Mathematical Modelling of DNA
Prof. John Maddocks
Session 11: A. Patelli

For exercise 1 and 2 consider again the dataset given in Session 10 exercise 2.

1 Kullback-Liebler divergence between : $\rho_{obs}(S)$, $\rho_{band}(S)$, $\rho_{cgDNA}(S, \mathcal{P})$

In this exercise we want to study the Kullback-Leibler divergence between the different steps of the approximation of the stiffness and the mean for a given sequence. We will use the following notation: $\rho_{obs}(S)$ is the Gaussian distribution which parameters are the ensemble mean and the ensemble covariance, $\rho_{band}(S)$ is the Maximum Entropy Gaussian distribution fitted to the ensemble mean and covariance, and $\rho_{cgDNA}(S, \mathcal{P})$ is the cgDNA reconstruction of S using the parameter set $\mathcal{P}$ (check that you are using the cgDNAparamset2).

1. Write a MATLAB script implementing the Kullback-Leibler divergence for Gaussian distributions (see Session 9, exercise 2.2). Allow your script to output the value of the Kullback-Leibler divergence, the value of the stiffness part, and the value of the mean part (Mahalanobis distance).

2. Using your script do the following computations:

i) $D(\rho_{band}(S), \rho_{obs}(S))$,
ii) $D(\rho_{cgDNA}(S, \mathcal{P}), \rho_{band}(S))$,

Compare the values of the Kullback-Leibler divergence as well as the values coming from the stiffness part and the mean part. What can you say?

2 Estimate of mean and stiffness from MD simulation data

Download from `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/plotMatrix2D.m` a visualization tool for cgDNA matrices : `plot2DMatrix`.

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. The cgDNA inter and intra rotations are rescaled by $\frac{1}{5}$ during the parameter extraction procedure, thus

$$\frac{1}{5} u^{cgDNA} = u^{ob},$$
$$\frac{1}{5} \eta^{cgDNA} = \eta^{ob}.$$

where u_i^{ob}, η_i^{ob} are the rotations observed from MD data. This scaling has been adopted in order to have all the entries of the stiffness matrix and of the mean within a similar range. Rescale correctly only the diagonal entries of the 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 ground state `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 Session 10 , exercise 1.1]

3 On the computation of marginals of the cgDNA probability distribution

Given a sequence S and a parameter set $\mathcal{P}$, the cgDNA model is the Gaussian distribution:

$$\rho(\mathbf{x}; S, \mathcal{P}) = \frac{1}{Z}\exp\left\{-\beta(\mathbf{x} - \widehat{\mathbf{x}}(S,\mathcal{P}))\cdot K(S,\mathcal{P})(\mathbf{x} - \widehat{\mathbf{x}}(S,\mathcal{P}))\right\} \tag{1}$$

where $\widehat{\mathbf{x}}(S,\mathcal{P})$ and $K(S,\mathcal{P})$ are respectively the mean and the stiffness matrix. We recall that the covariance is $\Sigma = K^{-1}(S,\mathcal{P})$. In this exercise we want to do two different possible marginals of a cgDNA Gaussian distribution.

3.1 Marginalise over intra-base-pair variables

In the first part of this exercise we will focus on the marginalisation of the intra variables. For the computation consider the R. E. Dickerson palindromic dodecamer S_D = CGCGAATTCGCG. Write a code for the two different method explained hereafter.

1. Consider the stiffness matrix $K_D = K(S_D, \mathcal{P})$.

i) Recombine the stiffness matrix $K(S_D, \mathcal{P})$ in the following form:

$$\widetilde{K}_D = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

where $A = A^T \in \mathbb{R}^{6(n-1)\times 6(n-1)}$, $B \in \mathbb{R}^{6n\times 6(n-1)}$ and $C = C^T \in \mathbb{R}^{6n\times 6n}$. The block A is the block associated to the inter variables, B is the block related to the coupling between inter and intra variable, while C is the block associated to the intra variable. What is the pattern of $\widetilde{K}_D$?

ii) Apply now the formula obtained in Exercise 1 Session 9, to compute the marginal stiffness matrix (noted $K_1^{(u,v)}$) for the inter variables. Is the matrix dense?

2. Consider the covariance $\Sigma_D = K_D^{-1}$. We stress on the fact that the stiffness matrix K_D as a specific pattern and is sparse while Σ_D is dense.

i) Compute $\widetilde{\Sigma}_D$ in such a manner that it has the same block structure as $\widetilde{K}_D$. The obtained matrix has a specific pattern?

ii) Invert the block corresponding to the inter variable to obtain the marginal stiffness matrix (denoted $K_2^{(u,v)}$) of the inter. Is the matrix dense?

Compare the two obtained marginal stiffness matrices. Which of the two method is faster? Test the performance of the two methods with longer sequences.

[Note: The above way of marginalise leads to a DNA model base only of inter coordinates. This kind of model is called a rigid-basepair model of DNA.]

3.2 A localized cgDNA model: marginalise over the configurations of the flanking sequences

The following marginalisation could be useful to study the statistical mechanics property of a small segment of a potentially very long fragment of DNA. Begin by adding randomly 100 basepairs at each end of the Dickerson dodecamer, i.e, define $\tilde{S} = S_1 S_D S_2$ where S_i are randomly chosen (but then fixed) 100 basepair long sequences. Do the following steps:

i) Reconstruct the stiffness matrix and the ground-state for $\tilde{S}$ using cgDNA.

ii) Invert the reconstructed stiffness matrix and extract the entries of the covariance that correspond to S_D. Invert them to obtain the marginalised Dickerson dodecamer.

iii) Extract the entries of the ground-state corresponding to S_D.

What is the sparsity pattern of the marginalised stiffness? Compare the marginal stiffness and marginal ground-state with the corresponding cgDNA reconstruction of S_D (for example compute the Kullback-Leibler divergence between the two distributions). What happen if you change the flanking sequences?

[Note: Based on above method one can also marginalise over flanking sequences. By considering the following ensemble $\mathcal{S}(S_D) = \{S|S = S_1 S_D S_2,\ S_1, S_2 \text{ flanking sequences }\}$, on can compute the marginal of S_D over flanking sequences as the ensemble average of all the localized marginals of S_D computed for all $S \in \mathcal{S}$.]