1 Relative entropy for Gaussians II

1. Let $M \in \mathbb{R}^{N\times N}$ and X_1 and X_2 the normal random variables with respectively probability density functions p and q. Define the change of variable $Y_i = MX_i$, $i = 1,2$.

i) A linear transformation of a normally distributed random variable is another normally distributed random variable. Using the property of the expected value one can show that

$$\begin{aligned} E\left[MX_i\right] &= ME\left[X_i\right] = M\widehat{x}_i \\ E\left[(MX_i - E\left[MX_i\right])(MX_i - E\left[MX_i\right])^T\right] &= ME\left[(X_i - \widehat{x}_i)(X_i - \widehat{x}_i)^T\right]M^T = MK_i^{-1}M^T. \end{aligned}$$

for $i = 1,2$.

ii) A direct computation gives:

$$\begin{aligned} D(\tilde{p},\tilde{q}) &= \frac{1}{2}\left[\operatorname{tr}(M^{-T}K_2M^{-1}MK_1^{-1}M^T) - \ln\frac{|M^{-T}K_2M^{-1}|}{|M^{-T}K_1M^{-1}|}\right] \\ &+ \frac{1}{2}M(\widehat{x}_1 - \widehat{x}_2)\cdot M^{-T}K_2M^{-1}M(\widehat{x}_1 - \widehat{x}_2) \\ &= \frac{1}{2}\left[\operatorname{tr}(M^{-T}K_2K_1^{-1}M^T) - \ln\frac{|K_2|}{|K_1|}\right] \\ &+ \frac{1}{2}(\widehat{x}_1 - \widehat{x}_2)\cdot M^{-T}M^{-T}K_2M^{-1}M(\widehat{x}_1 - \widehat{x}_2) \\ &= \frac{1}{2}\left[\operatorname{tr}(K_2K_1^{-1}M^TM^{-T}) - \ln\frac{|K_2|}{|K_1|}\right] + \frac{1}{2}(\widehat{x}_1 - \widehat{x}_2)\cdot K_2(\widehat{x}_1 - \widehat{x}_2) \\ &= \frac{1}{2}\left[\operatorname{tr}(K_2K_1^{-1}) - \ln\frac{|K_2|}{|K_1|}\right] + \frac{1}{2}(\widehat{x}_1 - \widehat{x}_2)\cdot K_2(\widehat{x}_1 - \widehat{x}_2) \\ &= D(p,q). \end{aligned}$$

2. The generalised eigenvalue problem can be rewritten as

$$K_2v_i = \mu_iK_1v_i \Rightarrow K_1^{-1}K_2v_i = \mu_iv_i \Rightarrow \operatorname{tr}(K_2K_1^{-1}) = \sum_{i=1}^N \mu_i\ ,\ \ln|K_2K_1^{-1}| = \sum_{i=1}^N \ln\mu_i \tag{1}$$

We can also rewrite the definition of $D^\dagger$ as

$$D^\dagger(K_1,K_2) = \frac{1}{2}\left[\operatorname{tr}\left(K_2K_1^{-1}\right) - \ln|K_2K_1^{-1}| - N\right]. \tag{2}$$

By combining the consequences of the generalized eigenvalues problem and the latter formula we obtain:

$$D^\dagger(K_1,K_2) = \frac{1}{2}\sum_{i=1}^N(\mu_i - \ln\mu_i - 1). \tag{3}$$

3. i) By using the fact that $\ln\frac{|K_2|}{|K_1|} = \ln|K_2| - \ln|K_1|$, we get that the symmetrized version of the stiffness part of the relative entropy between two Gaussians is :

$$D^\dagger_{sym} = \frac{1}{4}\operatorname{tr}\left(K_2K_1^{-1} + K_1K_2^{-1}\right) - \frac{N}{2}. \tag{4}$$

ii) By using the same argument used in part 2 of this exercise we find that

$$D^{\dagger}_{sym} = \frac{1}{4}\sum_{i=1}^{N}\left(\sqrt{\mu_i} - \frac{1}{\sqrt{\mu_i}}\right)^2. \tag{5}$$

2 Jensen's inequality - Part 2

1. In Figure 1 we show the graph of $\phi(x) = x\ln(x)$.

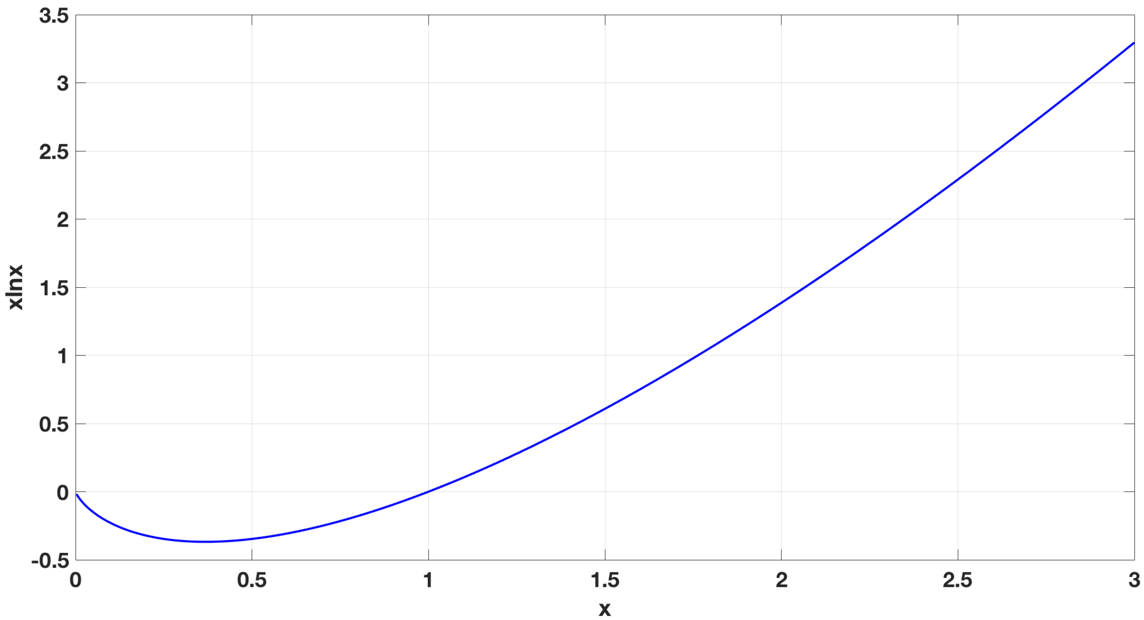

Figure 1: Graph of $\phi(x) = x\ln(x)$.

We prove now the hint. Consider $\phi(x) = x\ln(ax)$, for $a \in \mathbb{R}$. We have that

$$\phi'(x) = \ln(ax) + 1$$
$$\phi''(x) = \frac{1}{x}.$$

Let now $f(x) = \ln(ax)$ and $g(x) = \frac{1}{x}$, we have that $\lim_{x\to 0}|f(x)| = \lim_{x\to 0}|g(x)| = +\infty$. Therefore we can use L'Hopital's rule for computing the limit $\lim_{x\to 0}\phi(x)$:

$$\lim_{x\to 0}\frac{f'(x)}{g(x')} = \lim_{x\to 0}\frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x\to 0} -x = 0.$$

2. Let $\phi(p) = p\ln(|\Omega|p)$, for any probability measure p. By using the Jensen's inequality (see statement of the exercise).

$$\int_\Omega \phi(p)d\mu(x) = \int_\Omega p\ln(|\Omega|p)d\mu(x) \geq |\Omega|\phi\left(\frac{1}{|\Omega|}\int_\Omega pd\mu(x)\right) = |\Omega|\phi\left(\frac{1}{|\Omega|}\right) = 0,$$

where we used the fact that $\int_\Omega pd\mu(x) = 1$.

3. We have that

$$\int_\Omega p\ln(p)d\mu(x) \geq |\Omega|\phi\left(\frac{1}{|\Omega|}\int_\Omega pd\mu(x)\right) = \left(\int_\Omega d\mu(x)\right)\phi\left(\frac{1}{|\Omega|}\right) = \int_\Omega \frac{1}{|\Omega|}\ln\left(\frac{1}{|\Omega|}\right)d\mu(x) \quad (6)$$

Finally we can conclude that the distribution that minimizes the entropy is the uniform distribution $c(x) = \frac{1}{|\Omega|}$.

4. Let p and q two probability density functions. Let us define the following integral

$$D(p,q) = \int_\Omega \frac{p}{q}\ln\left(\frac{p}{q}\right)qd\mu(x).$$

By using the Jensen's inequality we obtain that

$$\begin{aligned} D(p,q) &\geq \phi\left(\int_\Omega \frac{p}{q}qd\mu(x)\right) = \phi\left(\int_\Omega pd\mu(x)\right) \\ &\geq \phi(1) = 0. \end{aligned}$$

3 Estimate of mean and stiffness from MD simulation data

1. One can observe that the raw stiffness matrix has the most of the non zero entries near the diagonal, in fact by using `plot2Dmatrix` one can observe that the most of the non zero entries are in the stencil. On the contrary the raw covariance matrix is dense and do not show any specific pattern around the diagonal.

2. In general, in order to get the right scaling you should multiply the rotation-rotation blocks by 25, the rotation-translation block by 5 and the translation-translation by 1.

4 Palindromic symmetry of a shape vector and stiffness matrix

1. The shape vector and covariance matrix do not satisfy palindromic symmetry conditions: the biggest difference between the shape elements is 0.1508 and the biggest difference between the elements of covariance matrices is 0.8124.

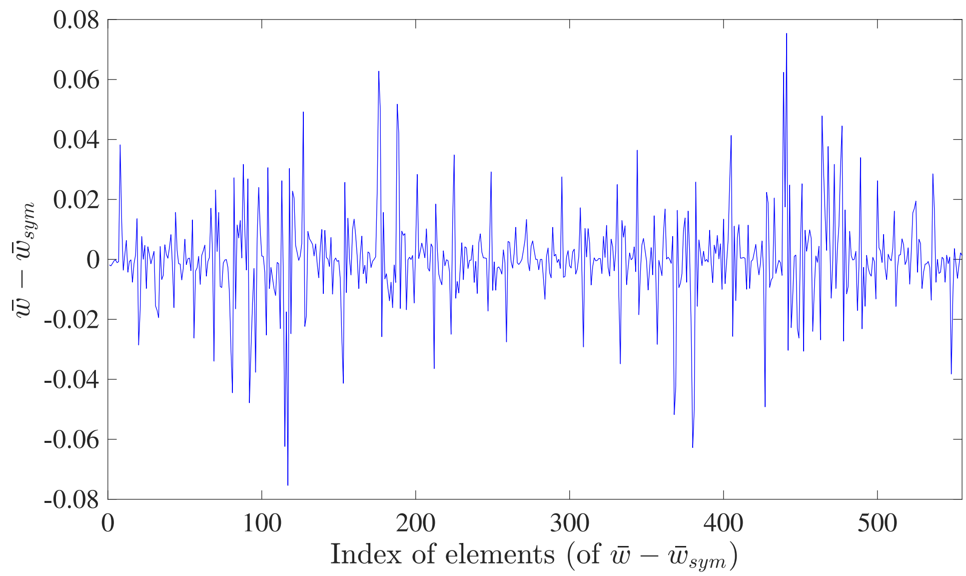

Figure 2: Plot of $\bar{\mathbf{w}} - \bar{\mathbf{w}}_{sym}$.

2. By looking at the plot in Figure 2 we can see that most of differences between symmetrized and observed shape elements are actually quite small, under 0.04 in absolute value.

3. First compute $D = C + \bar{\mathbf{w}} \otimes \bar{\mathbf{w}}$ and then symmetrize it.

4. Use the script `computeMaxEntropy.m` to get the maximum entropy fit to the symmetrized covariance computed in the previous point. You can see the matlab script `main.m` in which the script `computeMaxEntropy.m` has been used.

5. In the following table we reported the result of the computations:

	KLD	stiffness part	mean part
$D(\rho_{obs}^{sym}(S), \rho_{obs}(S))$	0.0039	0.0038	0.0001
$D(\rho_{band}^{sym}(S), \rho_{obs}^{sym}(S))$	0.0065	0.0065	0
$D(\rho_{cgDNAp}(S), \rho_{band}^{sym}(S))$	0.0223	0.0209	0.0014

5 From atomistic representation to cgDNA internal coordinates

Download the solution file here (`http://lcvmwww.epfl.ch/teaching/modelling_dna/protected_files/codes_exercises/frames_embedding_soln.zip`) for this exercise which contains all the useful matlab scripts required for the solutions of different parts of this exercise.

1. Load pdb file as `molviewer('Palin_141_3.1.ions.pdb')` in matlab and visualise the variation between different snapshots (you can visualise individual snapshots or all 6 snapshots at a time). In pdb file we have coordinates for all the atoms and these coordinates are with respect to the simulation cell origin, which is point (0,0,0). In Figure 3 we showed the atomistic view in x-y plane with cartoon of backbone for 2 snapshots(1st and 6th) and one can notice the variation in the atomic coordinates between two snapshots. In Figure 4 we showed the cartoon view of all 6 snapshots in x-y plane and one can observe that all 6 backbones are distinct due to the fluctuation in atomic coordinates.

2. In order to understand the frame fitting, first recall from Qu3 corr 5 that for given vectors $\boldsymbol{\alpha}_i \in \mathbb{R}^3$ & $p_i \in \mathbb{R}^3$. The values of $\mathbf{r}$ & R are the best fit frame to a given set of atomic coordinates of atoms $\mathbf{p}_i$, $i = 1, \dots, M$, whose idealized coordinates in the frame $(\mathbf{r}, R)$ are known to be $\boldsymbol{\alpha}_i$, $i = 1, \cdots, M$. In the context of this exercise we have data for atomic coordinates (coming from MD snapshots) of atoms in a base in file `PDBAtoms_p24.mat`. Basically we have $\mathbf{p}_i$, $i = 1, \cdots, M$ and M is the number of atoms in a base which is different in different bases (A, T, C, G), for example, A has 11, T has 10, C has 9 and G has 12 atoms in the bases (you should check this in the data file `PDBAtoms_p24.mat`). And we also have idealized coordinates of all the atoms in each base and for all the bases (i.e. we have $\boldsymbol{\alpha}_i$, $i = 1, \cdots, M$ for all the bases (A, T, C, G)) in the data file `Ideal_Bases.mat`. Just to conclude, now we have $\mathbf{p}_i$ (for all the base in reading and complimentary strand of given sequence) and idealized coordinates $\boldsymbol{\alpha}_i$ (for A, T, C, G). Notice that the atomic coordinates (coming from a MD snapshot) of atoms in base will be different in each base in the strand (for example, we have given sequence $GCCCTTGGCGATATCGCCAAGGGC$, in this sequence we have A at many places and atomic coordinates of atoms in each A will be different, similarly for C, G and T). However, idealise coordinates of each atoms in each base are unique.

   Now from Qu5 corr 5, we have explicit expression for $\mathbf{r} \in \mathbb{R}^3$ and $R \in SO(3)$, which represent the the the best fit frame to a given set of atomic coordinates $(\mathbf{p}_i)$ and idealized coordinates

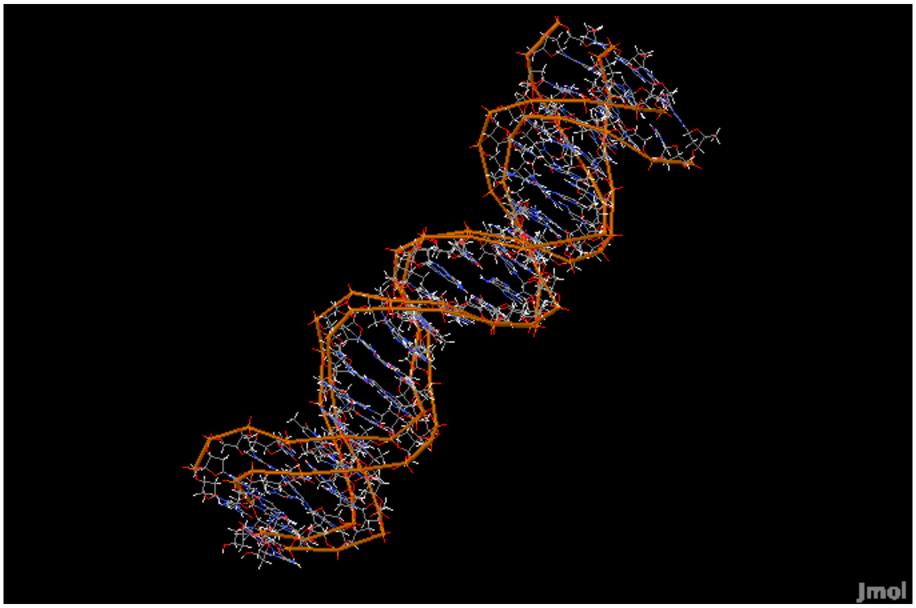

Figure 3: Cartoon of backbone and base atoms for two snapshots (1st and 6th) having 10 nanoseconds of time difference.

($\boldsymbol{\alpha}_i$) for atoms ($i = 1, \ldots, M$). The expression for $\mathbf{r}$ is

$$\mathbf{r} = \frac{1}{M}\sum_i(\mathbf{p}_i - R\boldsymbol{\alpha}_i) = \frac{1}{M}\sum_i \mathbf{p}_i - \frac{R}{M}\sum_i \boldsymbol{\alpha}_i \tag{7}$$

and R is the polar factor (see notes on matrix factorisation) of the following matrix

$$\Big[(\sum_i \mathbf{p}_i) \otimes (\sum_i \boldsymbol{\alpha}_i) - M \sum \mathbf{p}_i \otimes \boldsymbol{\alpha}_i\Big]. \tag{8}$$

Above expressions have been coded in matlab function `rigidmotion_using_leastsquare.m` which computes the frame from set of atomic coordinates and then this function is used in the scrip `frames_fitting.m` to compute frames to all the bases in a given sequence. One can directly run this script to compute the solution for this part of the exercise. Then use cgDNAviewer to visualise these frames. In Figure5 we have shown the frames embedded in each bases of reading and complimentary strand of the given sequence for two snapshots (1st and 6th). In order to plot Figure5, we first aligned all snapshots around the reference frame (we choose frame of Watson base in first basepair as a reference frame) and transform all the frames in a sequence using rigid body transformation. Then use cgDNAviewer (see Qu2 Serie 6) to visualize the transformed frames along the sequence. You can choose any base frame in the sequence as a reference frame and then do the rigid body transformation of all the frames of all the snapshots with respect to the chosen reference frame. In Figure5 one can observe the noise between frames of two different snapshots. Also, one can notice that molecule for one snapshot is a bit bended (due to the noise in snapshot) in comparison to other snapshot. Do not confuse between Figure 3 and Figure 5 since both represents snapshot1 and snapshot6. In Figure3, just the atomistic coordinates (in x-y plane) have been shown. However, in Figure5, frames fitted to the base atoms and then after doing alignment of snapshots with respect to chosen reference point have been shown, which is not the case with Figure3 because atomistic coordinates (in file `Palin_141_3.1.ions.pdb`) are not aligned in same sense as in Figure5. So, Figure3 and Figure5 are not directly comparable.

[ht]

Figure 4: Visualising 6 consecutive snapshots (with 2 nanosecond of time difference) with cartoon of backbone.

3. From the computation done in previous part of this exercise now we have found the the best fit frames $(\mathbf{r}, R)$ for each base of the given sequence. The fitting error for each base will then simply be the value of following expression (value of the function which we are minimising in order to get optimal frames) computed for each bases:

$$\sum_{i=1}^{M} \|\mathbf{r} + R\boldsymbol{\alpha}_i - \mathbf{p}_i\|^2, \tag{9}$$

here, M represent the number of atoms in base, $\mathbf{p}_i$ are atomic coordinates of the base atoms on which frames $(\mathbf{r}, R)$ have been fitted and α_i are the idealised coordinates for atoms of that base. In the Matlab script `frames_fitting.m` you will find the code for computing the above expression for each bases.

4. First recall about relative coordinates between pairs of rigid bodies from lecture 5. Basically the relative coordinates (u, v) between two rigid body frames $g1 = (R1, r1)$ to $g2 = (R2, r2)$ can be computed using the explicit formulas. The rotational coordinate u (Cayley vector) is $Cay(Q)$, where $Q = {R_1}^T R_2$ and translational coordinate $v = (\sqrt{Q})^T q$, where $q = {R_1}^T (r_2 - r_1)$. To get u and v from given g_1 and g_2 as an input, is coded in matlab script `getCoord.m`. Now compute the mid frame and junction frame (see Qu 2 corr 4 and Qu 4 corr 6) using the frames found in previous part (3) of this exercise. And then, think of two rigid bodies g_1 and g_2 as two consecutive mid frames in order to get intra variable of cgDNA model. Further consider g_1 and g_2 as two consecutive junction frames which will give you inter variable of cgDNA model. These steps are coded in matlab script `cgDNAcoord.m` and you can directly use this script to compute cgDNA model coordinates.

Finally, in Figure 6 we have shown the plot of each intra and inter coordinates along the molecule for all 6 snapshots. In Figure 6, notice that internal coordinates are in dimensional form. One can also observe the fluctuations between different snapshots and notice that range for same kind of inters and intras are similar for all 6 snapshots but the qualitative values at basepair level are different. For example, values of buckle (also for shear) at the 15th basepair for all 6 snapshots are different.

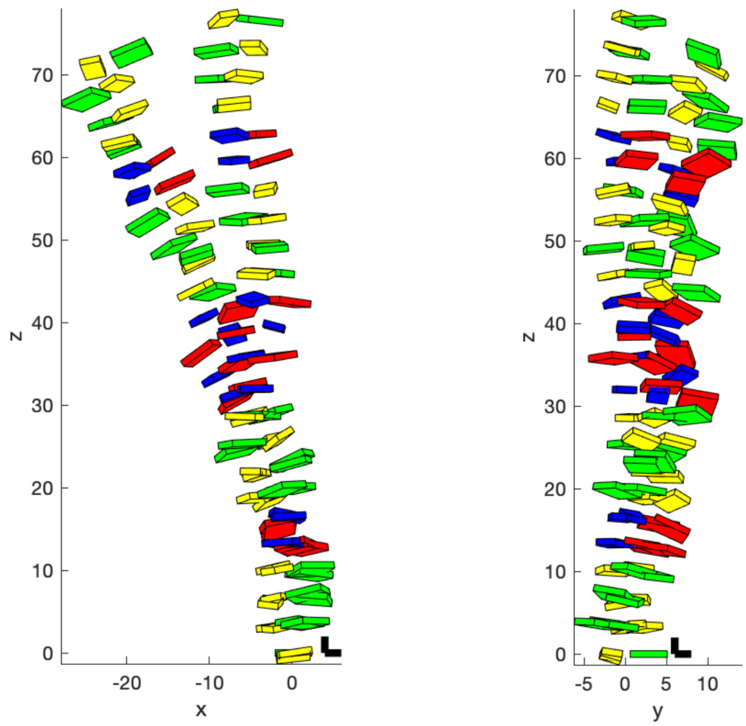

Figure 5: Visualising noise in frames betweem two (1st and 6th) snapshots. Left figure represent the x-z view and right figure represent the y-z view. Color scheme : yellow - C , green - G , blue - T , red - A.

5. In the following table we have shown the value of $\|w_i - E_N w_i\|$ for three individual snapshots. One can notice that individual snapshots do not follow palindromic symmetry since value of $\|w_i - E_N w_i\|$ is not zero. Check with many different snapshots and you should always find non zero values of $\|w_i - E_N w_i\|$ for any i. You can use matlab script `Crick_Watson_sym.m` (provided in the solution files) for the computation of this part of the exercise.

i	$\|w_i - E_N w_i\|$
1	13.33
2	12.83
5	14.03

In table below we have shown the value of $\|\bar{w} - E_N \bar{w}\|$, where $\bar{w}$ is computed by averaging the w_i over four different M.

M	$\|\bar{w} - E_N \bar{w}\|$
25	8.33
50	7.67
100	7.53
200	7.26

One can observe that as the value of M is increasing then corresponding value of $\|\bar{w} - E_N \bar{w}\|$ is decreasing, which means that when M will increase sufficiently then associated Crick-Watson symmetry relation will satisfy. In reality, if M approaches to infinity then $\|\bar{w} - E_N \bar{w}\|$ will

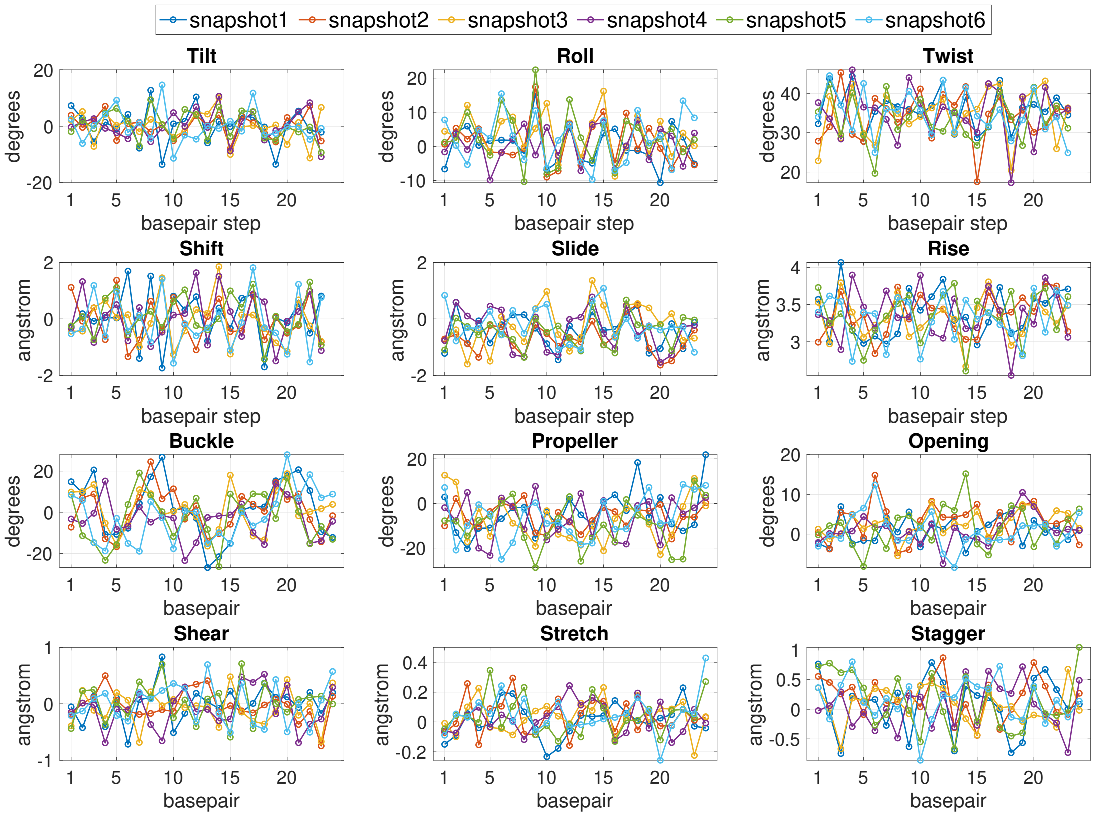

Figure 6: Plot of cgDNA model coordinates (intra and inter) along the molecule for 6 snapshots. Here, intras and inters are plotted in a sense of standard helical parameters (see Figure (1) Serie 5 for the definition of helical parameters). The rows 1 and 2 are the components of the inters while rows 3 and 4 are components of the intras. Rows 1 and 3 are the rotation parts while rows 2 and 4 are the translation parts.

approach to zero which means that the equilibrium statistics will satisfy an associated Crick-Watson symmetry relation. In cgDNA model parameter set estimation process range of M is in the order of millions.