1 Monte Carlo simulation with the cgDNA model

**Remark:** We let available a few MATLAB scripts that can be helpful for this session and for the session 7. You can download the scripts at [http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/Useful_function_Session7](http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/Useful_function_Session7). Moreover you can download the modified version of run_cgDNAmc.cpp at [http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/run_cgDNAmc_mod.cpp](http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/run_cgDNAmc_mod.cpp).

1.1 cgDNA reconstruction of the 6 distinct poly–dinucleotide

i) Here we compare only the poly–dinucleotide AT with 300 base-pairs and we focus only on some entries in the middle of the shape vectors reconstructed with cgDNAparamset 1 and cgDNAparamset 2. In Fig. (1) one can observe that the two vectors are close. We recall that the entries of the ground state vector represent the intra and inter relative rotations and orientation and that small changes in the relative coordinates could lead to important differences in the 3D reconstruction. However we invite to check that the 3D reconstructions, using cgDNAparamset 1 and cgDNAparamset 2, of the six distinct poly–dinucleotides are relatively close.

![Figure 1: Comparison of some on the coordinates of the ground state vector of poly(AT)150 reconstructed using cgDNAparamset1(solid red) and cgDNAparamset2 (dashed blue).](image)

We stress also on the fact that the six poly–dinucleotides have relevant differences in the shape. For example we have reconstructed the six oligomers using the cgDNAparamset2 and we have selected 18 entries corresponding to (intra,inter,intra), see Fig. (2). Is interesting to notice that even if the six poly–dinucleotides have considerable differences in the entries of the shape vector, the 3D reconstructions are all super-helical.
Figure 2: Comparison of eighteen entries (intra, inter, intra) of the ground states of the six poly(\(\alpha\beta\)).

ii) As briefly said in the previous part, the six ground states are different but all six present a super-helical structure. In Fig. 3 for each poly–dinucleotide we plot the xyz position of each base-pair. The obtained helices have clearly a different radius and a different pitch. These two features will become important for the tangent–tangent correlation, see part 1.4 of this exercise.

Figure 3: Left: xyz position of each base–pair for all the six poly-dinucleotides. Right: Zoom to check that each obtained helices have a different radius.

1.2 Monte Carlo simulation of the cgDNA model with MATLAB

Let \( \mathcal{P} \) be a cgDNAparamset and \( S \) a \( n \) base-pair long sequence. The cgDNA package reconstructs then the ground-state \( \hat{\mathbf{x}} \equiv \hat{\mathbf{x}}(S, \mathcal{P}) \in \mathbb{R}^{12n-6} \) and the stiffness \( K \equiv K(S, \mathcal{P}) \in \mathbb{R}^{12n-6 \times 12n-6} \), thus it
predicts the probability density function

\[
\rho(x; S, P) = \frac{1}{Z} \exp \left\{ \frac{1}{2} (x - \hat{x}) \cdot K(x - \hat{x}) \right\}.
\]

Using the latter probability density function we can explore the set of possible configurations for the given sequence \(S\) and parameter set \(P\). The Monte Carlo method, of the cgDNA model, consist in evaluating a deterministic function on an ensemble of configurations \(\{x_i\}_{i=1}^M\) sampled from \(\rho(x; S, P)\). For sampling from a cgDNA probability density function we can take advantage from the sparsity pattern of the cgDNA stiffness matrix \(K\). In fact one can observe that the Cholesky factorisation of a block diagonal matrix preserve the block structure, i.e, let \(K = L^T L\) the Cholesky factorisation of the cgDNA stiffness matrix \(K\), where \(L\) is a upper triangular sparse matrix which sparsity pattern correspond to the half, with respect to the main diagonal, of the one of \(K\). Please check by your self the latter statement. We can use now the Cholesky factorisation to perform the following change of variable:

\[
\rho(x; S, P) = \frac{1}{Z} \exp \left\{ \frac{1}{2} (x - \hat{x}) \cdot K(x - \hat{x}) \right\}
\]

\[
= \frac{1}{Z} \exp \left\{ \frac{1}{2} (x - \hat{x}) \cdot L^T L(x - \hat{x}) \right\}
\]

\[
= \frac{1}{Z} \exp \left\{ \frac{1}{2} L(x - \hat{x}) \cdot L(x - \hat{x}) \right\},
\]

define \(y = L(x - \hat{x}) \in \mathbb{R}^{12n-6}\), to get the following distribution

\[
\rho(y; S, P) = \frac{1}{Z} \exp \left\{ \frac{1}{2} y \cdot y \right\} = \frac{1}{Z} \prod_{i=1}^{12n-6} \exp \left\{ \frac{1}{2} y_i^2 \right\} \tag{1}
\]

Finally we can sample \(y\) component wise from an unidimensional standard normal distribution \(\mathcal{N}(0,1)\) and recompute the configuration \(x\) by solving the sparse system

\[
x = L^{-1} y + \hat{x}. \tag{2}
\]

i) In Fig. (4) we plot only 250 configuration sample from the cgDNA distribution for poly\((AT)_{150}\) using the MATLAB function \texttt{mvnrnd}. We stress that even if the ground state is remarkably straight, few sampled configurations are incredibly bended.

![3D reconstruction of 250 sampled configurations of poly\((AT)_{150}\) (blue) and the ground state (red).](image)

ii) Just by plotting the last base-pair position of the 250 sample configuration of Fig. (4) we can observe that the cloud of points seems define a half sphere like shape around the ground state. It is now interesting to ask our self if this cloud of point is "converged" in the sense that the
real distribution of this point is uniform on a half sphere which ray, roughly speaking, is the ground state. For doing that we need to sample more configuration and to reconstruct the last base–pairs position, thus we need to use the cgDNAmc code because is much faster then using MATLAB.

![Cloud of points (blue) defined by the last base–pair position of 250 sampled configurations](image)

**Figure 5:** Cloud of points (blue) defined by the last base–pair position of 250 sampled configurations

### 1.3 The cgDNAmc code

**Modified run_cgDNAmc**

By a minor modification of the script run_cgDNAmc one can get all the last base–pairs positions of all the sampled configurations. Using cgDNAmc one can draw $10^5$ configurations of a 300mers in a couple of minutes. In Fig. (6) one can observe that the cloud of points of last base–pair positions form almost a sphere which "ray" is the ground state. The isotropic behaviour of the poly($AT$)$_{150}$ is due to his high twisted structure and his almost straight ground state. The two side views, right column in Fig. (6), show a more concentrated region of point which in this two dimensional view seems like a banana shaped region.

This banana shape region is more visible when considering less repetitions of the dinucleotide $AT$, for example in Fig. (7) we show the cloud of $10^5$ points obtained for poly($AT$)$_{50}$ and the over all shape looks more like an umbrella, this is due to the fact that poly($AT$)$_{50}$ is more rigid than poly($AT$)$_{150}$. In the two side views in Fig. (7) the banana shape regions is more evident.
1.4 Compute the persistence length using the cgDNAmc code

**Convergence study of the Monte Carlo simulation**

In this part we present briefly the result of the convergence test for the tangent–tangent correlation (ttc) that have been done in "Sequence-dependent persistence lengths of DNA", J. S. Mitchell et al., JCTC, 2016, on a sequence, which is a part of the lambda phage genome, called λ₃ (see Session 8 for more details about the lambda phage genome). Ten independent Monte Carlo runs have been done for three different total numbers of wanted configurations: $10^4$, $10^5$, and $10^6$. The results are reported in Fig. 8. The error bars are plotted every 5 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.
The tangent–tangent correlation

i) Using the cgDNAmc code we obtained the following figure, Fig. (9), 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. (3). This relation will become more clear in the Exercise Session 11.

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 $\ln(t_i / t_0)$. For the six poly–dinucleotide we have computed the following values:

- $\text{poly}(AA)_{150}$: 218 bp,
- poly\((AT)_{150}\): 146 bp,
- poly\((AC)_{150}\): 171 bp,
- poly\((AG)_{150}\): 193 bp,
- poly\((GC)_{150}\): 166 bp,
- poly\((GG)_{150}\): 175 bp.

**The Flory vector**

i) In Fig.(10) we plotted all the six Flory persistence vector computed with cgDNAmc 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 to consecutive crosses is decreasing. For simple model of polymer chain, as the Wormlike chain model or the Freely rotating chain model or the random-\(\phi\) 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.

![Figure 10: The Flory persistence vector computed using \(10^5\) configurations sampled with cgDNAmc. The crosses are plotted each 25 base-pairs.](image)

ii)&iii) For sake of brevity, here we show just the result for poly\((AT)_{600}\). In Fig. [11] 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 cgDNAparamset1 and cgDNAparamset2. The most important difference between the parameter sets is the prediction of the persistence length, in fact cgDNAparamset1 predict an higher persistence length for the DNA then the cgDNAparamset2 that predict a persistence length for the DNA that is closer to the accepted value. For that reason the cgDNAparamset2 is better then the cgDNAparamset1.

Figure 11: Left: The Flory persistence vector for poly(\textit{AT})_{600} computed using $10^5$ configurations sample with cgDNAmc. 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 cgDNAparamset1, black, and cgDNAparamset2, purple