Mathematical Modelling of DNA
Prof. John H. Maddocks
Summary all exercises spring 2020 (followed by 2019) A. Patelli, R. Singh

Spring 2020

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. First version of Euler-Rodrigues formula which gives a rotation matrix in terms of a unit rotation axis and a rotation angle in $[0, \pi]$ using the right handed rule (regle du tir bouchon). Moreover for each rotation through an angle less than π there is a unique unit rotation axis and angle.

3. Computing matrix exponential using Cayley-Hamilton: Using Taylor series definition of matrix exponential applied with a skew-symmetric matrix and characteristic polynomial coming from Cayley-Hamilton, it is showm that exponential of skew-symmetric matrix is of the form of Euler-Rodrigues formula which gives a rotation matrix in terms of a unit rotation axis and a rotation angle.

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 the special cases for SO(3) and SE(3), which form the basis for the rotational coordinates in the cgDNA model.

Session 4

1. Additional properties (change of basis and inverse) of the Cayley transform of a proper rotation matrix.

2. The change of reading strand transformation – Part 1: In the first part we prove two useful properties of the Cayley transformation. In the second part we show that for two rigid-bodies there is a linear relationship between their internal coordinates and the internal coordinates of a transformation (on the right) by an orthogonal matrix on both rigid–body orientations. The fact that the relation is linear is due to the choice of writing the relative translation in the mid–frames.

3. More on square roots of matrices: To have the simple symmetry rule in exercise 1), 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). In the second part of the exercise we show why the square root of a transformation in SE(3) is **not** used.

4. Review of various (standard) Matrix Factorisations. In particular we will use the Cholesky factorisation of a symmetric positive definite matrix (in our Monte Carlo code) and the polar decomposition (in fitting frames to atomistic data).

Session 5

1. cgDNAweb : A website for visualizing ground-states of the cgDNA model.
2. The cgDNA matlab package : Full implementation on the cgDNA model.
3. Optimisation in SE(3): From atomic coordinates of a given base we associated a rigid-body computed by solving an optimization problem in SE(3).
4. Completing the square in vector quadratic forms: In this exercise we explicitly compute the stiffness matrix, the shift, and the constant term of a quadratic form defined as sum of (local) quadratic contributions. Explains why the ground state in the cgDNA model has nonlocal sequence dependence even though the parameter set is dimer dependent, i.e. very localised sequence dependence.

Session 6

1. 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.
2. A MATLAB cgDNA viewer : Programming this viewer yourself ensure you understand the cgDNA coordinates systems.
3. Scaling and the Cayley transform: We introduce a scaled version for the Cayley transformation and its inverse. Used later to have stiffness matrices whose diagonal entries are of the same order of magnitude for rotational and translational degrees of freedom.
4. This exercise reveals the close connection between the Cayley transform and the matrix exponential by comparing the two respective matrix Taylor series with the choice of scaling $\alpha = 2$ in the Cayley transform (see Session 3). For small rotations (expressed in radians) the components of the vector generating the skew matrix in both the exponential parametrization of SO(3) and the Cayley transformation (with $\alpha = 2$) are close to rotation angles about the coordinate axes.
5. Proof of the change of reading strand transformation – Part 2: Mathematical proof of the change of reading strand transformation that is numerically illustrated in Qu 6.
6. 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.

Session 7

1. Understanding ground state and stiffness for palindromic and non-palindromic sequences. Moreover understanding the symmetries in its ground state and observing set of zero entries in its stiffness matrices.
2. 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.
3. Using the cgDNA web for visualising the variation in shape with different parameter set for a given sequence. Further using the cgDNA web for visualising left-handed and right-handed superhelices for given interesting sequences.
4. Various simulations with the cgDNA model such as looking for the differences in ground-states of cgDNA for parameter set 1 and 2. Notice that all the 6 distinct poly$(\alpha\beta)_N$ ground-states are helical but with noticeably different pitch and radii.
5. Downloading and compiling the cgDNAmc code : Preparation for the next exercise session.

Session 8

1. Monte Carlo simulation with the cgDNA model (using both Matlab and cgDNAmc) I: Monte Carlo sampling of frame configurations drawn from a cgDNA Gaussian pdf in internal coordinates **w**. Six poly-dinucleotide sequences have been considered. Note that for all the poly dimers the cloud of points produced by the n-th base-pair position look symmetric. Moreover the cloud gets broader for larger values of n.

2. Relative speeds of the MC sampling in Matlab versus compiling cgDNAmc code.

3. Monte Carlo simulation with the cgDNA model II : Same as in point 1 &2 of this session but for biological sequences (lambda phage) which exhibit strongly different behaviours compared to the poly dimer sequences considered in point 1 & 2 of this session.

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

Session 9

1. As discussed in the video lecture this week, one of the classical quantity in polymer physics is the Flory vector. And second classic quantity is the tangent-tangent correlation (ttc) and the associated persistence length. This exercise is mainly focus on computing these two classical quantity using the cgDNAmc code.

2. Computing persistence length (using cgDNAmc) I: Convergence study for the tangent–tangent correlation (ttc) using the Monte Carlo simulations. Numerical computation of ttc and the associated persistence length for six poly-dinucleotide. Computation of Flory persistence vector for six poly-dinucleotide sequences and comparison between parameter sets cgDNAps1 and cgDNAps2.

3. Explicit computation of apparent persistence length for a tractable probability density function (the HWLC) : Comparison of numerics with analytics (analytical formula) explained in lecture. Also one can see the document in the week-by-week correspondence of week 8.

4. Computing persistence length (using cgDNAmc) II: Computations of tangent–tangent correlation (ttc), factorized tangent-tangent correlation and Flory vectors for biological sequences (lambda phage).

Session 10

1. Banded matrices and their inverses : Characterising matrices with banded inverses, and computing banded matrices whose inverses are partially prescribed. A computational trick that is very important in being able to explicitly compute best estimates of stiffness matrices with a prescribed banded sparsity pattern.

2. Definition of Entropy and Relative entropy for continuous PDFs (ie for probability density functions on continuous, not discrete, state spaces). Explicit formulas in the case of Gaussians: Both entropy and relative entropy (or Kullback-Liebler divergence) as introduced here are used later in parameter estimation.

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.

4. Jensen's inequality - Part 1 : Classic estimate that we use later in maximum entropy estimation. Only applies for integrals with respect to a measure for which the volume of the domain is finite.

Session 11

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

2. Jensen's inequality - Part 2: This exercise is divided into four parts and uses Jensen's inequality to understand various properties of entropy and relative entropy (defined with respect to a measure $d\mu(x)$). Part 1 just confirms that the entropy integrand is a convex function (on R^+). Part 2 then applies Jensen to prove that the (scaled) entropy is non-negative for any domain with finite measure. Part 3 then rewrites the result of Part 2 in the form which states that the minimal entropy distribution is uniform. If the domain Ω is bounded this result applies to the standard Riemann measure dx. Part 4 shows that the relative entropy is always non-negative and is minimised when the two PDFs coincide. This result applies with the standard Riemann measure dx even if the domain Ω is unbounded.

3. 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 9. 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 as the translation-translation diagonal entries, which is convenient. However the invariance of the KL divergence means that the parameter fitting is fact unaffected.

4. Palindromic symmetry of a shape vector and stiffness matrix: How to use palindromic symmetry to improve estimates of oligomer-based Gaussian PDF. Uses relative entropy/KL divergence to quantify various errors/differences.

5. From atomistic representation to cgDNA internal coordinates: Part 1) visualising the variation in atomic coordinates between six snapshots of full atomistic 24 mer palindromic sequence. Part 2) frames fitting on each base of the sequence using the set of atomic coordinates corresponding to the snapshots. Part 3) computing the fitting error for each bases. Part 4) computation of the mid-frame and the junction frame using the frames found in part 2. Then computation of the cgDNA model internal coordinates for all the snapshots and visualising the noise between snapshots through plot of inters and intras. Part 5) checking that individual snapshots do not follow palindromic symmetry even though the sequence is palindrome. Further computation of equilibrium statistics (average of cgDNA model coordinates using very small number of snapshots and then checking the palindromic symmetry) shows that equilibrium statistics will satisfy the palindromic symmetry if the ensemble size is big. Overall, the aim of this exercise is to understand each steps in the process of computing cgDNA model internal coordinates from the real time series of MD snapshots.

Session 12

1. Gaussian integrals II : Computing the explicit formula for the marginals of a general Gaussian. It will be used later to get a cgDNA model restricted to a local subsequence.

2. On the computation of marginals of the cgDNA probability distribution : Part 1.1) The rigid basepair marginal of the cgDNA rigid base model are Gaussians, but with non banded stiffness matrices. Part 1.2) However, the localised cgDNA marginal over flanking sequences is itself a banded Gaussian. Marginalising over flanking sequence gives a localised version of the cgDNA pdf to a given window size which can be used to scan genomes for statistical mechanically exceptional regions.

3. Gaussian Integral III : Computing the explicit formula for the conditional of a general Gaussian. It is used in the next exercise of this session.

4. 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, where in the part of the DNA double helix bound to a protein the values of the cgDNA coordinates are prescribed.

Session 13

Spring 2019

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. First version of Euler-Rodrigues formula which gives a rotation matrix in terms of a unit rotation axis and a rotation angle in $[0, \pi]$ using the right handed rule (regle du tir bouchon). Moreover for each rotation through an angle less than π there is a unique unit rotation axis and angle.

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 the special cases for SO(3) and SE(3), which form the basis for the rotational coordinates in the cgDNA model.

Session 4

1. Additional properties (change of basis and inverse) of the Cayley transform of a proper rotation matrix.

2. The change of reading strand transformation – Part 1: In the first part we prove two useful properties of the Cayley transformation. In the second part we show that for two rigid-bodies there is a linear relationship between their internal coordinates and the internal coordinates of a transformation (on the right) by an orthogonal matrix on both rigid–body orientations. The fact that the relation is linear is due to the choice of writing the relative translation in the mid–frames.

3. Completing the square in vector quadratic forms: In this exercise we explicitly compute the stiffness matrix, the shift, and the constant term of a quadratic form defined as sum of (local) quadratic contributions. Explains why the ground state in the cgDNA model has nonlocal sequence dependence even though the parameter set is dimer dependent, i.e. very localised sequence dependence.

4. More on square roots of matrices: To have the simple symmetry rule in exercise 1), 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). In the second part of the exercise we show why the square root of a transformation in SE(3) is **not** used.

5. Optimisation in SE(3): From atomic coordinates of a given base we associated a rigid-body computed by solving an optimization problem in SE(3).

6. Review of various (standard) Matrix Factorisations. In particular we will use the Cholesky factorisation of a symmetric positive definite matrix (in our Monte Carlo code) and the polar decomposition (in fitting frames to atomistic data).

Session 5

1. cgDNAweb : A web site for visualizing 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.

Session6

1. Scaling and the Cayley transform: We introduce a scaled version for the Cayley transformation and its inverse. Used later to have stiffness matrices whose diagonal entries are of the same order of magnitude for rotational and translational degrees of freedom.
2. This exercise reveals the close connection between the Cayley transform and the matrix exponential by comparing the two respective matrix Taylor series with the choice of scaling $\alpha = 2$ in the Cayley transform (see exercise 1) For small rotations (expressed in radians) the components of the vector generating the skew matrix in both the exponential parametrization of SO(3) and the Cayley transformation (with $\alpha = 2$) are close to rotation angles about the coordinate axes.
3. 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.
4. Proof of the change of reading strand transformation – Part 2: Mathematical proof of the change of reading strand transformation that is numerically illustrated in Qu 5.
5. 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.
6. Downloading and compiling the cgDNAmc code : Preparation for the next exercise session.

Session 7

1. Various Monte Carlo simulations with the cgDNA model: Purposes 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 gets broader for larger values of n.
   - A second classic quantity is the tangent-tangent correlation and the associated persistence length (as will also be 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 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 session 6 and session 7 is discussed in the solution sheet of this exercise.

2. Monte Carlo simulation with the cgDNA model II : Same as exercise 1) session 7 but for biological sequences which exhibit strongly different behaviours compared to the poly dimer sequences considered session 7.

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

Session 9

1. Banded matrices and their inverses : Characterising matrices with banded inverses, and computing banded matrices whose inverses are partially prescribed. A computational trick that is very important in being able to explicitly compute best estimates of stiffness matrices with a prescribed banded sparsity pattern.

2. Definition of Entropy and Relative entropy for continuous PDFs (ie for probability density functions on continuous, not discrete, state spaces). Explicit formulas in the case of Gaussians: Both entropy and relative entropy (or Kullback-Liebler divergence) as introduced here are used later in parameter estimation.

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.

4. Jensen's inequality - Part 1 : Classic estimate that we use later in maximum entropy estimation. Only applies for integrals with respect to a measure for which the volume of the domain is finite.

Session 10

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

2. Jensen's inequality - Part 2: This exercise is divided into four parts and uses Jensen's inequality to understand various properties of entropy and relative entropy (defined with respect to a measure $d\mu(x)$). Part 1 just confirms that the entropy integrand is a convex function (on R^+). Part 2 then applies Jensen to prove that the (scaled) entropy is non-negative for any domain with finite measure. Part 3 then rewrites the result of Part 2 in the form which states that the minimal entropy distribution is uniform. If the domain Ω is bounded this result applies to the standard Riemann measure dx. Part 4 shows that the relative entropy is always non-negative and is minimised when the two PDFs coincide. This result applies with the standard Riemann measure dx even if the domain Ω is unbounded.

3. 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 9. 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 as the translation-translation diagonal entries, which is convenient. However the invariance of the KL divergence means that the parameter fitting is fact unaffected.

4. Palindromic symmetry of a shape vector and stiffness matrix: How to use palindromic symmetry to improve estimates of oligomer-based Gaussian PDF. Uses relative entropy/KL divergence to quantify various errors/differences.

Session 11

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

2. 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) However the localised cgDNA marginal over flanking sequences is itself a banded Gaussian. Marginalising over flanking sequence gives a localised version of the cgDNA pdf to a given window size which can be used to scan genomes for statistical mechanically exceptional regions.

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

4. 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, where in the part of the DNA double helix bound to a protein the values of the cgDNA coordinates are prescribed. But we do not discuss this idea any further in the lectures (only fourteen weeks ...).

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 prescribing all first moments, and second (centred) moments within a block band. Same answer as the max likelihood case done in class, but for max ent the fact that the optimal pdf is a banded Gaussian is a conclusion not an assumption.

2. 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.

3. Zero entries in the cgDNA parameter set: The relation discussed in ex2 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.

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

Session 13

This session just completes various points that arose during the lectures of the semester

1. On the average of rotation matrices sharing a common (deterministic) axis : Except for this very special ensemble $||\langle Q \rangle|| < 1$. Clarifies that the average of an ensemble of rotation matrices is NOT a rotation matrix unless all members of the ensemble share a common rotation axis

2. On the parametrization of base pair frames using quaternions (a final exercise pertaining to efficiency in the cgDNAmc code). Quaternions (also known as Euler parameters, which are different from the much more commonly encountered Euler angles) are a four parameter set of coordinates on the proper rotation group SO(3). They are closely related to the Cayley vector coordinates we use for junctions. There is a large literature on quaternions which we do not enter into. Briefly they eliminate the singularity that arises at rotations through π but at the price of introducing a double covering of the group. In the context of cgDNAmc quaternions are used because there is a composition rule directly on the coordinates of two rotation matrices to obtain the coordinates of the product of the two matrices. When multiplying lots of rotation matrices (such as in cgDNAmc when reconstructing base pair frames along the chain, or in computer graphics games) it is significantly faster to multiply the quaternion coordinates than to multiply the matrices themselves.

3. In the last lecture we discussed that Kullback-Leibler divergence could be used as an objective function in fitting PDFs in two ways depending on the order in which the arguments model-pdf/target-pdf are taken in the KL divergence. Because the KL divergence is asymmetric you get different best fits. This exercise is an example taken from the cgDNA model which shows that the two fits can be really quite different in physically significant ways.