For the exercises 3 and 4 download the dataset `http://lcvmwww.epfl.ch/teaching/modelling_dna/protected_files/codes_exercises/palin_md_data.mat`. The dataset consist in an oligomer-based statistics of a 24 basepairs long Palindrome. The matlab structure contains the following fields:

- `seq` : sequence,
- `nbp` : number of base pair,
- `nsnap` : total number of accepted snapshots from the MD simulation (= M).
- `shape` : ensemble mean ($\bar{\mathbf{w}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]}$),
- `c1b` : ensemble covariance, ($C = \frac{1}{M}\sum_{j=1}^{M}(\mathbf{w}^{[j]} - \bar{\mathbf{w}}) \otimes (\mathbf{w}^{[j]} - \bar{\mathbf{w}})$)
- `stiff_me` : maximum entropy fit to `c1b`.

1 Relative entropy for Gaussians II

In session 10, exercise 2.2, we showed the following formula for the relative entropy between two Gaussian density functions:

$$D(p,q) = \frac{1}{2}\left[\mathrm{tr}\left(K_2K_1^{-1}\right) - \ln\frac{|K_2|}{|K_1|} - N\right] + \frac{1}{2}(\widehat{x}_1 - \widehat{x}_2)\cdot K_2(\widehat{x}_1 - \widehat{x}_2), \tag{1}$$

where

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{N/2}|K_1^{-1}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\widehat{x}_1)\cdot K_1(x-\widehat{x}_1)\right\},\ x\in\mathbb{R}^N,\\ q(x) &= \frac{1}{(2\pi)^{N/2}|K_2^{-1}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\widehat{x}_2)\cdot K_2(x-\widehat{x}_2)\right\},\ x\in\mathbb{R}^N, \end{aligned}$$

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

i) What are the distributions of Y_1 and Y_2? Give explicitly the mean value and the covariance for each random variable.

ii) By denoting by $\tilde{p}$ and $\tilde{q}$ the density functions of, respectively, Y_1 and Y_2, show that $D(\tilde{p},\tilde{q}) = D(p,q)$.

2. In exercise 2.2 of the exercise session 10 we also introduced:

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

for K_i, symmetric positive defined matrices, $i = 1, 2$. Define now the following generalised eigenvalue problem

$$K_2v_i = \mu_iK_1v_i,\ i = 1,\dots,N. \tag{3}$$

Show that using the generalised eigenvalue problem (3), equation (2) becomes

$$D^{\dagger}(K_1, K_2) = \frac{1}{2}\sum_{i=1}^{N}(\mu_i - \ln \mu_i - 1). \tag{4}$$

3. The stiffness part of the relative entropy (2) is in general not symmetric, i.e, $D^{\dagger}(K_1, K_2) \neq D^{\dagger}(K_2, K_1)$ for $K_1 \neq K_2$. We then define the symmetrized version of (2) in the following way:

$$D^{\dagger}_{sym}(K_1, K_2) := \frac{1}{2}\left(D^{\dagger}(K_1, K_2) + D^{\dagger}(K_2, K_1)\right). \tag{5}$$

i) Find an explicit form for $D^{\dagger}_{sym}(K_1, K_2)$.

ii) Using the generalized eigenvalue problem (3) find the corresponding eigenvalue version form for $D^{\dagger}_{sym}(K_1, K_2)$.

[Note: Equation (4) can be useful to prove part 1, ii)]

2 Jensen's inequality - Part 2

We recall the main result of Jensen's inequality Part 1 exercise (see Qu 4 Serie 10). Let $u : \Omega \subset \mathbb{R}^N \to \Gamma \subset \mathbb{R}^N$ with $|\Omega| = \int_\Omega d\mu(x) < \infty$, for a measure $\mu(x)$ and ϕ convex from $R(u(\Omega))$ to $\mathbb{R}$. We have that

$$\frac{1}{|\Omega|}\int_\Omega \phi(u(x))d\mu(x) \geq \phi\left(\frac{1}{|\Omega|}\int_\Omega u(x)d\mu(x)\right) \tag{6}$$

with equality if and only if u is constant on Ω. If Ω is a bounded set then $d\mu(x) = dx$ is one possible choice.

1. Take $\phi(x) = x\ln(ax)$ for $x, a \in \mathbb{R}^+$. Sketch $\phi(x)$. Hint: compute $\phi(x)'$, $\phi(x)''$, and show that $\lim_{x\to 0}\phi(x) = 0$.

2. Suppose $|\Omega| = \int_\Omega d\mu(x) < \infty$. Prove that for all probability density functions $p(x)$ on Ω with respect to $d\mu(x)$, i.e, such that $\int_\Omega p(x)d\mu(x) = \int_\Omega q(x)d\mu(x) = 1$,

$$\int_\Omega p\ln(|\Omega|p)d\mu(x) \geq 0. \tag{7}$$

3. Show that (7) can be rewritten as

$$\int_\Omega p\ln(p)d\mu(x) \geq \int_\Omega \frac{1}{|\Omega|}\ln\left(\frac{1}{|\Omega|}\right)d\mu(x), \tag{8}$$

i.e, the minimal entropy distribution is uniform with respect to the measure $\mu(x)$. If Ω is bounded in $\mathbb{R}^N$ then $d\mu(x) = dx$ is one possible choice of measure.

4. Suppose p and q are two probability density functions on Ω with respect to a measure $d\mu(x)$. Show that

$$\int_\Omega p\ln\left(\frac{p}{q}\right)d\mu(x) \geq 0, \text{ with equality iff } p = q \text{ a.e .} \tag{9}$$

Not that here $\int_\Omega d\mu(x) < \infty$ is *not* required as a hypothesis, so that $d\mu(x)$ is a possible choice even if Ω is not bounded in $\mathbb{R}^N$.

3 Estimate of mean and stiffness from MD simulation data

Download a visualization tool for cgDNA+ matrices: `plot2DMatrix.m` `http://lcvmwww.epfl.ch/teaching/modelling_dna/protected_files/codes_exercises/plot2DMatrix.m`.

1. Compute the inverse of the raw covariance `c1b` to obtain the raw stiffness `s1b`. Visualize both matrices using `plot2DMatrix`. What can you say about the two plots (differences, etc..) ?

2. In the cgDNA+ model internal coordinates, inter, intra and phosphates rotations (Cayley vectors) are rescaled by $\frac{1}{5}$ during the parameter extraction procedure, thus

$$\frac{1}{5}u_{cgDNA+} = u_{ob}, \quad \frac{1}{5}\eta_{cgDNA+} = \eta_{ob}, \quad \frac{1}{5}\eta^{pC}_{cgDNA+} = \eta^{pC}_{ob}, \quad \frac{1}{5}\eta^{pW}_{cgDNA+} = \eta^{pW}_{ob},$$

   where $u, \eta, \eta^{pC}, \eta^{pW}$ (see Qu 4 corr 6 for notations) are the rotations (Cayley vectors) corresponding to inter, intra, Crick and Watson phosphates. Variable with subscript *ob* are the rotations observed from MD data. This scaling has been adopted in order to have all the diagonal entries of the stiffness matrix and all the components of the mean within a similar range. Rescale correctly only the diagonal entries of the raw stiffness matrix `s1b`. Extract and then plot the diagonal entries of `s1b` and the diagonal entries of the rescaled matrix. Do the same for the entries of the groundstate `shape`. What can you say?

[Note: The invariance of the Kullback-Leibler divergence means that the divergence between two Gaussians p and q does not change for different scaling, see exercise 1.1 of this session.]

4 Palindromic symmetry of a shape vector and stiffness matrix

Download `http://lcvmwww.epfl.ch/teaching/modelling_dna/protected_files/codes_exercises/MaxEntropy` a file containing useful matlab scripts required for this exercise.

As seen in Exercise 5 of session 6, the mean vector $\mu(S)$ and the stiffness matrix $K(S)$ of the cgDNA+ Gaussian distribution $\rho(\mathbf{w}, S)$, satisfy the following symmetric properties: $\mu(S) = E_N\mu(\overline{S})$, and $K(S) = E_N K(\overline{S})E_N$, where $\overline{S}$ is the complementary sequence to S and the matrix E_N is defined in Qu 5 session 6. For palindromic sequences, i.e. when $S = \overline{S}$, these properties become $\mu(S) = E_N\mu(S)$ and $K(S) = E_N K(S)E_N$.

1. Load the file `palin_md_data.mat` with the DNA 24-mer data. Construct the matrix E_N for $N = 24$, and check if the palindromic properties of the shape vector and the covariance matrix are satisfied.

2. Before using the MD data to fit a cgDNA+ parameter set, in order to minimize modelling errors, it is useful to first get shape and stiffness estimates satisfying palindromic symmetry conditions, if a sequence is palindromic. Symmetrize the shape estimate by computing $\bar{\mathbf{w}}_{sym} = \frac{1}{2}(\bar{\mathbf{w}} + E_N\bar{\mathbf{w}})$. Plot the difference between $\bar{\mathbf{w}}$ and $\bar{\mathbf{w}}_{sym}$.

3. The covariance matrix was obtained as $C = \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]}\otimes\mathbf{w}^{[j]} - \bar{\mathbf{w}}\otimes\bar{\mathbf{w}}$. Symmetrize $D := \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}^{[j]}\otimes\mathbf{w}^{[j]}$ and then compute the symmetrized covariance matrix using D_{sym} and $\bar{\mathbf{w}}_{sym}$. Compute the symmetrized stiffness K_{sym} as the inverse of C_{sym}. Using `plot2DMatrix.m`, plot the differences between C and C_{sym} and between K and K_{sym}.

4. Use the script `computeMaxEntropy.m` to get the maximum entropy fit to the symmetrized covariance computed in the previous point.

5. Compute the following Kullback-Leibler divergences (per degree of freedom):

   i) $D(\rho_{obs}^{sym}(S), \rho_{obs}(S))$,
   ii) $D(\rho_{band}^{sym}(S), \rho_{obs}^{sym}(S))$,
   iii) $D(\rho_{cgDNAp}(S), \rho_{band}^{sym}(S))$.

   Compare this error with the modelling errors (Kullback-Leibler divergences computed in exercise 3 of session 10). Is the convergence error big or small compared to the modelling errors?

5 From atomistic representation to cgDNA internal coordinates

The aim of this exercise is to understand the procedure of computing the internal coordinates from time series of atomistic coordinates of snapshots of MD simulations. In order to make this particular exercise slightly shorter we only work with cgDNA model internal coordinates (intra & inter) and we ignore the phosphates coordinates, which could nevertheless be treated in a similar way.

Remark: In extracting cgDNA parameter sets there are a number of steps to pass from a time series of atomistic coordinates of snap shots of MD simulations, to a time series of snapshots of frames fit to atomistic coordiates, and then to a time series of snapshots of cgDNA internal coordinates. The time series of internal coordinate snapshots is then used as an ensemble from which to estimate first and second moments, ie Gaussian statistics for that oligomer. For actual computations this estimation process has to be implemented quite efficiently as very large data sets are involved, with specialised data file formats, and we do not discuss the actual code that is used. Instead in this exercise in order that you understand quite concretely what is entailed we give you an artificially small data set, taken from an actual MD simulation, and you then implement each step in the pipeline using pre-prepared Matlab scripts. In particular note that despite the sequence being a palindrome, individual snapshots of configurations need not satisfy palindromic symmetry. It is only equilibrium statistics for a palindromic oligomer that need satisfy an associated Crick-Watson symmetry relation.

For this exercise download the zipped file here (`http://lcvmwww.epfl.ch/teaching/modelling_dna/protected_files/codes_exercises/frames_embedding_qus.zip`). This contains the following files:

- `Palin_141_3.1.ions.pdb` : pdb structure with 6 snapshots (each snapshot is saved consecutively with 2 nanosecond of time difference) of full atomistic 24 mer (for palindromic sequence $GCCCTTGGCGATATCGCCAAGGGC$). This pdb structure is taken from real MD simulation which has been used for computing cgDNA parameter sets.

- `PDBAtoms_p24.mat` : coordinates of atoms in base (for each bases in both reading and complimentary strand) corresponding to the 6 snapshots mentioned in above point (for each of these snapshots pdb structures are given in the file `Palin_141_3.1.ions.pdb`).

- `Ideal_Bases.mat` : coordinates of atoms in ideal base (for all A, T, C and G bases) as per Tsukuba convention.

- `cgDNA_200snap_coord.mat` : cgDNA model internal coordinates corresponding to 200 individual snapshots of MD (each snapshot is saved consecutively after 1 picosecond of time

difference so in total this file represent the data from 200 picoseconds of MD simulation) i.e., we have $w_i \ \forall \ i = 1:200$. So, these 200 different internal coordinates have been computed using 200 distinct MD snapshot. We use this data to show that even though the sequence is palindrome there is no reason why individual snapshots should be palindromic. However, the equilibrium statistics should satisfy the palindromic symmetry.

Now do the following:

1. Load pdb structure (`Palin_141_3.1.ions.pdb`) using molviewer in matlab and visualise the variation in atomic coordinates between six snapshots. What can you say?

2. Write a matlab script to fit a frame $(\mathbf{r}, R)$ on each bases of reading and complementary strand. Input for this script should be the atomic coordinates (data in file `PDBAtoms_p24.mat`) and idealized coordinates (data in file `Ideal_Bases.mat`) and output should be frame $(\mathbf{r}, R)$. Compute the frames for all the bases in reading and complimentary strand for all 6 snapshots. Then visualise these frames for two or more snapshots in one figure using `cgDNAviewer`. What can you say?

Note: Recall from the analysis done in Qu3 corr5 that we have explicit expression for the best fit frame $(\mathbf{r}, R)$ to a given set of atomic coordinates of atoms $\mathbf{p}_i$, $i = 1, \ldots, M$, whose idealized coordinates in the frame $(\mathbf{r}, R)$ are known to be $\boldsymbol{\alpha}_i$, $i = 1, \cdots, M$.

3. What's the fitting error (see equation 3 of Qu3 Serie 5 for mathematical details)?

4. Now we want to compute cgDNA model internal coordinates (intra & inter) using the frames found in previous point (2). In order to do that, perform the following steps (recall about mid frame, junction frame and cgDNA coordinates from lecture of weeks 5 & 6):

i) Write a matlab script to compute relative coordinates (u, v) between two rigid body frames $g_1 = (R_1, r_1)$ to $g_2 = (R_2, r_2)$. Here, $g_1, g_2 \in \mathrm{SE}(3)$ and u is rotational coordinates (Cayley vector) and v is translational coordinates.

ii) Compute cgDNA model internal coordinates (intra and inter) using the script written in above point and using the frames found in part 2 of this exercise.

iii) Plot the internal coordinates for all 6 snapshots in one figure. What can you say?
[Hint: You can use script `cgDNA2dplot.m` provided in cgDNA package, which will create the plot of each intra and inter basepair coordinate along the molecule.]

5. Use data `cgDNA_200snap_coord.mat` and compute $\|w_i - E_N w_i\|$ for i =1, 2 ..., some random snapshot. Here for cgDNA model E_N is a $(12N-6) \times (12N-6)$ matrix defined as

$$E_N = \begin{bmatrix} 0 & 0 & \cdots & E \\ 0 & \cdots & E & 0 \\ \cdots & \iddots & \cdots & 0 \\ E & \cdots & 0 & 0 \end{bmatrix},$$

where $E = diag(-1, 1, 1, -1, 1, 1)$. Notice that as a consequence of only treating cgDNA coordinates in this exercise, E_N here is different from E_N in Qu 4. Then compute the average $\bar{w} = \frac{1}{M} \sum_{i=1}^{M} w_i$ for $M = 25, 50, 100, 200$ and use it for computing the value of $\|\bar{w} - E_n \bar{w}\|$. What can you say?