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

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

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

1. Here (`http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/KL_Div.m`) you can find a possible way of coding the Kullback-Leibler divergence. We stress here that the formula of the Kullback-Leibler divergence for Gaussian implies the computation of the log of a determinant. In general in MATLAB one has to avoid the computation of the determinant of large positive definite matrix as the value could easily hit the MATLAB infinity. As the determinant of a matrix is equal to the product of its eigenvalues, the log of the determinant can de computed as the sum of the logs of the eigenvalues. The latter strategy is implemented in `KL_Div.m`.

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

	KLD	stiffness part	mean part
$D(\rho_{band}(S), \rho_{obs}(S))$	11.6101	11.6101	0
$D(\rho_{cgDNA}(S, \mathcal{P}), \rho_{band}(S))$	5.5201	3.4391	2.0811

2 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 present any specific pattern around the diagonal.

2. 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.

3 On the computation of marginals of the cgDNA probability distribution

3.1 Marginalise over intra-base-pair variables

1. i) See Figure (1).
ii) The marginal stiffness matrix $K_1^{(u,u)}$ is dense.

2. i) $\tilde{\Sigma}_D$ does not have any specific pattern as the covariance matrix Σ_D is dense.
ii) The marginal stiffness matrix $K_2^{(u,u)}$ is dense.

The method 1) should be faster as it involve only a for loop over all the 18×18 blocks of the stiffness matrix, while for the method 2) two for loop over all the 6×6 blocks of the covariance matrix are needed, thus for the recombination of the matrices method one is faster when considering long sequences. Moreover for the method 2) we need to invert a dense matrix of dimension $6(N-1)$, where N is the number of base pair considered while in the method 1) we invert a sparse matrix of dimension $6N$. We can then conclude that method 1) is faster then method 2).

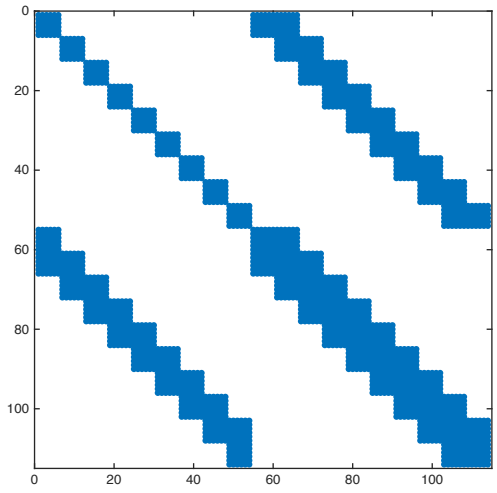

Figure 1: Sparsity of $\tilde{K}_D$

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

The sparsity pattern of the marginalised stiffness matrix is the same as the cgDNA one, i.e, 18 times 18 blocks with 6 times 6 overlaps. Just be aware that if you want to use the MATLAB function `spy` to visualize the sparsity pattern of the matrix you should use the following combination : `spy(abs(K_marginal)> 1e-10)`. You can now use your Kullback-Leibler divergence to estimate the difference between the marginal and the cgDNA reconstruction of S_D, or you can visualize the difference using the function `plotMatrix2D` given in exercise 2 of this sheet. It is possible to prove why the localized marginal still banded, but this prove is beyond the purpose of this course as it involves the prove for the Maximum Entropy fit of banded Gaussian. For the motivated students we refer to the Phd thesis of Jaroslaw Glowacki, "Computation and Visualization in Multiscale Modelling of DNA Mechanics", 2016, EPFL Thesis #7062, Chapter P1.1 where you can find the prove of the Maximum Entropy fit and make yourself and idea about the prove of the localized marginal.

Here (`http://lcvmwww.epfl.ch/teaching/modelling_dna/public_files/LocalizedCgDNA.m`) you can download the code for computing the marginal over the configurations of the flanking sequence.