1 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.

1.1 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.

[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.]

1.2 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 $S \in \mathcal{S}$.]

2 Gaussian Integral III

Let $\hat{\mathbf{x}} \in \mathbb{R}^n$ and a symmetric, positive - definite matrix $K = \Sigma^{-1} \in \mathbb{R}^{n \times n}$ be given. Show that a conditional of a Gaussian distribution is also a Gaussian distribution: if $\mathbf{x} \sim N(\hat{\mathbf{x}}, \Sigma)$,

$$\begin{aligned}
\mathbf{x} &= \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \quad \hat{\mathbf{x}} = \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \end{bmatrix}, \quad \mathbf{x}_1, \hat{\mathbf{x}}_1 \in \mathbb{R}^k, \quad \mathbf{x}_2, \hat{\mathbf{x}}_2 \in \mathbb{R}^m, \\
\Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}, \quad \Sigma_{11} = \Sigma_{11}^T \in \mathbb{R}^{k \times k}, \quad \Sigma_{12} \in \mathbb{R}^{k \times m}, \\
& \Sigma_{22} = \Sigma_{22}^T \in \mathbb{R}^{m \times m}, \quad \text{and} \quad k + m = n,
\end{aligned}$$

then $(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{a}) \sim N(\overline{\mathbf{x}}, \overline{\Sigma})$, where

$$\overline{\mathbf{x}} = \hat{\mathbf{x}}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{a} - \hat{\mathbf{x}}_2), \tag{2}$$

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}. \tag{3}$$

[Hint: Use the definition for the conditional density function for continuous random variable and the solution of Exercise 1 of Session 9.]

3 On the computation of conditionals of the cgDNA probability distribution

For this exercise you should download the following scripts: `http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/Conditional_scripts.zip`.

The aim of this exercise is to develop a basic statistical model for modelling the interaction between DNA and proteins. DNA-binding proteins are proteins which have an affinity with DNA (see for more details the Wikipedia article : DNA-binding protein), which can bind to the DNA in either the major or minor groove. In the context of the cgDNA model one can see the protein that binds to a molecule of DNA as a constrains on some of the internal coordinates describing the DNA segment. Thus, from a statistical mechanics point of view the interaction between DNA and protein can be modelled as a conditional distribution of the density function related to the DNA fragment, which in the cgDNA land is Gaussian. Thanks to the previous exercise we know that a conditional distribution of a Gaussian distribution still be a Gaussian. Let us assume that the interaction DNA-protein is reduced to a change in only an intra coordinate. Let $\mathbf{w} = (y_1, x_1, y_2, \ldots, x_i, y_{i+1}, x_{i+1}, \ldots, y_n) = (w_1, y_i, w_2) \in \mathbb{R}^{12n-6}$ where $y_j \in \mathbb{R}^6$ are the intras and $x_k \in \mathbb{R}^6$ are the inters, and

$$\begin{aligned} \rho(\mathbf{w}; S, \mathcal{P}) &= \frac{1}{Z} \exp\left\{-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}}(S,\mathcal{P})) \cdot K(S,\mathcal{P})(\mathbf{w} - \hat{\mathbf{w}}(S,\mathcal{P}))\right\} \\ &= \frac{1}{Z} \exp\left\{-\frac{1}{2}\begin{bmatrix} w_1 - \hat{w}_1 \\ y_i - \hat{y}_i \\ w_2 - \hat{w}_2 \end{bmatrix}\begin{bmatrix} A & B & 0 \\ B^T & C & D^T \\ 0 & D & E \end{bmatrix}\begin{bmatrix} w_1 - \hat{w}_1 \\ y_i - \hat{y}_i \\ w_2 - \hat{w}_2 \end{bmatrix}\right\}, \end{aligned} \tag{4}$$

a cgDNA Gaussian for the sequence S and the parameter set $\mathcal{P}$. Imagine now that a protein is binding to the i–th basepair, thus it constraints $y_i = \mathbf{a} \in \mathbb{R}^6$.

1. By using the Exercise 2 of this sheet find the conditional mean $\overline{\mathbf{w}} = (\overline{w}_1, \overline{w}_2)$.

2. Complete the lines 53 and 54 in `calc_conditional_shapes.m` with your findings. and run it with the following input arguments :

- sequence : ATCGCGAATGCGAGCCTGTA ;
- cond_index : 10 ;
- cond : [0.1 0.5 0.5 0.3 0.6 0] (= $\delta\mathbf{a}$) . In the following code we consider $\mathbf{a} = \hat{y}_i + \delta\mathbf{a}$.

Be aware that you have also to complete lines 19 and 20 by adding to the path your cgDNA folder and your cgDNAviewer.

[Hint: Conditioning one intra (example on the left) leads to a specific decomposition of the stiffness matrix that can be seen in the matrix on the left (each little block is a 6 times 6 matrix). This implies that the conditional stiffness can be seen as a block diagonal matrix.]