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

Marius BEAUD, Assistant: Raushan SINGH, Professor: John H. MADDOCKS

Modelling DNA and DNA minicircles

DNA modeling. Depending on the sequence, the shape of a DNA molecule can be completely straight or, on the opposite, have a strong intrinsic bend. To better understand DNA, it is important to be able to predict the shape and the stiffness of a DNA fragment based on its sequence. Models for DNA can be grouped in two categories: the ***atomic level*** models and the ***coarse-grain*** models. We focus on coarse-grain models. They introduce simplifying assumptions to reduce complexity: the ***rigid base*** or ***rigid base pair*** assumptions. Each base or base pair is represented by a rigid body to reduce the number of free variables. Due to their higher versatility and lower computational cost, coarse-grain models are often preferred when possible.

DNA minicircles. One experimentally important motif is the formation of DNA closed loops, called DNA ***minicircles***. Using an adapted version of a coarse-grain model combined with a discretization of a continuum model, one can predict the shape of a DNA minicircle. For this thesis, the main goal was to adapt the existing algorithm for discrete DNA minicircles to use the latest cgDNA+ coordinates. We tested our implementation with different case studies coming from the experimental literature.

Discrete model: cgDNA [Gonzalez, 2013] [Petkevičiūtė, 2014]

Rigid base. Each base is represented by a frame in $SE(3)$, this is a vector in $\mathbb{R}^3$ and a rotation matrix in $SO(3)$. They represent the absolute position and rotation of the frame.

Gaussian pdf. For a given sequence S and parameter set $\mathcal{P}$, the model predicts a Gaussian probability density function

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

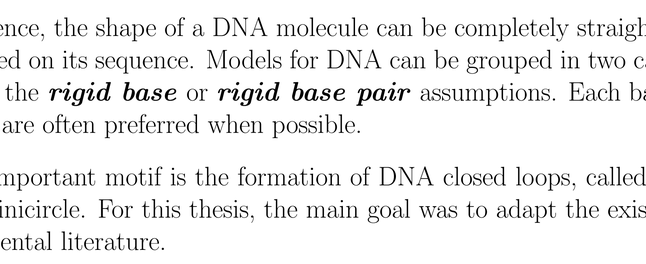
U is the molecule energy, K is the stiffness matrix and μ is the ground-state.

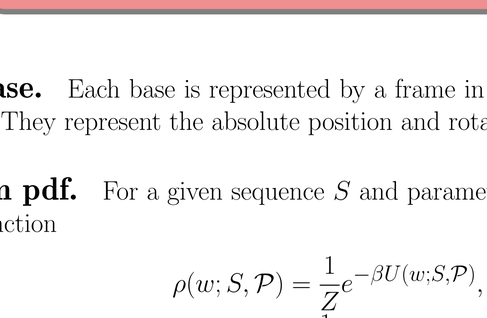
cgDNA coordinates.

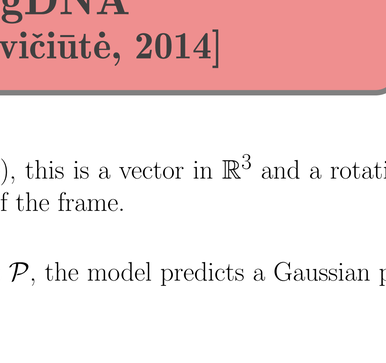
$w = (x_1, y_1, x_2, y_2, \ldots, x_{n-1}, y_{n-1}, x_n)$,

$x_i \in \mathbb{R}^6$: ***intras***, relative displacement between base frames,

$y_i \in \mathbb{R}^6$: ***inters***, relative displacement between base pair frames.





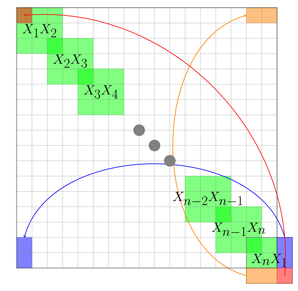


On the left: chain structure of the cgDNA coordinates, in blue: intras, in red: inters coordinates.
On the right: construction of the stiffness matrix K from the parameter blocks of $\mathcal{P}$. End blocks are different to account for the end effects in the model.

Adaptations of cgDNA

periodic cgDNA [Glowacki, 2016]

For periodicity, one has to add an extra set of inter coordinates and a block to model the interaction between the first and last base pairs. The extra block is split in the four corners of the periodic stiffness matrix. We no longer use different parameter blocks for the ends.
NOTE: periodicity in cgDNA coordinates implies a helical structure, but **do not** imply periodicity in absolute coordinates.



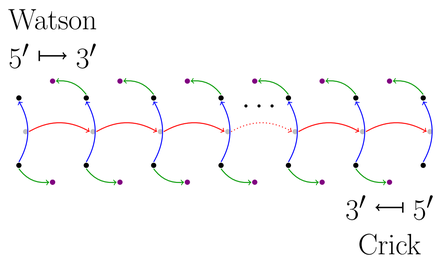
New construction of the stiffness matrix.

cgDNA+ [Patelli, 2019]

Add representation of phosphate groups, each one associated to a base. Their coordinates are the relative displacement between the base and the phosphate frames. The new base pair level coordinates becomes:

$$\tilde{x}_i = [p_i^+, x_i, p_i^-] \in \mathbb{R}^{18}.$$

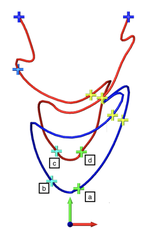
$p_i^\pm$ are the Watson (+) and Crick (-) phosphate coordinates and x_i is the standard intra. The inter coordinates remain unchanged.



cgDNA+ coordinates chain structure.

Continuum model [Grandchamp and Glowacki, 2016]

For the continuum model, we rely on the birod model. It models DNA strands as two interacting continuum rods.
We use the bBDNA software to obtain equilibrium configurations for the continuum birod model. Such configurations are then discretized and used as initial guesses for the discrete energy minimization. The family of solutions is shown as a bifurcation diagram where crosses represent configurations where the two backbones align to form DNA minicircles.



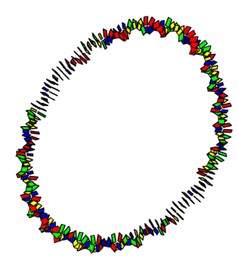
Example of Bifurcation diagram.

cgDNA minicircles: cgDNAmin [Manning, 2017]

Two step pipeline. In order to obtain a discrete configuration for a DNA minicircle, one has to combine the continuum and discrete models. We initially run the continuum model to obtain equilibrium configuration, we discretize it to get the initial configuration for a final discrete energy minimization.

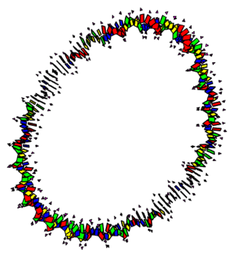
Quaternions. In order to implement the closure assumption in the simplest way, we replace the inter coordinates by absolute coordinates for the positions of the base pair frames. They are of the form of a vector in $\mathbb{R}^3$ for the translation and a quaternion for the absolute rotation. This induces a change in the energy function that has to be minimized.

cgDNA output



Example of discrete DNA minicircle configuration obtained with original (NCC) cgDNAmin.

cgDNA+min



Example of a result of cgDNA+min.

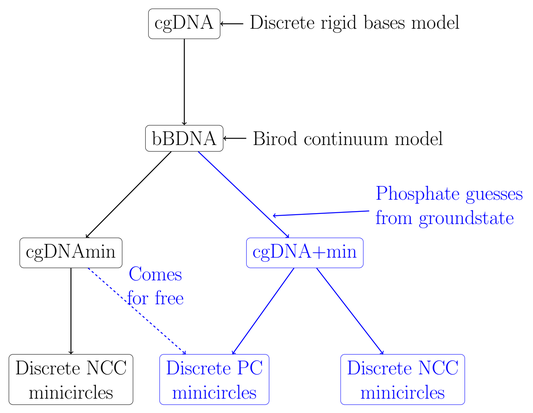
Four discrete energies. In my thesis, I implemented 3 different adaptations of the original cgDNAmin code. I adapted the model to support cgDNA+ coordinates as they are believed to be physically more accurate than cgDNA. And I implemented two types of closure assumptions: Non-Continuous (NCC) and Periodic (PC) closure.

NCC vs PC. Non-continuous closure models the formation of DNA minicircles whereas periodic closure computes the shape of a fully formed minicircle. They are solving distinct problems and are not interchangeable.

Five case studies. We tested our implementation on five sequences,with length ranging from 94 to 339 base pairs, coming from the experimental literature.

Results. We observed that cgDNA+min generally behaves similarly to the original cgDNAmin algorithm. However certain sequences produced very different results between the cgDNA and cgDNA+ cases. We believe this is due to the stability of the discretized continuum equilibrium. Changing the energy may change the solution when starting from an unstable configuration.

Model flowchart



New implementations are shown in blue.

Webpage of the thesis.

Further details. The reader is encouraged to read the full thesis for more details on the constructions of cgDNA+min. Along with the paper, we built a webpage containing all our results as well as the MATLAB scripts for the four different models: https://lcvmwww.epfl.ch/research/cgDNA/beaud/index.html.

References. All detailed references for this work can be found in the thesis provided online.