DNA Modelling

1 What Can Be Done With the cgDNA Model?

These notes can be regarded as a polycopie for the lectures given on 3.11.17 and 10.11.17. Thank you to Alastair Flynn for typing a first draft.

The long-term goal (ideally of any mathematical model) is to predict or explain/interpret experimental data or observations. We will discuss (briefly due to available time in the semester) two experimental contexts in which the *cgDNA* model can be applied:

A Looping or cyclisation experiments

B The notion of persistence lengths

The length scale of experiments of type A is usually 100-500 basepairs or so, while the scale of type B is usually 100-1500 basepairs, both of which are well beyond atomistic MD simulations even for short durations in time, so a coarse-grain model is required, and we will of course use the *cgDNA* model to run some simulations. Modelling looping involves estimating probabilities, which we will consider later, while modelling persistence lengths involves estimating expectations, which we will consider first.

Let $\langle\cdot\rangle$ denote expectation with respect to the *cgDNA* pdf (for a given sequence S and parameter set $\mathcal{P}$) which we recall takes the form:

$$\rho(\mathbf{w};S,\mathcal{P}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})K(\mathbf{w}-\boldsymbol{\mu})\right\}.$$

Explicitly the expectation of any given function $\phi(\mathbf{w})$ is

$$\langle\phi(\mathbf{w})\rangle := \frac{1}{Z}\int\cdots\int \phi(\mathbf{w})\exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})K(\mathbf{w}-\boldsymbol{\mu})\right)\,\mathrm{d}\mathbf{w}.$$

Then (standard Gaussian integral formulas as described in Series 1 state that) $\langle\mathbf{w}\rangle = \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is just the minimum, or ground state, of the 'free energy'

$$\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})K(\mathbf{w}-\boldsymbol{\mu}).$$

We will discuss units and scalings in this free energy during the parameter estimation part of the course later in the semester.

Experimental data from both X-ray crystallography and NMR (nuclear magnetic resonance), which is usually at the scales of 10-20 bp, provides structural data in the form of PDB files of atomic Cartesian coordinates of an "average" structure that can be fit to a *cgDNA* configuration $\mathbf{w}^*$. It then seems reasonable to directly compare the configuration coordinates $\mathbf{w}^*$ of such an experimentally observed "average" structure with the *cgDNA* ground state $\boldsymbol{\mu} = \langle\mathbf{w}\rangle$. The resulting comparison seems quite encouraging. In particular the observed errors seem to be consistently smaller than the variation of

the data with local sequence composition, see for example data presented in [1]. In particular in this structural, short length scale, experimental data it is possible to make a full comparison between experimental and coarse grain predictions for the expectations of both the intra and inter base parts of the *cgDNA* configuration variables. However it is unclear how to compare experimental data of this type with *cgDNA* predictions of the stiffness matrix K.

In contrast to X-ray crystallography and NMR data, both looping and persistence length experiments only involve the rigid basepair frame orientation and position $(R_i, \mathbf{r}_i) \in SO(3) \times \Re^3$ $i = 2, \ldots, n$, which in turn can be reconstructed as a (highly non-linear and non-local) function of only the inter variables in $\mathbf{w}$, with one six vector for each of the $(n-1)$ junctions, plus the values of $(R_1, \mathbf{r}_1)$. We will therefore compute expectations $\langle R_i, \mathbf{r}_i \rangle$ by construction of the corresponding functions $\phi(\mathbf{w})$, where in fact ϕ depends on only the inter part of $\mathbf{w}$. Within the *cgDNA* model these expectations will be computed numerically via Monte Carlo sampling, while in a simplified version of the *cgDNA* model that corresponds to a version of the classic worm like chain, we will approximate the expectations analytically to reveal connexions with the entries in the stiffness matrix for the simplified model that is analogous to, but much simpler than, the *cgDNA* stiffness matrix K.

The fact that the *cgDNA* pdf ρ is Gaussian implies that we can efficiently draw an ensemble of configurations $\left\{\mathbf{w}^{[m]}\right\}_1^M$ from the pdf. Such an algorithm is implemented in the *cgDNAmc* code, which e.g. can generate a million or so samples (i.e. $M = 10^6$) in only a few minutes on a contemporary laptop for a 300 bp fragment (of a given sequence S and for a given parameter set $\mathcal{P}$). This level of efficiency is only possible for the part of the *cgDNAmc* code (the only part that will be used in the course) that implements a direct Monte Carlo method which uses linear algebra to reduce the multivariate *cgDNA* Gaussian to a product of scalar normal (i.e. one-dimensional Gaussian) distributions that can be sampled by general purpose, highly optimised, random number generators. The efficiency of the direct sampling of *cgDNAmc* also depends crucially on the fact that the *cgDNA* stiffness matrix K is sparse, and, more specifically, banded around the principal diagonal. There will be some further comments about this feature later.

Once we have a (sufficiently large) ensemble $\left\{\mathbf{w}^{[m]}\right\}_1^M$ of configurations drawn from the *cgDNA* pdf, we can easily approximate both probabilities (as necessary for modelling looping experiments, essentially by counting how many configurations satisfy some event, more comments later), and expectations. An expectation $\langle \phi(\mathbf{w}) \rangle$ is estimated as:

$$\langle \phi(\mathbf{w}) \rangle \approx \frac{1}{M} \sum_{j=1}^{M} \phi(\mathbf{w}^j) \tag{1}$$

where $\mathbf{w}^j : j = 1, \ldots, M$ are the samples. Two classic choices of ϕ leading to the notion of persistence length are:

1. $\phi(\mathbf{w}) = R_1^{\mathsf{T}}(\mathbf{r}_n - \mathbf{r}_1)$ i.e. the components of the end-to-end chord vector with respect to the first basepair frame. These triples of numbers are called Flory vectors. Note that for components with respect to the lab (or simulation cell) frame, we have automatically that $\langle \mathbf{r}_n - \mathbf{r}_1 \rangle = 0$ by invariance of the free energy of the structure with coordinates $\mathbf{w}$ under an overall rotational symmetry in the absence of any external loading, while $(\mathbf{r}_n - \mathbf{r}_1)$ changes under such a rotation. In other words there is a uniform pdf to overall rotations, and therefore $\langle \mathbf{r}_n - \mathbf{r}_1 \rangle = 0$. In contrast the triple $\phi(\mathbf{w})$ are the components of the relative end-to-end displacement in the first basepair frame, which are unchanged by an overall rotation, so their average can be non vanishing.

2. $\phi(\mathbf{w}) = R_1^{\mathsf{T}} R_n$, which is the relative rotation (on the right) from the first base-pair orientation R_1 to the nth base-pair orientation R_n. As before the average $\langle R_n \rangle$ of the absolute orientation of the nth base pair frame must vanish, while (for fixed n) the expectation of the relative rotation $\langle R_1^{\mathsf{T}} R_n \rangle$ need not *a priori* vanish. However we will see that as $n \to \infty$ $\langle R_1^{\mathsf{T}} R_n \rangle$ does in fact vanish in a very specific manner.

The functions ϕ given in 1 & 2 comprise a 3-vector and a 3×3 matrix for each base pair $n=2,\ldots,N$, and $\langle R_1, \mathbf{r}_1\rangle$ can be regarded as fixed to eliminate the overall translation and rotation. The functions ϕ can be evaluated relatively efficiently using the linear algebra associated with $SE(3)$ that was introduced earlier. Specifically let

$$X_n = \begin{pmatrix} R_n & \mathbf{r}_n \\ \mathbf{0} & 1 \end{pmatrix} \in SE(3) \qquad n=2,\ldots N,$$

be the absolute orientation and position of the nth base pair and

$$A_n = \begin{pmatrix} Q_n & \mathbf{q}_n \\ \mathbf{0} & 1 \end{pmatrix} \in SE(3) \qquad n=1,\ldots N-1,$$

be the junction displacements so that we have the recursion

$$X_n = X_{n-1}A_{n-1}, \qquad n=2,\ldots N.$$

Then $X_n = X_1 A_1 \cdots A_{n-1}$ so

$$X_1^{-1}X_n = \prod_{k=1}^{n-1} A_k.$$

And finally for any given configuration $\mathbf{w}$ we have (for $n\geq 3$ with $n=2$ being a simple special case in which the summation notation for δ_{n-1} degenerates)

$$\begin{bmatrix} R_1^\mathsf{T}R_n & R_1^\mathsf{T}(\mathbf{r}_n-\mathbf{r}_1) \\ \mathbf{0} & 1 \end{bmatrix} = X_1^{-1}X_n = \prod_{k=1}^{n-1} A_k = \begin{bmatrix} \prod_{k=1}^{n-1} Q_k(\mathbf{u}_k) & \delta_{n-1} \\ \mathbf{0} & 1 \end{bmatrix}$$

where

$$\begin{aligned}\delta_{n-1} &= \mathbf{q}_1(\mathbf{u}_1,\mathbf{v}_1) + \sum_{k=1}^{n-2}\left\{\prod_{j=1}^{k} Q_j(\mathbf{u}_j)\right\}\mathbf{q}_{k+1}(\mathbf{u}_{k+1},\mathbf{v}_{k+1}) \\ &= \mathbf{q}_1(\mathbf{u}_1,\mathbf{v}_1) + Q_1(\mathbf{u}_1)\,\mathbf{q}_2(\mathbf{u}_2,\mathbf{v}_2) + \cdots + \{Q_1(\mathbf{u}_1)\cdots Q_{n-2}(\mathbf{u}_{n-2})\}\,\mathbf{q}_{n-1}(\mathbf{u}_{n-1},\mathbf{v}_{n-1}),\end{aligned}$$

and we have explicitly added the dependence of each quantity on the configuration variable $\mathbf{w}$ with the notation that $\mathbf{u}_i$ is the Cayley vector in the ith junction and $\mathbf{v}_i$ is the translation coordinates in the ith junction i.e. the six inter variables in the ith junction. Each matrix $Q_k(\mathbf{u}_k)\in SO(3)$ depends only on its junction Cayley vector, while each vector $\mathbf{q}_k(\mathbf{u}_k,\mathbf{v}_k)$ depends on all six junction coordinates. Note also that multiplication of rotation matrices does not in general commute (in fact only if there is a common rotation axis). This means that the product notation $\prod_{k=1}^{n-1} Q_k$ has to imply a prescribed ordering, which we take to mean indices increasing from left to right, so $\prod_{k=1}^{n-1} Q_k = Q_1 Q_2 \ldots Q_{n-1}$. These formulas are easily verified by induction on the number of basepairs using block matrix multiplication.

We want to compute

$$\begin{bmatrix} \langle R_1^\mathsf{T}R_n\rangle & \langle R_1^\mathsf{T}(\mathbf{r}_n-\mathbf{r}_1)\rangle \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \langle \prod_{i=k}^{n-1} Q_k(\mathbf{u}_k)\rangle & \langle\delta_{n-1}\rangle \\ \mathbf{0} & 1 \end{bmatrix}, \tag{2}$$

where

$$\langle\delta_{n-1}\rangle = \langle\mathbf{q}_1(\mathbf{u}_1,\mathbf{v}_1)\rangle + \sum_{k=1}^{n-2}\left\langle\left\{\prod_{j=1}^{k} Q_j(\mathbf{u}_j)\right\}\mathbf{q}_{k+1}(\mathbf{u}_{k+1},\mathbf{v}_{k+1})\right\rangle$$

for expectations $\langle\cdot\rangle$ with respect to a given pdf (and the locations of the expectation brackets $\langle\cdot\rangle$ should be noted carefully, they distribute over + but not over any form of matrix-matrix or matrix-vector multiplication). For the *cgDNA* model pdf the distribution of each set of junction variables $(\mathbf{u}_k,\mathbf{v}_k)$ is not independent of all the other junction variables. Consequently the only currently known possibility of evaluating these formulas is to approximate the expectations numerically using the *cgDNAmc* code

to a) generate a sample ensemble $\left\{\mathbf{w}^{[m]}\right\}_1^M$, b) for each $\mathbf{w}^{[m]}$ carry out all the matrix and vector multiplications to reconstruct the frames $(R_i, \mathbf{r}_i) \in SO(3) \times \Re^3$, c) average over all reconstructions, i.e. implement the approximation (1). The *cgDNAmc* code is all set up to do this automatically, and the approximate benchmarks stated above include the reconstruction computations. In the exercises you will run *cgDNAmc* and plot as output $\ln\langle R_1^\mathsf{T} R_n\rangle_{(3,3)}$ and $\langle R_1^\mathsf{T}(\mathbf{r}_n - \mathbf{r}_1)\rangle$ as functions of n. You will observe that for many sequences the former scalar semi-log plot is close to a linearly decreasing function (which provides an exponential rate at which $\langle R_1^\mathsf{T} R_n\rangle_{(3,3)}$ vanishes), while the latter vector tends to a fixed point $\mathbf{0} \neq \delta \in \Re^3$, both as $n \to \infty$ (which means in practice for n sufficiently large).

It turns out that these two behaviours are features of linear chains of rigid body frames that are quite robust to changes in the precise model, i.e. to the specific pdf that is assumed to describe the statistics of the polymer chain, and which is used in the evaluation of the expectations $\langle\cdot\rangle$. To obtain more insight into why the behaviours arise, we can actually derive formulas implying both observations, but only in the context of a simpler model, which is finally a version of the Helical Worm Like Chain (or HWLC) of Yamakawa [2] which is itself a slightly more detailed version of the Twisted Worm Like Chain (TWLC) or just the Worm Like Chain (WLC) models that are discussed in many places in the polymer physics literature. The following discussion is largely based on Chapters 2 and 7 of [3].

The main simplifying assumption that can be made (and which was proposed by Schellman and Flory if not earlier) is that the statistics of each junction are independent one from another. In other words each junction is independently distributed or i.d. When the junctions are i.d. (which we stress is not the case for the *cgDNA* model pdf) then (2) simplifies (greatly) to assume the form

$$\begin{bmatrix} \langle R_1^\mathsf{T} R_n\rangle & \langle R_1^\mathsf{T}(\mathbf{r}_n - \mathbf{r}_1)\rangle \\ \mathbf{0} & 1 \end{bmatrix} = \left\langle \prod_{k=1}^{n-1} A_k \right\rangle = \prod_{k=1}^{n-1} \langle A_k\rangle = \begin{bmatrix} \prod_{i=1}^{n-1}\langle Q_i\rangle & \langle\delta_{n-1}\rangle \\ \mathbf{0} & 1 \end{bmatrix}$$

where now

$$\langle\delta_{n-1}\rangle = \langle\mathbf{q}_1\rangle + \sum_{k=1}^{n-2}\left\{\prod_{j=1}^{k}\langle Q_j\rangle\right\}\langle\mathbf{q}_{k+1}\rangle\,.$$

If in addition the chain is *uniform* (which we again emphasize is also not the case for the *cgDNA* model pdf), the statistical behaviours of each junction are both independent and identical. That is the junction statistics are i.i.d., which means that there is a single matrix $\langle Q\rangle$ and vector $\langle\mathbf{q}\rangle$ such that $\langle Q_i\rangle = \langle Q\rangle$ and $\langle\mathbf{q}_i\rangle = \langle\mathbf{q}\rangle$, $\forall i$. Now (2) reduces to

$$\begin{bmatrix} \langle R_1^\mathsf{T} R_n\rangle & \langle R_1^\mathsf{T}(\mathbf{r}_n - \mathbf{r}_1)\rangle \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \langle Q\rangle^{n-1} & \langle\delta_{n-1}\rangle \\ \mathbf{0} & 1 \end{bmatrix}$$

where now $\langle\delta_{n-1}\rangle = \left(\sum_{k=1}^{n-1}\langle Q\rangle^{k-1}\right)\langle\mathbf{q}\rangle$ (and where we have now adopted the notation $\langle Q\rangle^0 = I$). Using the matrix version of the summation formula for a finite geometric series we have

$$\langle\delta_{n-1}\rangle = (I - \langle Q\rangle)^{-1}(I - \langle Q\rangle^{n-1})\langle\mathbf{q}\rangle.$$

The proof of the matrix geometric series summation formula is identical to the better known scalar case that you see in secondary school. Its validity depends only on the assumption that the matrix $(I - \langle Q\rangle)$ is invertible. We note that for any $Q \in SO(3)$, $(I - Q)$ is certainly not invertible, because Q has $\lambda = 1$ as an eigenvalue so that $(I - Q)$ has a one-dimensional null space. However $Q \in SO(3) \not\Rightarrow \langle Q\rangle \in SO(3)$. In fact if any two Q_i in the ensemble generating $\langle\cdot\rangle$ have different axes of rotation then $||\langle Q\rangle|| < 1$ (where $||\cdot||$ is the spectral radius norm) so that $(I - \langle Q\rangle)^{-1}$ exists and entry-wise $\langle Q\rangle^k \to 0$ as $k \to \infty$. The norm inequality can be seen from the following argument. Suppose $\langle Q\rangle = \sum_1^M Q_i/M$. Then for any unit vector $\mathbf{x}$, $\langle Q\rangle\mathbf{x} = \sum_1^M Q_i\mathbf{x}/M = \sum_1^M \mathbf{y}_i/M$ where for each i, $\mathbf{y}_i := Q_i\mathbf{x}$ is a unit vector, because it is a rotation of a unit vector. But the unit sphere is convex in the sense that any convex linear combination of two or more distinct unit vectors on the surface of the sphere lies strictly in the interior of the unit ball. And there will be at least two (and usually many) distinct vectors $\mathbf{y}_i$ unless

all of the Q_i have $\mathbf{x}$ as a common rotation axis. Consequently except in this degenerate case $\langle Q\rangle\mathbf{x}$ is strictly in the interior of the unit ball for any unit vector $\mathbf{x}$, so that $||\langle Q\rangle|| = \sup_{\mathbf{x}} ||\langle Q\rangle\mathbf{x}||/||\mathbf{x}|| < 1$ as desired. Just with this bound on $||\langle Q\rangle||$ we may therefore conclude:

$$\begin{bmatrix} \langle Q\rangle^{n-1} & \langle \delta_{n-1}\rangle \\ \mathbf{0} & 1 \end{bmatrix} \to \begin{bmatrix} 0 & \delta \\ \mathbf{0} & 1 \end{bmatrix}$$

as $n \to \infty$ where $\delta := (I - \langle Q\rangle)^{-1}\langle \mathbf{q}\rangle$ is called the Flory vector.

The above conclusions are rather strong. They are based on the rather strong hypothesis that the junction statistics are i.i.d. In fact similar results are obtainable even for non-uniform chains, i.e. the case where the junction statistics are independent but not identical. But we will not pursue that line further. It is also the case that the conclusions are quite general in the sense that they assume almost nothing about the actual i.i.d. junction distribution, only the very weak restriction that $||\langle Q\rangle|| < 1$ which is essentially a non-degeneracy condition. In particular it is not assumed that the junction statistics are of any particular functional form, e.g. Gaussian.

We next show that if the junction statistics are in fact Gaussian, and the chain is sufficiently stiff (in a sense that will be made precise below), then we can approximate the entries in $\langle Q\rangle$ (and therefore the decay rates in $\langle Q\rangle^n$) as an explicit function of the coefficients in the Gaussian pdf. This will make a connexion to the classic notion of persistence length in a polymer.

We will assume that the junction displacements $A \in SE(3)$ are identically and independently distributed in the form

$$A(\mathbf{u}, \mathbf{q}) = \begin{bmatrix} Q(\mathbf{u}) & \mathbf{q} \\ \mathbf{0} & 1 \end{bmatrix}$$

where $\mathbf{q} \in \Re^3$ is constant and

$$Q(\mathbf{u}) = \frac{1 - \|\mathbf{u}\|^2}{1 + \|\mathbf{u}\|^2} I + \frac{2}{1 + \|\mathbf{u}\|^2}\mathbf{u}^\times + \frac{2}{1 + \|\mathbf{u}\|^2}\mathbf{u} \otimes \mathbf{u} \tag{3}$$

where $\mathbf{u}$ is drawn from the multivariate, but decoupled or factorisable or diagonal, Gaussian distribution

$$\rho(\mathbf{u}) = \frac{1}{Z} \exp\left(-\frac{1}{2}(K_1 u_1^2 + K_2 u_2^2 + K_3(u_3 - \hat{u}_3)^2)\right). \tag{4}$$

We allow a non-zero shift in u_3 because DNA is highly twisted, ad we will study the consequences of the presence of $\hat{u}_3$, which need not be small. On the other hand DNA is only believed to be slightly bent, so in this simple model the shifts in u_1 and u_2 are in the first instance assumed to vanish. This is just the Euler-Rodrigues formula (of Series 3) for the rotation matrix Q as a function of the Cayley vector $\mathbf{u}$ for the junction, where $\mathbf{u}$ is distributed according to the Gaussian (4). This model has seven constant parameters K_1, K_2, K_3, $\hat{u}_3$ and the three components of $\mathbf{q}$. It falls in the family of Helical Worm Like Chains, as its ground state has the points $\mathbf{r}_i$ lying on a circular helix. The model is not an immediate special case of *cgDNA* in part because assuming $\mathbf{q}$ constant breaks the Crick-Watson symmetry that is carefully built in to *cgDNA*. Assuming $\mathbf{q} = Q(\mathbf{u})^{\frac{1}{2}}\hat{\mathbf{v}}$, for $\hat{\mathbf{v}}$ constant would restore Crick-Watson symmetry and would correspond to freezing the *cgDNA* inter translation variables to be constant. It is possible to make analogous computations to what follows in such more complicated cases, including allowing a nondiagonal Gaussian with a more general shift instead of the specific case (4) (see for some examples [3]) but we will restrict attention to the simpler case, as it is quite standard in the literature.

We also concentrate on computing the specific matrix entry $\langle R_1^\mathsf{T} R_n\rangle_{3,3} = \langle \mathbf{d}_3^{[1]} \cdot \mathbf{d}_3^{[n]}\rangle$, which is called the tangent-tangent correlation, because it is assumed that the $\mathbf{d}_3^{[n]}$, i.e. the third columns of the rotation matrices $\mathbf{R}_n$ have been chosen in such a way that they are an approximation to the tangents of a curve interpolating the points $\mathbf{r}_n$. Because the $\mathbf{d}_3^{[n]}$ are unit vectors for all n, the expectation $\langle \mathbf{d}_3^{[1]} \cdot \mathbf{d}_3^{[n]}\rangle$ can be re-written as $\langle \cos\theta_{1,n}\rangle$, where $\cos\theta_{1,n} := \mathbf{d}_3^{[1]} \cdot \mathbf{d}_3^{[n]}$ is just the cosine of the angle between

the two unit vectors. Numerical simulations with *cgDNAmc* reveal that for the *cgDNA* model pdf of many sequences S, semi-log plots in the form of $\ln\langle\cos\theta_{1,n}\rangle$ versus n are reasonably close to linear. We will show that such a relation is exact in the HWLC model described above, and compute an approximation for the slope of the line in terms of the model parameters, provided that the stiffnesses K_i are sufficiently large.

We first show that in our simplified HWLC model the sparsity pattern of $\langle Q\rangle$ is of the form

$$\langle Q\rangle = \begin{pmatrix} * & * & 0 \\ * & * & 0 \\ 0 & 0 & * \end{pmatrix},$$

which is sufficient to show that $\langle Q\rangle^n$ is of the form

$$\langle Q\rangle^n = \begin{pmatrix} * & * & 0 \\ * & * & 0 \\ 0 & 0 & \langle Q\rangle_{3,3}^n \end{pmatrix}.$$

And this implies that

$$\ln\langle\cos\theta_{1,n}\rangle = n\ln\langle Q\rangle_{3,3}$$

which is an exactly linear function of n with a negative slope provided that $0 < \langle Q\rangle_{3,3} < 1$.

To prove the sparsity pattern we observe that

$$\mathbf{u}^\times = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}, \qquad \mathbf{u}\otimes\mathbf{u} = \begin{pmatrix} u_1^2 & u_1u_2 & u_1u_3 \\ u_1u_2 & u_2^2 & u_2u_3 \\ u_1u_3 & u_2u_3 & u_3^3 \end{pmatrix}.$$

Considering first the entry $\langle Q\rangle_{1,3}$ we see from (3) that we have to evaluate the sum of the two expectations $\left\langle \frac{u_2}{1+\|\mathbf{u}\|^2}\right\rangle$ and $\left\langle \frac{u_1\,u_3}{1+\|\mathbf{u}\|^2}\right\rangle$. Explicitly

$$\left\langle \frac{u_2}{1+\|\mathbf{u}\|^2}\right\rangle = \frac{1}{Z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\underbrace{\int_{-\infty}^{\infty}\frac{u_2}{1+\|\mathbf{u}\|^2}\exp\left(-\frac{1}{2}K_2u_2^2\right)\,\mathrm{d}u_2}_{=0}\exp\left(-\frac{1}{2}(K_1u_1^2+K_3(u_3-\hat{u}_3)^2)\right)\,\mathrm{d}u_1\,\mathrm{d}u_3,$$

where we have used the fact that the diagonal Gaussian pdf (4) factors so that the expectation can be written as nested single integrals in whatever order we choose. And for this ordering of the integration, the interior $\mathrm{d}u_2$ integral vanishes because (for each u_1 and each u_3) the integrand is the product of an even and odd function of u_2 and the range of integration is all of $\Re$. Similarly $2\left\langle \frac{u_1\,u_3}{1+\|\mathbf{u}\|^2}\right\rangle = 0$ by an even-odd argument applied to the ordering in which the interior one-dimensional integral is with respect to $\mathrm{d}u_1$. Similar odd-even arguments imply that both contributions to the $\langle Q\rangle_{2,3}$ entry vanish. Then trivially the arguments extend to the $\langle Q\rangle_{3,1}$ and $\langle Q\rangle_{3,2}$ entries, so that we may conclude the desired sparsity pattern of $\langle Q\rangle$. Regarding $\langle Q\rangle_{1,2}$ we note that odd-even arguments also imply that $2\left\langle \frac{u_1\,u_2}{1+\|\mathbf{u}\|^2}\right\rangle = 0$. However the contribution $2\left\langle \frac{u_3}{1+\|\mathbf{u}\|^2}\right\rangle$ from $\mathbf{u}^\times$ does not vanish whenever $\hat{u}_3 \neq 0$, because then there is no single integrand that is a product of an even and odd function. The value of this entry can be approximated when the stiffnesses are large by the method we introduce next, but we do not pursue that computation. In fact a slightly more refined version of the above argument shows that the sparsity pattern of $\langle Q\rangle$ actually does not depend on the assumed diagonal form of the stiffness matrix K in (4); the sparsity pattern only depends on the vanishing of the offsets or shifts in the Gaussian free energy $\hat{u}_1 = \hat{u}_2 = 0$. However in the next step our formulas would be slightly more complicated in the non-diagonal case.

Now we compute $\langle Q\rangle_{3,3}$. We will make use of:

Remark. (One-dimensional Gaussian expectation formulae)

$$\frac{1}{Z}\int_{-\infty}^{\infty} u^p \exp\left(-\frac{1}{2}Ku^2\right)\,\mathrm{d}u = \begin{cases} 0 & p \text{ is odd} \\ \frac{(2j)!}{2^j\,j!}K^{-j} & p = 2j \end{cases}$$

where the partition function is (see Series 1)

$$Z = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}Ku^2\right)\,\mathrm{d}u = \frac{\sqrt{2\pi}}{\sqrt{K}}.$$

We will use the above formulas to conclude that for $K \gg 1$, and $p \geq 3$, $\langle u^p \rangle$ is negligible compared to $\langle u^2 \rangle$. In fact the (multi-variate) version of the cases $p = 1, 2$ already appeared in Series 1, and for all p odd the integrals vanish by even-odd arguments. Formulas for the multi-variate case for even p also exist for $p \geq 4$, but the notation becomes a little intricate, and we only need the uni-modal versions. We will also introduce the technique of parameter differentiation (which was apparently a favourite of Richard Feynman) to derive the expressions. Explicitly the expression for the partition function can be differentiated in two ways

$$\begin{aligned}\sqrt{2\pi}\left(-\frac{1}{2}\right)\frac{1}{K^{\frac{3}{2}}} &= \frac{\partial Z}{\partial K} \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial K}\exp\left(-\frac{1}{2}Ku^2\right)\,\mathrm{d}u \\ &= -\frac{1}{2}\int_{-\infty}^{\infty} u^2 \exp\left(-\frac{1}{2}Ku^2\right)\,\mathrm{d}u\end{aligned}$$

Therefore after multiplying both sides by -2, and dividing both sides by Z (using its explicit expression) we find the case $p = 2$ or $j = 1$:

$$\frac{1}{Z}\int_{-\infty}^{\infty} u^2 \exp\left(-\frac{1}{2}Ku^2\right)\,\mathrm{d}u = \frac{1}{K}.$$

The general formula for $j+1$ follows by induction after differentiating with respect to K the identity for the case j in the form:

$$\frac{(2j)!\,\sqrt{2\pi}}{2^j\,j!}K^{-(j+\frac{1}{2})} = \int_{-\infty}^{\infty} u^{2j}\exp\left(-\frac{1}{2}Ku^2\right)\,\mathrm{d}u,$$

and simplifying (including dividing both sides by Z using its known expression).

Remark. We will also make use of the trivial algebraic identity for any two vectors $\mathbf{u}$ and $\hat{\mathbf{u}}$:

$$\|\mathbf{u}\|^2 = \|(\mathbf{u}-\hat{\mathbf{u}}) + \hat{\mathbf{u}}\|^2 = \|\hat{\mathbf{u}}\|^2 + 2\hat{\mathbf{u}}\cdot(\mathbf{u}-\hat{\mathbf{u}}) + \|\mathbf{u}-\hat{\mathbf{u}}\|^2.$$

We are now ready to compute an approximate expression for $\langle Q(\mathbf{u})\rangle_{3,3}$. From (3) we have that

$$\begin{aligned}Q(\mathbf{u})_{3,3} &= \frac{1-\|\mathbf{u}\|^2}{1+\|\mathbf{u}\|^2} + \frac{2}{1+\|\mathbf{u}\|^2}u_3^2 \\ &= \frac{1+u_3^2-u_1^2-u_2^2}{1+u_3^2+u_1^2+u_2^2} \\ &= \frac{1+\hat{u}_3^2+2\hat{u}_3(u_3-\hat{u}_3)+(u_3-\hat{u}_3)^2-u_1^2-u_2^2}{1+\hat{u}_3^2+2\hat{u}_3(u_3-\hat{u}_3)+(u_3-\hat{u}_3)^2+u_1^2+u_2^2} \\ &= \frac{1+\beta(\overbrace{2\hat{u}_3(u_3-\hat{u}_3)+(u_3-\hat{u}_3)^2-u_1^2-u_2^2}^{a_-})}{1+\beta(\underbrace{2\hat{u}_3(u_3-\hat{u}_3)+(u_3-\hat{u}_3)^2+u_1^2+u_2^2}_{a_+})},\end{aligned} \tag{†}$$

where

$$0 < \beta := \frac{1}{1+\hat{u}_3^2} \leq 1.$$

Now if $K_i \gg 1$ for $i = 1, 2, 3$ then the general idea of the Laplace Method for approximating integrals with an exponential in the integrand is that the only significant contribution to $\langle Q(\mathbf{u})\rangle_{3,3}$ will be when a_+, and therefore also a_- are small, because then the integration variable will be close to the maximum of the pdf (4) and there is rapid decay away from the peak. This means that it is plausible (and error bounds can be constructed) that we can approximate the integrand in $\langle Q(\mathbf{u})\rangle_{3,3}$ by making a Taylor expansion in a_+ of the denominator in (†) to obtain

$$\begin{aligned}(\dagger) &\approx (1+\beta a_-)(1-\beta a_+ + \beta^2 a_+^2 - \beta^3 a_+^3 + \cdots) \\ &= 1 + \underbrace{\beta(a_- - a_+)}_{-2\beta(u_1^2+u_2^2)} + \underbrace{\beta^2(a_+^2 - a_- a_+)}_{\text{cubic + H.O.T.}} - \underbrace{\beta^3(\cdots)}_{\text{cubic + H.O.T.}} + \cdots\end{aligned}$$

where we note that in the β term all dependence on $(u_3 - \hat{u}_3)$ cancels because of the particular expressions for a_+ and a_-, to leave the simple explicit expression that is given. Similarly in the β^2 term the dependence on $(u_3 - \hat{u}_3)^2$ cancels because of the particular expressions for a_+ and a_- to leave only cubic and higher order terms (or H.O.T.) where we count the total degree of the multi-variate polynomials in $(u_3 - \hat{u}_3)$, u_1 and u_2 that arise. The β^3 and higher terms are cubic plus H.O.T. by definition.

Now we can evaluate the approximation to $\langle Q\rangle_{3,3}$ using the factored form of the Gaussian (4) and the explicit formulas for univariate expectations to obtain

$$\begin{aligned}\langle Q\rangle_{3,3} &\approx \langle 1 - 2\beta(u_1^2+u_2^2)\rangle + \mathcal{O}\left(\frac{1}{K_{min}^2}\right) \\ &= 1 - 2\beta\left(\frac{1}{K_1}+\frac{1}{K_2}\right) + \mathcal{O}\left(\frac{1}{K_{min}^2}\right) \\ &= 1 - \underbrace{2\frac{1}{1+\hat{u}_3^2}\left(\frac{1}{K_1}+\frac{1}{K_2}\right)}_{=:\alpha} + \mathcal{O}\left(\frac{1}{K^2}\right),\end{aligned}$$

where the two quadratic expectations take the explicit values shown, and we use the facts that any term with an odd degree factor in $(u_3 - \hat{u}_3)$, u_1 or u_2 integrates to zero for the diagonal Gaussian (4), and any remaining fourth-order terms are $\mathcal{O}\left(\frac{1}{K_{min}^2}\right)$ where K_{min} is the smallest of the $K_i \gg 1$ for $i = 1, 2, 3$. We further note that the quantity α defined here is also small being $\mathcal{O}\left(\frac{1}{K_{min}}\right)$ (and in fact independent of K_3). Therefore we can Taylor expand $\ln(1-\alpha) = -\alpha + \mathcal{O}\left(\alpha^2\right)$ to conclude that

$$\ln\langle Q\rangle_{3,3} = -\alpha + \mathcal{O}\left(\frac{1}{K_{min}^2}\right),$$

so that finally we find that

$$\ln\langle \mathbf{t}^{[1]} \cdot \mathbf{t}^{[n]}\rangle \equiv \ln\langle \cos\theta_{1,n}\rangle \overbrace{=}^{\text{by sparsity}} n \ln\langle Q\rangle_{3,3} \overbrace{\approx}^{\text{by stiff Gaussian}} -n\alpha + \mathcal{O}\left(\frac{n}{K_{min}^2}\right).$$

The tangent-tangent persistence length (here measured in units of number of base pairs) is defined to be the negative reciprocal of the gradient of this straight line, or $\ell_\mathsf{p} = \frac{1}{\alpha}$. It is a measure of the rate of the exponential decay in the correlations $\langle \mathbf{t}^{[1]} \cdot \mathbf{t}^{[n]}\rangle$. For DNA its value is accepted to be approximately 150bp. which is dimensionless.

In the case of a stiff Gaussian, all of the entries of $\langle Q\rangle$ can be similarly (and now straight forwardly) approximated (merely by permuting indices in the case $\hat{u}_3 = 0$), so that the Flory vector $\delta := (I -$

$\langle Q\rangle)^{-1}\langle \mathbf{q}\rangle$ can similarly be approximated. We will not do this explicitly, but the case $\mathbf{q} = (0, 0, \ell)$ is a standard one in which we can immediately conclude that $\delta = (0, 0, \ell/\alpha) = (0, 0, \ell\,\ell_p)$. Here ℓ is the distance, in some units of length, between two adjacent basepair origins. Therefore the product $\ell\,\ell_p$ is just the length of one step times a number of steps, that is it is the arc length along the polymer chain, which shows that the Flory persistence vector has expectation $\delta = (0, 0, \ell/\alpha) = (0, 0, \ell\,\ell_p) = (0, 0, \ell_F)$, where ℓ_F is a persistence length with dimension of length and measured in whatever units ℓ is measured in. For DNA $\ell \approx 0.33$nm so $\ell_F \approx 150 \times 0.33$nm or $\ell_F \approx 50$nm. The tangent-tangent decay can also be written in terms of the dimensionfull ℓ_F

$$\ln\langle \mathbf{t}(0) \cdot \mathbf{t}(s)\rangle \approx -s/\ell_F$$

where $s = n\ell$ is the arc-length along the polymer chain in n steps, and now $\mathbf{t}^{[n]} \approx \mathbf{t}(s)|_{s=n\,\ell}$. This notation and approximation naturally leads to a continuum limit where the discrete chain of base pair frames is replaced by a curve

$$X(s) = \begin{pmatrix} R(s) & \mathbf{r}(s) \\ \mathbf{0} & 1 \end{pmatrix} \in SE(3) \qquad s \in (0, L),$$

but we do not consider that model here. In fact some authors start from a continuous model for the WLC and discretize it in arc-length to obtain a discrete chain.

Bibliography:

(1) **A sequence-dependent rigid-base model of DNA**, O. Gonzalez, D. Petkeviit, and J. H. Maddocks, Journal of Chemical Physics 138, no. 5 (2013), p. 055122 1-28, DOI:10.1063/1.4789411.

(2) **Helical Wormlike Chains in Polymer Solutions**, H. Yamakawa, Springer-Verlag, 1997.

(3) **On the statistical physics of chains and rods, with application to multi-scale sequence dependent DNA modelling**, A. Grandchamp, 2016, PhD dissertation #6977 (EPFL).