent–tangent correlation (ttc) for poly$(AT)_{150}$.

1. Solo version: Compute ttc for the following number of samples $N_s = 250, 10^3, 10^4, 10^5, 10^6$, and for each N_s repeat the experiment 10 times. For each N_s save all the computed ttc, average them and compute the standard error at some fixed base–pairs. Visualize them by plotting the log of ttc against the number of base-pairs.
2. Cooperative version: Work in a group (in Covid pandemic situation we encourage to work online) to do the same process explained above where each person do a certain number of repetition, try to have between 30 and 50 repetition for each N_s.

For which number of samples can the values of the ttc be considered as converged? Denote this number by N_s^c.

Note: Now for all the questions in this session use $N_s^c = 10^5$ (for this number ttc is considered as converged, see solution of Qu 1.1) number of samples in MC simulation.

1.2 The tangent–tangent correlation

Use parameter set `cgDNA+ps1.txt` and N_s^c number of samples for the following simulations of all the six poly$(\alpha\beta)_{150}$.

1. In a single figure plot the ln of the ttc values versus base–pair index for each distinct poly-dinucleotide. To what are related the wiggles?

2. The persistence length is computed as the negative reciprocal of the slope of the least squares fit to the plot of ln of the ttc. In matlab you can use the "backslash" to solve least square problems. Compute the persistence length for each poly-dinucleotide.

1.3 The Flory vector

Use parameter set `cgDNA+ps1.txt` and N_s^c number of samples for the following simulations of all the six poly $(\alpha\beta)_{150}$.

1. Plot in a single figure (use `plot3` and plot only each base–pair position) the groundstate and the Flory persistence vector, 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).

2. Using the N founded in the previous part, redo the Monte Carlo simulation using the parameter set `cgDNA+ps2.txt`. Compare the result with the Flory persistence vector computed with the parameter set `cgDNA+ps1.txt`.

3. (Optional) Do the study of the convergence of the Monte Carlo simulation of the Flory persistence vector for poly$(AT)_{500}$. Is N_s^c the same as for ttc?

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

Note: In this question we work with simple model where matrix K and vectors u and $\hat{u}$ appears. The entries of K are random numbers and one can choose any number. Also u and $\hat{u}$ can be random. But we choose value of $\hat{u}_3$ to be 36° because approximately the twist in DNA is about 360° after each 10.5 base pairs. Also, in one question we choose each entries of K to be 600 because this will predict persistence length ≈ 150 (when $\hat{u}$ is 0 vector) which is in some sense also the persistence length of the DNA.

During the Lecture of week 9 a simplified model of DNA is introduced in order to compute explicitly the expectation of the tangent-tangent correlation. In this exercise we will numerically check this result.

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

2. Take 10 random vectors as u and compute corresponding $Q(u)$ using your implementation of Euler-Rodrigues formula in matlab and then compute and plot eigenvalues for each of them. Further compute the average of all ten $Q(u)$ and compute the eigenvalues of averaged quantity. What you will observe?

3. 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 simulations the values of the following expectations $\langle Q(u)\rangle_{(1,3)}$ and $\langle Q(u)\rangle_{(2,3)}$, for $N = 10^5$ samples. What do you obtain and why?

4. 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 lecture for $\langle Q(u)\rangle_{(3,3)}$, i.e,

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

with the result obtained via Monte Carlo simulation (using $N = 10^5$ samples) for the following cases:

i) For $\hat{u}: 0°$ tilt, $0°$ roll, $36°$ twist , $K = \text{diag}(\text{range of 10-600}, 600, 600)$, show in one plot the variation in the value of $\langle Q(u)\rangle_{(3,3)}$ with $K_{1,1}$. What can you say?

ii) For $\hat{u}: 0°$ tilt, $0°$ roll, $36°$ twist , $K = \text{diag}(600, 600, \text{range of 10-600})$, show in one plot the variation in the value of $\langle Q(u)\rangle_{(3,3)}$ with $K_{3,3}$. What can you say?

iii) Consider $K = \text{diag}(600, 600, 600)$ with $\hat{u}$: $0°$ tilt, $0°$ roll, range of $0°$ to $150°$for twist. Show in one plot the variation in the value of $\langle Q(u)\rangle_{(3,3)}$ with twist $(\hat{u}_3)$. What can you say?

3 Computing persistence length (using cgDNApmc) part-2

In this exercise you will work with the five sequences, called $\texttt{lambda}_{1\ldots5}$ having 300 base-pairs and are all part of the λ phage genome (see Qu2 Serie 8 for matlab file of these sequences).

3.1 The tangent–tangent correlation

Use parameter set `cgDNA+ps1.txt` for the following Monte Carlo simulations with λ phage sequences and 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 groundstate. For all the five lambda sequence and all six distinct poly-dinucleotide considered in the Qu 1, 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.

3.2 The Flory vector

Again use λ phage sequences for the following Monte Carlo simulations and use 10^5 number of samples. Plot in a single figure (use `plot3` and plot only each base–pair position) the groundstate and the Flory persistence vector, 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).