Using the cgDNA+ model to compute sequence-dependent shapes of DNA minicircles

Marius BEAUD

Assistant: Raushan SINGH
Professor: John H. MADDOCKS

Master thesis at EPFL

July 20, 2021

Abstract

cgDNA+ is an enhanced version of a sequence-dependent coarse-grain model of double-stranded DNA called cgDNA. The cgDNA model has previously been used to compute sequence-dependent equilibrium shapes of closed loops of DNA called minicircles, which are an experimentally important motif. The goal of this thesis is to adapt the existing computational approach from the cgDNA model to the cgDNA+ model, which is more physically accurate. The computational issue is that the number of unknowns is roughly doubled due the explicit description of phosphate groups in cgDNA+ added to the rigid base description adopted in cgDNA. We successfully extend the existing algorithmic approach, and also add some enhancements associated with the precise boundary conditions used to model minicircles. Our code is validated by computing equilibria of a number of minicircle fragments of length 94 to 339 base pairs whose sequences have been published in the experimental literature.

Contents

List of Figures

List of Tables

Part I
Introduction

0.1 Interest for DNA

The existence of deoxyribonucleic acid, or DNA, was discovered in 1869 by Swiss researcher Friedrich Miescher [8]. However, it is Crick and Watson that first described the double helix structure of DNA in 1953 [24]. Since then, DNA has been a subject of great interest for scientific research. It has been observed to be the molecule responsible for much of the functioning of organic cells. Therefore, studying the behaviour of DNA is a key aspect to fully understand the molecular processes happening in the cells of any living organism [3].

0.2 Structure of DNA

DNA is a molecule with a double strand structure. Each strand is composed of nucleic *bases* connected to a sugar-phosphate backbone. Each base is one of four types: adenine (A), thymine (T), guanine (G) and cytosine (C). The backbones are linked together through hydrogen bonds between two complementary bases. A always forms two hydrogen bonds with T while C forms three bonds with G. A combination of two complementary bases is called a *base pair*. The two backbones have orientations, specified by the detailed chemical structure of the sugar rings. DNA is such that the two strands have an *anti-parallel* double helical structure meaning that the reading directions on each strand are opposite. The conventional reading direction for a single strand is referred to as the $5' \to 3'$ direction.

0.3 DNA modelling

A key feature of DNA is its intrinsic or ground-state shape. It has been shown that the average overall shape of a DNA double helix is strongly modulated by its *sequence* [3, 14, 22]. Some specific sequences of bases have been observed to have a strong intrinsic bend, whereas other sequences have a tendency to stay in an exceptionally straight double helix. Another important feature is the dependence of the local stiffness of the molecule on the base sequence. Combining both properties, one can characterise the overall sequence-dependent deformations and fluctuations of the DNA molecule. It is therefore important to be able to model and predict the shape and local stiffness sequence-dependence. Models for DNA structure can be separated in two main groups: *atomic level models* and *coarse-grain models*. Atomic models are developed to obtain the finest description of the molecule, they model the atoms directly. An issue with this kind of model is that they are computationally extremely demanding and they are therefore limited to rather short sequences (less than 100 base pairs) and relatively few simulations. Even for 10 base pairs, there exist $\frac{4^{10}}{2} \sim \frac{10^6}{2}$ different sequences. Molecular dynamic simulations cannot compute them all. The alternative is coarse-grain models. In order to reduce the computation costs, atoms are grouped in chosen rigid bodies and interactions are then assumed to happen between the rigid bodies. These models rely on one of two main simplifying assumptions: *rigid base-pair* or *rigid-base*. Furthermore, they often assume that the molecule energy is the sum of local energies. This further simplifies the model and allows for efficient sequence-dependent modelling of a large number of sequences. These models contain less information than atomic alternatives (they only model rigid bases instead of atoms) but,

Type of model	Advantages	Inconveniences
Atomic-level	Finer description of the molecule	Computationally expensive: - **low variety of sequences** - limited to short sequences - Small simulation durations
Coarse-grain	Less expensive, allows bigger simulations: - **multitude of sequences** - longer sequences - longer simulation duration	Loss of information: - Group atoms as rigid bases.

Table 1: Strengths and Weaknesses of Atomic-level and Coarse-grain models for DNA.

from all tested cases, e.g. Patelli's thesis [21], they are believed to provide a good approximation to efficiently and closely represent sequence-dependence of shape and stiffness, the key properties of DNA. They are therefore preferred over atomic level models if the atom positions are not specifically needed. Table 1 summarises the strengths and weaknesses of both types of model.

Here we adopt the cgDNA model [14, 22], a coarse-grain model for DNA. It is based on the rigid base assumption and works for sequences of a few hundred base pairs on a conventional laptop or via the cgDNAweb interface[1]. Recently, Patelli [21] proposed a new adaptation of the cgDNA model that treats the backbones explicitly: the cgDNA+ model. In addition to the bases, it models phosphate groups as rigid objects. It is now believed to be more accurate than the standard cgDNA model.

0.4 Minicircles

An important study case in DNA modelling is the formation of DNA minicircles [1, 2, 6, 7, 9, 10, 13, 15, 16, 17, 23][2]. Therefore it is important to be able to model the sequence-dependent shape and stiffness of such DNA loops to understand the processes under way. We hence have to adapt the existing coarse-grain models to take into account the cyclized shape of the DNA strands. There are two ways of modelling a DNA minicircle depending on the question we are interested to solve. If we are interested in the formation of DNA minicircles, we are interested in the probability that both ends of the fragment are sufficiently close to each other to bond. This is done without taking into account any periodicity within the interactions between the two ends. We will call it non-continuous closure (NCC) minicircles. The other question we may be interested in is the shape of said DNA circles after their formation. In this case we have to take into account the interactions between the two end base-pairs to form a continuous loop. We will call this periodic closure (PC) minicircles.

Energy minimization for minicircles have been done in the past within the cgDNA model by Manning [15]. He combined continuum DNA modelling with a discrete energy minimization to obtain discrete DNA minicircle shapes. For continuum modelling, he used the birod model [12] together with the bBDNA software [10] that allows computation and visualization of families of equilibria

[1] https://cgdnaweb.epfl.ch/

[2] These are only a few examples amongst the mass of articles.
EPFL library online search engine outputs about 5000 articles related to DNA minicircles

for the birod model. He used discretizations of continuum equilibria as initial guesses for a cgDNA energy minimization algorithm called cgDNAmin. However, the discrete energy minimization step has not yet been done for the cgDNA+ model. Hence the goal of this thesis is to adapt the minicircle energy minimization procedure to include the cgDNA+ coordinates, and the more accurate cgDNA+ energy.

0.5 Plan for this thesis

The main goal of this thesis is to implement a cgDNA+ adaptation of cgDNAmin. Here, we focus mainly on the practical implementation of cgDNA+min. We test our code implementation of the enhanced algorithm on sequences that have been published in the experimental literature, but a full discussion of the experimental consequences of our simulations is outside the scope of this Masters thesis.

In order to explain the cgDNA+ minicircle energy minimization, we have to set up the basics of DNA modelling. In the first part of this thesis, we detail the existing work on DNA modelling. We start by describing a discrete coarse-grain model of DNA: the cgDNA model [14, 22]. We detail its main assumptions and characteristics and we also explain the adaptation that exists to account for full periodicity [10]. Furthermore, we will highlight the key properties of the cgDNA+ model that takes into account the contributions of the phosphate groups to the local energy of the molecule [21]. We then follow with a description of a continuum model: the birod model and the bBDNA software that allows one to find continuum energy equilibria for minicircle configurations [10, 12]. This code is used to obtain initial guesses for the cgDNA and cgDNA+ minicircle energy minimization. We describe the existing work of Manning [15] about the minimization of the discrete energy for DNA minicircles starting from a continuum solution. In the second part of the thesis we detail the construction of the minicircle energy minimization algorithm for the cgDNA+ model. We present different case studies with observations and conclusions from our implementation. This includes comparison of solutions between cgDNA and cgDNA+ minicircle energy minimization.

Along with this thesis we provide a webpage[3] where all results and figures can be found. We also provide the 2D coordinate plots for a more precise analysis of the results. Finally, we provide the MATLAB scripts to run both periodic and non-continuous closure cgDNA and cgDNA+ energy minimization (also found on the webpage).

[3] If the hyperlink is not supported, the full link can be found in Appendices A.

Part II
Previous Work

In this first Chapter, we will explain the existing results in DNA modelling. This will be separated into three main parts: the discrete model, the continuum model and the bridge between the two provided by cgDNAmin. This chapter establishes the starting point for our work. We detail the standard cgDNA model, its adaptation to support periodic sequences and its evolution into cgDNA+ to take the phosphate groups into consideration. We follow with the continuum model, the birod model. We also describe the bBDNA software, used to obtain minicircle continuum energy equilibrium. Finally we go through cgDNA minicircle energy minimization, a discrete energy minimization procedure starting from an equilibrium configuration of the continuum model.

1 Discrete Model

1.1 cgDNA, rigid-base coarse-grain DNA model

This section explains the cgDNA model, a rigid base pair model detailed in [14, 22].[4] For a given sequence and parameter set, the cgDNA model is a coarse-grain model that predicts a probability density function for the configuration of the molecule. Formally, the output of the cgDNA model is

$$\rho(w; S, \mathcal{P}),$$

where w is the cgDNA coordinates, S is the sequence of DNA to be modelled and $\mathcal{P}$ is the parameter set. The parameter set is built such that we can use it to reconstruct μ, the *shape* or *ground-state*, and K, the cgDNA *stiffness matrix* corresponding to the given sequence S.

1.1.1 Main assumptions

The cgDNA model relies on three major assumptions: it considers *Rigid bases*, it models a *Gaussian probability density* and the molecule energy is assumed to have *double local dependence*, i.e. locality in sequence dependence, and each base is assumed to only interact with its five nearest neighbours.

Rigid Bases. As mentioned in Section 0.3, the cgDNA model is a rigid base coarse-grain model of DNA. This means that each base is represented by a rigid body that is best fit to the atoms of the base. The interactions inside the molecule are then simplified to only consider the contributions of the rigid bases. This limits the number of free variables and allows for better computational efficiency.

Gaussian Probability Density Function. As said earlier, the cgDNA model predicts the relative position and orientation of the bases within the molecule in the form of a probability density function (pdf). The natural choice for this pdf is a *Gaussian distribution*. The probability density depends on the *ground-state* and the *stiffness matrix* of the molecule. The ground-state is the most

[4] This material is also taught by J.H.Maddocks in the "Mathematical modelling of DNA" course at EPFL, https://lcvmwww.epfl.ch/teaching/modelling_dna/.

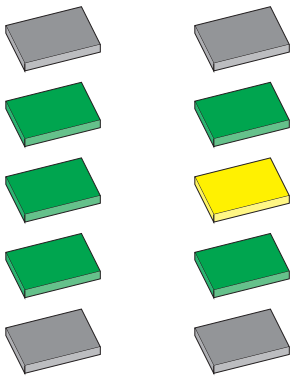


Figure 1: Representation of the nearest neighbour energy contribution. For the yellow base, we only consider its interactions with the five green bases. Any interaction between the gray bases and the yellow one are assumed to be negligible.

likely shape of the molecule under no external constraints or loadings, and the stiffness represents the resistance to change. Formally, we have

$$\rho(w; S, \mathcal{P}) = \frac{1}{Z} e^{-\beta U(w;S,\mathcal{P})}, \tag{1}$$

$$U(w; S, \mathcal{P}) = \frac{1}{2}\big(w - \mu(S,\mathcal{P})\big)^T K(S,\mathcal{P})\big(w - \mu(S,\mathcal{P})\big). \tag{2}$$

Here, Z is the normalization constant, β is inverse temperature energy scale, and U is the shifted quadratic cgDNA energy of configuration w. Again, w are the configuration coordinates, S is the sequence, $\mathcal{P}$ the parameter set and μ and K are respectively the ground-state and the stiffness matrix.

Local Energy. We have seen that the output of the cgDNA model depends on the molecule energy $U(w; S, \mathcal{P})$, which itself depends on the configuration w. In order to limit computational costs and facilitate parameter estimation, we only consider local contributions of the energy. This means that a base is assumed to only interact with its nearest neighbours. Figure 1 represents the nearest-neighbour interactions assumption: when considering the interactions of the yellow base, the contributions to the energy will only come from the interactions with the green bases, and we assume that the interactions of the gray with the yellow bases are negligible. We also assume that the energy depends locally on the sequence. It follows that the stiffness matrix $K(S, \mathcal{P})$ is a banded matrix with overlapping blocks along the diagonal. The blocks represent all the interactions between two consecutive base pairs and the overlaps represent the fact that each base pair is interacting with both the previous and the following base pairs. The interactions between non-neighbouring base pairs are assumed to vanish. Hence the banded diagonal structure.

1.1.2 cgDNA coordinates

Frame embedding. For any sequence, we have to define coordinates to specify the position and orientation of the bases. As said in Section 1.1.1, we embed a frame in each base. This frame is

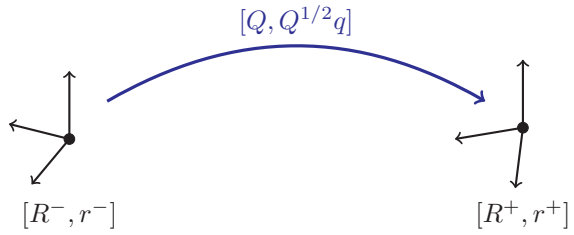


Figure 2: Illustration of the relative displacement between two frames: Equation (3). Here, the frames are only represented by a right handed set of 3 orthonormal vectors, showing the relative translation and the relative rotation between the two frames.

represented by an element in SE(3)[5]. Therefore, for each base we have an associated frame $[r, R]$ with $r \in \mathbb{R}^3$ and $R \in SO(3)$.

cgDNA coordinates can be separated in two types: intra base pair (or *Intras*) and inter base pairs (or *Inters*). Intras define the relation between the frames of the two bases within base pairs and Inters specify the relation between two consecutive base pair frames. The final coordinate vector w, consists of alternating Intras and Inters. For a n base pair sequence

$$S = X_1 X_2 \dots X_n,$$

we have the coordinate vector

$$w = (x_1, y_1, x_2, y_2, \dots, x_{n-1}, y_{n-1}, x_n),$$

where x_i is the intra coordinate vector for base pair i, i.e. base X_i and its Crick-Watson Complement $\overline{X_i}$, and y_i is the inter coordinate vector between base pairs $(X_i, \overline{X_i})$ & $(X_{i+1}, \overline{X_{i+1}})$.

Intras. For a base pair $(X_i, \overline{X_i})$, we define the intra coordinates as follows. Let us first consider the two frames associated to both bases: $[r_i^+, R_i^+], [r_i^-, R_i^-] \in SE(3)$. We define the relative displacement $[q_i, Q_i] \in SE(3)$ such that

$$\begin{bmatrix} R_i^+ & r_i^+ \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} R_i^- & r_i^- \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} Q_i & Q_i^{1/2} q_i \\ \mathbf{0} & 1 \end{bmatrix}. \tag{3}$$

Hence we have $Q_i = (R_i^-)^T R_i^+, \quad q_i = (R_i^- Q_i^{1/2})^T (r_i^+ - r_i^-)$. Note that the relative translation is expressed in the mid frame $R_i = R_i^- Q_i^{1/2}$ for symmetry. A graphical representation of Equation (3) is shown in Figure 2.

The final intra coordinates for base i are then the Cayley vector[6] of the relative rotation Q_i and the relative translation q_i expressed in the mid frame,

$$x_i = [Cay(Q_i), q_i] \in \mathbb{R}^6. \tag{4}$$

Figure 3 illustrate the six degrees of freedom for the relation between two base frames. The top line shows the relative rotations and the bottom lines shows the relative translations.

[5]See Appendices C for more detail on the SE(3) group.
[6]See Appendices D for definition of the Cayley vector.

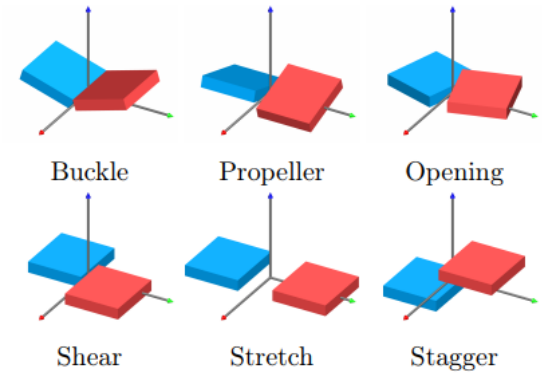


Figure 3: Graphical representation of the 6 degrees of freedom for the relative displacement between two frames: the intra coordinates. Figure taken from Głowacki [10].

Inters. The inter coordinates between base pairs $(X_i, \overline{X_i})$ and $(X_{i+1}, \overline{X_{i+1}})$ need a bit more work. We start by defining the i-th base pair frame as the mid frame:

$$[r_i, R_i] \in SE(3), \qquad R_i = R_i^- Q_i^{1/2}, \qquad r_i = \frac{1}{2}(r_i^+ - r_i^-).$$

Let $[j_{i+1/2}, J_{i+1/2}]$ be the relative displacement between the base pair mid frames $[r_i, R_i]$ and $[r_{i+1}, R_{i+1}]$.[7] The inter coordinate corresponding to junction i is then:

$$y_i = [Cay(J_{i+1/2}), j_{i+1/2}] \in \mathbb{R}^6. \tag{5}$$

Figure 4 illustrate the six dimensions of the relative displacement between two base pairs.

3D Reconstruction. For the reconstruction, we fix the first base pair frame:

$$[r_1, R_1] = [(0, 0, 0), I_3].$$

The base pair frames of the other bases are then obtained following a tree construction starting from the first base pair frame. Each set of inter coordinates y_i is converted to a corresponding relative displacement in $SE(3)$ using Euler-Rodrigues formula[8], the later can be used to construct every base pair frame with a chain construction. The base frames are obtained from the base pair frames and the intra coordinates. The base pair frame is the mid frame between the two bases. Hence each base frame is obtained from the base pair frame using half of the intra relative displacement. Figure 5 shows the chain structure of the reconstruction.

1.1.3 Energy

As specified in Section 1.1.1, the local molecule energy is assumed to be quadratic and we consider only the local interactions between the bases. Furthermore, we suppose that the total energy of the

[7] The inter relative displacement is obtain using the same procedure as the intra displacement, see Equation (3) and Figure 2 where we adapt the two compared frames.

[8] See Appendices D

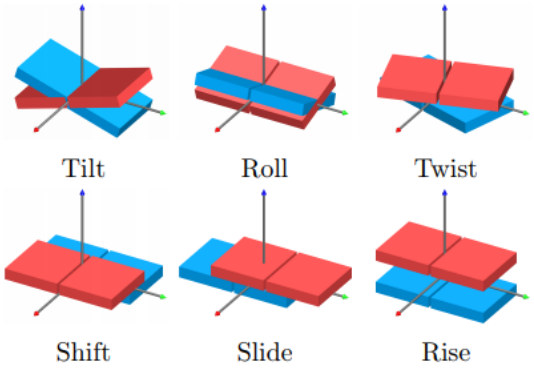


Figure 4: Graphical representation of the six cgDNA inter coordinates. Figure taken from Głowacki [10].

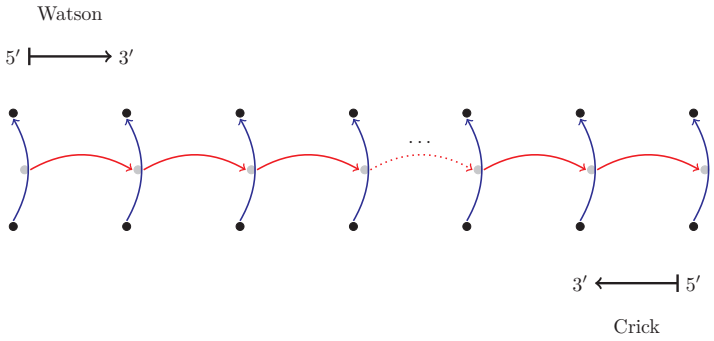
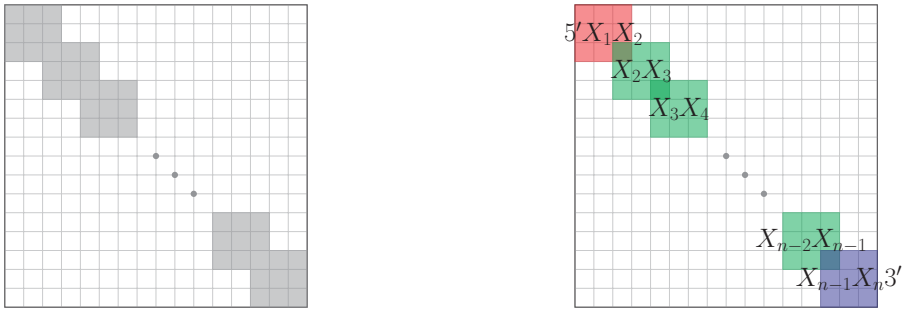


Figure 5: Chain structure of the cgDNA coordinates. The black dots represent the base frames, the gray dots the base pair mid frames, the red arrows represent the inter coordinates and the blue ones represent the intra displacement.



(a) SParsity pattern (b) Contruction of K

Figure 6: Sparsity pattern and construction of the cgDNA stiffness matrix K. Each grid delimitation represent a 6×6 block. The two end blocks are shown in a different color in the right panel to emphasise the fact that they differ from the identical dimer block in the interior.

DNA molecule is the sum of the local energies. Combining these assumptions we obtain that the energy for a n base pairs sequence is of the form:

$$U(w; S, \mathcal{P}) = \sum_{a=1}^{n-1} U_a, \tag{6}$$

where U_a is the local energy:

$$U_a = \frac{1}{2}(w_a - \mu_a)^T K_a (w_a - \mu_a), \qquad w_a = (x_a, y_a, x_{a+1}). \tag{7}$$

Here, μ_a and K_a are respectively the local ground-state and the local stiffness. A sum of shifted quadratic forms can itself be written as a shifted quadratic form. Hence, we obtain a total energy :

$$U(w; S, \mathcal{P}) = \frac{1}{2}(w - \mu)^T K (w - \mu) + c, \tag{8}$$

with $K \in \mathbb{R}^{(12n-6)\times(12n-6)}$ a banded matrix with $(n-1)$ 18×18 blocks with 6×6 overlaps, as shown in Figure 6a, $\mu \in \mathbb{R}^{12n-6}$ is the molecule's ground-state. The constant c plays no role and is therefore not taken into account.

1.1.4 Parameter set

The cgDNA model relies on the use of a pre-computed parameter set. All these parameters are obtained by fitting the model to match molecular dynamics simulations, see [11] for more details. The final probability distribution function ρ depends on both the sequence S and the parameter set $\mathcal{P}$. This parameter set is used to reconstruct the ground-state and stiffness matrix corresponding to the given sequence S. Through time, the model was improved and the parameter set refined to be more accurate. Here, we will focus on the latest development of cgDNA and we only give the description of the cgDNA 2.0 parameter set.

cgDNA 2.0 parameters. For all possible sixteen oligomer steps XY, the cgDNA 2.0 parameter set contains three types of 18×18 symmetric positive definite blocks: K^{XY}, $K^{5'XY}$ and $K^{XY3'}$ and three 18×1 types of vectors: σ^{XY}, $\sigma^{5'XY}$ and $\sigma^{XY3'}$. The K^{XY} stiffness block is for the case where both X and Y are in the interior of the sequence, the $K^{5'XY}$ and $K^{XY3'}$ stiffness blocks are respectively for the cases where X is the 5' end or Y is the 3' end of the sequence. The same principle apply to the three σ vectors. Parameter block symmetry comes from the fact that we model a quadratic energy. The energy, the stiffness and the ground-state have to be independent of the choice of reading strand. This gives us more constraints on the parameters, see [22] for more details.

For the reconstruction of the stiffness matrix and the ground-state, we start with the stiffness matrix K. We read the sequence S and build the banded stiffness matrix by adding the dimer 18×18 parameter blocks in the correct places. Each block has a 6×6 overlap with the previous and the following block. The shape vector σ is computed with the same method. Each 18×1 block is composed of the vector $\sigma^{X_i X_{i+1}}$ and the contributions of the previous and next dimers for the 6×1 overlap. Figure 6b illustrates this construction. The ground-state is then obtained using the relation $\mu = K^{-1}\sigma$. We construct the shape vector before the ground-state because σ only has a local dependence on the sequence whereas μ has not. It is therefore necessary to compute the ground-state from a different variable with a local dependence on the sequence.

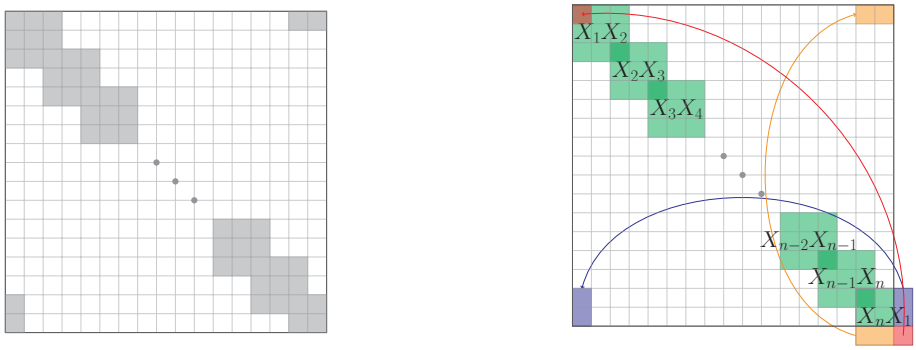
1.2 cgDNA for Periodic Sequences

This section highlights Głowacki's construction for the periodic ground-state and stiffness matrix for the cgDNA model. For a more detailed description, we refer to his PhD thesis [10].

1.2.1 Interest in Periodicity

In Nature, we observed that some DNA sequences are composed by multiple, end-to-end repetitions (more than 2) of a smaller sequence, the *base sequence* [20]. Such sequences are called *Tandem repeats* (even when there are more than two repeats). In order to improve computations for this particular type of sequence, Głowacki [10] and Grandchamp [12] proposed an adaptation of the cgDNA ground-state and stiffness matrix. They created the so-called *periodic* ground-state and stiffness matrix for the base sequence. This allows to characterize infinite tandem repeats using only one period of the base sequence. As we saw in Section 1.1, the cgDNA reconstruction procedure has a physically justified end effect. However, in the case of a periodic sequence, the end effects can be removed at the junctions between two repeats of the base sequence. This method turns out to be very helpful in our case as it can also be used to reconstruct the shape and stiffness matrix of a closed DNA loop. Indeed a closed loop can be seen as an infinite repeat of the same sequence. The main difference between periodic and standard ground-state and stiffness lies on the extremities. In order to account for periodicity, we have to consider the interactions between the first and last base pairs.

We note that sequence periodicity does not imply periodicity in the positions of the bases in general. Sequence periodicity only signifies that the sequence repeats itself a high number of times. The overall shape of a periodic sequence is therefore an infinite helix. In the special case of minicircles, the sequence appears only once but we enforce a cyclized shape. The base coordinates are then periodic when repeating the sequence. But, all periodic coordinates certainly do not correspond to closed loops.



(a) Block structure.

(b) Construction of K_p.

Figure 7: Structure (on the left) and construction (on the right) of the cgDNA periodic Stiffness matrix K_p. Each square of the grid is of size 6×6. The highlighted blocks on the right panel are the blocks representing the interaction between the first and last base pairs.

1.2.2 Construction of periodic stiffness matrix

The construction of the periodic stiffness K_p of a base sequence S is similar to the construction of the standard stiffness matrix of cgDNA model explained in Section 1.1. The difference is that we need to replace the contributions of the end base pairs to consider them as interior bases and we have to add the contribution of the relation between the first and last base pairs. For that purpose, the periodic stiffness contains an extra set of inter coordinates. For a sequence of length n, K_p is of size $12n \times 12n$ with overlapping 18×18 blocks and 6×6 overlaps. The last block is truncated to be 12×12. The extra entries are added to the first 6×6 block, the 6×12 upper right and 12×6 bottom left corners. The extra blocks in the anti-diagonal corners represent the interactions between the last and first base pairs and are taken from the $K^{X_nX_1}$ parameter block. The schematic construction is shown in Figure 7.

1.3 The explicit phosphate model, cgDNA+

In 2019, Patelli proposed a new version of the cgDNA model with explicit treatment of the backbones [21]. They consist of alternating phosphates groups and sugar rings. The new model, called cgDNA+, adds the contribution of the phosphate groups to the molecule energy while the contributions of the sugar rings remain implicit. This adaptation is currently believed to be more accurate than the standard cgDNA model [21]. In this section we explain the basics of the cgDNA+ model using Patelli's ideas. For more details, we refer to Patelli's thesis [21].

1.3.1 Explicit Backbones

The rigid base assumption of the cgDNA model is carried over to the phosphate groups. For a sequence $S = X_1X_2 \dots X_n$, we represent each base by a frame $[r_i^{\pm}, R_i^{\pm}] \in SE(3)$. Each phosphate group is also represented by a frame $[p_i^{\pm}, P_i^{\pm}] \in SE(3)$. The result of such a parametrization is illustrated in Figure 8. The coordinates of each phosphate are a relative $SE(3)$ displacement from the associated base.

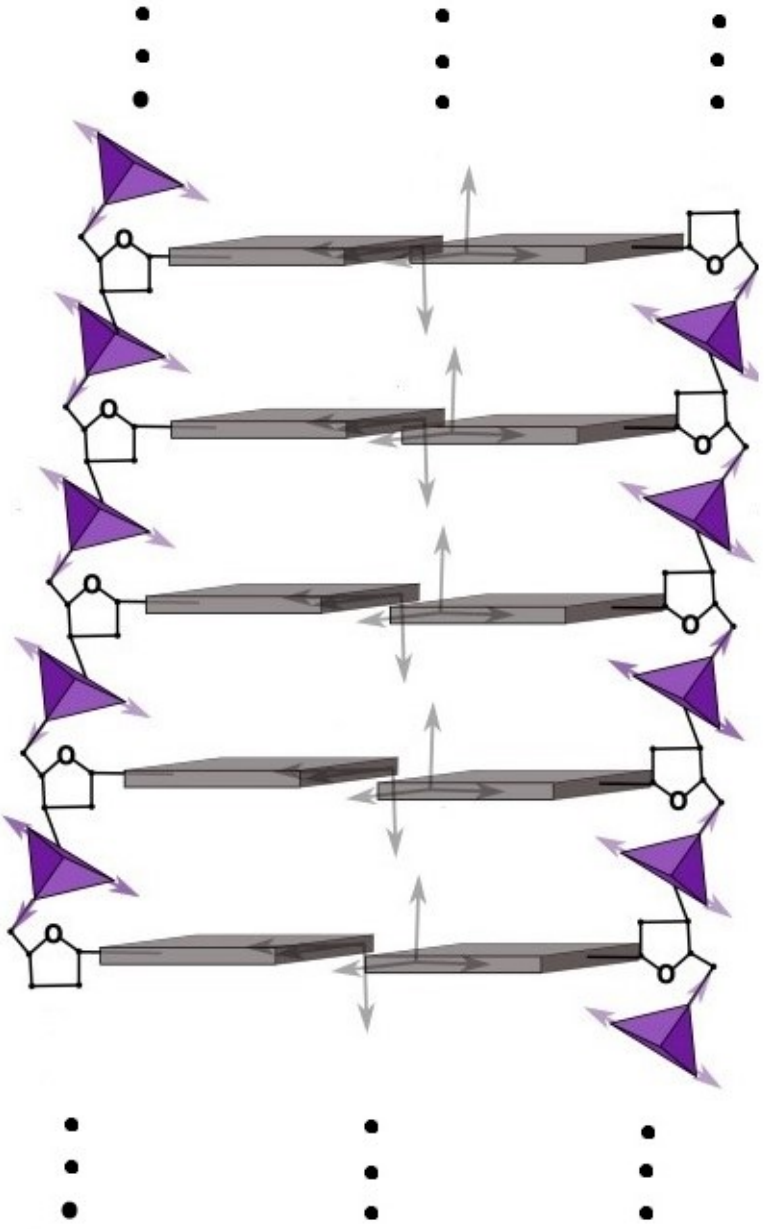


Figure 8: Double strand representation with explicit backbones of five base pairs within a sequence. The gray frames are the bases and the purple shapes are the explicit phosphates groups. The sugar rings are also shown in this figure even if they are not explicit in the cgDNA+ model. This figure is taken from Patelli's thesis [21].

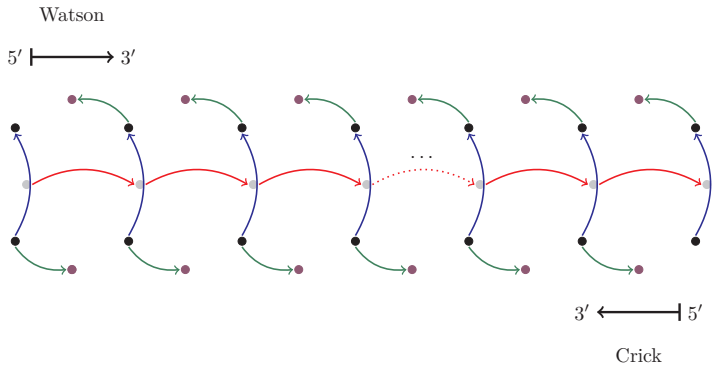


Figure 9: Chain structure of the cgDNA+ coordinates. We reuse the color code of Figure 5 for the cgDNA coordinates: Intras in blue, Inters in red, base frames in black and base pair frames in gray. We add the phosphate coordinates in green and the phosphate frames in purple dots.

cgDNA+ coordinates. We recall that we used the relative displacement between the two base frames in a base pair as intra coordinates in the standard cgDNA model. The inter base pairs coordinates were defined as the relative displacement between the two base pair frames $[r_i, R_i]$ and $[r_{i+1}, R_{i+1}]$. Both sets of coordinates remain unchanged in the cgDNA+ model. In order to implement the phosphates frames, we construct the base pair level coordinates by adding the relative displacements between the base frames and both Crick and Watson phosphate groups to the intra coordinates. Let $[m_i^{\pm}, M_i^{\pm}]$ be the relative displacements between the Crick (-) or Watson (+) phosphate group $[p_i^{\pm}, P_i^{\pm}]$ and the base frame $[r_i^{\pm}, R_i^{\pm}]$, expressed in the corresponding base frames.[9]

$$\begin{bmatrix} R_i & r_i \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} M_i^{\pm} & m_i^{\pm} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} P_i^{\pm} & p_i^{\pm} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{9}$$

We note that the Crick phosphate is associated to the unflipped Crick frame. The base pair level coordinates at base pair i then become

$$\tilde{x}_i = [\underbrace{Cay(M_i^+), m_i^+}_{\text{Watson Phosphate}}, \underbrace{Cay(Q_i), q_i}_{\text{Standard intras}}, \underbrace{Cay(M_i^-), m_i^-}_{\text{Crick Phosphate}}] \in \mathbb{R}^{18}, \tag{10}$$

where $Cay(A)$ is the Cayley vector of A detailed in Appendices D and $x_i = [Cay(Q_i), q_i]$ is the standard cgDNA intra coordinate at base pair i. We can adapt Figure 5 to cgDNA+ coordinates and we obtain the new tree structure shown in Figure 9.

End bases. When modelling the explicit phosphates groups, we have to be particularly careful when modelling the ends of the sequence. Each phosphate group is associated to a base pair, however phosphates are not exactly aligned with the bases. They sit in between two base pairs. This induces a problem at the ends: we cannot model accurately the position of a phosphate group that is not lying in between two bases using a Gaussian probability density function. Therefore we remove the extra phosphate groups at the end bases only to keep the phosphates that are inside junctions of

[9] The relative translation between a phosphate group and its associated base is expressed in the base frame. This is similar to the cgDNA relative translation being expressed in the mid frame.

the sequence. This correspond to so-called blunt ends in the chemistry literature. Then, the end base pair level coordinates are:

$$\tilde{x}_1 = [Cay(Q_1), q_1, Cay(M_1^-), m_1^-] \in \mathbb{R}^{12} \text{ and}$$
$$\tilde{x}_n = [Cay(M_n^+), m_n^+, Cay(Q_n), q_n] \in \mathbb{R}^{12}.$$

Figure 10 shows the deleted end phosphate groups. The dimension of the cgDNA+ stiffness is then $18(n-2) + 6(n-1) + 12 \cdot 2 = 24n - 18$.

Parameter set. The cgDNA+ parameter set is very similar to the standard cgDNA 2.0 parameter set. For each dimer XY, it contains three stiffness blocks and three shape vector blocks. In the case of cgDNA+ the size of the interior stiffness blocks is 42×42 and the end blocks are 36×36 due to the outer phosphates not being modelled. Similarly, the shape vector blocks are 42×1 for the interior and 36×1 for the ends. The parameter set is obtained with a similar construction to the standard cgDNA construction detailed in Section 1.1.4 with the difference lying in the dimensions of the blocks. The diagonal blocks of the cgDNA+ stiffness matrix are now 42×42 (except for the end blocks that are 36×36) and the overlap becomes 18×18. We refer to Patelli [21] for more details on parameter estimation.

cgDNA+ energy. In the explicit backbone model the assumptions about local dependence on the sequence and local contributions for the total energy remain unchanged. Hence the only difference is coming from the change in the local energies that now include the contributions of the phosphate groups. Since each phosphate group is associated with a base pair, the nearest neighbour assumption is still valid. The phosphate groups interact with its associated base as well as its five nearest neighbour bases and their phosphate groups. Hence the definition of the energy does not change fundamentally between the cgDNA and cgDNA+ models, we still model with a quadratic energy.

1.3.2 Periodic cgDNA+

Similarly to the periodization of the cgDNA model in Section 1.2, Patelli's work can be adapted to support periodic sequences. The construction is the same with the only difference being the size of the base pair level coordinates. The periodic stiffness for cgDNA+ is then a $24n \times 24n$ matrix with 42×42 overlapping blocks and 18×18 overlaps. There are also an added 24×18 block in the upper right corner and its transpose in the bottom left corner. The new sparsity pattern and construction of the periodic cgDNA+ stiffness matrix is shown in Figure 11. We highlighted the parameter blocks differing from the standard cgDNA+ stiffness matrix construction.

2 Continuum Model

2.1 Birod DNA

This section highlights the key aspects of the continuum birod model for DNA [12]. Since we use the birod model as a black box together with the bBDNA software, we do not develop its detailed construction here. We only mention the guidelines for birod modelling of DNA. The DNA birod model was first introduced by Moakher and Maddocks as a continuum model for DNA molecules [19] and was then refined by Grandchamp [12]. In perticular he established a procedure for passing from

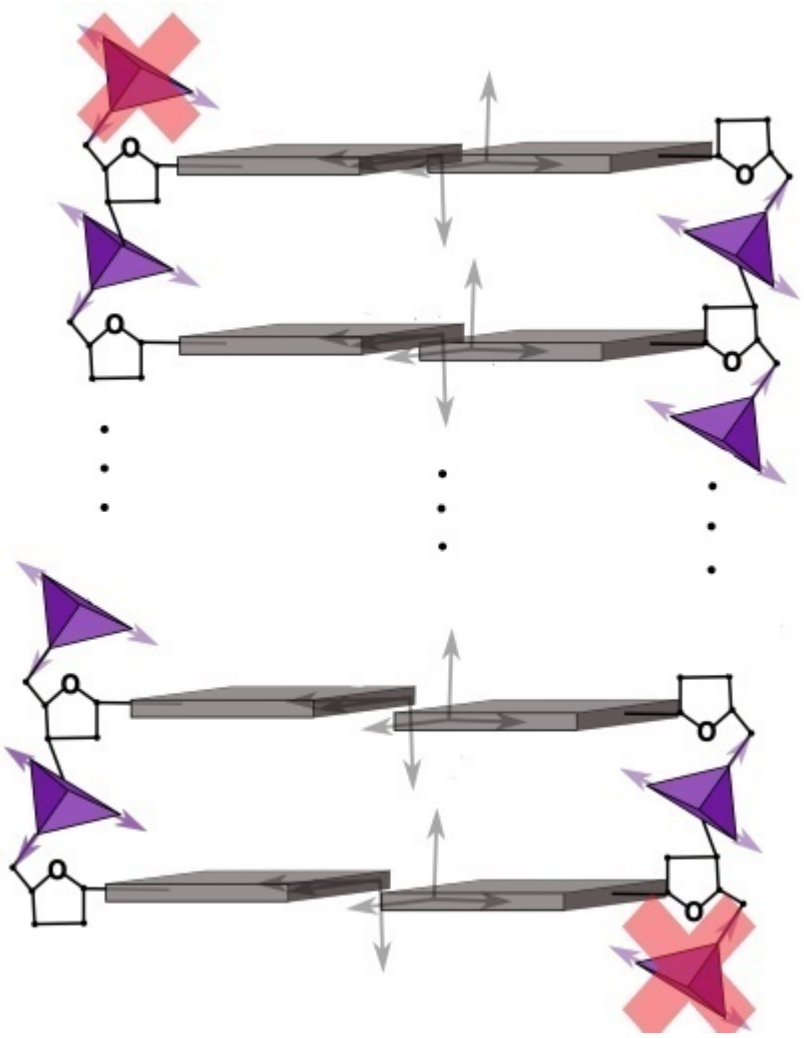
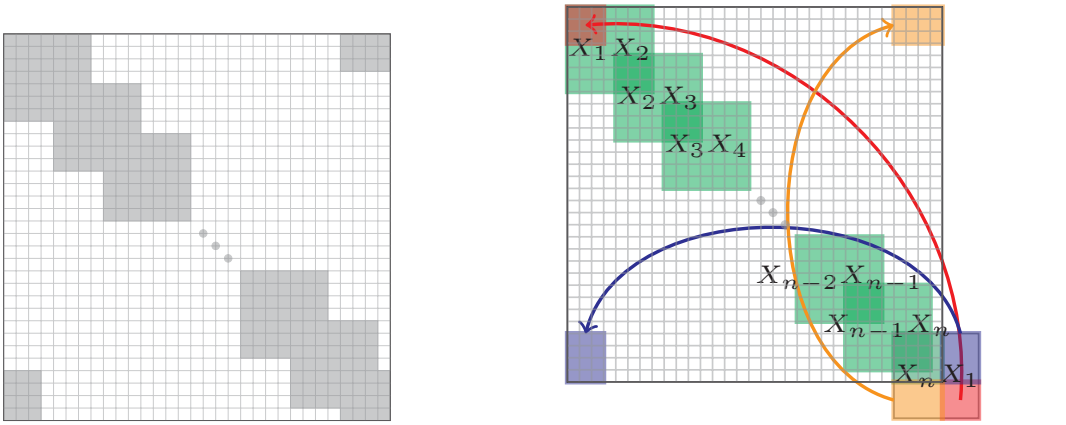


Figure 10: Ends of the cgDNA+ representation with dropped outer phosphate groups (marked with red crosses). This figure is a modified version of Figure 8.2 of Patelli's thesis [21].



(a) Block structure.

(b) Construction of the periodic cgDNA+ stiffness matrix.

Figure 11: Structure (on the left) and construction (on the right) of the cgDNA+ periodic Stiffness matrix. Each square of the grid is of size 6×6. The highlighted blocks on the right panel are the blocks representing the interaction between the first and last base pairs. They are the blocks changing from the standard cgDNA+ stiffness matrix.

discrete cgDNA energy to continuum birod coefficients. We then refer to these original publications for a more precise description.

2.1.1 Birod configuration

We start by describing a single *rod*. A rod is a continuum path $\mathbf{g}(s) = (R(s), r(s)) \in SE(3)$ for $s \in]0, L[$. The rotation matrix $R(s)$ represents the orientation of the rod cross sections and the vector $r(s) \in \mathbb{R}^3$ represents the position of the rod centerline.[10]

A birod configuration $(\mathbf{g}, \mathcal{P})$ is described by two structures at two different levels. First, $\mathbf{g}(s)$, the *macrostructure*, is a continuum rod configuration that represents the average of the double rod configuration. Second, $\mathcal{P}(s)$, the *microstructure*, represents the relative rotation and the relative translation and allows the reconstruction of the double rods $\mathbf{g}^{\pm}$ from the macrostructure. Grandchamp [12] adapted the proposed model of Moakher and Maddocks [19] and parametrized the microstructure $\mathcal{P}$ using the discrete cgDNA model detailed in Section 1.1. This allowed him to obtain a continuum sequence-dependent birod model for DNA.

2.1.2 Internal energy and equilibrium

Continuum rod configurations adapt to the application of local couples and forces. An equilibrium configuration for a rod is a configuration where the total couple and force densities acting on each cross section balance. The formal rod equilibrium conditions can be found in [12, Equation 3.2.1]. Similarly to the discrete case discussed in Section 1.1, the internal energy for the rod is also assumed to be local, see [12] for more details.

[10] See Appendices C for the description of $SE(3)$.

The equilibrium condition for a double rod model is obtained by considering each strand as a single continuum rod in an external field. The birod equilibrium configuration is then obtained by requiring both individual rod equilibrium conditions. The interactions of a rod with the other is treated as external couple and force from the perspective of the second rod. Similarly to the discrete case, we assume double locality of the internal energy (local interactions, local dependence on the sequence). With some work, one can show that the equilibrium conditions for the birod model satisfy a Hamiltonian structure, we refer to Grandchamp's thesis [12] for further details. Finding equilibrium configurations comes back to solving the Hamiltonian system. In his thesis, Głowacki [10] proposed a procedure to efficiently solve such problems: the bBDNA software.

2.2 The bBDNA software

In this section, we present the bBDNA software developed Głowacki as an important part of his thesis [10]. This software is "an interactive parameter continuation and visualization tool for the continuum birod model of DNA." In particular it provides a graphical user interface (GUI) to visualize families of solutions for the birod model in a bifurcation diagram. It also provides a 3D reconstruction of solutions in the diagram through probes.

The complete pipeline consists of MATLAB scripts to prepare the continuum input coefficients for the particular sequence, before solving the Hamiltonian system of ODE of the birod model from Section 2.1. The equilibrium configurations are thereby computed. Finally, the family of solutions is presented through the bBDNA GUI. For this thesis, we decided to use the bBDNA software as a black box. We only use the computed equilibrium configurations to construct initial guesses for the discrete energy minimization. For that reason, we do not present the full internal mechanisms of the software and equilibrium computations. We just present the MATLAB preprocessing script and bBDNA GUI so that the reader can repeat our procedure.

2.2.1 MATLAB preprocessing bBDNA script

The MATLAB scripts prepare the sequence-dependent birod DNA coefficients. It creates the Hamiltonian system of the birod model[11] via the Lagrangian coefficients[12] using the predefined cgDNA parameter set[13].

The MATLAB script to run is the `setupcomputations.m` script located in the `matlab` subfolder. In that file, one can choose the desired sequence under the `seq` variable and the name of the folder in which the parameters will be generated under the `name` variable. Finally, one can choose the maximal energy and the step size. For the two last variables, Głowacki gives examples of values for different length of sequences. More advanced parameters can then be adjusted, but we do not change any of them so we do not detail them here.

The MATLAB script then produces files for the coefficients and outputs a command to past in the command prompt to run bBDNA with the correct files. Once run, bBDNA computes the family of solutions and launches the GUI.

[11] See Section 2.1

[12] The stiffness matrix $\mathbf{K}(s)$, its boundary conditions and the shape internal parameters. See [10, 12] for the definitions of the Lagrangian coefficients.

[13] Note: the parameter set currently used is a cgDNA 1.0 version of the parameters, refer to [10] for the definition of the parameters

2.2.2 bBDNA GUI

The solutions of the Hamiltonian system are presented as a bifurcation diagram that shows curves of equilibrium configurations. In the default projection, the higher the solution the higher the internal energy and the abscissa represent the force which is constant along the equilibrium and the internal twist. Examples of bifurcation diagrams can be found on Figures 13, 17, 25 and 31. The 2D projections show Energy versus the twist. To visualize the 3D reconstructions one has to create a new probe. In our case, we are interested in solutions where the two ends of the sequence are correctly aligned in the sense that both strands are aligned to form a correctly closed loop[14]. These particular solutions are highlighted with crosses in the bifurcation diagram. Those are the solutions we discretize to use as initial guesses for cgDNAmin. Note that the colors of the crosses represent the link, roughly the number of twists of the coupled backbones, of the configurations. For more details on the GUI and for the user commands we refer to Głowacki's thesis [10].

3 From Continuum to Discrete

This Section describes the work of Manning [15] which we later extend from cgDNA to cgDNA+ model. Manning was interested in the formation of DNA minicircles. As we said in Part I, DNA minicircles are an important study case in DNA modelling. Amongst the possibilities to model such configurations, Manning decided to start with a continuum equilibrium obtained through the bBDNA software. This equilibrium is then discretized to provide an initial configuration and finally a minimization procedure is done to find the discrete energy minimizer.

3.1 bBDNA initial run

As mentioned before, we start with a run of the bBDNA software detailed in Section 2.2. We input the sequence and the chosen parameters to build the birod model. We then run bBDNA to find equilibrium configurations and generate the bifurcation diagram. We finally choose a continuum energy equilibrium configuration through the GUI. We have to be careful to select a configuration that is correctly closed (crosses) to be as close as possible to a valid discrete configuration.

Once the continuum configuration is chosen, we discretize it, as mentioned in Section 2.2. We are now ready to proceed to the next step, the discrete energy minimization.

3.2 cgDNAmin, the Discrete Energy Minimization

After obtaining the continuum energy equilibrium configuration and its discretization, we need a discrete energy minimization step. Indeed, the assumptions for the continuum model are slightly different than the discrete case. Moreover, during the discretization process, errors are induced. Therefore, we apply a cgDNA minicircle energy minimization procedure to ensure that the final configuration is a proper minimizer for the discrete energy.

We recall that the standard inter cgDNA coordinates represent the relative displacement between two base pairs. It is then difficult to enforce a closure condition with such coordinates. Closure expressed in inter coordinates is a highly non-linear and non-local constraint. Therefore, Manning proposed to change representation and use absolute coordinates for each base pair frame [15]. The absolute rotation of each base pair frame is represented by a quaternion. The closure assumption

[14] Both strands must be closed on themselves for a valid configuration.

then becomes a local condition, the last base pair must be close to the first one. This comes at the cost of extra computations for computing the relative displacement between base pairs.

3.2.1 Coordinate vector

After the discretization, for each base pair i we have intra coordinates $x_i \in \mathbb{R}^6$ and base pair coordinates $(o_i, q_i) \in \mathbb{R}^7$, where $o \in \mathbb{R}^3$ is the absolute translation between base pair frame i and the origin and $q \in \mathbb{R}^4$ is the absolute rotation of the base pair frame expressed with a quaternion. The standard cgDNA inter coordinate y_i can then be obtained as a function of (o_i, q_i) and (o_{i+1}, q_{i+1}). For the minimization procedure, we have to create the vector z containing all coordinates.

$$z = (x_1, o_1, q_1, x_2, o_2, q_2, \dots, x_{n-1}, o_{n-1}, q_{n-1}, x_n) \in \mathbb{R}^{13n-14}, \tag{11}$$

where n is the length of the sequence. Since the first inter coordinate is our reference point in space, it is fixed to $o_1 = (0, 0, 0)$ and $q_1 = (0, 0, 0, 1)$. Therefore we do not need to keep track of it during the energy minimization procedure.

As we saw in Section 3.2.1, Manning used quaternions to represent the rotation between base pairs. However, the standard cgDNA energy uses relative displacement expressed in $\mathbb{R}^3$ with the help of Cayley vectors.[15] This means that he had to first convert the quaternion coordinates into relative rotation and then into vectors in $\mathbb{R}^3$. Manning also detailed the explicit formulas for the gradient and the Hessian matrix of the energy with respect to the vector of unknowns z [15]. The standard i-th inter coordinate is obtained from the absolute coordinates of the base pair frames (o_i, q_i) and (o_{i+1}, q_{i+1}). The relative rotation between the two base pair frames is defined by the rotation matrix represented by $(q_i^{-1} \circ q_{i+1})$[16] The relative translation between base pair frames is then $(o_i + o_{i+1})$ expressed in the junction frame. The later can be obtained from the absolute rotation induced by the quaternion $(q_i \circ \sqrt{q_i^{-1} \circ q_{i+1}}) = q_i + q_{i+1}$.

3.2.2 Energy

We here detail the description of the molecule configuration energy with its dependence on the coordinate vector z. We recall Equation (8)and write the energy as

$$U(w) = \frac{1}{2}(w - \mu)^T K (w - \mu),$$

with $\mu \in \mathbb{R}^{(12n-6)}$ being the ground-state and $K \in \mathbb{R}^{(12n-6)\times(12n-6)}$ the banded stiffness matrix. With the transformation of quaternions to Cayley vectors, the energy is then

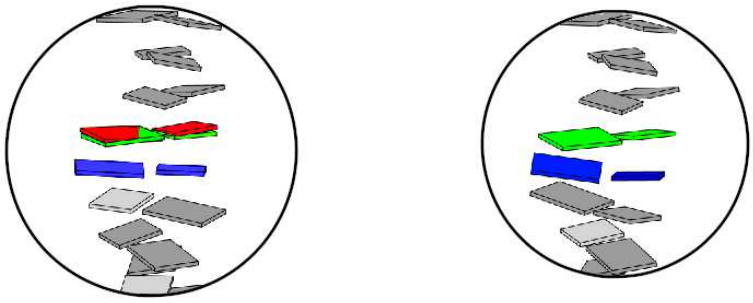
$$U(w) = U(F(z)) = \frac{1}{2}(F(z) - \mu)^T K (F(z) - \mu), \tag{12}$$

where, $F(z) = w$ is the function that recovers the original cgDNA coordinates from the vector z of Equation (11). Intras remain unchanged and inters are obtained with the construction explained in Section 3.2.1, see Appendices E for the explicit formula.

For the purpose of enforcing $\|q_i\| = 1,\ \forall i = 1 \dots, n$, we add a penalty factor

$$E = p \sum_{i=1}^{n} (\|q_i\|^2 - 1)^2, \tag{13}$$

[15] See Appendices D for the relation between rotations and $\mathbb{R}^3$.
[16] Here, $\circ$ represents quaternion multiplication, see [15] for the detailed formula.



(a) Non-continuous closure (NCC) (b) Periodic closure (PC)

Figure 12: Differences between the non-continuous (NCC) and periodic (PC) closure assumptions. On the left: the non-continuous case, we add an extra phantom base pair (in red) at the coordinates of the first base pair (in green) ($o_1 = o_{n+1}$ and $q_1 = \pm q_{n+1}$). The junction is modelled through the interactions between the last (blue) and the phantom (red) base pairs. The last base pair is shown in blue. On the right: the periodic case, we treat the first (green) and last (blue) base pairs as if they were interior bases.

where p is a constant scalar penalty weight. In our case, we use $p = 100$. Since the energy is invariant under a scaling of the q_i, Equation (13) enforces the unit norm condition. The function to be minimized is then

$$\tilde{U}(z) = \frac{1}{2}(F(z) - \mu)^T K(F(z) - \mu) + p\sum_{i=1}^{n}(\|q_i\|^2 - 1)^2. \tag{14}$$

In order to speed up the process, we feed explicit expressions for the gradient and the Hessian matrix to the minimization algorithm. Detailed expressions for the gradient and the Hessian due to Manning can be found in Appendices F.

3.3 Periodic cgDNAmin

In his work, Manning was interested in the formation of minicircles. Hence, he modelled non-continuous closure (NNC) DNA minicircles. He did this by introducing an additional phantom base pair representing the first base pair at the end of the sequence and requiring that the first and last base pair frames overlap. However, another interesting question is to model the shape of an already fully formed DNA minicircle. We can do this using Głowacki's work, see Section 1.2. Using the periodic cgDNA construction, we can adapt NNC cgDNAmin to model a periodic interaction between the two end base pairs. We will denote it PC cgDNAmin for "Periodic Closure". The difference in the relation between the first and last base pairs is shown in Figure 12.

The main adaptation appears in the construction of the gradient and the Hessian. We need to adapt the closed form expression used in our algorithm and derived by Manning to work with the

extra upper-right and bottom-left blocks of the periodic stiffness matrix. Fortunately, the entries are computed similarly to the other diagonal entries. Each part of the block is an intra-intra or intra-inter relation between the last and first base pairs. Hence we can use the same formulas, being careful to place the entries at the correct positions. We again refer to Appendices F for the detailed computations.

To compare the effects of closure assumptions, we refer to the results for the Kahn & Crothers sequence, Section 3.4.1, Figures 15 and 16. They show the solutions for four different initial guesses for respectively non-continuous and periodic closure assumptions. By comparing the 3D views of the solutions, we observe that changing the closure assumption can change the solution slightly. However, the final configurations remain relatively close to each others. We measure the relative $\mathcal{L}^2$ norm of the difference in cgDNA coordinates to obtain a quantitative comparison. For the four initial guesses, we get

$$\begin{aligned}
\frac{\|w_{\mathrm{NCC}}(\text{K-C}, a) - w_{\mathrm{PC}}(\text{K-C}, a)\|_2}{\|w_{\mathrm{NCC}}(\text{K-C}, a)\|_2} &= 1.42\%, \\
\frac{\|w_{\mathrm{NCC}}(\text{K-C}, b) - w_{\mathrm{PC}}(\text{K-C}, b)\|_2}{\|w_{\mathrm{NCC}}(\text{K-C}, b)\|_2} &= 1.41\%, \\
\frac{\|w_{\mathrm{NCC}}(\text{K-C}, c) - w_{\mathrm{PC}}(\text{K-C}, c)\|_2}{\|w_{\mathrm{NCC}}(\text{K-C}, c)\|_2} &= 1.43\%, \\
\frac{\|w_{\mathrm{NCC}}(\text{K-C}, d) - w_{\mathrm{PC}}(\text{K-C}, d)\|_2}{\|w_{\mathrm{NCC}}(\text{K-C}, d)\|_2} &= 1.48\%,
\end{aligned} \tag{15}$$

where $w_b(g)$ is the cgDNA coordinate vector of the converged configuration from the minimization of initial guess g and boundary condition b. Note that in order to compare the results, one has to remove the last intra and inter coordinates from the NCC solutions and the last inter coordinates form the PC solutions. This is done to remove the phantom base pair of NCC algorithm and remove the extra inter coordinates introduced by the periodic closure assumption. The computations of the relative norm of the differences tell us that the solutions are not changing drastically when changing the closure assumption. We can compute the relative difference between the coordinates of the NCC and PC closure assumptions. We find that it is about 1.5%, this is sufficiently big to be non negligible but we can say that the PC solutions remain relatively close to the original NCC configurations. If we compare the energies of NCC and PC cgDNAmin solution configurations, we see that they are very close to each others. Because the results are not fundamentally different, we will only focus on the non-continuous case (NCC) from now on. This will enhance the comparison against Manning's work [15].

3.4 Examples

Now that we saw the process behind cgDNAmin, we show some examples of cgDNA discrete energy minimization for two particular sequences: the Kahn & Crothers c11t15 sequence already studied in [6, 7, 18] and the Poly A 158bp sequence. Both sequences have very different ground-states. The first one is very bent due to the presence of so-called A-tracts, whereas the second one is completely straight. It is then interesting to compare the behaviour of cgDNAmin with both sequences. We show the 3D views of 4 initial guesses and their corresponding results for both sequences. For all 3D

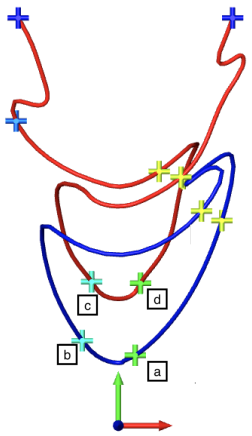


Figure 13: bBDNA bifurcation diagram for the Kahn & Crothers sequence. The energies of the initial configurations are computed and shown together with the 3D views in Figure 14. The crosses are all the computed closed equilibria along the branches. We note that there could be additional branches that are entirely missed by the computation.

views, we specify the link[17] (number of twists in the coupled backbones) and the cgDNA energy[18] of the configuration for further interpretation of the results. We recall that the energy is obtained using Equation (8). Finally, for all final configurations we compute the smallest eigenvalue of the Hessian matrix to ensure that the configuration is a minimizer for the discrete energy. We discuss our observations on the results in Section 3.4.3.

3.4.1 Kahn & Crothers c11t15 sequence

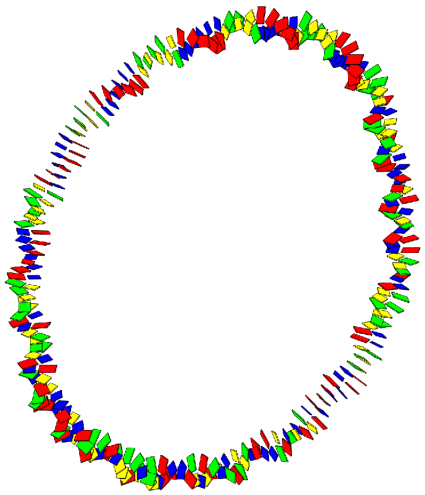
We start with the sequence c11t15 with a strong intrinsic bend. Its strong intrinsic bend may reduces the stress on the molecule when forced into a minicircle shape, which facilitates the formation of such loops.

As we explained in Section 3.1, we start by computing equilibria for the continuum model with the help of bBDNA. The bifurcation diagram is shown in Figure 13. As detailed earlier, it represents what are believed to be all closed continuum equilibrium configurations. The crosses represent the configurations where the backbones are correctly aligned to form a correctly closed minicircle. We recall that, in the default case which is used here, the height (Y-axis, in green) represents the energy of the solution and the abscissa represent the constant force along the equilibrium on the Z-axis (in blue, out of the plane) and the internal twist on X-axis (in red). The color of the crosses shows the link of the configuration. Refer to [10, 12] for more details on the continuum birod model and the bBDNA software.

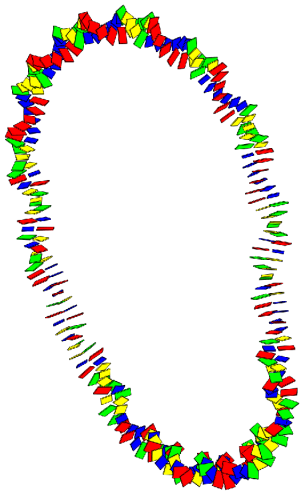
For the sake of comparison, we show the 3D views of the final solution as well as the configurations of the initial guesses. The initial guesses are shown in Figure 15 and the corresponding outputs

[17] Link is counted by hand, this may induce error.
[18] Rounded to the closest integer for clarity. The unity is $k_B T$

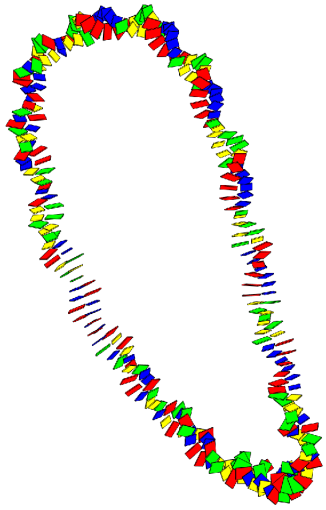
of NCCcgDNAmin can be seen in Figure 15. For more detailed comparison, we plot the cgDNA coordinates of the same initial guesses in the dedicated webpage, see Appendices A Finally, we show the result for the periodic closure assumptions in Figure 16. This allows us to observe the effect of changing the closure assumption. We refer to Section 3.3 for the differences between the NCC and PC closure assumptions.



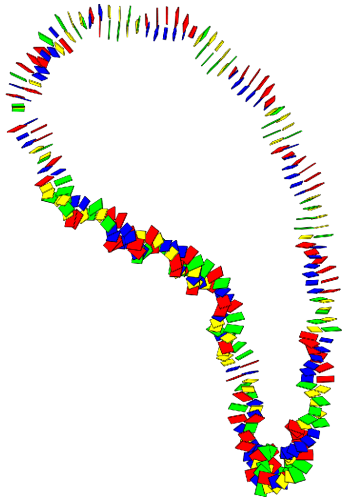
(a) Link: 15, Energy: 418



(b) Link: 14, Energy: 482

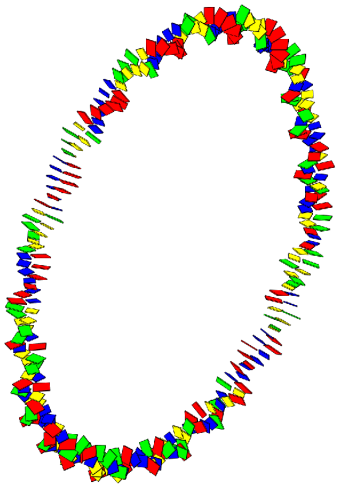


(c) Link: 14, Energy: 481

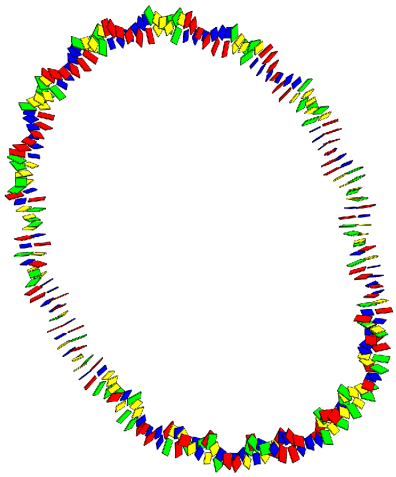


(d) Link: 15, Energy: 699

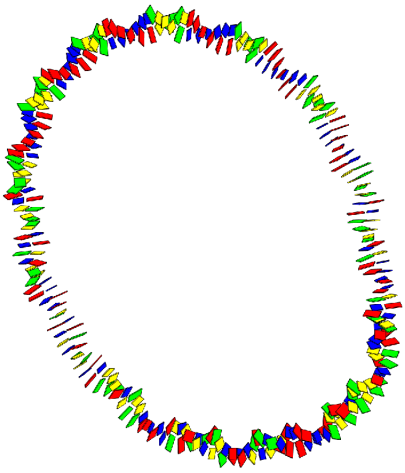
Figure 14: 3D views of **initial guesses** for the **Kahn & Crothers** sequence. Each initial guess is the indicated cross in Figure 13. The Link (number of twists of the coupled backbones) is counted by hand and the energy is obtained using Equation (8).



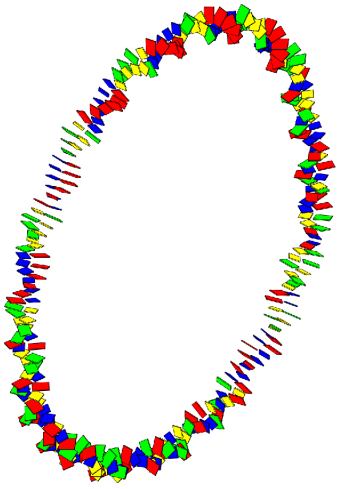
(a) Link: 15, Energy: 35,
minimum eigenvalue: 1.36e-6



(b) Link: 14, Energy: 84,
minimum eigenvalue: 2.42e-6

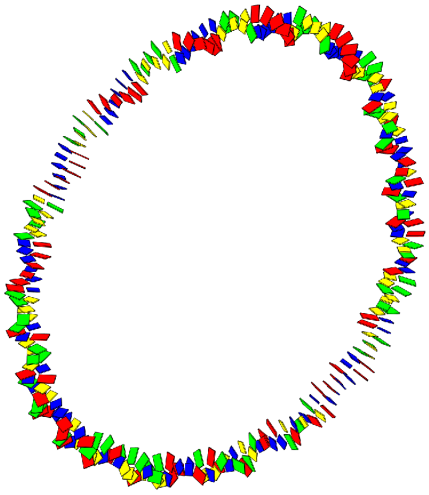


(c) Link: 14, Energy: 84,
minimum eigenvalue: 2.41e-6

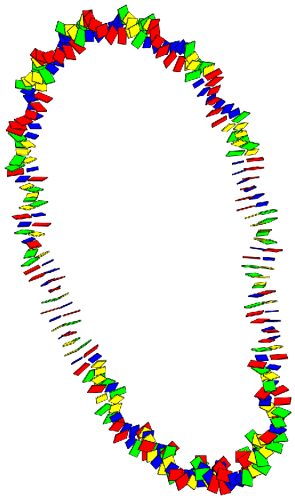


(d) Link: 15, Energy: 35,
minimum eigenvalue: 1.40e-6

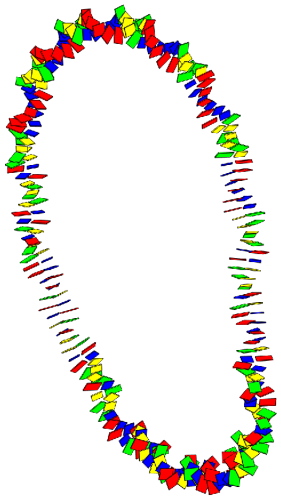
Figure 15: 3D views of the **NCCcgDNAmin** solutions for the **Kahn & Crothers** sequence. Each panel is the solution given by the corresponding initial guess from Figure 14. The Link (number of turns of the intertwined backbones) is counted by hand. Solutions a) and d) represent the same configuration and solution b) and c) are similar as well. All four configurations have positive definite Hessian, hence they all seem to be local minimizers of the energy.



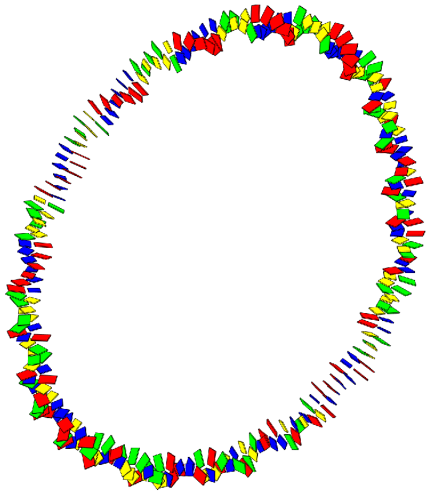
(a) Link: 15, Energy: 35,
minimum eigenvalue: 1.48e-6



(b) Link: 14, Energy: 84,
minimum eigenvalue: 2.23e-6



(c) Link: 14, Energy: 84,
minimum eigenvalue: 2.26e-6



(d) Link: 15, Energy: 35,
minimum eigenvalue: 1.46e-6

Figure 16: 3D views of the **PCcgDNAmin** solutions for the **Kahn & Crothers** sequence. From the 2D coordinates inspection, we have that solutions a) and d) represent the same configuration and solution b) and c) are similar as well. Compared to the NCCcgDNAmin solutions of Figure 15, they only differ by 1.5%, which is significant but relatively small.

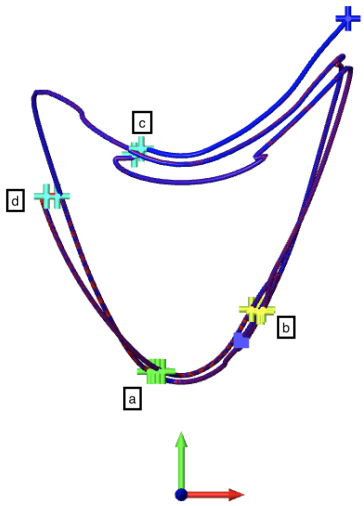
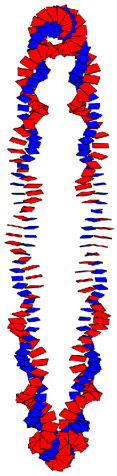


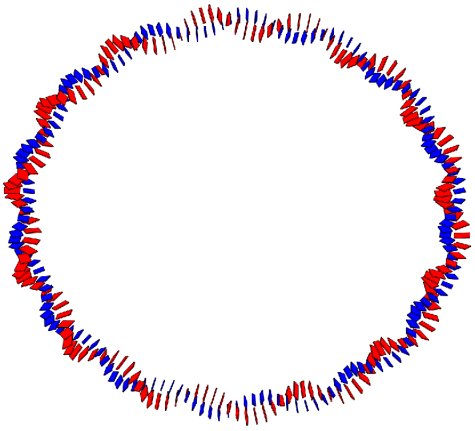
Figure 17: bBDNA bifurcation diagram for Poly A 158bp. Note all guesses within the small groups are physically the same up to a symmetry. We also see that the curve is at least double covered. This may be a result of the solver running twice or more on the same curve but this is not entirely clear.

3.4.2 Poly A, 158 bp

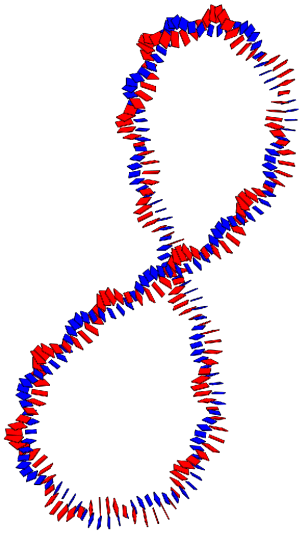
We repeat the procedure with the Poly A sequence of 158 base pairs. This is a very stiff sequence with a straight ground-state. The formation of a minicircle require a much higher energy than the previous example. Once again, we start with an initial run of bBDNA and the bifurcation diagram is shown in Figure 17. We immediately see that the bifurcation diagram is exceptional. Due to the high symmetry of the sequence, we can observe four groups of solutions. This explains the special form of the diagram. All crosses are valid initial guesses for the cgDNAmin energy minimization. We show four discretized bBDNA equilibrium in Figure 18 and their corresponding NCCcgDNAmin minimal energy configuration in Figure 19. The 2D plots of the cgDNA coordinates are shown on the webpage. The straight shape of the ground state is highlighted in the 2D plots and we can observe the periodicity in the coordinates induced by the cyclization.



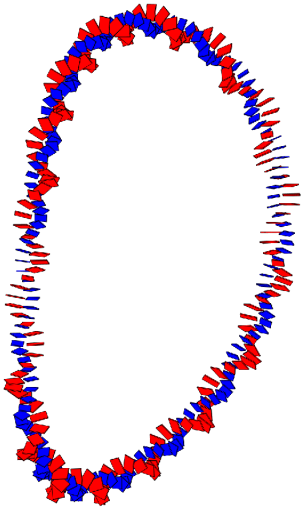
(a) Link: 15, Energy: 132



(b) Link: 14, Energy: 212

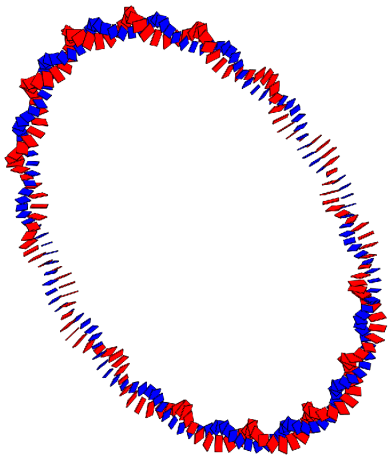


(c) Link: 16, Energy: 321

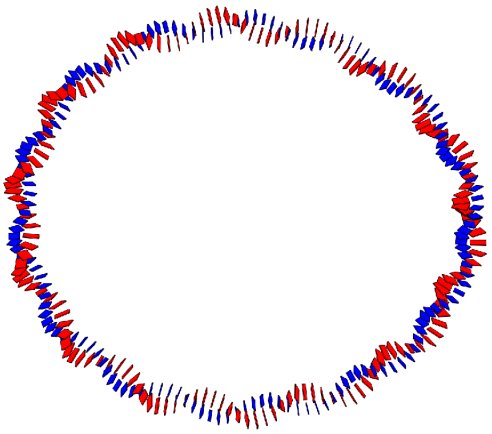


(d) Link: 17, Energy: 219

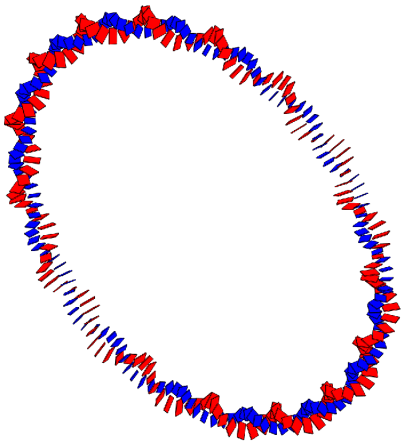
Figure 18: 3D views of **Poly A 158bp initial guesses**. Each configuration is represented by a cross in Figure 17.



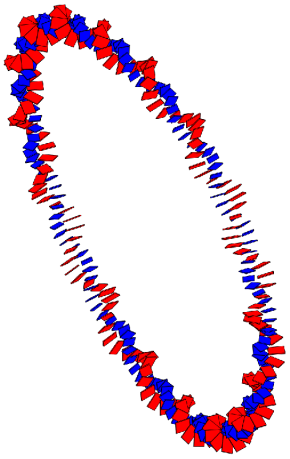
(a) Link: 15, Energy: 59,
minimum eigenvalue: 4.44e-7



(b) Link: 14, Energy: 140,
minimum eigenvalue: 3.48e-7



(c) Link: 16, Energy: 59,
minimum eigenvalue: 5.67e-7



(d) Link: 17, Energy: 143,
minimum eigenvalue: 6.26e-7

Figure 19: 3D views of **Poly A 158bp NCCcgDNAmin** solutions. Each panel is the solution given by the corresponding initial guess in Figure 18. From 2D coordinates and link inspection, we find that all four solutions are different. We verify the positive definiteness of the Hessian matrix and all four configurations appear to be local minima.

3.4.3 Observations on cgDNAmin

The first observation to make is that the obtained solutions are shapes close to circle. We do not see any highly writhed or non-planar equilibrium configuration regardless of the link of the initial guess.

The second striking observation from both examples is that two different initial guesses can lead to the same minimization output. For the Kahn & Crothers sequence, initial guesses a) and d) yield almost the same results and guesses b) and c) are very similar as well. If we compute the relative $\mathcal{L}^2$ norms of the differences, we obtain that

$$\frac{\|w(\text{K-C}, a) - w(\text{K-C}, d)\|_2}{\|w(\text{K-C}, a)\|_2} = 0.19\%, \tag{16}$$

and

$$\frac{\|w(\text{K-C}, b) - w(\text{K-C}, c)\|_2}{\|w(\text{K-C}, b)\|_2} = 0.07\%, \tag{17}$$

where $w(\text{K-C}, g)$ is the configuration of the solution given by the Kahn & Crothers sequence and initial guess g. We immediately see that the difference is relatively small and solutions a) and d) can be considered the same and configurations b) and c) are also similar. We refer to the 2D coordinate plots on the webpage for a more detailed inspection of the coordinates. We note that this similarity is independent of the choice of closure assumption. For both the NCC and PC cgDNAmin minimizations, pairs of initial guesses yielded the same final configuration.

For the Poly A sequence, we observe a similar behaviour. Guesses a) and c) output configurations that look very similar in the 3D viewer. However, the relative norm of the difference shows the opposite result:

$$\frac{\|w(\text{Poly A}, a) - w(\text{Poly A}, c)\|_2}{\|w(\text{Poly A}, a)\|_2} = 10.34\%. \tag{18}$$

We can use this example as a warning, we should not only make our observations on 3D views alone. We also need to interpret the 2D coordinates plots as they are much more precise. Another good way to compare solutions is to count the link of the molecule, i.e. the number of twists of the coupled backbones. It is interesting to see that the cgDNAmin solutions preserve the number of twists of the initial guess. That means, from our observations, that initial solutions with different link will not converge to the same final configuration. This need not be the case because the cgDNA model (and by extension the cgDNA+ model) is a so-called phantom chain model, i.e. remote parts of the backbones can pass through each others leading to change in link.

Different guesses leading to the same solution can be explained by the fact that some initial guesses are likely to be unstable equilibria. The minimization procedure then converges probably to a different stable configuration that can also be obtained from other initial guesses. We also observe that two initial guesses can seemingly converge to the same solution but the coordinates are not exactly the same, one possible explanation is that the two energy minimization procedures converge to the same actual solution but stop at slightly different configurations. Since the minimization algorithms stopping criteria is based on the objective function improvement[19], the step can be considered too small and the algorithm stop at a sightly different configuration.

[19] In our case the energy as a function of the configuration.

Part III
cgDNA+ Minicircles

This chapter is focused on the adaptation of the minicircle energy minimization to cgDNA+ coordinates. In other words the adaptation of Manning cgDNAmin algorithm to support Patelli's cgDNA+ model. We detail both non-continuous and periodic minicircle configurations. The idea is to change the coordinates and adapt the discrete energy minimization step to produce minimized cgDNA+ energy minicircle configurations starting from a discretization of a bBDNA equilibrium. We present different case studies for sequences from the literature. We use the non-continuous closure (NCC) cgDNA+min algorithm to validate our model. In order to observe the effect of the addition of the phosphate groups to the model, we also present a short comparison between the cgDNAmin and cgDNA+min procedures. Finally, we go through the provided MATLAB package so that the reader can repeat our results.

4 cgDNA+ Non-Continuous Minicircles

In this section we focus strictly on adapting cgDNAmin[20] to support cgDNA+ coordinates. We keep the discontinuity at the junction between the two ends (NCC condition). The procedure is similar to the original cgDNAmin pipeline. We start with the continuum minicircle energy minimization, we discretize the solution and we perform a discrete energy minimization. We note that we do not change the continuum estimation of coefficients that are input to bBDNA. Thus the initial guesses for the intra and inter coordinates are the same as in the case of cgDNAmin and initial guesses for the phosphate coordinates have to be added as the discretization of the continuum equilibrium does not directly provide information about the positions of the phosphates.

4.1 To cgDNA+ coordinates

As we saw in Section 1.3, the difference between cgDNA and cgDNA+ coordinates lies at the base pair level. Each set of intra base pair coordinates is now associated to two phosphate groups to form the new base pair level coordinates. We recall Equation (10) with the special cases $i = 1, n$:

$$\tilde{x}_i = [p_i^+, x_i, p_i^-], \quad i = 2, \dots, n-1,$$
$$\tilde{x}_1 = [x_1, p_1^-], \quad \tilde{x}_n = [p_n^+, x_n],$$

where x_i is the standard cgDNA intra coordinate of Equation (4) and $p_i^\pm$ are the coordinates of the Crick and Watson phosphates associated to base pair i.

The cgDNAmin energy is therefore modified to include the phosphate coordinates. As we see in Appendices F, the principal difficulty in the discrete energy minimization is the computation of derivatives with respect to the inter coordinates due to the use of quaternions. In the cgDNA+min case, the inter coordinates remain unchanged and the difference lies in the dimension of the base pair level coordinates. The new stiffness matrix is now a $(24n-18)\times(24n-18)$ matrix: $18(n-2)+2\cdot 12$ entries for the base pair level coordinates and $6(n-1)$ entries corresponding to inters. The subdivision of diagonal blocks shown in Equation (42) is slightly modified. The diagonal blocks are now

[20] detailed in Section 3.

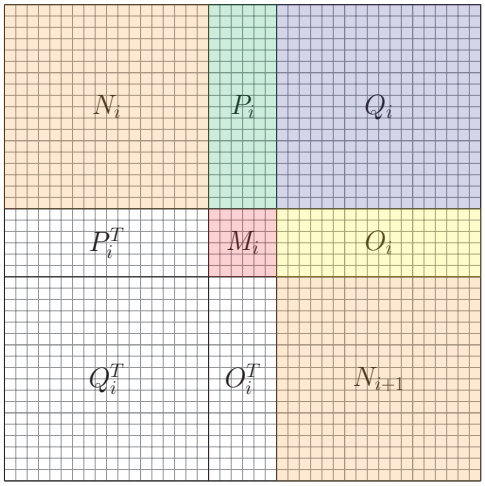


Figure 20: Sub-block division of the i-th 42×42 diagonal block of the cgDNA+ stiffness matrix K. Adaptation of Figure 36 to cgDNA+ coordinates. The full block is the diagonal sub-block of the stiffness matrix with indices ranging from $1 + 24(i-1)$ to $42 + 24(i-1)$ in both dimensions.

42×42 blocks and the sub-division is as follows: the intra-intra blocks N_i and Q_i become 18×18 blocks and the intra-inter and inter-intra blocks P_i and O_i change respectively to 18×6 and 6×18 blocks. The adaptation of the sub-division is shown in Figure 20. Once all dimensions are adapted to add the phosphate components within the intra coordinates, all derivatives are computed using the cgDNAmin formulas explained in Appendices F.

4.2 Phosphate initial guesses

As mentioned in the beginning of this section, the discretization of the continuum bBDNA solution does not give any information about the position of the phosphate groups. Hence we have to choose a valid initial guess. We need to start from coordinates that are sufficiently close to the final solution to ensure that the energy minimization procedure will converge.

The natural choice is to use the ground-state coordinates for the phosphates as initial guess. This choice has two motivations. First taking the ground-state phosphate coordinates as initial guess is the simplest thing to do. Since we work with the shifted coordinates $(w - \mu)$ the initial guess for the shifted phosphates coordinates is simply zero. This reduces the amount of computations before the minimization procedure. Second, this choice works. The minimization procedure converges with this choice for initial phosphate coordinates, hence we keep it.

5 cgDNA+ Periodic Minicircles

The adaptation of the periodic closure (PC) cgDNAmin to the cgDNA+ coordinates is straight forward. As we mentioned in Section 4, the only difference in the coordinates is the dimension of the base pair level coordinates. They lie in $\mathbb{R}^{18}$ instead of $\mathbb{R}^6$ to include phosphates coordinates. Hence, the periodic cgDNA+ stiffness matrix dimension is $24n \times 24n$. The issue about initial guesses for

the phosphate coordinates is exactly the same for both periodic and non-continuous cases. Hence we use the same technique to get the initial phosphates coordinates[21].

To represent the interactions between the first and last base pairs, we use the periodic cgDNA+ ground-state and stiffness matrix explained in Section 1.3.2 which are themselves adaptations to cgDNA+ coordinates of Głowacki's work shown in Section 1.2. Similarly to the PCcgDNAmin detailed in Section 3.3, the extra corner blocks introduce new components in the minimization vector z and the gradient and Hessian matrix. These extra components are computed following the same construction as the other blocks in the diagonal. Compared to NCCcgDNA+min, the periodic case has the advantage that there is no external phosphate groups. The intra coordinates of both end bases are not truncated. This slightly facilitates the implementation as we do not have to take care of the smaller base pair level coordinates for the two end base pairs.

6 cgDNA+min Case Studies

In this section, we provide different examples of results obtained with the non-continuous closure (NCC) cgDNA+ discrete energy minimization. We start with the same two examples shown in the cgDNA case in Section 3.4: the Kahn & Crothers c11t15 and the Poly A 158bp sequences. After these two examples, we turn to recently studied Pyne et al. sequences [23]. The idea is to compare the results and see if our final configurations are similar to the ones obtained in the Pyne et al. study. Finally, we also compute the cgDNA+min result for the well-studied Widom 601 sequence [4]. We will here only focus on the non-continuous closure assumption (NCC) as the difference between periodic and non-continuous closures are relatively small and analogous to the standard cgDNAmin case detailed in the examples of Section 3.4.

For each sequence, we show two figures: the 3D views of both chosen initial guesses and their corresponding NCC cgDNA+min solution. Again, we show the link (hand counted), the energy of each configuration and the smallest eigenvalue of the Hessian matrix. The idea of showing the initial guesses configurations is to observe the effect of the discrete cgDNA+ energy minimization. We also provide the 2D plots of the cgDNA and phosphate coordinates for the ground-state, the initial guesses and the final solutions on the thesis webpage.[22]

[21] We use the ground-state coordinates, see Section 4.2

[22] We recall that the full link to the webpage can be found in Appendices A.

6.1 Kahn & Crothers c11t15 sequence

As we do not change the intra and inter coordinates for the initial guesses between the cgDNA and cgDNA+ cases, we reuse the bifurcation diagram from Figure 13. The only difference is the presence of phosphate groups in the cgDNA+ case. We recall that the initial guess for the phosphate coordinates is the ground-state coordinates. Similarly to the standard cgDNA case in Section 3.4.1, we show the four different initial guesses in Figure 21 and their corresponding solution given by NCCcgDNA+min in Figure 22.

Similarly to the standard cgDNA case, we observe that initial guesses b) and c) yield similar configurations and a) and d) also give comparable results. In order to measure the difference between the two configurations, we compute the relative $\mathcal{L}^2$ norm of their coordinate vectors difference. For the a)-d) difference, we obtain

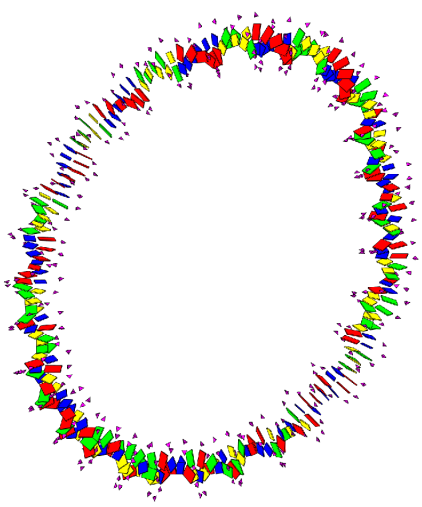
$$\frac{\|w(\text{K-C}, a) - w(\text{K-C}, d)\|_2}{\|w(\text{K-C}, a)\|_2} = 0.008\%, \tag{19}$$

and for the b)-c) comparison we have

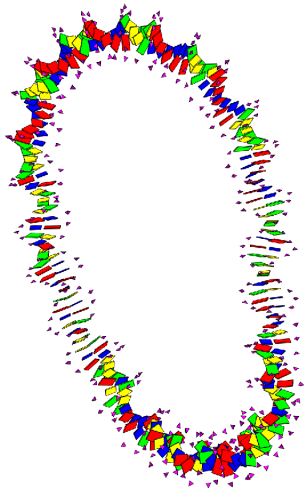
$$\frac{\|w(\text{K-C}, b) - w(\text{K-C}, c)\|_2}{\|w(\text{K-C}, b)\|_2} = 0.19\%. \tag{20}$$

Hence we can say that the solutions obtained represent the same configurations.

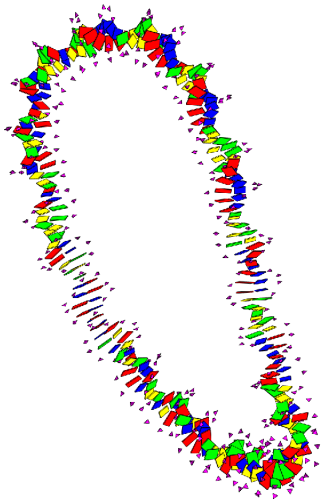
One striking observation is that the link is not conserved. For both guesses b) and c) the link is not the same between the initial guess and the final solution. This is striking since, from our observations, standard cgDNAmin appears to conserve link. Even if the count of the link is done by hand, the number of divergences between initial and final links is too big to be a counting mistake. We then have to ask ourselves the following question: what causes cgDNA+min solutions to change link? As we mentioned before in Section 3.4.3, both cgDNA and cgDNA+ are phantom chain models in the sense that the backbones can pass through each others without energy penalty. Therefore, there is no contradiction in the links of initial and final configuration differing. However, we noted that the initial and final links were always the same in all cgDNA simulations that have been carried out. On the contrary, for the cgDNA+ simulations of the same sequences, link does change in two cases. We note that it would be of interest to understand this phenomenon in greater details. But such study is beyond the scope of this thesis.



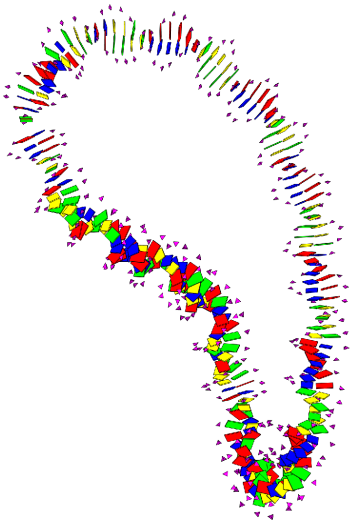
(a) Link: 15, Energy: 4015



(b) Link: 14, Energy: 4970

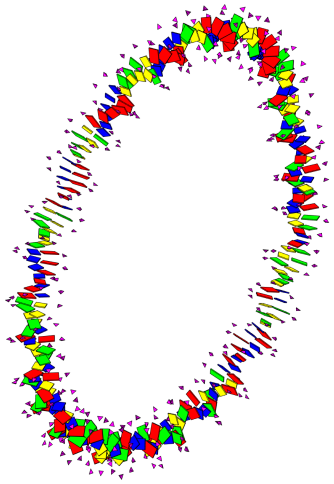


(c) Link: 14, Energy: 5051

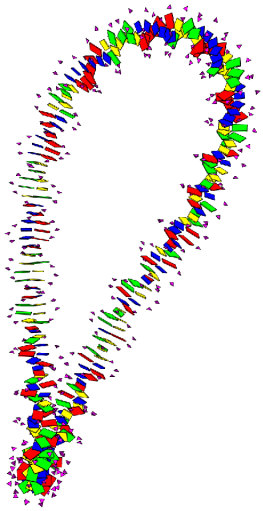


(d) Link: 15, Energy: 7094

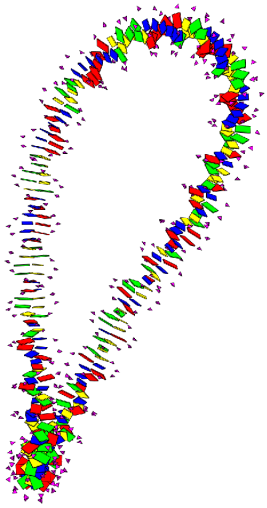
Figure 21: 3D views of the **Kahn & Crothers** sequence **initial guesses** with the added phosphate groups. Each configuration is represented by a cross in Figure 13. Note: the cgDNA coordinates (intras and inters) are identical to the ones of Figure 14.



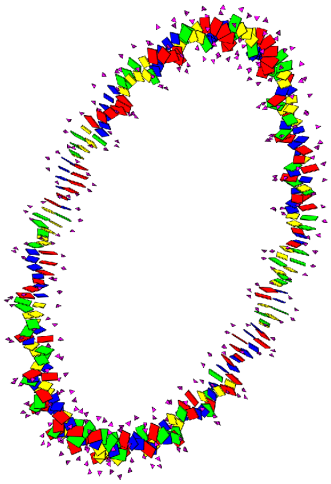
(a) Link: 15, Energy: 46,
minimum eigenvalue: 3.57e-6



(b) Link: 16, Energy: 110,
minimum eigenvalue: 2.46e-6



(c) Link: 16, Energy: 110,
minimum eigenvalue: 2.58e-6



(d) Link: 15, Energy: 46,
minimum eigenvalue: 3.63e-6

Figure 22: 3D views of the **Kahn & Crothers** solutions with **NCCcgDNA+min**. Each panel is the solution given by the corresponding initial guess in Figure 21. After 2D coordinates verifications, panels a) and d) show the same configuration and panels b) and c) also represent the same solution. After verification, all four configurations seem to be local minimizers.

6.2 Poly A, 158 bp

For the Poly A 158bp sequence, we again use the same bBDNA bifurcation diagram as in the cgDNAmin case, see Figure 17. As done before, we provide 3D MATLAB figures of the initial guesses in Figure 23. Those are the same as the standard cgDNAmin case in Figure 18 with added guesses for the phosphate groups. We then show the 3D views of the solutions in Figure 24.

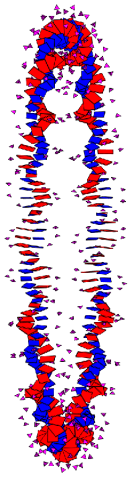
Again, we observe very similar results with different initial guesses. By computing the relative $\mathcal{L}^2$ norm of the difference between the coordinate vectors we get

$$\frac{\|w(\text{Poly A}, a) - w(\text{Poly A}, d)\|_2}{\|w(\text{Poly A}, a)\|_2} = 0.015\%, \tag{21}$$

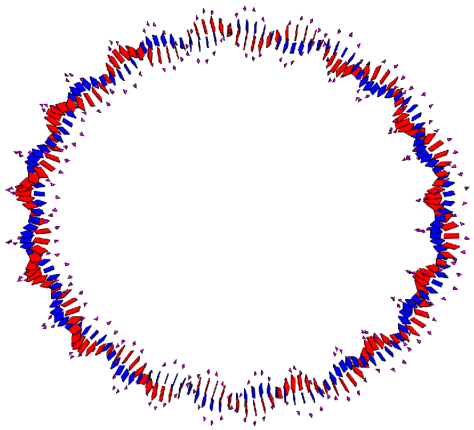
for the a)-d) and the b)-c) comparison gives

$$\frac{\|w(\text{Poly A}, b) - w(\text{Poly A}, c)\|_2}{\|w(\text{Poly A}, b)\|_2} = 0.75\%. \tag{22}$$

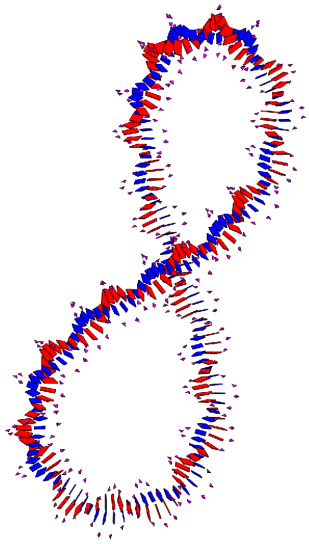
We consider the relative norms and the visual inspection of the 2D coordinates plots sufficient to conclude that both configurations are representing the same minimum energy configuration. This is surprising as the cgDNAmin algorithm outputs four different solutions whereas the cgDNA+min algorithm only outputs two different configurations for the same initial guesses. On the link conservation, we can again see that two initial guesses yield solutions with different links.



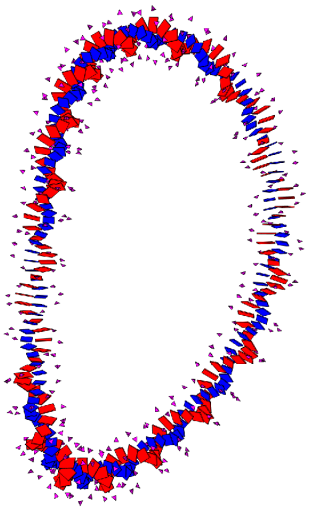
(a) Link: 15, Energy: 601



(b) Link: 14, Energy: 2003

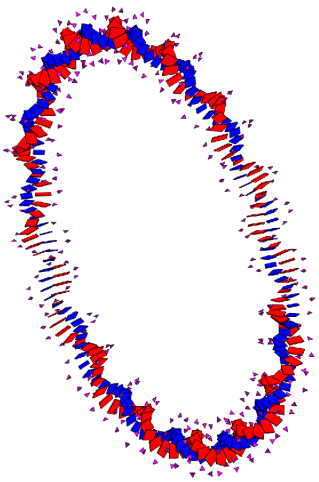


(c) Link: 16, Energy: 1427

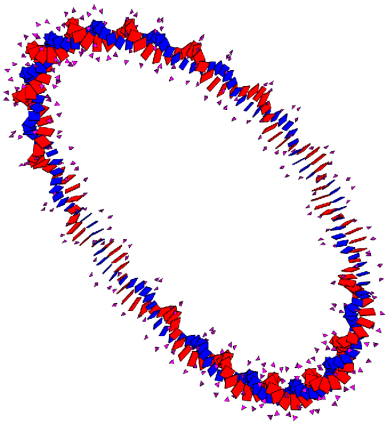


(d) Link: 17, Energy: 4220

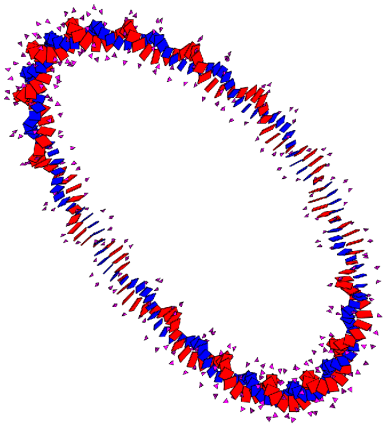
Figure 23: 3D views of **Poly A 158bp initial guesses** with the added phosphate groups. Each configuration is represented in Figure 17. Note: the coordinates for the bases are identical to the ones of Figure 18.



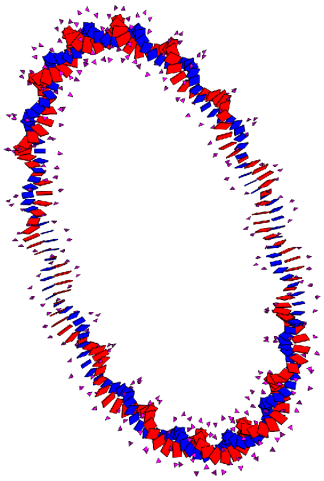
(a) Link: 15, Energy: 72,
minimum eigenvalue: 5.57e-7



(b) Link: 16, Energy: 112,
minimum eigenvalue: 8.84e-7



(c) Link: 16, Energy: 112,
minimum eigenvalue: 9.15e-7



(d) Link: 15, Energy: 72,
minimum eigenvalue: 5.72e-7

Figure 24: 3D views of **Poly A 158bp NCCcgDNA+min** solutions. Each panel is the solution given by the corresponding initial guess in Figure 23. Here we have panels a) and d) representing the same equilibrium configuration and panels b) and c) representing a second equilibrium. All four configurations have a positive definite Hessian, hence we assume they all are energy minimizer.

6.3 Pyne et al. sequences

After the two first examples, we focus on two sequences recently studied by Pyne et al. [23]. The interest to compute the cgDNA+min results for these sequences is to compare Pyne et al. molecular dynamics observations with our predictions. We show both bifurcation diagrams in Figure 25. The 251 base pairs sequence results can be found in Figures 27 and 28 while the results for the 339 base pairs sequence are shown in Figures 29 and 30. We provide an example of a 2D coordinate plot in Figure 26 and all other 2D coordinates plots can be found on the webpage.

For the 251 base pair sequence, we first observe that the result for the first initial guess is a particularly twisted loop. It is interesting as all previous results were mainly flat. All other three initial guesses yield more expected solutions, all three very close to each others. We can do the norm comparison to see that solution b) and d) are the same configuration with the relative norms of the differences of the coordinate vectors being

$$\frac{\|w(\text{Pyne } 251, b) - w(\text{Pyne } 251, d)\|_2}{\|w(\text{Pyne } 251, b)\|_2} = 0.013\%. \tag{23}$$

Solution c) is slightly more different, we verify this by computing the norms of the difference:

$$\frac{\|w(\text{Pyne } 251, b) - w(\text{Pyne } 251, c)\|_2}{\|w(\text{Pyne } 251, b)\|_2} = 0.24\%. \tag{24}$$

The relative norm is not sufficiently big to assert that initial guess c) gives a different solution from guesses b) and d). Hence we obtain the same minimum energy configuration for initial guesses b), c) and d) while a) gives an unexpectedly high energy configuration. When looking at the 2D cgDNA coordinate plots of Pyne 251 bp sequence in Figure 26, one can observe a particular behaviour: between approximatively bases 60 and 80 the coordinates vary less than in the rest of the sequence. This is seen for all guesses and in the cgDNA coordinates as well as in the phosphate coordinates (The other plots are provided on the webpage). We can therefore assume that this behaviour is induced by the sequence itself. The corresponding bases are alternating AG, see Appendices B. This sub-sequence is also seen in the 399 bp and the same effect can be observed.

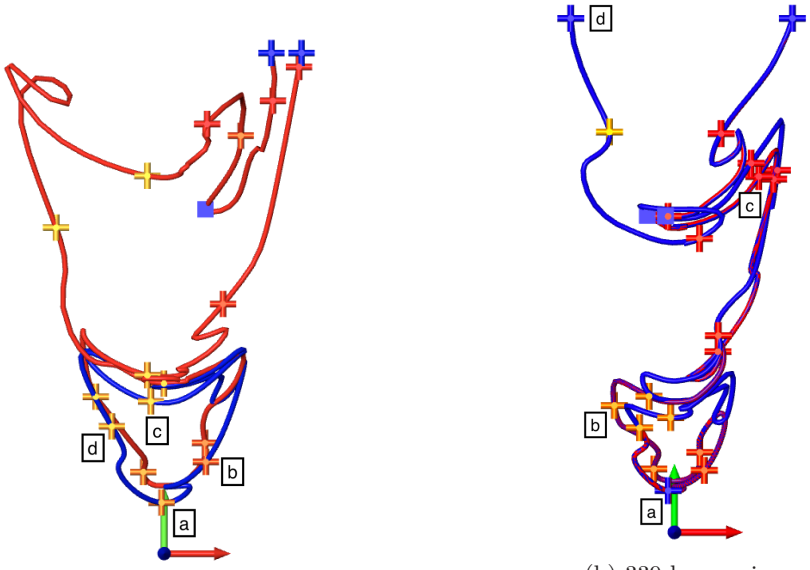
The 399 base pair results are similar. The first three initial guesses output standard minicircle shapes. Solutions b) and c) being very close to each others:

$$\frac{\|w(\text{Pyne } 251, b) - w(\text{Pyne } 251, c)\|_2}{\|w(\text{Pyne } 251, b)\|_2} = 0.32\%. \tag{25}$$

Again we observe an intriguing result, the solution obtained starting from initial guess d) is particular. The output forms some unexpected angles in some places in the sequence. I currently have no explanation for this phenomenon. This may be induced by A-tracts present close to the highly bent regions. This result is particularly unexpected and should require a deeper scientific analysis. On the link, for both sequences we see sequences changing link between the initial guess and the final solution. With all those observation with diverging links, we can assert that the cgDNA+min minimization does not conserve the link counter to original cgDNAmin.

We now compare solutions to the images presented in [23][23]. We can see that the solutions we obtain are less twisted than the images shown by Pyne et al. For both sequences, we obtain solutions with lightly twisted solutions in most cases. The overall energy of our results are, in most

[23] Fig.1 and Fig.2.a, link to the web article: https://www.nature.com/articles/s41467-021-21243-y.



(a) 251 base pairs

(b) 339 base pairs

Figure 25: Bifurcation diagrams for both Pyne sequences. Note that for the 339 bp sequence, we observe two superposed curves, this is due to the bBDNA solver.

cases, lower than the energy of the configurations seen in [23]. However, we still see configurations with higher energy in the bBDNA equilibriums. This shows that the cgDNA+ energy minimization converges to a very low energy configuration in most cases, even if the initial guess has a high energy.

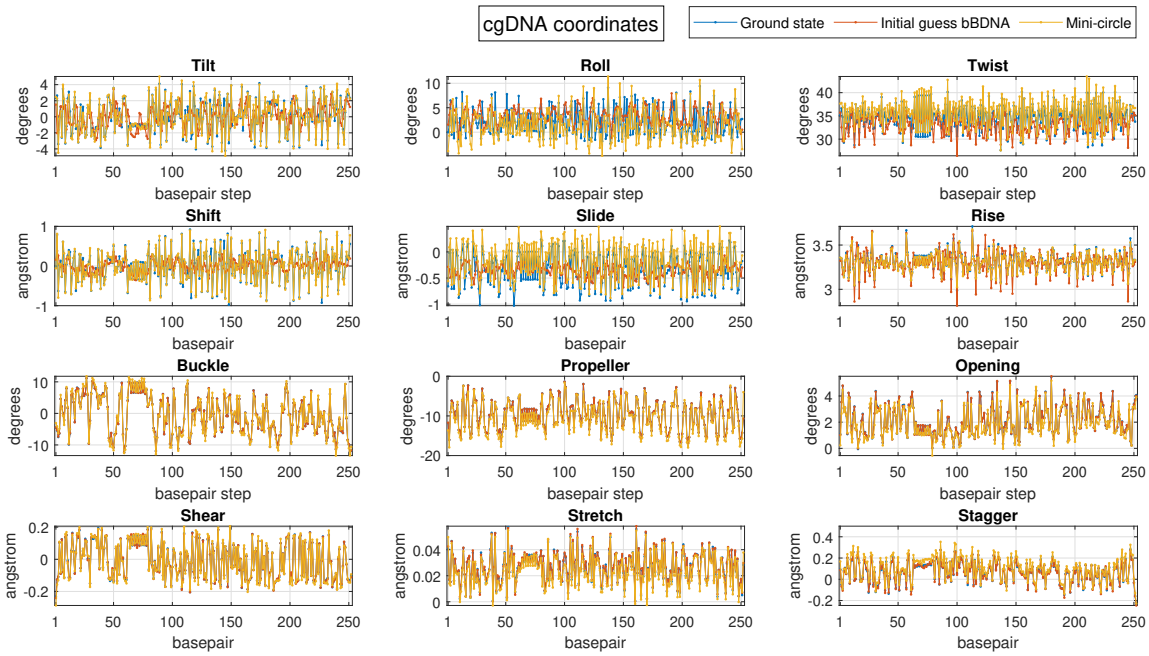


Figure 26: 2D plot of the cgDNA coordinates of the **NCCcgDNA+min** results for the **Pyne 251** sequence, initial guess a). We note the intriguing behaviour between base 60 and 80 which is the location of a sub-sequence of alternating AG.

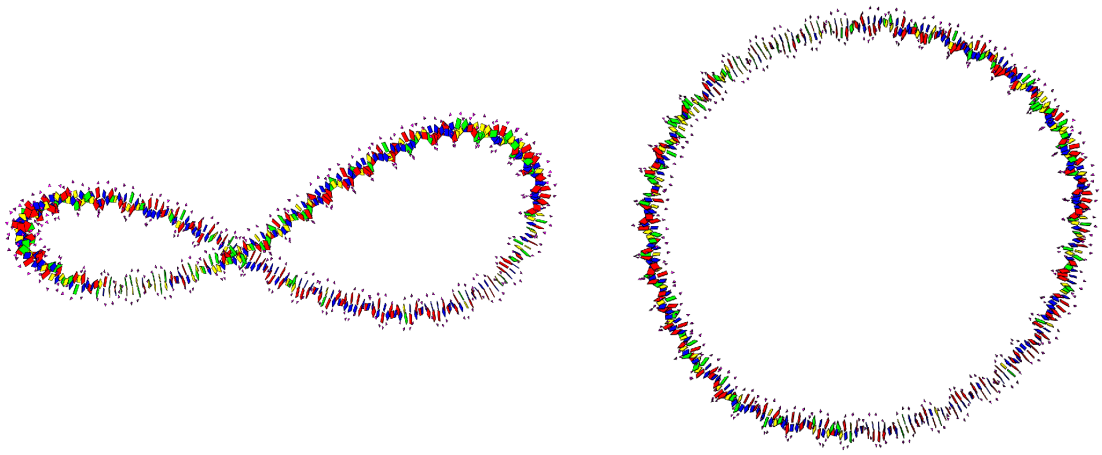
(a) Link: 23, Energy: 7356

(b) Link: 24, Energy: 7131

(c) Link: 22, Energy: 8900

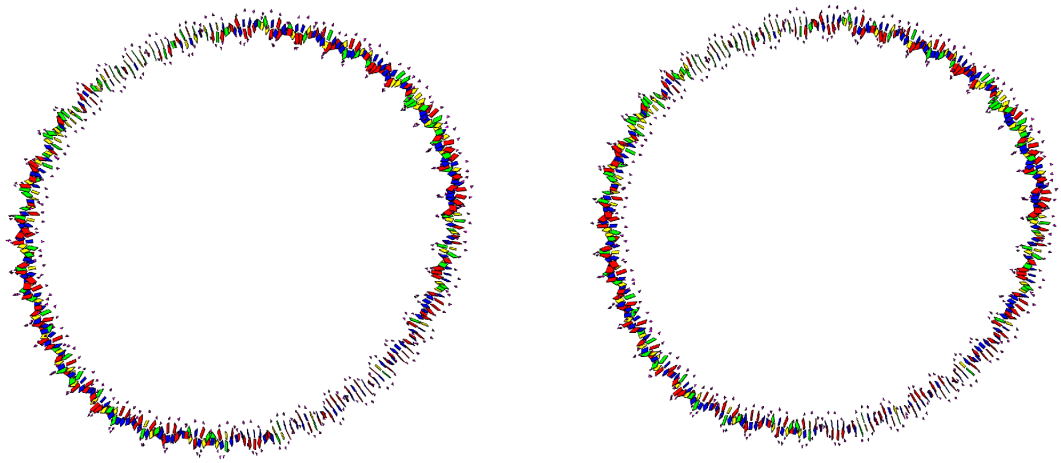
(d) Link: 24, Energy: 7941

Figure 27: 3D views of **Pyne 251bp** sequence **initial guesses** for cgDNA+min. Each configuration is represented in Figure 25a.



(a) Link: 23, Energy: 77,
minimum eigenvalue: 2.14e-7

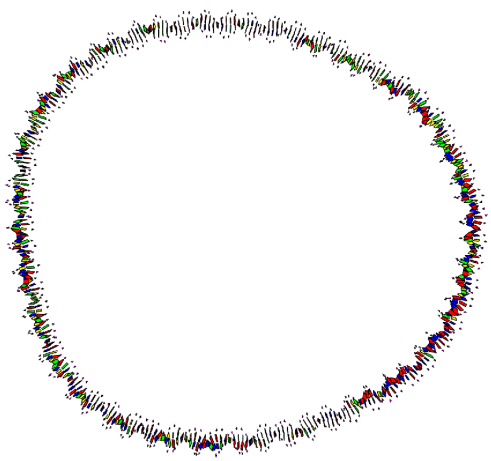
(b) Link: 24, Energy: 33,
minimum eigenvalue: 2.87e-7



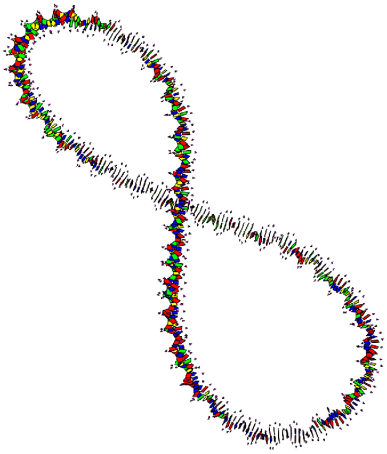
(c) Link: 24, Energy: 33,
minimum eigenvalue: 2.63e-7

(d) Link: 24, Energy: 33,
minimum eigenvalue: 2.87e-7

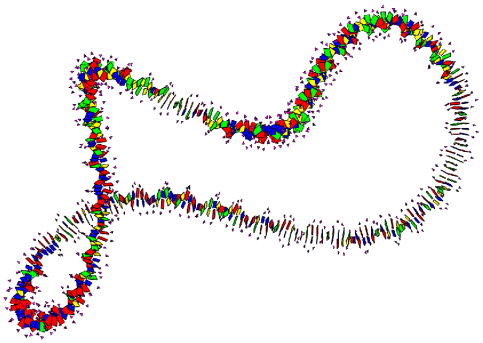
Figure 28: 3D views of **Pyne 251bp** sequence **NCCcgDNA+min** solutions. Each panel is the solution given by the corresponding initial guess in Figure 27. Solutions b), c) and d) are the all represent the same configuration. All four configurations have positive definite Hessian, hence they all seem to be local minimum energy configurations.



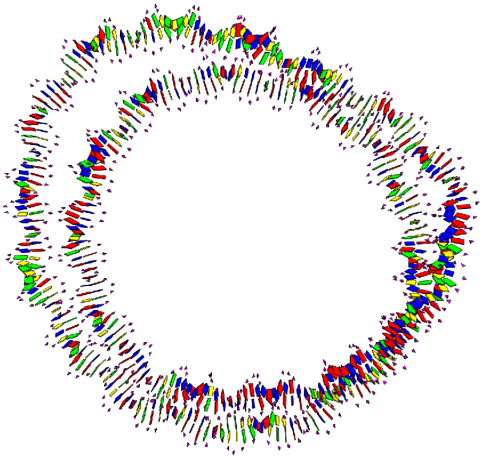
(a) Link: 31, Energy: 9867



(b) Link: 31, Energy: 10361

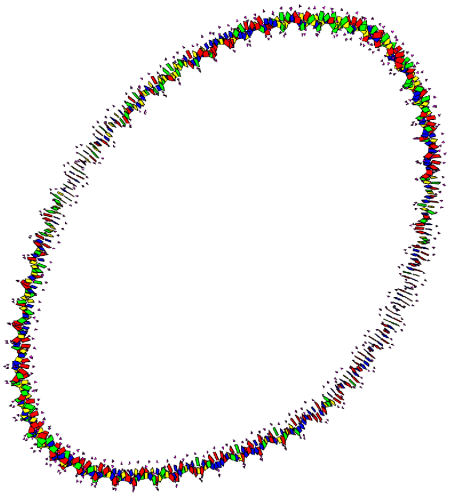


(c) Link: 34, Energy: 10654

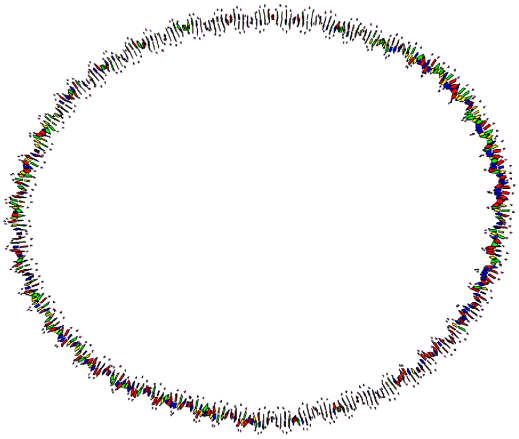


(d) Link: Too hard to count, we assume it to be 31 as in a) because of the color code in Figure 25b. Energy: 40691

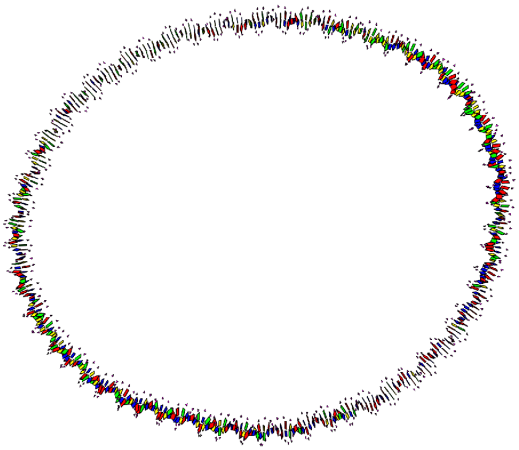
Figure 29: 3D views of **Pyne 339bp** sequence **initial guesses** for cgDNA+min. Each configuration is represented in Figure 25b.



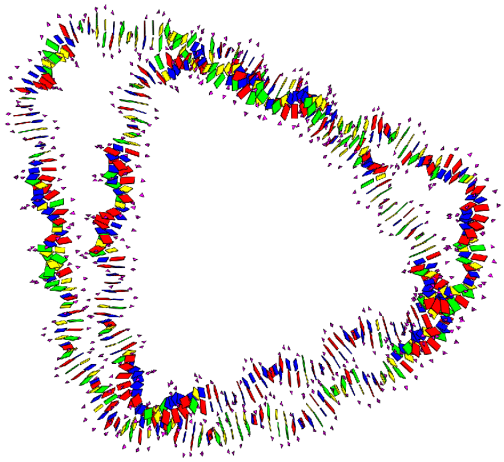
(a) Link: 33, Energy: 31,
minimum eigenvalue: 8.16e-8



(b) Link: 32, Energy: 40.5,
minimum eigenvalue: 1.09e-7



(c) Link: 32, Energy: 40.5,
minimum eigenvalue: 9.09e-8



(d) Link: Too hard to count! Energy: 955,
minimum eigenvalue: 5.39e-5

Figure 30: 3D views of **Pyne 339bp** sequence **NCCcgDNA+min** solutions. Each panel is the solution given by the corresponding initial guess in Figure 29. Solutions b) and c) both show the same configuration while a) is different. Panel d) shows a particularly interesting solution. All configurations have positive definite Hessian and are believed to be local energy minimizers. The equilibrium shown in panel d) is nevertheless quite suspicious.

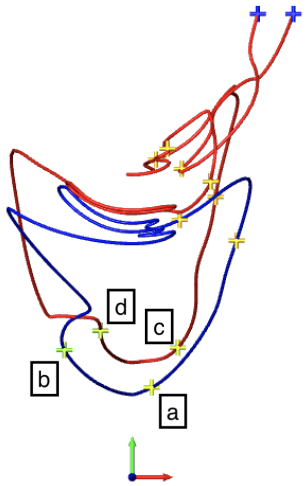


Figure 31: bBDNA bifurcation diagram for the Widom 601 sequence.

6.4 Widom 601 sequence

Finally, we present another well studied sequence: the Widom 601 94bp sequence [4, 13]. This is a small sequence with a high tendency to form minicircles. As usual, we present the bifurcation diagram, in Figure 31, and the 3D views of the initial guesses and their corresponding results in Figures 32 and 33 respectively.

We can see that the solutions given by initial guesses a) and c) are similar. We compute the relative norm of the difference of the coordinate vectors to verify this:

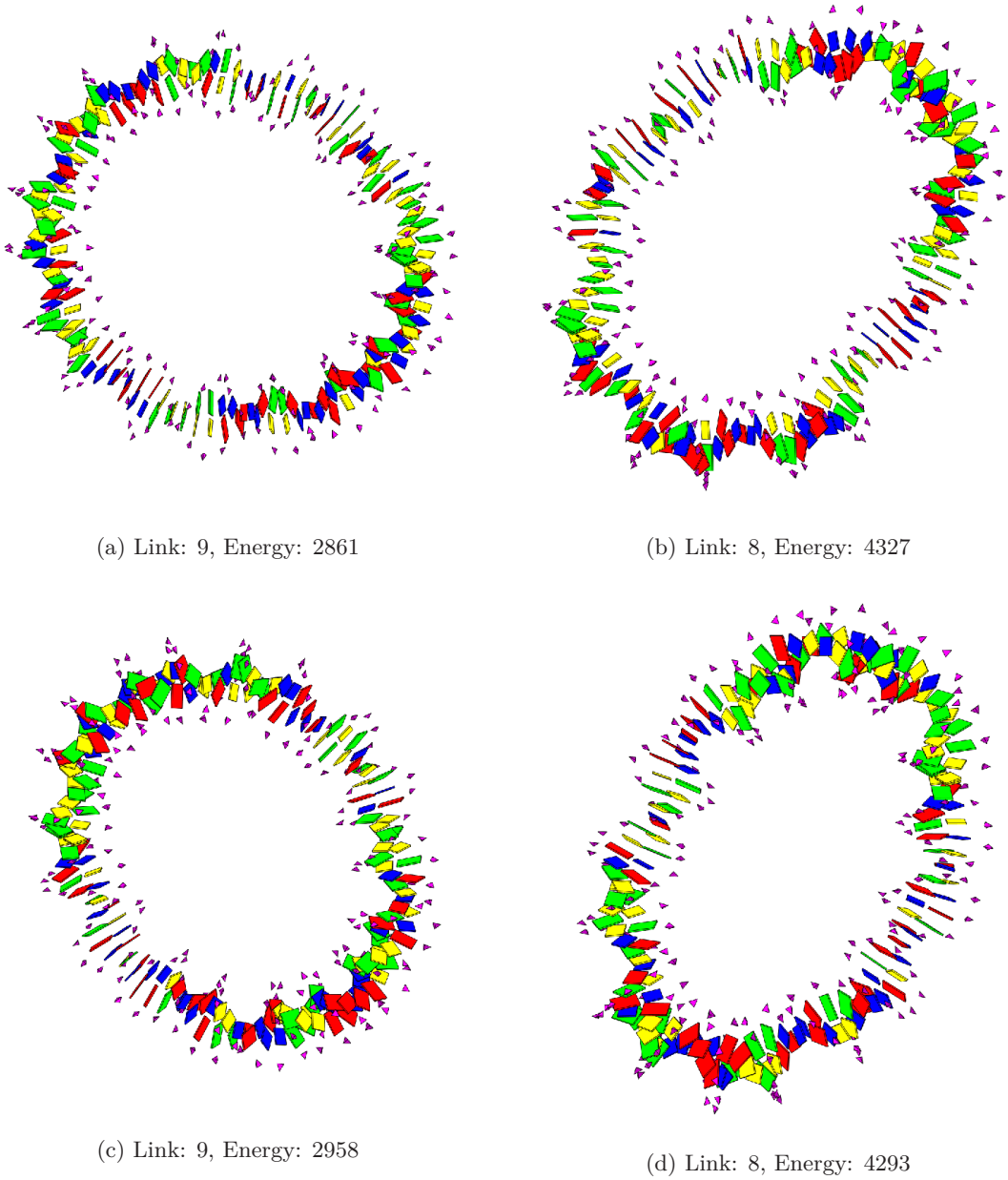
$$\frac{\|w(\text{Widom}, a) - w(\text{Widom}, c)\|_2}{\|w(\text{Widom}, a)\|_2} = 0.14\%. \tag{26}$$

From the norms, we conclude that both solutions are the same. Solution d) has a shape that look similar to solutions a) and c), however the solution is different. This can be seen directly by looking at the link. Once again, the link are not always conserved during the cgDNA+ energy minimization. Considering the b) and d) comparison, the relative norm is

$$\frac{\|w(\text{Widom}, b) - w(\text{Widom}, d)\|_2}{\|w(\text{Widom}, b)\|_2} = 8.84\%. \tag{27}$$

Hence, we reject the hypothesis that both b) and d) solutions represent the same configuration.

We remark that initial guess b) yields an interesting result. By its position on the bifurcation diagram, we would expect that the continuum equilibrium is stable, however, it seems not to be the case. The discrete energy minimization yields a highly twisted minicircle configuration, which is unexpected with a small sequence.



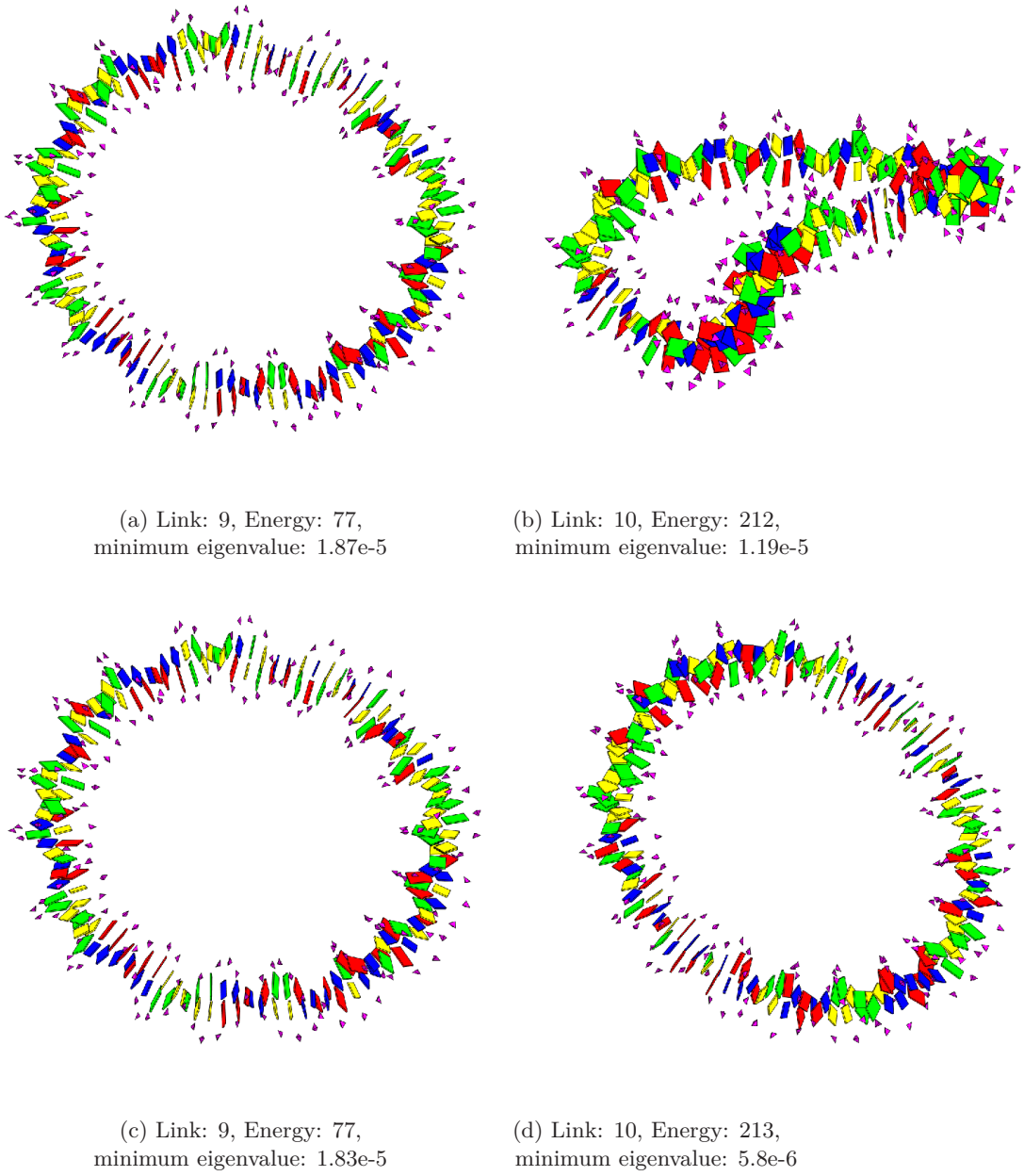
(a) Link: 9, Energy: 2861

(b) Link: 8, Energy: 4327

(c) Link: 9, Energy: 2958

(d) Link: 8, Energy: 4293

Figure 32: 3D views of **Widom 601** sequence **initial guesses** for cgDNA+min. Each configuration is represented in Figure 31.



(a) Link: 9, Energy: 77, minimum eigenvalue: 1.87e-5

(b) Link: 10, Energy: 212, minimum eigenvalue: 1.19e-5

(c) Link: 9, Energy: 77, minimum eigenvalue: 1.83e-5

(d) Link: 10, Energy: 213, minimum eigenvalue: 5.8e-6

Figure 33: 3D views of **Widom 601** sequence **NCCcgDNA+min** solutions. Each panel is the solution given by the corresponding initial guess in Figure 32. Panels a) and c) show the same solution while b) and d) are both different. After verification via the Hessian matrix, all four configurations are believed to be local energy minima.

7 Comparison between cgDNA+min vs cgDNAmin

In this section, we are interested to compare the results obtained with the cgDNA+ adaptation of the cgDNAmin software to Manning's original work.[24] We follow the same examples detailed in Section 3.4 to be able to compare the adaptation to cgDNA+ coordinates. We recall that the intra and inter coordinates of the initial guesses obtained from the discretization of the bBDNA equilibrium solution remain unchanged in the cgDNA+min adaptation. Hence we refer to Figure 13 for the bifurcations diagram. We note that we compare cgDNA and cgDNA+ energies despite them having different numbers of degrees of freedom.

7.1 Examples

Looking only at the cgDNA coordinates (ignoring the phosphate groups), the results are, at first glance, very similar. When looking a bit more in detail, we can observe a few differences in the coordinates. This is observable in the 2D plots and the differences are shown in Figure 34. These differences are normal since the cgDNA+ model adds the phosphate groups interactions. However, for certain initial guesses, the final result obtained with the cgDNA+ adaptation is very different from the one obtained in the cgDNA case, see Figure 35.

In both examples (Figures 34 and 35), the 2D plots of the coordinate difference show that the difference is relatively big for the inter rotations, medium for the inter translations, small for the intra rotations and tiny for the intra translations. This shows that adding the phosphates has a bigger impact on the inter coordinates and on the relative rotations between frames.

In the first case, the relative norm of the difference is

$$\frac{\|w_{\text{cgDNA}}(\text{K-C}, a) - \tilde{w}_{\text{cgDNA+}}(\text{K-C}, a)\|_2}{\|w_{\text{cgDNA}}(\text{K-C}, a)\|_2} = 5.30\%, \tag{28}$$

which is relatively low. The difference in the second case is much bigger:

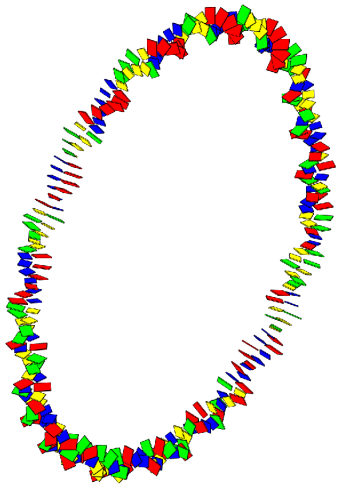
$$\frac{\|w_{\text{cgDNA}}(\text{K-C}, c) - \tilde{w}_{\text{cgDNA+}}(\text{K-C}, c)\|_2}{\|w_{\text{cgDNA}}(\text{K-C}, c)\|_2} = 15.75\%. \tag{29}$$

In both equations $\tilde{w}_{\text{cgDNA+}}$ is the marginalized coordinate vector of the cgDNA+ solution. It is the cgDNA+ configuration vector where the phosphate coordinates are removed. As we can see, in both case the difference is significant. This is due to the fundamental difference between cgDNA and cgDNA+. In the first case, this difference is sufficiently small to be seen only as the effect of the cgDNA+ phosphate coordinates. However, in the second case the difference is too big to be only the effect of the phosphates.

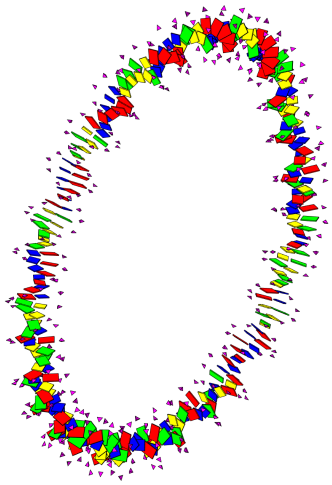
7.2 Suppositions for explanation

When comparing the results of cgDNA and cgDNA+ energy minimization, some initial guesses from bBDNA may yield very different results. Additionally, we saw in Section 3.4 that different initial guesses may lead to similar results after standard cgDNA energy minimization. The supposed explanation is that the discretization of the bBDNA minimum energy configuration may be unstable and the discrete energy minimzation will lead to a different stable configuration. Therefore, the

[24] see Section 3 for the standard cgDNAmin and Sections 4 and 5 for the cgDNA+ adaptations



(a) cgDNA 3D view



(b) cgDNA+ 3D view

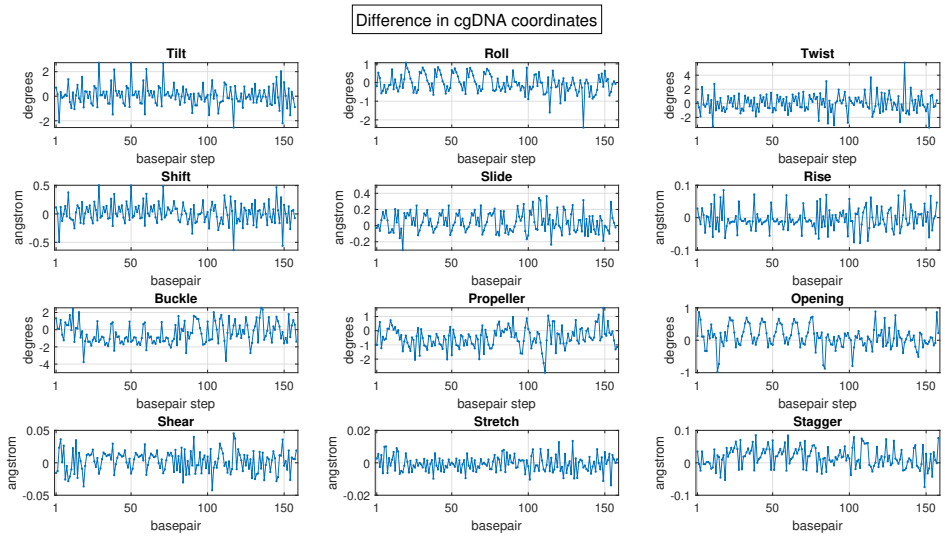


Figure 34: 3D view and 2D coordinates plots of the difference of the cgDNA coordinates between the cgDNA and cgDNA+ minimization routines, i.e. this is the difference between coordinates of converged minimizer. Note, we do not plot the phosphate coordinates as we cannot compare them with the results of cgDNAmin. Usual case: both results are similar, sequence: Kahn & Crothers c11t15, initial guess a) in Figures 14 and 21. The relative difference between the cgDNA coordinates of the two configurations is only 5.30%.

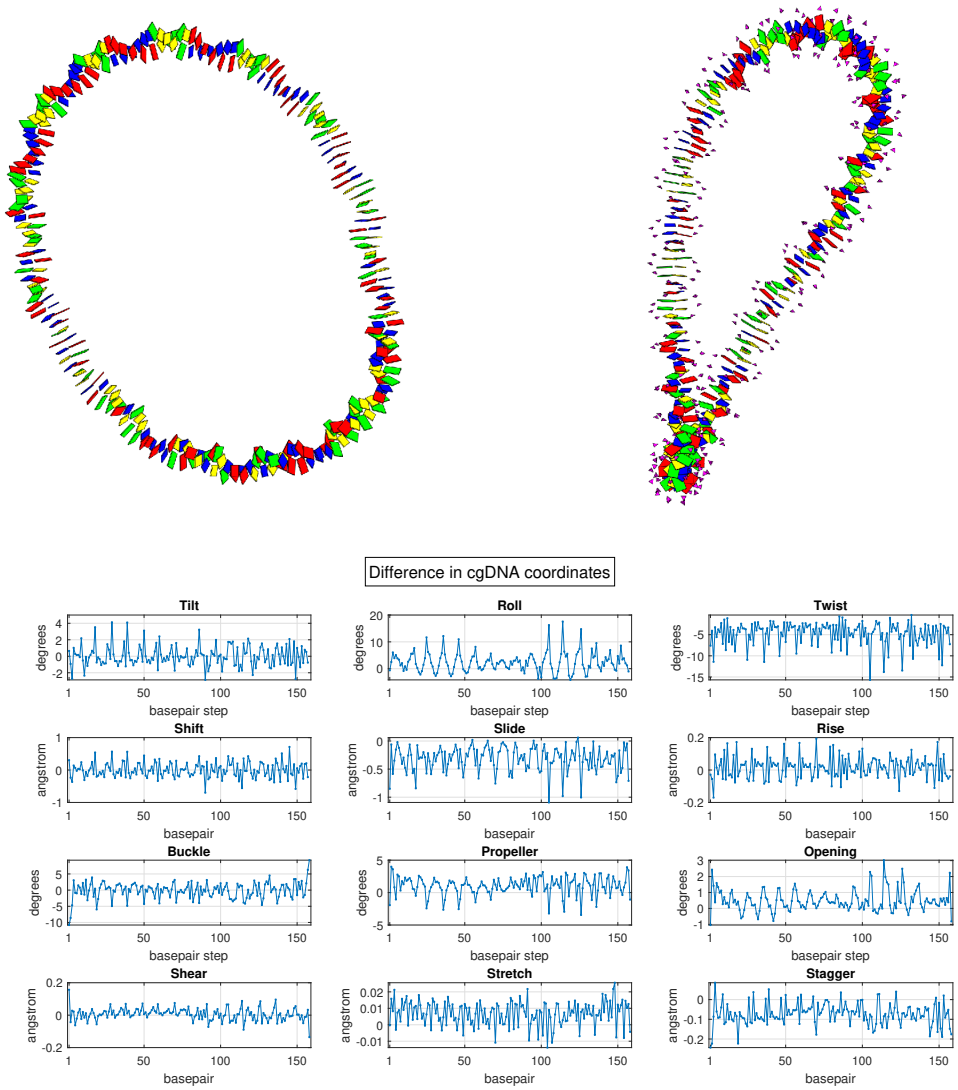


Figure 35: 3D views and 2D plots of the difference of the cgDNA coordinates between the cgDNA and cgDNA+ minimization routines. Again, the phosphate coordinates are not shown because we are interested in the comparison between cgDNAmin and cgDNA+min. Particular case: high differences, sequence: Kahn & Crothers c11t15 initial guess c) in Figures 21 and 14. The relative difference is 15.75%, which is huge.

unstable configuration may converge to a stable equilibrium that can be obtained from another initial guess. The minimization procedure then outputs configurations that can represent the same local minima.

We believe that the same phenomenon is behind the differences between the cgDNA and cgDNA+ results. The definition of the energy is different between the cgDNA and cgDNA+ cases. Hence an unstable equilibrium from bBDNA will not necessary converge to the same stable solution after the discrete energy minimization. If we start form an unstable configuration, slightly changing the definition of the energy might change the final solution after the energy minimization.

8 The MATLAB package

Together with this thesis, we provide MATLAB scripts to obtain the results shown above. The MATLAB files can be found on the thesis webpage. All results can be obtained by running the main script `Energy_min.m`. It allows to choose between cgDNA and cgDNA+ models and between non-continuous and periodic closure assumptions. This section present the method used to get our results.

Choose the model. The models are chosen through the `models` and `boundaries` variables: `models` is set to `"cgDNAmin"` for cgDNA and `"cgDNA+min"` for cgDNA+. The choices of boundary assumptions are `"NNC"` for non-continuous closure and `"PC"` for periodic assumption. Both variables support list format, therefore it is possible to request solutions for both models at the same time. The software will then loop through all specified combinations of models and boundary conditions.

The `find_min_e` function lying in the folder corresponding to the chosen model is called and starts the minimization. The objective function with its gradient and Hessian matrix is defined in the `discrete_DNA_penalty_en_grad_hess.m` file. Note, the cgDNA parameter sets are stored in the `ParameterSets` folder. The chosen sets for our results are `cgDNAps4` for the cgDNA model and `cgDNA+ps1` for cgDNA+.

Define the sequence. The sequence must be defined under the `seq` variable. For each sequence, a sequence name must be provided under the `seq_name` variable. This name must be used appropriately for the names of the initial guesses. Finally, the guess numbers must be provided under the `range_guesses` variable under a list format.

Read initial guesses. The script reads bBDNA initial guesses from the `Initial_guesses` folder. The initial guesses have to be chosen and saved manually from the bBDNA GUI before running the discrete energy minimization. Each initial guess must be saved under the format

<seq_name>_guess<guess_nb>.txt.

The sequence name and the guess number must then be given to the main script in the definition of the sequence.

Outputs. The main scripts then outputs both the 3D view and the 2D coordinates plots of the solution of the chosen minimization problem. Figures are respectively saved in the `Results/3D_views`

and `Results/2D_plots` folders with the following nomenclature:

`<seq_name>_2Dsol<guess_nb>_<boundary><model>.fig`,
`<seq_name>_3Dsol<guess_nb>_<boundary><model>.fig`.

Aside from the figures, the script saves the key variables of the procedure. They can be found inside the `Results/Variables` folder, under the name

`<seq_name>_sol<guess_nb>_<boundary><model>.mat`.

The structure of the saved data is shown in Table 2.

Entry of the data	stored variable
`data.model`	Model used
`data.seq`	Sequence tested
`data.guess`	Name of the Guess: `<seq_name>_guess<guess_nb>`
`data.init_inters`	Intras coordinates of the initial guess
`data.init_intras`	Inters coodrinates of the initial guess
`data.sol_inters`	Inters coordinates of the solution
`data.sol_intras`	Intras coordinates of the solution
`data.iter`	Number of iterations to converge
`data.time`	Running time for the minimization
`data.Energy`	Energy of the final configuration
`data.gradient`	Gradient at the final configuration
`data.Hessian`	Hessian at the final configuration

Table 2: Data stored from the main script `Energy_min.m`

Produce the figures. In order to generate the exact figures shown in this thesis[25], we use the `Visualize_solution.m` script. We can choose the model, closure assumption and the sequence using the nomenclature of `Energy_min.m`. However, we can only process one model at the time.

The script then reads the values of the solutions within the `Results/Variables` folder. The output is the different figures of the solutions: 3D view and 2D coordinates plots with reduced titles for better text incorporation. The figures are saved in the `Results/3D_views` and `Results/2D_plots` folders under the eps format. For the image names, we use the nomenclature from the main script. Note: this script is only used to generate the figures, the minimization solutions must be computed with the main script before.

[25]The main code generates figures with a slightly different layout and only saves them under the .fig format.

Part IV
Conclusion

9 On this thesis.

This thesis presents the adaptation of Manning's cgDNAmin code [15] to cgDNA+ coordinates. We also implemented periodic closure in addition to the original non-continuous closure. We started by detailing the different existing models such as the cgDNA model [14, 22] and its adaptation to periodic sequences [10]. We presented the latest cgDNA+ model that adds modelling of the phosphate groups to the standard cgDNA model [21]. Then, we presented the birod DNA model, a continuum model for DNA [12]. Along with the birod model, we quickly presented the bBDNA software that allows computation and visualization of continuum solutions [10]. Finally we presented the starting point of our work, the cgDNAmin software [15]. This combines the continuum model with a discrete energy minimization to obtain discrete configurations of DNA minicircles.

9.1 Existing model for DNA minicircle

The previously existing pipeline to obtain discrete DNA minicircles configurations comes from Manning [15]. Starting from a cgDNA parameter set, we compute continuum coefficients to prepare the computations for a continuum configuration. We use the bBDNA software with the continuum coefficients and we obtain equilibrium configurations for the continuum energy. These configurations are discretized to obtain initial guesses for a discrete energy minimization (cgDNAmin). The minimum discrete energy configuration then outputs the final discrete minicircle configuration.

9.2 New discrete energy minimization

The idea to improve cgDNAmin was to use the recent cgDNA+ coordinates instead of cgDNA as they represent the molecule more accurately. To do so, we had to change the dimensions of the coordinates in the discrete energy minimization procedure to include the phosphate representations. This also means adapting computations of the gradient and Hessian matrix of the energy as a function of the molecule's shape. Along with the adaptation to cgDNA+ coordinates, we implemented a new type of closure assumption. Manning was interested in DNA minicircle formation, hence he used a non-periodic closure (NCC). We propose a feature that allows to change between non-continuous (NCC) and periodic (PC) closures. Using the construction of periodic cgDNA stiffness matrix [10], we adapted cgDNAmin to model periodic closure. Note that the two methods model different questions. Manning's non-continuous closure assumption is focused on the probability of formation of minicircles whereas the periodic closure assumption is focused on modelling the shape of a fully formed DNA minicircle.

Along with this thesis we provide a MATLAB package that allows to compute both cgDNAmin and cgDNA+min solutions for non-continuous and periodic closure assumptions.

9.3 Results & Observations

In Section 6, we presented different cases of cgDNA+min solutions for well studied sequences such as the Kahn & Crothers [6, 7], Pyne et al. [23] and Widom [4] sequences.[26]. We observed that two different initial guesses may converge to the same stable configuration. This phenomenon was already observed with the original cgDNAmin minimization procedure. We believe that this phenomenon is related to the stability of the bBDNA equilibrium configuration used as initial guess, and that starting cgDNAmin close to an unstable equilibrium leads to cgDNAmin converging to a far away equilibrium that is a minimizer and so stable.

We also compared solutions obtained with the original cgDNAmin algorithm against our new cgDNA+min. We observe that the solutions are similar in most cases. However, we find some situations where the solutions differ a lot. Again, we assume that this is due to the fact that initial guesses from bBDNA are unstable equilibria for cgDNA+ but stable for cgDNA. The cgDNA and cgDNA+ minimization may then converge to different stable configurations. The biggest change using cgDNA+min is that the link of the original solution is sometimes not conserved. From our observations, the original cgDNAmin code do conserve link even if it mathematically does not need to be the case.

10 Further improvements

We think it can be possible to improve the bBDNA initial guesses and therefore reduce the number of iterations needed to converge in cgDNA+min. Instead of using cgDNA 1.0 parameters to generate the continuum coefficients used for bBDNA, we could try to use a cgDNA+ parameter set. However this is a challenging task. Discretizing the birod, or double chain, continuum model leads to a sparsity pattern in the stiffness matrix of 18×18 blocks with 6×6 overlaps, which precisely matches the block structure in the cgDNA model. As shown by Grandchamp [12] and Głowacki [10], the positive definite 18×18 cgDNA parameters blocks can then be used in an interpolation process to generate continuum birod coefficients. However, the block structure of cgDNA+ is 42×42 blocks with 18×18 overlaps, a change in dimension corresponding to the additional degrees of freedom of the phosphate groups. Consequently there is a mismatch with the sparsity pattern of a discretized birod. There are two possible ways to reconcile this mismatch. The first possibility is to switch from a birod to a tetrarod continuum model. The overlapping 42×42 block structure very probably corresponds to the discretization of a tetrarod, i.e. four interacting rods, one for each of the two chains of phosphate backbones and two for the chains of bases. However, the mathematical details of such theory have not been worked out, and even if it had, the analogous computational package to bBDNA for birods would still have to be written. We therefore did not attempt that route.

We did attempt the second possible route, which is for any given sequence to marginalize the phosphate degrees of freedom from the overlapping 42×42 stiffness matrix, and appropriately truncate to get the optimal overlapping 18×18 block approximation to the rigid-base marginal of the cgDNA+ model. This happens to be a well-defined and numerically easy procedure.[27] However, to compute continuum birod coefficients for bBDNA using the current methods, the necessary inputs are positive-definite 18×18 stiffness blocks for each junction. This is what is provided by cgDNA parameter set blocks. Accordingly we tried to tear apart overlapping 18×18 block of the cgDNA+ marginal stiffness matrices into individual 18×18 positive definite blocks for each junction.

[26] Full results can be found on the webpage.
[27] See [10]

Somewhat surprisingly we were unable to achieve this in a consistent way. Some 18×18 junction stiffness blocks were invariably indefinite. And finally we abandoned these efforts.

In summary a cgDNA model parameter set was used as the input necessary to generate continuum birod coefficients. A discretization of the bBDNA output, i.e. an equilibrium configuration, was then used as an initial guess for a minimization procedure applied to a cgDNA+ energy. This procedure, with its use of two different discrete models for input and output, is in some way mathematically inelegant and perhaps computationally sub-optimal. Nevertheless we observed it to be numerically robust. Every initial guess configuration of a cgDNA+ minicircle generated this way, always converged to a cgDNA+ minicircle configuration that (locally) minimized the associated cgDNA+ energy.

References

[1] Amzallag, A. “Effects of base-pair sequence, nicks and gaps on DNA minicircle shapes : analysis and experiment”. PhD thesis. EPFL, 2008.

[2] Amzallag, A. et al. “3-D reconstruction and comparison of shapes of DNA minicircles observed by cryo-electron microscopy”. In: *Nucleic Acids Research* 34.18 (2006), e125.

[3] Calladine, C. R. *Understanding DNA : the molecule & how it works*. Third edition. Elsevier Academic Press, 2004.

[4] Cloutier, T. E. and Widom, J. “Spontaneous Sharp Bending of Double-Stranded DNA”. In: *Molecular cell* 14.3 (2004), pp. 355–362.

[5] Cotta-Ramusino, L. “A path integral formalism of DNA looping probability”. PhD thesis. EPFL, 2008.

[6] Crothers, D. and Kahn, J. “Protein-Induced Bending and DNA Cyclization”. In: *Proceedings of the National Academy of Sciences - PNAS* 89.14 (1992), pp. 6343–6347.

[7] Crothers, D. M. et al. “DNA bending, flexibility, and helical repeat by cyclization kinetics”. In: *Methods in Enzymology* 212 (1992), pp. 3–29.

[8] Dahm, R. “Friedrich Miescher and the discovery of DNA”. In: *Developmental Biology* 278.2 (2005), pp. 274–288.

[9] Furrer, P. B., Manning, R. S., and Maddocks, J. H. “DNA Rings with Multiple Energy Minima”. In: *Biophysical Journal* 79 (2000), pp. 116–136.

[10] Głowacki, J. “Computation and Visualization in Multiscale Modelling of DNA Mechanics”. PhD thesis. EPFL, 2016.

[11] Gonzalez, O. et al. “Absolute versus Relative Entropy Parameter Estimation in a Coarse-Grain Model of DNA”. In: *Multiscale modeling & simulation* 15 (2017), pp. 1073–1107.

[12] Grandchamp, A. “On the statistical physics of chains and rods, with application to multi-scale sequence dependent DNA modelling”. PhD thesis. EPFL, 2016.

[13] Lankaš, F., Lavery, R., and Maddocks, J. H. “Kinking Occurs during Molecular Dynamics Simulations of Small DNA Minicircles”. In: *Structure (London)* 14.10 (2006), pp. 1527–1534.

[14] Lankas, F. et al. “On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations”. In: *Physical chemistry chemical physics : PCCP* 11.45 (2009), pp. 10565–10588.

[15] Manning, R. S. *Notes on cgDNAmin, Discrete-birod DNA Cyclization*. "unpublished". 2017.

[16] Manning, R. S. and Maddocks, J. H. “Symmetry breaking and the Twisted Elastic Ring”. In: *Computer Methods in Applied Mechanics and Engineering* 370 (1999), p. 313.

[17] Manning, R. S., Maddocks, J. H., and Kahn, J. D. “A continuum rod model of sequence-dependent DNA structure”. In: *The Journal of chemical physics* 105.13 (1996), pp. 5626–5646.

[18] Mitchell, J. S. et al. “Sequence-Dependent Persistence Lengths of DNA”. In: *Journal of chemical theory and computation* 13.4 (2017), pp. 1539–1555.

[19] Moakher, M. and Maddocks, J. H. “A Double-Strand Elastic Rod Theory”. In: *Archive for rational mechanics and analysis* 177.1 (2005), pp. 53–91.

[20] Padeken, J., Zeller, P., and Gasser, S. M. "Repeat DNA in genome organization and stability". In: *Current opinion in genetics & development* 31 (2015), pp. 12–19.

[21] Patelli, A. S. "A sequence-dependent coarse-grain model of B-DNA with explicit description of bases and phosphate groups parametrised from large scale Molecular Dynamics simulations". PhD thesis. EPFL, 2019.

[22] Petkevičiūtė, D. et al. "cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA". In: *Nucleic Acids Research* (2014).

[23] Pyne, A. L. B. et al. "Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides". In: *Nature communications* 12.1 (2021), pp. 1053–1053.

[24] Watson, J. D. and Crick, F. H. "Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid". In: *Nature*. 1953.

Part V
Appendices

A Webpage for supplementary material

All figures and results are provided online on the following webpage:

https://lcvmwww.epfl.ch/research/cgDNA/beaud/index.html.

One can find the 3D views and 2D coordinates plots of the results for the different studied sequences with both NCC and PC cgDNA+min algorithms. The MATLAB scripts can also be downloaded from the same page.

B Sequences used in examples

Poly A, 158 base pairs:

```
AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
AAAAAAAA
```

c11t15 sequence of Kahn & Crothers 1992 [6, 7]:

```
GATGAATTCA CGGATCCGGT TTTTTGCCCG TTTTTTGCCG TTTTTTGCCC
GTTTTTTGCC GTTTTTTGCC CGTTTTTTCC GGATCCGTAC AGGAATTCTA
GACCTAGGGT GCCTAATGAG TGAGCTAACT CACATTAATT GCGTTGCGCC
ATGGAATC
```

Noy, 251 base pairs [23]:

```
TTTATACTAA CTTGAGCGAA ACGGGAAGGT AAAAAGACAA CAAACTTTCT
TGTATACCTT TAAGAGAGAG AGAGAGAGAC GACTCCTGCG ATATCGCCTC
GGCTCTGTTA CAGGTCACTA ATACCATCTA AGTAGTTGAT TCATAGTGAC
TGCATATGTT GTGTTTTACA GTATTATGTA GTCTGTTTTT TATGCAAAAT
CTAATTTAAT ATATTGATAT TTATATCATT TTACGTTTCT CGTTCAGCTT
T
```

Noy, 339 base pairs [23]:

```
TTTATACTAA CTTGAGCGAA ACGGGAAGGG TTTTCACCGA TATCACCGAA
ACGCGCGAGG CAGCTGTATG GCGAAATGAA AGAACAAACT TTCTTGTACG
CGGTGGTGAG AGAGAGAGAG AGATACGACT ACTATCAGCC GGAAGCCTAT
GTACCGAGTT CCGACACTTT CATTGAGAAA GATGCCTCAG CTCTGTTACA
GGTCACTAAT ACCATCTAAG TAGTTGATTC ATAGTGACTG CATATGTTGT
GTTTTACAGT ATTATGTAGT CTGTTTTTTA TGCAAAATCT AATTTAATAT
ATTGATATTT ATATCATTTT ACGTTTCTCG TTCAGCTTT
```

Widom sequence, 94 bp [4]:

```
GGCCGGGTCG  TAGCAAGCTC  TAGCACCGCT  TAAACGCACG  TACGCGCTGT
CTACCGCGTT  TTAACCGCCA  ATAGGATTAC  TTACTAGTCT  CTAC
```

C The Special Euclidean group $SE(3)$

Elements of $SE(3)$ are composed of a rotation matrix in $SO(3)$ and a vector in $\mathbb{R}^3$.

$$SE(3) = \left\{ M \middle| M = \begin{bmatrix} R & r \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}, R \in SO(3), r \in \mathbb{R}^3 \right\} \tag{30}$$

Elements of $SE(3)$ are often denoted $M = [r, R]$ for simplicity.

The group multiplication is the standard matrix product. Let $[r, R], [q, Q] \in SE(3)$,

$$\begin{bmatrix} R & r \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} Q & q \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} RQ & Rq + r \\ \mathbf{0} & 1 \end{bmatrix} \in SE(3). \tag{31}$$

The $SE(3)$ group is used to represent frames in space. For a fixed origin $[\mathbf{0}, I_3]$, a frame in space can be represented by an element $[r, R] \in SE(3)$. R represents the rotation of the frame with respect to the origin and r correspond to the frame translation.

D Cayley vectors

For this section, we follow the EPFL lecture on DNA modelling given by J.H. Maddocks.[28] The Cayley transfom is a mapping $Cay : SO(3) \to \mathbb{R}^3$. Let $Q \in SO(3)$ with $tr(Q) \neq -1$,, i.e. Q is not a rotation through an angle π and $u \in \mathbb{R}^3$.

$$\begin{aligned} Q \to [u\times] &= \frac{1}{1 + tr(Q)}(Q - Q^T) \in \mathbb{R}^3, \\ u \to Q &= \frac{1 - \|u\|^2}{1 + \|u\|^2} I + \frac{2}{1 + \|u\|^2}[u\times] + \frac{2}{1 + \|u\|^2} u \otimes u \\ &= I + \frac{2}{1 + \|u\|^2}([u\times] + [u\times]^2), \end{aligned} \tag{32}$$

where

$$[u\times] = \begin{bmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{bmatrix}, \tag{33}$$

the skew matrix constructed from $u = [u_1, u_2, u_3]^T$. u is called the Cayley vector of Q. In order to avoid unnecessary notation, we will abuse the notation and write $u = Cay(Q)$. The formula relating a rotation to its Cayley vector is called the Euler-Rodrigues formula.

We note that the mapping is not exactly 1-to-1. Matrices with a rotation through π cannot be represented by a Cayley vector.

[28] "Mathematical modelling of DNA", https://lcvmwww.epfl.ch/teaching/modelling_dna/

E From Quaternions to rotations

E.1 About quaternions

Following Cotta-Ramusino's thesis for example for a given normalized quaternion $\tilde{q}$, any arbitrary quaternion q can be written as:

$$q = c_1 B_1 \tilde{q} + c_2 B_2 \tilde{q} + c_3 B_3 \tilde{q} + \sqrt{1 - c_1^2 - c_2^2 - c_3^2}\tilde{q}, \tag{34}$$

with

$$B_1 \equiv \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, B_2 \equiv \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, B_3 \equiv \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

Hence $\{\tilde{q}, B_1\tilde{q}, B_2\tilde{q}, B_3\tilde{q}\}$ is an orthonormal basis for R^4.

Each rotation matrix can be described by a quaternion $q = [a, b, c, d] \in \mathbb{R}^4$.

$$R(q) = \frac{1}{a^2 + a^2 + c^2 + d^2} \begin{bmatrix} a^2 + b^2 - c^2 - d^2 & 2bc - 2ad & 2ac + 2bd \\ 2bc + 2ad & a^2 - b^2 + c^2 - d^2 & 2cd - 2ab \\ 2bd - 2ac & 2ab + 2cd & a^2 - b^2 - c^2 + d^2 \end{bmatrix}, \tag{35}$$

which is invariant with respect to the norm of q. For the quaternion multiplication we define

$$q^a \circ q^b := \begin{bmatrix} q_1^a q_4^b + q_2^a q_3^b - q_3^a q_2^b + q_4^a q_1^b \\ -q_1^a q_3^b + q_2^a q_4^b + q_3^a q_1^b + q_4^a q_2^b \\ q_1^a q_2^b - q_2^a q_1^b + q_3^a q_4^b + q_4^a q_3^b \\ -q_1^a q_1^b - q_2^a q_2^b - q_3^a q_3^b + q_4^a q_4^b \end{bmatrix}.$$

We note that $R(q_a)R(q_b) = R(q_a \circ q_b)$. The inverse of a normalized quaternion $q = [q_1, q_2, q_3, q_4]^T$ is $q^{-1} = [-q_1, -q_2, -q_3, q_4]^T$ and its associated rotation matrix is $R(q^{-1}) = R(q)^T$. We also note that

$$q_a^{-1} \circ q_b = \begin{bmatrix} q_b^T B_1 q_a \\ q_b^T B_2 q_a \\ q_b^T B_3 q_a \\ q_a^T q_b \end{bmatrix}. \tag{36}$$

Finally we compute

$$q_i \circ \sqrt{q_i^{-1} \circ q_{i+1}} = \frac{q_a + q_b}{\|q_a + q_b\|} = \frac{q_a + q_b}{\sqrt{2 + 2q_a^T q_b}}. \tag{37}$$

E.2 Applied to cgDNAmin

In the cgDNAmin coordinates, inters are represented using quaternions. This means that a translation o_i and a quaternion q_i are associated to every base pair. We need to define the function f that recovers the standard inter coordinate vector y_i from (o_i, q_i) and (o_{i+1}, q_{i+1}), the quaternion parametrizations of the absolute base pair frames.

$$y_i = f(o_i, q_i, o_{i+1}, q_{i+1}) = (\theta_i^1, \theta_i^2, \theta_i^3, \zeta_1^1, \zeta_i^2, \zeta_i^3), \tag{38}$$

with ζ the relative translation part and θ the relative rotation part of the coordinates.

The relative rotation between two base pair frames is then the rotation defined by $q_i^{-1} \circ q_{i+1}$, and the junction frame is the halfway transformation $R(q_i \circ \sqrt{q_i^{-1} \circ q_{i+1}})$. This can be simplified to $R(q_{i+1} + q_i)$ using Equation (37). The relative translation expressed in the junction frame then reads:

$$\zeta_i^T = (o_{i+1} - o_i)R(q_{i+1} + q_i), \tag{39}$$

with $R(q_{i+1} + q_i)$, the rotation matrix induced by the quaternion $q_{i+1} + q_i$. For the rotation component, we use the Cayley transform from Appendices D to get

$$\theta_i^a = \frac{10 q_{i+1}^T B_a q_i}{q_{i+1}^T q_i}, (1 \leq i \leq n-1), a = 1, 2, 3. \tag{40}$$

We can therefore define the overall transformation from quaternions to standard cgDNA coordinates as:

$$\begin{aligned} w = F(z) = (x_1, f(o_1, q_1, o_2, q_2), x_2, \dots \\ \dots, x_i, f(o_i, q_i, o_{i+1}, q_{i+1}), \dots \\ \dots, x_{n-1}, f(o_{n-1}, q_{n-1}, o_n, q_n), x_n), \end{aligned} \tag{41}$$

with f from Equation (38).

F cgDNAmin Energy minimization

In this section, we derive the entries of the gradient following the work of Manning [15] for the cgDNAmin minimization procedure. The notation (mainly the indices and superscripts) change slightly.

F.1 Notation

Before entering the detailed computations of the gradient and the Hessian of the energy with respect to the cgDNA coordinates, we need to introduce some notation. Recall Equation (11): the minimization vector is

$$z = (x_1, o_1, q_1, x_2, o_2, q_2, \dots, x_{n-1}, o_{n-1}, q_{n-1}, x_n) \in \mathbb{R}^{13n-14},$$

where $x_i \in \mathbb{R}^6$ are the standard intra coordinates and $(o_i, q_i) \in \mathbb{R}^7$ are the absolute coordinates for the base pair frames, expressed with quaternions. We denote by $y_i = (\theta_i^1, \theta_i^2, \theta_i^3, \zeta_i^1, \zeta_i^2, \zeta_i^3)$, the standard inter coordinates in $\mathbb{R}^6$, ζ_i the translation component and θ_i the Cayley vector of the rotation. From the quaternions, we have $y_i = f(o_i, q_i, o_{i+1}, q_{i+1})$ with $f(\cdot)$ shown in Appendices E.

For more clarity, we introduce the notation $\hat{x}_i$ and $\hat{y}_i$ for the elements of the ground-state μ corresponding to the coordinates x_i or y_i respectively.

In order to work with inter and intra coordinates separately, we decompose the banded stiffness K as follow. Each 18×18 diagonal block is split into nine 6×6 sub-blocks, where each sub-block is

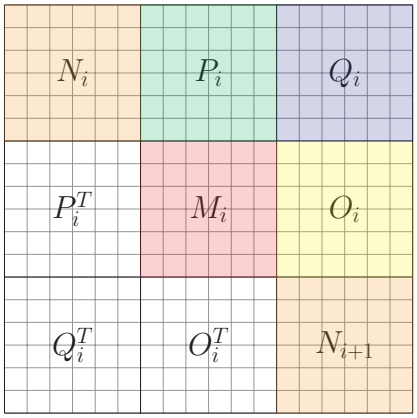


Figure 36: Separation of each 18×18 diagonal block of the cgDNA stiffness matrix. The grid represent the i-th diagonal 18×18 block of the stiffness matrix K. That is coordinates from $1 + 12(i-1)$ to $18 + 12(i-1)$ for both dimensions.

responsible for the interaction between two components of the coordinates vector (intras and inters). Let K_i be the i-th overlapping block of K. The block sub-division is then

$$\begin{aligned} N_i &= K_i[1:6, 1:6], \\ M_i &= K_i[7:12, 7:12], \\ Q_i &= K_i[13:18, 13:18], \\ P_i &= K_i[1:6, 7:12], \\ O_i &= K_i[7:12, 13:18]. \end{aligned} \tag{42}$$

The notation $K[a:b, c:d]$ means the sub-block of rows a to b and columns c to d of K. The separation is shown in Figure 36.

F.2 Gradient

In order to speed the process of the energy minimization, we need to provide an explicit formula for the gradient with respect to the vector z. This is obtained through a chain rule,

$$\frac{d}{dz}U(F(z)) = \frac{d}{dw}U(w) \cdot \frac{d}{dz}F(z). \tag{43}$$

Intra entries. The gradient with respect to the intras is not impacted by the chain rule and is straight forward.

$$\begin{aligned} \frac{\partial}{\partial x_i}U = {} & N_i(x_i - \hat{x}_i) + Q_i(x_{i+1} - \hat{x}_{i+1}) + Q_{i-1}^T(x_{i-1} - \hat{x}_{i-1}) \\ & + P_i(y_i - \hat{y}_i) + O_{i-1}^T(y_{i-1} - \hat{y}_{i-1}), \qquad i = 2, \ldots, n-1, \end{aligned} \tag{44}$$

with special cases for $i = 1, n$

$$\frac{\partial}{\partial x_1}U = N_1(x_1 - \hat{x}_1) + Q_1(x_2 - \hat{x}_2) + P_1(y_1 - \hat{y}_1), \tag{45}$$

$$\frac{\partial}{\partial x_n}U = N_n(x_n - \hat{x}_n) + Q_{n-1}^T(x_{n-1} - \hat{x}_{n-1}) + O_{n-1}(y_{n-1} - \hat{y}_{n-1}). \tag{46}$$

Inter entries. In the minimization vector we use quaternions for inter rotations whereas the energy is defined using Cayley vectors. Hence we have to use the chain rule. We start by preparing some preliminary derivatives,

$$\frac{\partial}{\partial q_{i+1}}\left(\frac{1}{q_{i+1}\cdot q_i}\right) = -(q_{i+1}\cdot q_i)^{-2}q_i, \tag{47}$$

$$\frac{\partial}{\partial q_i}\left(\frac{1}{q_{i+1}\cdot q_i}\right) = -(q_{i+1}\cdot q_i)^{-2}q_{i+1}. \tag{48}$$

Following this and based on the transformation formula from quaternions to Cayley vectors[29], we have,

$$\begin{aligned}\frac{\partial\theta_i^a}{\partial q_{i+1}} &= \left[10(q_{i+1})^T B_a q_i\right]\left[-(q_{i+1}\cdot q_i)^{-2}\right]q_i + (q_{i+1}\cdot q_i)^{-1}10B_a q_i \\ &= (q_{i+1}\cdot q_i)^{-1}(10B_a - \theta_i^a I)q_i.\end{aligned} \tag{49}$$

And similarly, we have

$$\begin{aligned}\frac{\partial\theta_i^a}{\partial q_i} &= \left[10(q_{i+1})^T B_a q_i\right]\left[-(q_{i+1}\cdot q_i)^{-2}\right]q_{i+1} + (q_{i+1}\cdot q_i)^{-1}[-10B_a q_{i+1}] \\ &= (q_{i+1}\cdot q_i)^{-1}(-10B_a - \theta_i^a I)q_{i+1}.\end{aligned} \tag{50}$$

For the translation coordinates, we have

$$\frac{\partial\zeta_i^a}{\partial o_i} = -d_i^a(q_{i+1} + q_i), \tag{51}$$

$$\frac{\partial\zeta_i^a}{\partial o_{i+1}} = d_i^a(q_{i+1} + q_i), \tag{52}$$

$$\frac{\partial\zeta_i^a}{\partial q_i} = \frac{\partial\zeta_i^a}{\partial q_{i+1}} = \left(\frac{\partial d_i^a}{\partial q}(q_{i+1} + q_i)\right)^T(o_{i+1} - o_i). \tag{53}$$

[29] See Appendices E

Here d_i^a represent the a-th column of the i-th junction frame. With these, we can compute the derivative with respect to o_i and q_i using the chain rule.

$$\begin{aligned}\frac{\partial U}{\partial o_i} &= \sum_{a=1}^{3}\left[\frac{\partial U}{\partial \zeta_{i-1}^a}\frac{\partial \zeta_{i-1}^a}{\partial o_i} + \frac{\partial U}{\partial \zeta_i^a}\frac{\partial \zeta_i^a}{\partial o_i}\right] \\ &= \left(\frac{\partial \zeta_{i-1}}{\partial o_i}\right)^T \frac{\partial U}{\partial \zeta_{i-1}} + \left(\frac{\partial \zeta_i}{\partial o_i}\right)^T \frac{\partial U}{\partial \zeta_i}\end{aligned}, \tag{54}$$

$$\begin{aligned}\frac{\partial U}{\partial q_i} &= \sum_{a=1}^{3}\left[\frac{\partial U}{\partial \zeta_{i-1}^a}\frac{\partial \zeta_{i-1}^a}{\partial q_i} + \frac{\partial U}{\partial \zeta_i^a}\frac{\partial \zeta_i^a}{\partial q_i} + \frac{\partial U}{\partial \theta_{i-1}^a}\frac{\partial \theta_{i-1}^a}{\partial q_i} + \frac{\partial U}{\partial \theta_i^a}\frac{\partial \theta_i^a}{\partial q_i}\right] \\ &= \left(\frac{\partial \zeta_{i-1}}{\partial q_i}\right)^T \frac{\partial U}{\partial \zeta_{i-1}} + \left(\frac{\partial \zeta_i}{\partial q_i}\right)^T \frac{\partial U}{\partial \zeta_i} + \left(\frac{\partial \theta_{i-1}}{\partial q_i}\right)^T \frac{\partial U}{\partial \theta_{i-1}} + \left(\frac{\partial \theta_i}{\partial q_i}\right)^T \frac{\partial U}{\partial \theta_i}.\end{aligned} \tag{55}$$

Combining the two results we obtain

$$\begin{aligned}\frac{\partial U}{\partial (o_i, q_i)} &= \begin{bmatrix} 0 & \left(\frac{\partial \zeta_{i-1}}{\partial o_i}\right)^T \\ \left(\frac{\partial \theta_{i-1}}{\partial q_i}\right)^T & \left(\frac{\partial \zeta_{i-1}}{\partial q_i}\right) \end{bmatrix} \begin{bmatrix} \frac{\partial U}{\partial \theta_{i-1}} \\ \frac{\partial U}{\partial \zeta_{i-1}} \end{bmatrix} + \begin{bmatrix} 0 & \left(\frac{\partial \zeta_i}{\partial o_i}\right)^T \\ \left(\frac{\partial \theta_i}{\partial q_i}\right)^T & \left(\frac{\partial \zeta_i}{\partial q_i}\right) \end{bmatrix} \begin{bmatrix} \frac{\partial U}{\partial \theta_i} \\ \frac{\partial U}{\partial \zeta_i} \end{bmatrix} \\ &= \begin{bmatrix} 0 & \frac{\partial \zeta_{i-1}}{\partial o_i} \\ \frac{\partial \theta_{i-1}}{\partial q_i} & \frac{\partial \zeta_{i-1}}{\partial q_i} \end{bmatrix}^T \frac{\partial U}{\partial x_{i-1}} + \begin{bmatrix} 0 & \frac{\partial \zeta_i}{\partial o_i} \\ \frac{\partial \theta_i}{\partial q_i} & \frac{\partial \zeta_i}{\partial q_i} \end{bmatrix}^T \frac{\partial U}{\partial x_i}\end{aligned} \tag{56}$$

We can obtain the explicit inter derivative by

$$\frac{\partial U}{\partial y_i} = P_i^T(x_i - \hat{x}_i) + M_i(y_i - \hat{y}_i) + O_i(x_{i+1} - \hat{x}_{i-1}). \tag{57}$$

The two matrices in Equation (56) will recur in the Hessian matrix computations, we therefore name them

$$\frac{\partial x_{i-1}}{\partial (o_i, q_i)} \equiv \begin{bmatrix} 0 & \frac{\partial \zeta_{i-1}}{\partial o_i} \\ \frac{\partial \theta_{i-1}}{\partial q_i} & \frac{\partial \zeta_{i-1}}{\partial q_i} \end{bmatrix}, \qquad \frac{\partial x_i}{\partial (o_i, q_i)} \equiv \begin{bmatrix} 0 & \frac{\partial \zeta_i}{\partial o_i} \\ \frac{\partial \theta_i}{\partial q_i} & \frac{\partial \zeta_i}{\partial q_i} \end{bmatrix}. \tag{58}$$

Finally, we need to add the derivatives of penalty term from Equation (13).

$$\frac{\partial U}{\partial q_i} = 4p(\|q_i\|^2 - 1)q_i. \tag{59}$$

F.3 Hessian

When it comes to the Hessian matrix, entries can be separated into three different groups, intra-intra, intra-inter and inter-inter derivatives. We treat each case separately.

Intra-intra entries. Similarly to the gradient, the intra-intra entries of the Hessian are straight forward

$$\frac{\partial^2 U}{\partial (x_i)^2} = N_i, \quad 1 \leq i \leq n, \tag{60}$$

$$\frac{\partial^2 U}{\partial x_i \partial x_{i+1}} = Q_i, \quad 1 \leq i \leq n-1, \tag{61}$$

$$\frac{\partial^2 U}{\partial x_i \partial x_{i+1}} = Q_{i-1}^T, \quad 2 \leq i \leq n. \tag{62}$$

Intra-Inter & Inter-Intra entries. We start from Equation (44). We rewrite it omitting the intra dependence.

$$\frac{\partial}{\partial x_i} U = P_i(y_i - \hat{y}_i) + O_{i-1}^T (y_{i-1} - \hat{y}_{i-1}) + \text{function of Intra.}$$

We use the general relation

$$\frac{\partial}{\partial \alpha}[Af(\alpha)] = A \frac{\partial \alpha}{\partial x}.$$

This yields

$$\frac{\partial^2 U}{\partial x_i \partial (o_{i+1}, q_{i+1})} = P_i \frac{\partial y_i}{\partial (o_{i+1}, q_{i+1})}, \tag{63}$$

$$\frac{\partial^2 U}{\partial x_i \partial (o_{i-1}, q_{i-1})} = O_{i-1}^T \frac{\partial y_{i-1}}{\partial (o_{i-1}, q_{i-1})}, \tag{64}$$

$$\frac{\partial^2 U}{\partial x_i \partial (o_i, q_i)} = P_i \frac{\partial y_i}{\partial (o_i, q_i)} + O_{i-1}^T \frac{\partial y_{i-1}}{\partial (o_i, 1_i)}, \tag{65}$$

with the definitions of the matrices in Equation (58). The other order of partial derivation yields the transposes of these results:

$$\frac{\partial^2 U}{\partial (o_{i+1}, q_{i+1}) \partial x_i} = \left(\frac{\partial y_i}{\partial (o_{i+1}, q_{i+1})} \right)^T P_i^T, \tag{66}$$

$$\frac{\partial^2 U}{\partial (o_{i-1}, q_{i-1}) \partial x_i} = \left(\frac{\partial y_{i-1}}{\partial (o_{i-1}, q_{i-1})} \right)^T O_{i-1}, \tag{67}$$

$$\frac{\partial^2 U}{\partial (o_i, q_i) \partial x_i} = \left(\frac{\partial y_i}{\partial (o_i, q_i)} \right)^T P_i^T + \left(\frac{\partial y_{i-1}}{\partial (o_i, q_i)} \right)^T O_{i-1}. \tag{68}$$

Inter-Inter entries. These are the most complex entries of the Hessian matrix. Again we start with some preliminary derivatives:

$$\begin{aligned}\frac{\partial^2\theta_i^a}{\partial q_{i+1}^2} &= \Big[-(q_{i+1}\cdot q_i)^{-2}\Big](10B_a-\theta_i^a I)q_i(q_i)^T-(q_{i+1}\cdot q_i)^{-1}q_i\left(\frac{\partial\theta_i^a}{\partial q_{i+1}}\right)^T\\ &= -(q_{i+1}\cdot q_i)^{-2}(10B_a-\theta_i^a I)q_i(q_i)^T-(q_{i+1}\cdot q_i)^{-2}q_i(q_i)^T(-10B_a-\theta_i^a I)\\ &= (q_{i+1}\cdot q_i)^{-2}\big[2\theta_i^a q_i(q_i)^T-10(B_aq_i)(q_i)^T-10q_i(B_aq_i)^T\big],\end{aligned} \tag{69}$$

$$\begin{aligned}\frac{\partial^2\theta_i^a}{\partial q_i^2} &= \Big[-(q_{i+1}\cdot q_i)^{-2}\Big](-10B_a-\theta_i^a I)q_{i+1}(q_{i+1})^T-(q_{i+1}\cdot q_i)^{-1}q_{i+1}\left(\frac{\partial\theta_i^a}{\partial q_i}\right)^T\\ &= -(q_{i+1}\cdot q_i)^{-2}(-10B_a-\theta_i^a I)q_{i+1}(q_{i+1})^T-(q_{i+1}\cdot q_i)^{-2}q_{i+1}(q_{i+1})^T(10B_a-\theta_i^a I)\\ &= (q_{i+1}\cdot q_i)^{-2}\big[2\theta_i^a q_{i+1}(q_{i+1})^T+10(B_aq_{i+1})(q_{i+1})^T+10q_{i+1}(B_aq_{i+1})^T\big],\end{aligned} \tag{70}$$

$$\begin{aligned}\frac{\partial^2\theta_i^a}{\partial q_i\partial q_{i+1}} &= \Big[-(q_{i+1}\cdot q_i)^{-2}\Big](10B_a-\theta_i^a I)q_i(q_{i+1})^T-(q_{i+1}\cdot q_i)^{-1}q_i\left(\frac{\partial\theta_i^a}{\partial q_i}\right)^T\\ &\qquad+(q_{i+1}\cdot q_i)^{-1}(10B_a-\theta_i^a I)\\ &= -(q_{i+1}\cdot q_i)^{-2}(10B_a-\theta_i^a I)q_i(q_{i+1})^T-(q_{i+1}\cdot q_i)^{-2}q_i(q_{i+1})^T(10B_a-\theta_i^a I)\\ &\qquad+(q_{i+1}\cdot q_i)^{-1}(10B_a-\theta_i^a I)\\ &= (q_{i+1}\cdot q_i)^{-2}\big[2\theta_i^a q_i(q_{i+1})^T-10(B_aq_i)(q_{i+1})^T+10q_i(B_aq_{i+1})^T\big]\\ &\qquad+(q_{i+1}\cdot q_i)^{-1}(10B_a-\theta_i^a I),\end{aligned} \tag{71}$$

$$\frac{\partial^2\theta_i^a}{\partial q_{i+1}\partial q_i} = \left(\frac{\partial^2\theta_i^a}{\partial q_i\partial q_{i+1}}\right)^T = (q_{i+1}\cdot q_i)^{-2}\big[2\theta_i^a q_{i+1}(q_i)^T-10q_{i+1}(B_aq_i)^T+10(B_aq_{i+1})(q_i)^T\big] + (q_{i+1}\cdot q_i)^{-1}(-10B_a-\theta_i^a I). \tag{72}$$

We also compute the remaining derivatives of ζ_i^a.

$$\frac{\partial^2\zeta_i^a}{\partial o_{i+1}\partial q_{i+1}} = \frac{\partial^2\zeta_i^a}{\partial o_{i+1}\partial q_i} = \left[\frac{\partial d_a}{\partial q}(q_{i+1}+q_i)\right]^T \tag{73}$$

$$\frac{\partial^2\zeta_i^a}{\partial o_i\partial q_{i+1}} = \frac{\partial^2\zeta_i^a}{\partial o_i\partial q_i} = -\left[\frac{\partial d_a}{\partial q}(q_{i+1}+q_i)\right]^T \tag{74}$$

$$\frac{\partial^2\zeta_i^a}{\partial(q_{i+1})^2} = \frac{\partial^2\zeta_i^a}{\partial(q_i)^2} = \frac{\partial^2\zeta_i^a}{\partial q_{i+1}\partial q_i} = \sum_{k=1}^{3}(o_{i+1}^k)+o_i^k\frac{\partial^2 d_{ak}}{\partial\mathbf{q}^2}(q_{i+1}+q_i). \tag{75}$$

Again, the derivatives in the opposite order are obtained through transposition. Here, we need the $\mathbf{d}$ derivatives:

$$\frac{\partial\mathbf{d}_1}{\partial\mathbf{q}} = \frac{2[\mathbf{d}_2(B_3\mathbf{q})^T-\mathbf{d}_3(B_2\mathbf{q})^T]}{\|\mathbf{q}\|^2}, \tag{76}$$

$$\frac{\partial\mathbf{d}_2}{\partial\mathbf{q}} = \frac{2[\mathbf{d}_3(B_1\mathbf{q})^T-\mathbf{d}_1(B_3\mathbf{q})^T]}{\|\mathbf{q}\|^2}, \tag{77}$$

$$\frac{\partial\mathbf{d}_1}{\partial\mathbf{q}} = \frac{2[\mathbf{d}_1(B_2\mathbf{q})^T-\mathbf{d}_2(B_1\mathbf{q})^T]}{\|\mathbf{q}\|^2}. \tag{78}$$

To move toward second derivatives we note that Equations (76)-(78) imply, for each $k = 1, 2, 3$,

$$\frac{\partial d_{1k}}{\partial \mathbf{q}} = \frac{2[d_{2k}(B_3\mathbf{q}) - d_{3k}(B_2\mathbf{q})]}{\|\mathbf{q}\|^2}, \tag{79}$$

$$\frac{\partial d_{2k}}{\partial \mathbf{q}} = \frac{2[d_{3k}(B_1\mathbf{q}) - d_{1k}(B_3\mathbf{q})]}{\|\mathbf{q}\|^2}, \tag{80}$$

$$\frac{\partial d_{3k}}{\partial \mathbf{q}} = \frac{2[d_{1k}(B_2\mathbf{q}) - d_{2k}(B_1\mathbf{q})]}{\|\mathbf{q}\|^2}. \tag{81}$$

Using

$$\frac{\partial}{\partial \mathbf{q}}\left(\frac{1}{\|\mathbf{q}\|^2}\right) = -(\|\mathbf{q}\|^2)^{-2}(2\mathbf{q}) = \frac{1}{\|\mathbf{q}\|^2}\left(-\frac{2\mathbf{q}}{\|\mathbf{q}\|^2}\right), \tag{82}$$

we have

$$\frac{\partial^2 d_{1k}}{\partial \mathbf{q}^2} = \frac{2[d_{2k}B_3 - d_{3k}B_2]}{\|\mathbf{q}\|^2} + \frac{2\left[(B_3\mathbf{q})\left(\frac{\partial d_{2k}}{\partial \mathbf{q}}\right)^T - (B_2\mathbf{q})\left(\frac{\partial d_{3k}}{\partial \mathbf{q}}\right)^T\right]}{\|\mathbf{q}\|^2} - \frac{2\frac{\partial d_{1k}}{\partial \mathbf{q}}\mathbf{q}^T}{\|\mathbf{q}\|^2}. \tag{83}$$

The entries of d_{2k} and d_{3k} are obtained by cycling the indices $1 \to 2 \to 3 \to 1$.

From the gradient entries, we have that

$$\begin{aligned}\frac{\partial U}{\partial(o_i, q_i)} =& \left(\frac{\partial y_{i-1}}{\partial(o_i, q_i)}\right)^T [M_{i-1}(y_{i-1} - \hat{y}_{i-1}) + \text{intra dependence}] \\ &+ \left(\frac{\partial y_i}{\partial(o_i, q_i)}\right)^T [M_i(y_i - \hat{y}_i) + \text{intra dependence}],\end{aligned} \tag{84}$$

where y_i is a function of $(o_i, q_i, o_{i+1}, q_{i+1})$. We use the relation

$$\frac{\partial}{\partial \alpha}[M(\alpha)\mathbf{v}] = \frac{\partial M}{\partial \alpha}\mathbf{v} \tag{85}$$

to get the Hessian entries:

$$\frac{\partial^2 U}{\partial(o_{i-1}, q_{i-1})\partial(o_i, q_i)} = \frac{\partial[\partial y_{i-1}/\partial(o_i, q_i)]^T}{\partial(o_{i-1}, q_{i-1})}\frac{\partial U}{\partial y_{i-1}} + \left[\frac{\partial y_{i-1}}{\partial(o_i, q_i)}\right]^T M_{i-1}\frac{\partial y_{i-1}}{\partial(o_{i-1}, q_{i-1})}, \tag{86}$$

$$\frac{\partial^2 U}{\partial(o_{i+1}, q_{i+1})\partial(o_i, q_i)} = \frac{\partial[\partial y_i/\partial(o_i, q_i)]^T}{\partial(o_{i+1}, q_{i+1})}\frac{\partial U}{\partial y_i} + \left[\frac{\partial y_i}{\partial(o_i, q_i)}\right]^T M_i\frac{\partial y_i}{\partial(o_{i+1}, q_{i+1})}, \tag{87}$$

$$\begin{aligned}\frac{\partial^2 U}{\partial(o_i, q_i)\partial(o_i, q_i)} =& \frac{\partial[\partial y_{i-1}/\partial(o_i, q_i)]^T}{\partial(o_i, q_i)}\frac{\partial U}{\partial y_{i-1}} + \left[\frac{\partial x_{i-1}}{\partial(o_i, q_i)}\right]^T M_{i-1}\frac{\partial x_{i-1}}{\partial(o_i, q_i)} \\ &+ \frac{\partial[\partial y_i/\partial(o_i, q_i)]^T}{\partial(o_i, q_i)}\frac{\partial U}{\partial y_i} + \left[\frac{\partial y_i}{\partial(o_i, q_i)}\right]^T M_i\frac{\partial y_i}{\partial(o_i, q_i)}.\end{aligned} \tag{88}$$

All right hand terms are already defined in the Gradient computation except for the tensor terms highlighted in green. These are all of the form $\frac{\partial[\partial x_b/\partial(o_a, q_a)]^T}{\partial(o_c, q_c)}$ with $b = a, a-1$ and $c = a-1, a, a+1$. These tensor can be related to previous derivatives computed before. The details for these computations can be found in [15].