1 Computing persistence length (using cgDNApmc) part-1

1.1 Convergence study of the Monte Carlo simulation

In Fig. (1) we present the result of the convergence test for the tangent–tangent correlation (ttc) for poly$(AT)_{150}$. Ten independent Monte Carlo runs have been done for five different total numbers of wanted configurations: 250, 10^3, 10^4, 10^5, and 10^6. The error bars are plotted every 2 base-pairs, and one can observe that the size of the error bars decrease drastically when passing from 10^4 to 10^5 number of sampled configurations. This tells us that in order to estimate the value of the ttc one has to sample at least 10^5 configuration in order to decrease the variability of the resulting value. For more details on convergence tests we refer to the article "Sequence-dependent persistence lengths of DNA", J. S. Mitchell et al., JCTC, 2016.

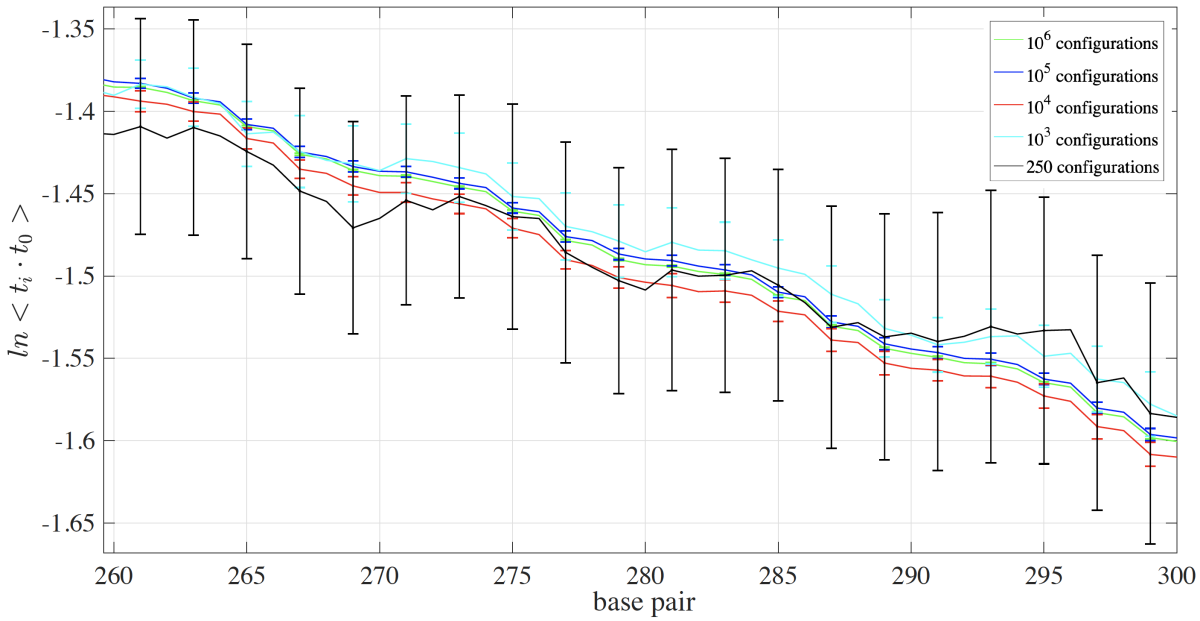

Figure 1: Convergence of the log of the ttc for different number of sampled configurations.

1.2 The tangent–tangent correlation

i) Using the cgDNApmc code we obtained the following Fig. (2), for the tangent–tangent correlation for each poly–dinucleotide. Here we sampled 10^5 configurations for each sequence. The wiggles are strictly related to the radius and pitch previously shown in Fig. (7) corr 7.

ii) The persistence length, in this case, is computed as the number of base-pairs equal to the negative reciprocal of the slope of the straight line through the origin that is the least square fit of the plot of $\ln\langle t_i \cdot t_0\rangle$. For the six poly–dinucleotide we have computed the following values:

- poly$(AA)_{150}$: 248 bp,

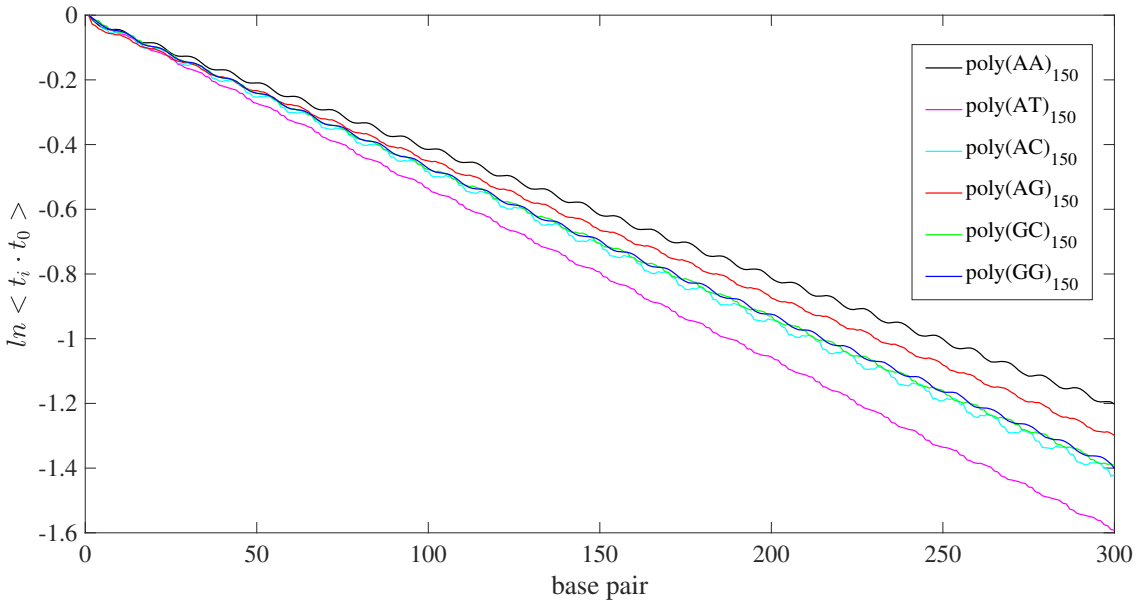

Figure 2: Tangent–tangent correlation for each poly–dinucleotide as function of the number of base–pair. 10^5 configuration have been sampled for each sequence.

- poly$(AT)_{150}$: 188 bp,
- poly$(AC)_{150}$: 211 bp,
- poly$(AG)_{150}$: 230 bp,
- poly$(GC)_{150}$: 214 bp,
- poly$(GG)_{150}$: 215 bp.

Remark: One can notice that depending on the sequence persistence length can vary up to 40% (persistence length for poly$(AA)_{150}$ is almost 1.4 times of the persistence length of poly$(AT)_{150}$).

1.3 The Flory vector

i) In Fig. (3) we plotted all the six Flory persistence vector computed with cgDNApmc for each poly-dinucleotide. Again we sample 10^5 configurations for each sequence. The crosses plotted on all the six Flory persistence vectors are plotted every 25 base-pairs and shows that they are converging in the sense that the Euclidean distance between two consecutive crosses is decreasing. For simple model of polymer chain, as the Wormlike chain model or the Freely rotating chain model or the random–ϕ model, it can be shown that the norm of the Flory persistence vector converge to a finite value for sufficiently long chains. We can expect a similar behaviour also for the DNA and for our model, the cgDNA+, that is rather complex.

ii) For sake of brevity, here we show just the result for ploy$(AT)_{600}$. In Fig. (4) one can observe the accumulation of the crosses plotted each 25 base–pairs, on the left, and the comparison between that Flory persistence vectors computed using the parameter sets cgDNA+ps1 and cgDNA+ps2 in the right. The most important difference between the parameter sets is the prediction of the persistence length, in fact cgDNA+ps1 predict an higher persistence length for the DNA then the cgDNA+ps2.

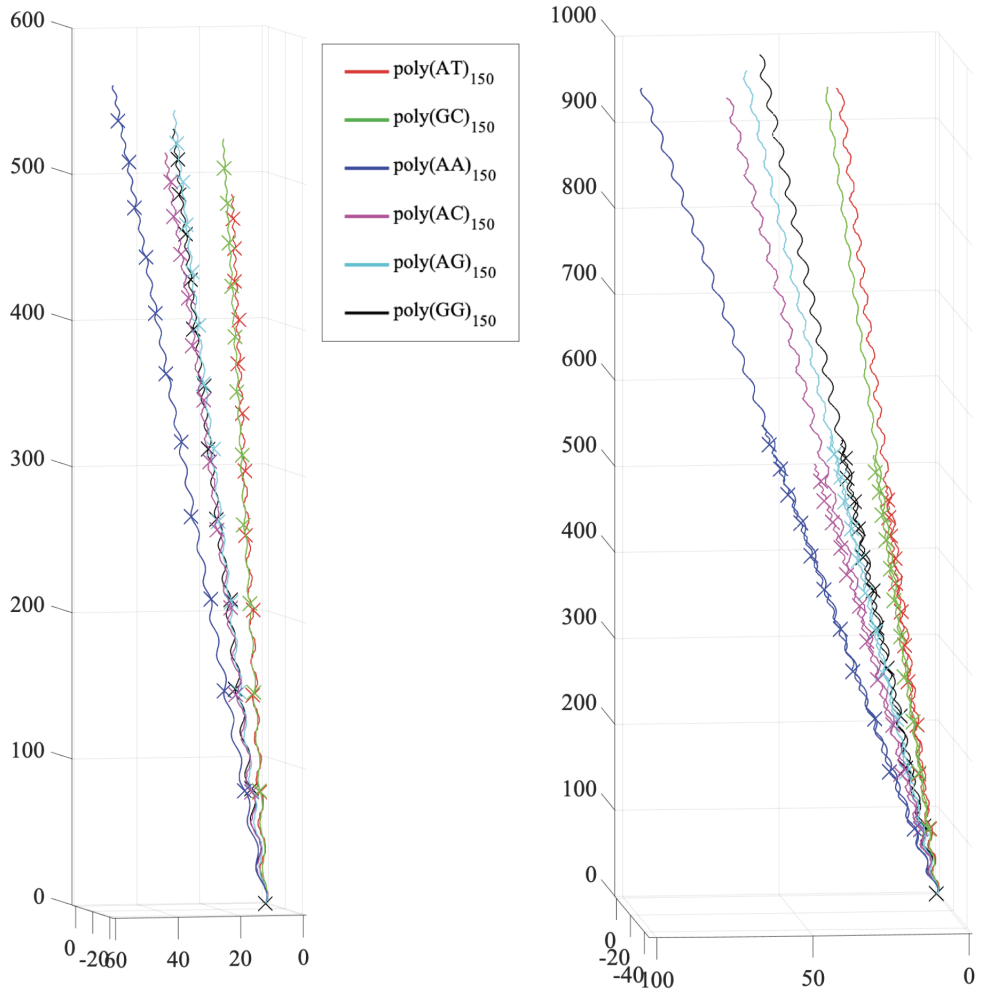

Figure 3: The Flory persistence vectors (left figure) for all the six poly-dinucleotide sequences computed with cgDNApmc and 10^5 configurations have been sampled. The crosses in the Flory persistence vectors have been plotted after each 25 base-pairs. However, in the right side we have shown the groundstate along with the Flory persistence vectors (same as in the left figure) for each poly-dinucleotide sequences.

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

1. Here you have an example of implementation of the Euler–Rodrigues formula.

```
function Q = Euler_Rodrigues( u )
% This Ensure that the vector u is a column vector
u = reshape(u , [ 3 1]) ;
% Compute norm of u
uu = u'*u ;
% Get the matrix [ u x ]
u_cross = [  0    -u(3)   u(2)
             u(3)   0     -u(1)
            -u(2)   u(1)   0     ] ;

```

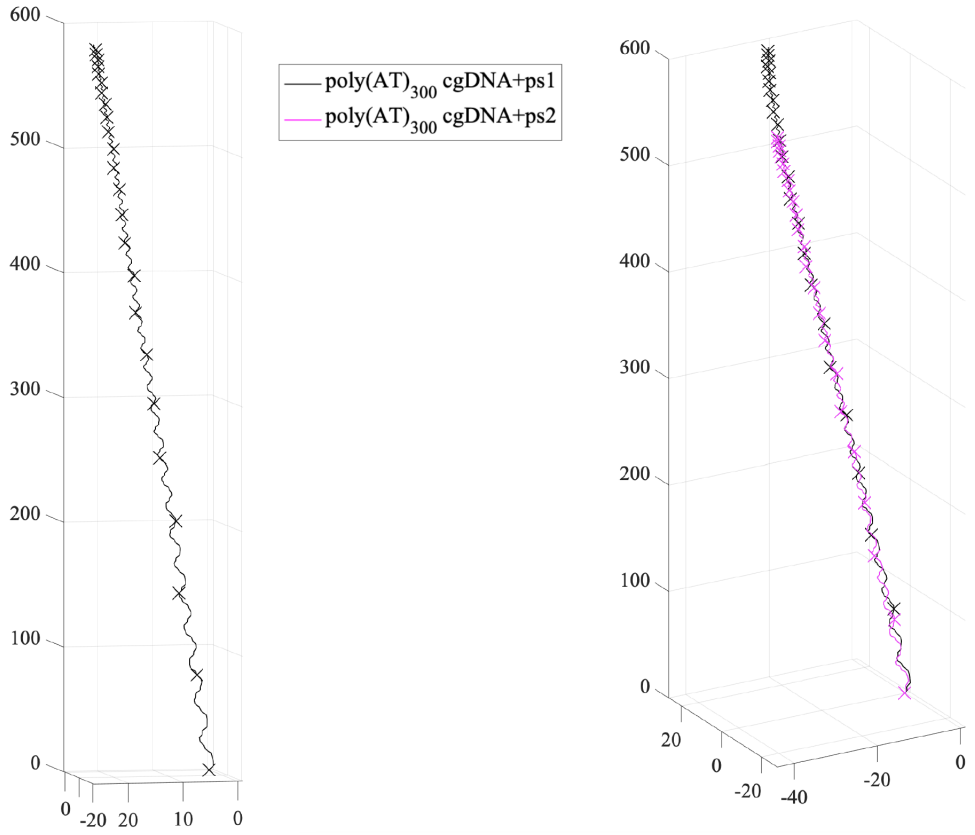

Figure 4: Left: The Flory persistence vector for poly$(AT)_{600}$ computed using 10^5 configurations sample with cgDNApmc. The crosses are plotted each 25 base-pairs. We can observe the accumulation point of crosses. The Flory persistence length is then the norm of the Flory persistence vector. Right: Comparison between the Flory persistence vector computed using cgDNA+ps1 (black) and cgDNA+ps2 (magenta).

```
Q = ( 1 - uu ) / ( 1 + uu ) * eye(3)   + ...
                2 / ( 1 + uu ) * u_cross  + ...
                2 / ( 1 + uu ) * (u*u')   ;
end
```

2. You will see that eigenvalues for each 10 cases will lie on the unit circle (see Qu2 Serie 2) and eigenvalues corresponding to averaged matrix will be within the unit circle.

3. You will obtain that the entries $(2,3)$ and $(1,3)$ tend to zero.

4. First you have to compute $\hat{u}$, the Cayley vector for given tilt, roll and twist angles. For 0° tilt, 0° roll and non zero twist angle Cayley vector $\hat{u}$ can be written as $\left(0\ 0\ \tan(\frac{\pi}{180} * \frac{\text{twist}}{2})\right)$. See Qu3 of Serie 6 for computing Cayley vector, actually scaling of Cayley vector doesn't matter for this question because there is no translation.

i) & ii) In Fig.(5) the variation in $\langle Q(u)\rangle_{(3,3)}$ with $K_{1,1}$ (left figure) and $K_{3,3}$ (right figure) have been shown. One can notice from left figure that as the value of K_{11} is decreasing the formula start showing the error. Also one can notice from right figure (see carefully the values on the Y-axis) that MC simulation matches to the formula.

Remark: One can also check that if $K_{11} << K_{22}$ then K_{22} will not play a role.

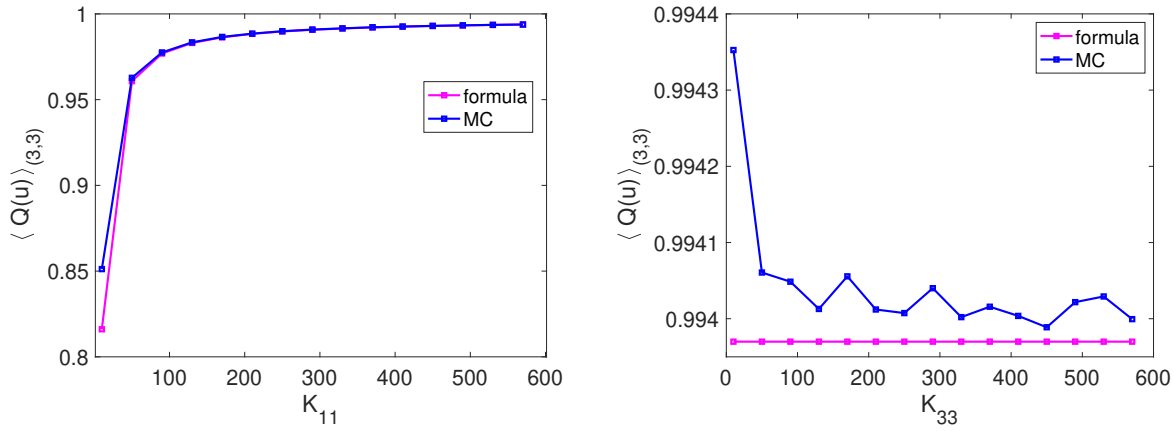

Figure 5: Comparing $\langle Q(u)\rangle_{(3,3)}$ computed with MC simulations and explicit formula for different values of K_{11} (left figure) and K_{33}(right figure). For MC simulations 10^5 samples have been used.

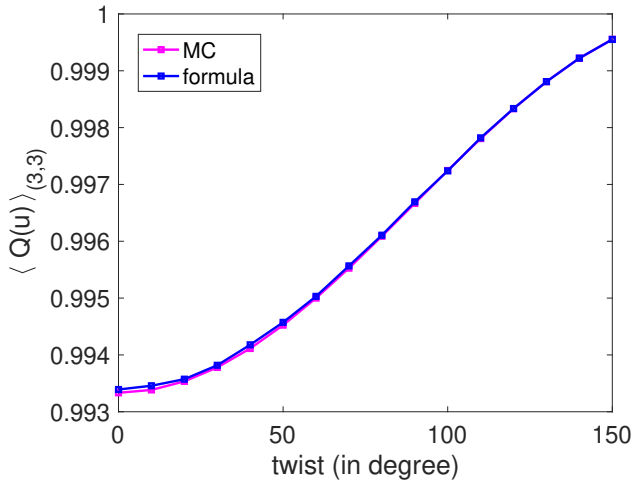

Figure 6: Comparing $\langle Q(u)\rangle_{(3,3)}$ computed with MC simulations and explicit formula for different values of twist. For MC simulations 10^5 samples have been used.

iii) In Fig.(6) we the variation in $\langle Q(u)\rangle_{(3,3)}$ with twist has been shown. One can observe that prediction through formula is in good aggrement with MC simulations (because value of entries in K is sufficiently large). One will find more deviation with formula for the cases in which entries of K_{11}, K_{22} or both is small (as it shown in solution of part (i) of this question).

3 Computing persistence length (using cgDNAmc) part-2

3.1 The tangent–tangent correlation

i) In Fig.(7) one can find the resulting curves for the five lambda sequences. One can observe that the curve for λ_2 has a different behaviour if compared to the other four. In fact starting from the base–pairs 140, more or less, the curve doesn't decrease exponentially and starts to vary a lot. This behaviour is all related to the fact that λ_2 is very bended, see for instance the groundstate in Fig. (5) in Corr 8.

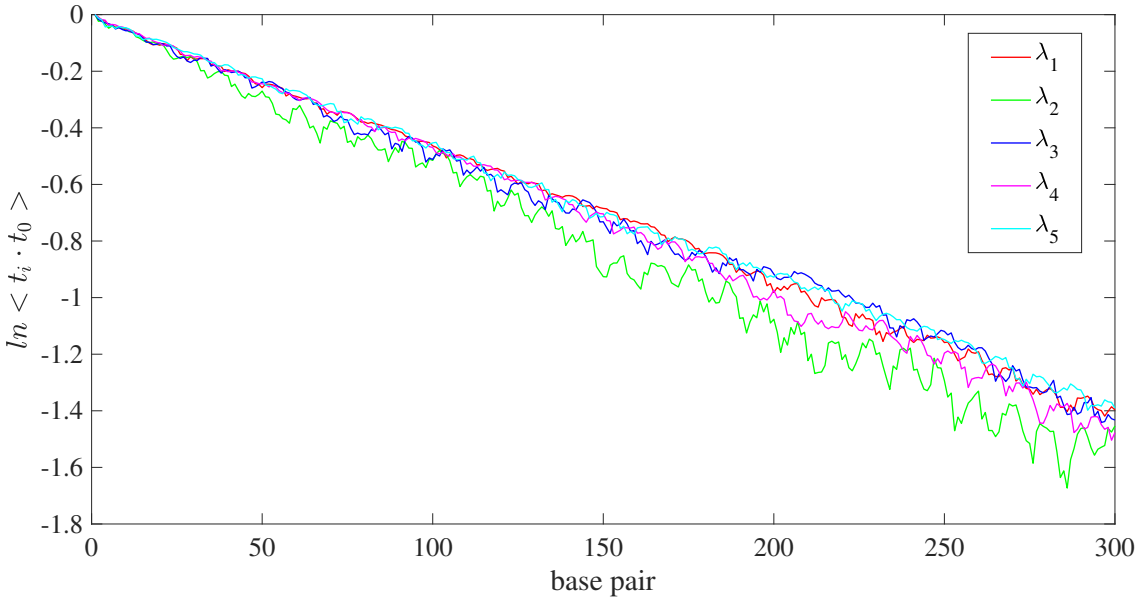

Figure 7: Tangent–tangent correlation for each λ_i as function of the number of base–pair. 10^5 configurations have been sampled for each sequence.

ii) With the method explained in the Qu1 of this Session, one can compute the persistence length for the five lambda sequences. Hereafter we report the computed values for the tangent–tangent correlation.

- λ_1: 212 bp,
- λ_2: 186 bp,
- λ_3: 213 bp,
- λ_4: 204 bp,
- λ_5: 216 bp.

Remark: Notice that it is not sensible to do least square fit to the plots corresponding to sequence λ_2 (green curve) & λ_4 (magenta curve) in Fig.(7) because these curves do not decrease exponentially and starts to vary a lot. And for these sequences notion of computing persistence length from ttc curve fails. Infact in reality all the λ sequences have very similar persistence length which you can see from above that $\lambda_{1,3,5}$ have similar number and $\lambda_{2,4}$ do not (that is because the notion of computing persistence length for bended sequences fails and it computes wrong persistence length). However one can use factorized tangent–tangent correlation for these cases which is the topic of next question.

iii) The factorisation decompose the tangent–tangent correlation into two terms, a deterministic one depending on the groundstate, and a stochastic one. In Fig.(8) we show the result of this factorization on the five lambda sequences while in Fig.(9) we plot the factorized tangent–tangent correlation for the six distinct poly-dinucleotide considered in the Qu 1 of this Session.

Remark: While comparing Fig.(7) with Fig.(8) one can observe that curves in Fig.(8) seems to decrease exponentially and do not vary a lot for all the λ sequences. Also, for the six distinct poly-dinucleotide one can notice that wiggling in the curves in Fig.(9) is gone which was there in Fig. (2).

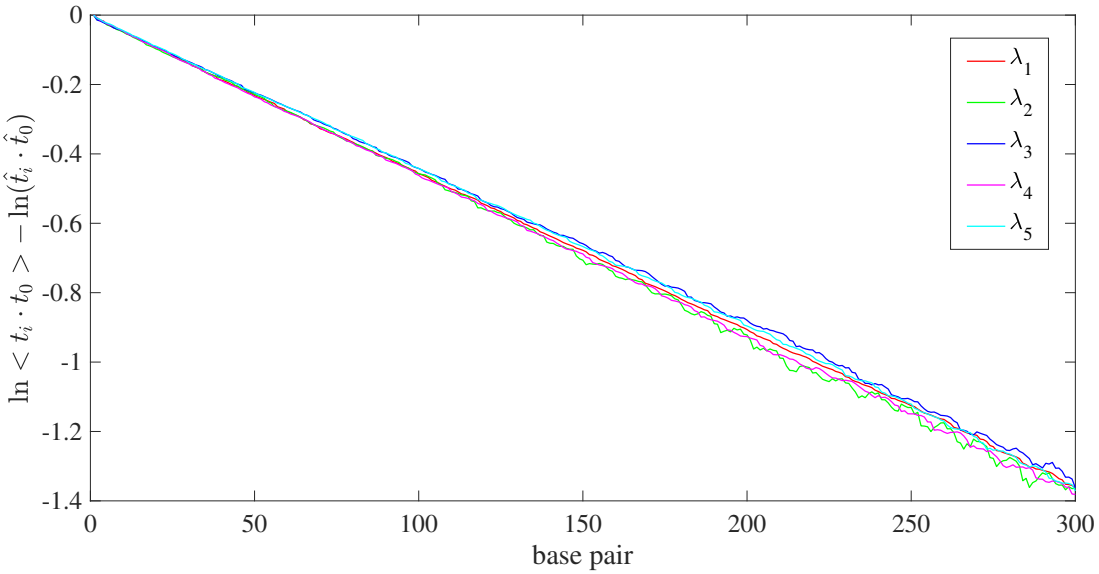

Figure 8: Factorised tangent–tangent correlation for each λ_i as function of the number of base–pair. 10^5 configurations have been sampled for each sequence.

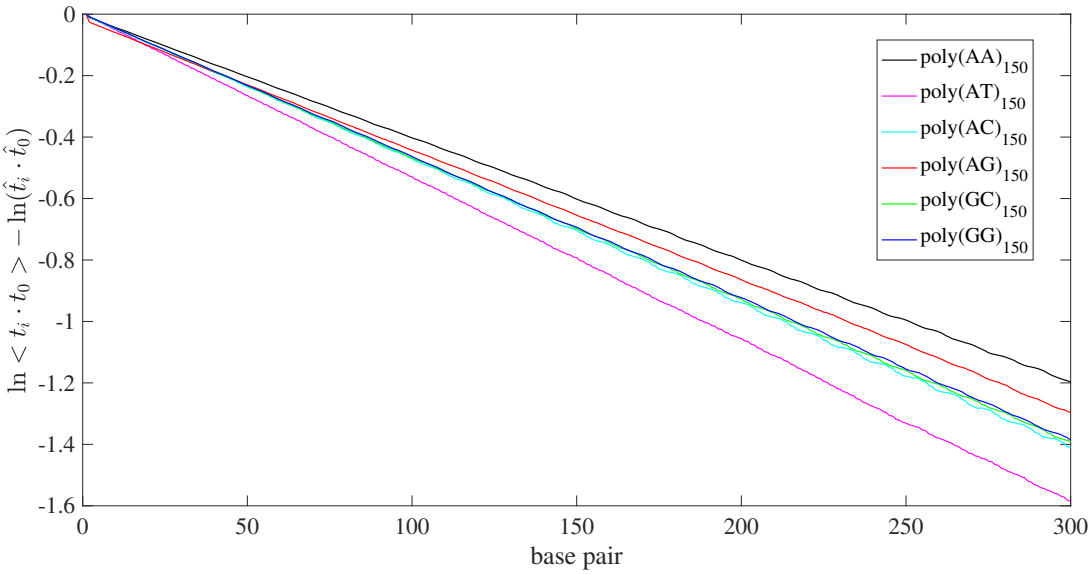

Figure 9: Factorised tangent–tangent correlation for each poly$(\alpha\beta)_{150}$ as function of the number of base–pair. 10^5 configurations have been sampled for each sequence.

3.2 The Flory vector

In the left side in Fig. (10) we have plotted all the five Flory persistence vectors computed with cgDNApmc for each lambda sequences. The crosses have been plotted on all the five Flory persistence vectors after every 25 base-pairs and one can notice that they are converging in the sense that the Euclidean distance between two consecutive crosses is decreasing, but not yet converged. However, in the right side in Fig. (10) we have shown the groundstate along with the Flory persistence vectors (same as in the left figure) for each lambda sequences.

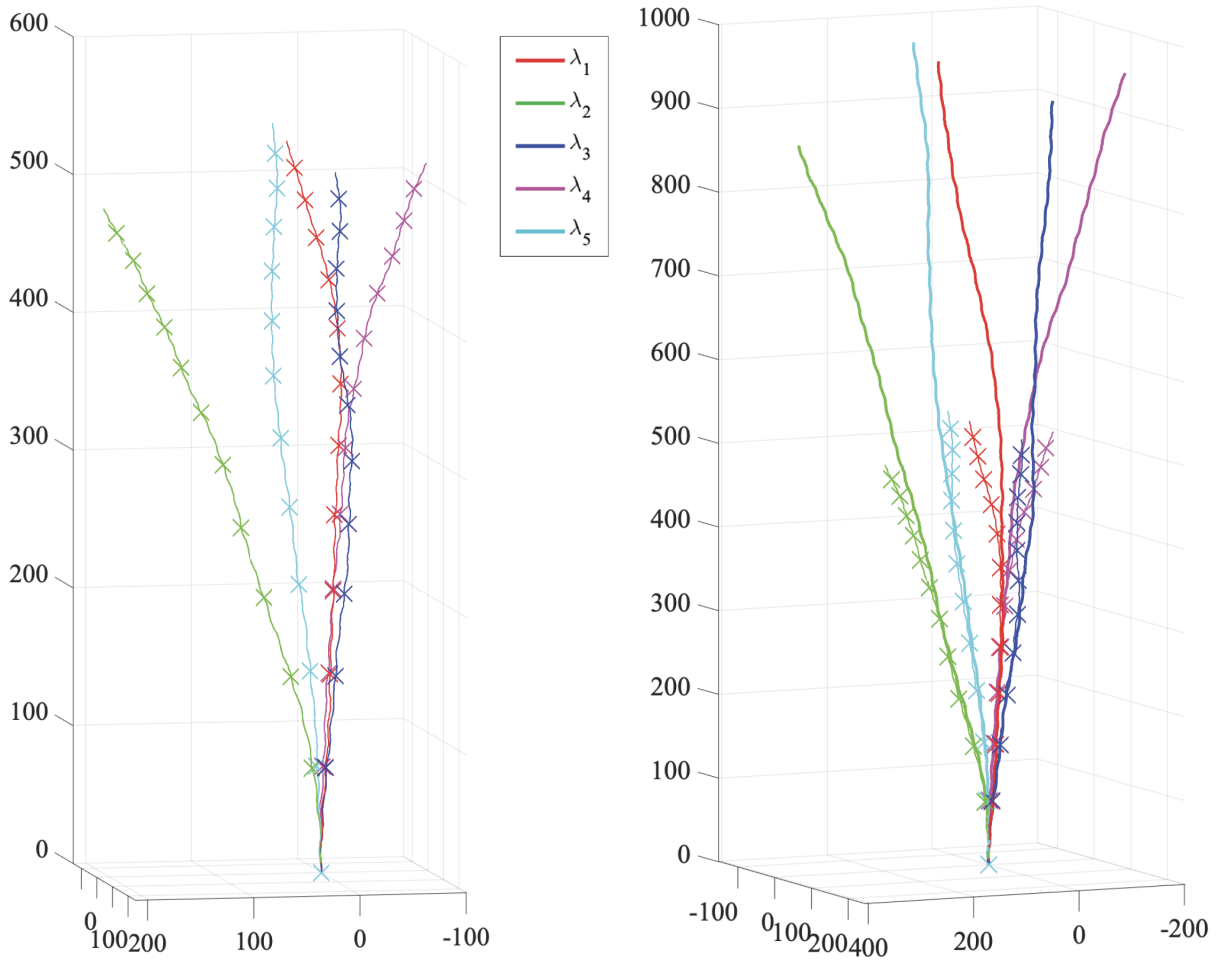

Figure 10: The Flory persistence vectors (left figure) for all the five lambda sequences computed with cgDNApmc and 10^5 configurations have been sampled. The crosses in the Flory persistence vectors have been plotted after each 25 base-pairs. However, in the right side we have shown the groundstate along with the Flory persistence vectors (same as in the left figure) for each lambda sequences.