Mathematical Modelling of DNA
Prof. John H. Maddocks
Summary all exercises fall 2017/2018 A. Patelli

Session 1

1. The length scales of DNA : basic background in length scale of DNA
2. Gaussian integrals I : Gaussian integral formulas that will be used later in the course

Session 2

1. Properties of the skew symmetric matrices : Basic properties to be used later on.
2. Rotations in three dimensions : Basic properties to be used later on.

Session 3

1. Cayley transforms : The Cayley transforms can be regarded as an algebraic approximation to the relation between the exponential and logarithm of a matrix. This exercise considers Cayley transform for general matrices and a special case for the SO(3), which forms the basis for the rotational coordinates in the cgDNA model.

Session 4

1. cgDNAweb : A web site for visualize ground-states of the cgDNA model.
2. The cgDNA matlab package : Full implementation on the cgDNA model.
3. On the main outputs of the cgDNA package : Examples of capabilities of the cgDNA package, in particular the frames.m subroutine for base, base-pair, and junction frames reconstructions.
4. A MATLAB cgDNA viewer : Programming this viewer yourself ensure you understand the cgDNA coordinates systems.

Session 5

1. Effect of a point mutation on the Shape and on the Stiffness : In the cgDNA model a local change in a sequence can cause a non local changes in the ground-state but only a local change in the stiffness.
2. On the symmetry of the coordinates system : Understanding numerically the consequences of exchanging roles of Crick and Watson (or reading) strand on the cgDNA coordinates.
3. Properties of SO(3) matrix representation and square roots : To have the simple symmetry rule in exercise 2), the base-pair frames must be symmetrically placed between the two base frames and the junction frame must be symmetrically placed between two successive base-pair frames. This involves computing the square root of a matrix in SO(3).
4. Proof of the change of reading strand transformation : Mathematical proof of the change of reading strand transformation previously numerically observed in question 2)
5. Downloading and compiling the cgDNAmc code : Preparation for the next exercise session.

Session 6

1. Monte Carlo simulation with the cgDNA model : Purpose of the exercise

   - Monte Carlo sampling of frame configurations drawn from a cgDNA Gaussian pdf in internal coordinates $\mathbf{w}$.
   - Differences in ground-states of cgDNA for parameter set 1 and 2. (differences in parameter set discussed more in Chap. 3)
   - Note that all the 6 distinct poly$(\alpha\beta)_N$ ground-states are helical but with noticeably different pitch and radii.
   - Relative speeds of the MC sampling in matlab and compiled code.
   - One of the classical quantity in polymer physics is the Flory vector (as will be discussed in chapter 2). Exercise 1.3 gives a numerical introduction to the Flory vector. Note that for all the poly dimers the cloud of points produced by the n-th base-pair position look symmetric. Moreover the cloud get broader for larger values of n.
   - A second classic quantity is the tangent-tangent correlation and the associated persistence length (as will be also discussed in chapter 2). Exercise 1.4 gives a numerical introduction to this quantities. Note that the ln(ttc) versus n plots are almost straight for poly dimers.
   - Numerical study of the Flory persistence vector and comparison between parameter set 1 and 2.

Session 7

1. On the average of rotation matrices sharing a common (deterministic) axis : Except for this very special ensemble $||\langle Q\rangle|| < 1$.

2. Monte Carlo simulation with the cgDNA model II : Same as exercise 1) session 6 but for biological sequences which exhibit strongly different behaviours compared to the poly dimer of session 6.

3. Effect of sparsity on Monte Carlo simulation efficiency : The fact that cgDNA has a banded stiffness matrix is computationally very important.

Session 8

1. Explicit computation of apparent persistence length for a tractable probability density function (the HWLC) : Comparison of numerics with analytics we made in class. See for instance the document in the week-by-week correspondence of week 8. The connection between this exercise and the session 6 and session 7 is discussed in the solution sheet of this exercise.

2. On the parametrization of junction displacement using quaternions (a final exercise pertaining to efficiency cgDNAmc)

Session 9

1. Gaussian integrals II : Computing the explicit formula for the marginals of a Gaussian. Used later to get a cgDNA model restricted to a local subsequence.

2. Entropy and Relative entropy formulas for Gaussians: Both entropy and relative entropy (or Kullback-Liebler divergence) as introduced here are used later in parameter estimation.

3. Jensen's inequality : Classic estimate that we use later in maximum entropy estimation.

Session 10

1. Relative entropy for Gaussians II : 1.1) shows that relative entropy is independent of rescaling of coordinates. 1.2) an alternative formula for part of relative entropy involving generalized eigenvalues of stiffness matrices. 1.3) another formula using generalized eigenvalue for the symmetrized Kullback-Liebler divergence.

2. Banded matrices and their inverses : Characterising matrices with banded inverses, and computing banded matrices whose inverses are partially prescribed.

3. Kullback-Liebler divergences between $\rho_{obs}(S), \rho_{band}(S), \rho_{cgDNA}(S,P)$: Evaluating relative entropies between different Gaussian approximations to oligomer-based pdf.

Session 11

1. On the computation of marginals of the cgDNA probability distribution : 1.1) The rigid basepair marginal of the cgDNA rigid base model are Gaussians but with non banded stiffness matrices. 1.2) The localized cgDNA marginal over flanking sequences is itself a banded Gaussian.

2. Gaussian Integral III : Computing the explicit formula for the conditional of a Gaussian. Used in the next exercise.

3. On the computation of conditionals of the cgDNA probability distribution : Special features of the conditional of the banded cgDNA Gaussian. Such conditionals can be used as a simple model of protein binding to DNA, but we will not discuss it further in the lectures.

Session 12

1. Principle of maximum entropy parameter estimation for banded stiffness matrices: Detailed computation of how to obtain a banded Gaussian as the solution to Jayne's maximum entropy principle with constraints on the admissible set of PDFs.

2. Estimate of mean and stiffness from MD simulation data: Part 1) observe that the raw covariance is dense while its inverse is almost banded. In fact the pattern of the raw stiffness matrix is close to the cgDNA overlapping 18x18 blocks (not by accident). Note that this question should be compared to exercise 3 session 10 but the ordering of the exercises had to be adapted at short notice because of the rescheduling of the lectures. Part 2) The rescaling by 1/5 means that the rotation-rotation diagonal entries of the stiffness matrix are of the same order of magnitude of the translation-translation blocks, which is convenient. However the invariance of the KL divergence means that the parameter fitting is fact unaffected.

3. Palindromic symmetry of a shape vector and stiffness matrix: How to use palindromic symmetry to improve estimates of oligomer-based Gaussian PDF.

Session 13

1. Positive Definiteness of the cgDNA stiffness matrix and Invariance of the parameter set: Part 1) In Paramset 2 (P_2) (which you have) the stiffness blocks $K^{\alpha\beta}$ are indefinite but any reconstructed stiffness matrix $K(S,P_2)$ is in fact positive definite for any sequence S. This exercise explains a sufficient set of conditions, satisfied by P_2, that guarantee this property. Part 2) The oligomer based PDFs for a sequence S, and its Crick Watson complementary sequence $\bar{S}$, are related by appropriate symmetry conditions on both means and stiffnesses. This exercise describes a sufficient set of relations on a cgDNA parameter set which guarantee the required relations between $\rho(S,P)$ and $\rho(\bar{S},P)$, for any S.

2. Zero entries in the cgDNA parameter set: The relation discussed in ex1 part 2 include sufficient conditions that the oligomer based PDF for palindromic sequences satisfy the necessary palindromic symmetry. This exercise discusses the consequences of this symmetries on the parameter set itself. In particular certain entries in the parameter set must vanish. We identify and count them.

3. Total number of unknowns in a cgDNA parameter set: Combine exercise 1 and exercise 2 for counting the total number of independent scalars of a cgDNA parameter set.

4. Square roots in SE(3): This exercise is a bit of an orphan. It is just a technical computation to show that while the square root of a rotation matrix in SO(3) plays a central role in the definition of cgDNA basepair and junction frames (ex 2-3-4 session 5) the square root of a transformation in SE(3) is **not** used.