Computation and Visualization in Multiscale Modelling of DNA Mechanics

THÈSE N^{O} 7062 (2016)

PRÉSENTÉE LE 19 JUILLET 2016

À LA FACULTÉ DES SCIENCES DE BASE

CHAIRE D'ANALYSE APPLIQUÉE

PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Jarosław GŁOWACKI

acceptée sur proposition du jury:

Prof. W. Zwaenepoel, président du jury
Prof. J. H. Maddocks, directeur de thèse
Prof. C. Schütte, rapporteur
Prof. R. C. Paffenroth, rapporteur
Prof. B. Moret, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

for Maria

Acknowledgements

First of all I would like to thank my advisor Prof. John H. Maddocks for sharing with me his immense knowledge, for his guidance through my PhD research and for his constructive criticism. Many of the results presented in this thesis are an effect of close collaboration with Alex Grandchamp to whom I am much obliged. I am also thankful to Prof. Gilbert Strang for his collaboration on Chapter P1.1 and to the members of the doctoral exam committee: Prof. Willy Zwaenepoel, Prof. Bernard Moret, Prof. Christof Schütte and Prof. Randy C. Paffenroth for their valuable comments.

Getting through the thicket of administration would be impossible without our precious secretary Carine Tschanz. Great thanks to all other present and former members of the Laboratory for Computation and Visualization of Mathematics and Mechanics (LCVMM), namely Philippe, Daiva, Ludovica, Henryk, Jonathan, Evgeny, Nadia, Pauline, Alessandro, Thomas, Lennart and Hanna for many fruitful scientific meetings and discussions (and for all the less scientific ones as well). I couldn't also omit here all the rest of the EPFL comrades (including my brother Przemek, Maciek and Maya) and other friends that made the life much more enjoyable.

My sincere gratitude goes to my parents whose constant support lead me throughout my formation until this point.

Last but not least I would like to thank my brilliant wife for her unflagging patience and enthusiasm, especially in the last weeks before the defence.

This work was supported by the Swiss National Science Foundation, project number 20021_126666, and project number 200020_143613.

Abstract

Amongst the substances found in any living organism one of the most central in the workings of each living cell is deoxyribonucleic acid (DNA). For that reason it has been the subject of research in many of its aspects. These range from DNA sequencing to atomistic-level structural analysis. The most well known property of DNA is its coding function – the fact that it is the carrier of genetic information. It is, however, only a small fraction of the whole DNA (1% in humans) that is responsible for coding for proteins. The other functions of DNA such as transcription, replication and recombination are apparently strongly influenced by mechanical properties of the molecule, *i.e.* its shape and flexibility at the length scale of several hundreds of base pairs. These, in turn, differ with the sequence. Our main goal here is to provide tools that facilitate the analysis of such sequence-dependent statistical mechanical properties of DNA.

Our considerations concern two recently introduced sequence-dependent models of DNA mechanics. The first one, called *cgDNA*, is a discrete, rigid base, nearest neighbour model with a shifted quadratic energy depending on the internal parameters of the 3D configuration. In this context we describe a method of maximum entropy fitting that can be applied in the procedure of extracting *cgDNA* parameters from molecular dynamics simulation data. We also introduce a formulation within the *cgDNA* framework that allows modelling of long, repeating sequences as well as closed loops of DNA. We apply this formulation to analyse superhelical structures of the intrinsic shape of such repeating sequences. Finally a technique for efficiently computing persistence lengths of short ($\sim 200\,\mathrm{bp}$ long) DNA oligomers using an optimized Monte Carlo code within the *cgDNA* model is presented.

The second mechanics model of DNA to be considered is the continuous elastic birod. In this setting the DNA is modelled as two long, thin elastic rods with local, elastic interactions. The associated sequence-dependent birod Hamiltonian system has coefficient functions extracted from the *cgDNA* model. For this model we address issues concerning the use of these coefficients in numerical computation. We then describe the *bBDNA* software, which provides a graphical user interface for running computations within the model. We end with a presentation of example results of numerical simulations obtained using *bBDNA* and numerical techniques adapted from elastic rod models.

Keywords: *DNA mechanics, coarse-grain models, maximum entropy fit, superhelix, persistence lengths, Monte Carlo, birods, parameter continuation*

Résumé

Parmi les substances qu'on trouve dans tout organisme vivant, l'une des plus importantes pour le fonctionnement de chaque cellule est l'acide désoxyribonucléique (ADN). Pour cette raison, de nombreux aspects de l'ADN sont étudiés. Ceux-ci vont de séquençage de l'ADN à l'analyse structurale au niveau atomique. Le mieux connu est sa fonction de codage – le fait qu'il est porteur d'information génétique. Mais seule une petite fraction de l'ADN total (1 % chez les humains) code les protéines. Les autres fonctions de l'ADN telles que la transcription, la réplication et la recombinaison sont apparemment fortement influencées par les propriétés mécaniques de la molécule, c'est-à-dire sa forme et sa flexibilité à l'échelle de plusieurs centaines de paires de bases de long. Celles-ci, pour leur part, dépendent de la séquence de paires de bases. Notre objectif principal ici est de fournir des outils qui facilitent l'analyse des propriétés mécaniques de l'ADN qui dépendent de la séquence.

Notre étude porte sur deux modèles récents de la mécanique de l'ADN dépendant de la séquence. Le premier, appelé *cgDNA*, est un modèle de bases rigides interagissant avec les bases voisines les plus proches, avec une énergie quadratique décalée en fonction des paramètres internes de la configuration 3D. Dans ce contexte, on décrit une méthode d'ajustement de l'entropie maximale qui peut être appliquée pour extraire les paramètres pour *cgDNA* à partir de données venant de simulations de dynamique moléculaire. On présente également une formulation qui permet de modéliser de longues séquences répétitives, ainsi que des boucles fermées de l'ADN, dans le cadre de *cgDNA*. On utilise cette formulation pour analyser les structures superhélicoïdal de la forme intrinsèque de telles séquences répétitives. Enfin, une technique efficace pour calculer la longueur de persistance des oligomères d'ADN courts ($\sim 200\,\mathrm{pb}$), en utilisant une méthode de Monte-Carlo optimisée pour le modèle *cgDNA* est présentée.

Le deuxième modèle de la mécanique de l'ADN qu'on considère, est la bi-tige élastique continue. Dans ce contexte, l'ADN est modélisé comme deux tiges élastiques longues et minces avec des interactions élastiques locales. Les coefficients du système hamiltonien dépendant de la séquence qui est associé à la bi-tige sont des fonctions extraites de modèle *cgDNA*. Pour ce modèle, on discute des problèmes concernant l'utilisation de ces coefficients dans les calculs numériques. On décrit ensuite le logiciel *bBDNA* qui fournit une interface graphique pour effectuer des calculs avec ce modèle. On termine par des exemples de résultats des simulations numériques obtenus à l'aide de *bBDNA* et des techniques numériques adaptées des modèles de tige élastique.

Contents

List of Figures

List of Tables

Introduction

Standard Watson-Crick B-form DNA molecules consist of two strands of bases or nucleotides: adenine (`A`), cytosine (`C`), guanine (`G`) and thymine (`T`). There is a correspondence between the bases, namely `A` pairs with `T` forming two hydrogen bonds, while `G` pairs with `C` with three hydrogen bonds. Bases are covalently bonded to one of the two antiparallel sugar-phosphate backbones. The backbones have a direction specified by the chemical structure of the sugars, referred to as $5' \to 3'$, which is by convention its reading direction.

Our efforts are focused on the question of the influence of the base pair composition of DNA oligomers on their mechanical properties such as flexibility or intrinsic shape. The importance of these relations at the length scales of tens to several hundreds of base pairs has been well recognized (see *e.g.* [BedFurKat1995; OlsGorLu1998; VolVol2002; VirBerHen2004]).

To put our research in a wider context we describe a hierarchy of available models of DNA mechanics that take into account the effects of sequence (see Figure 1) and present particular pertinent examples.

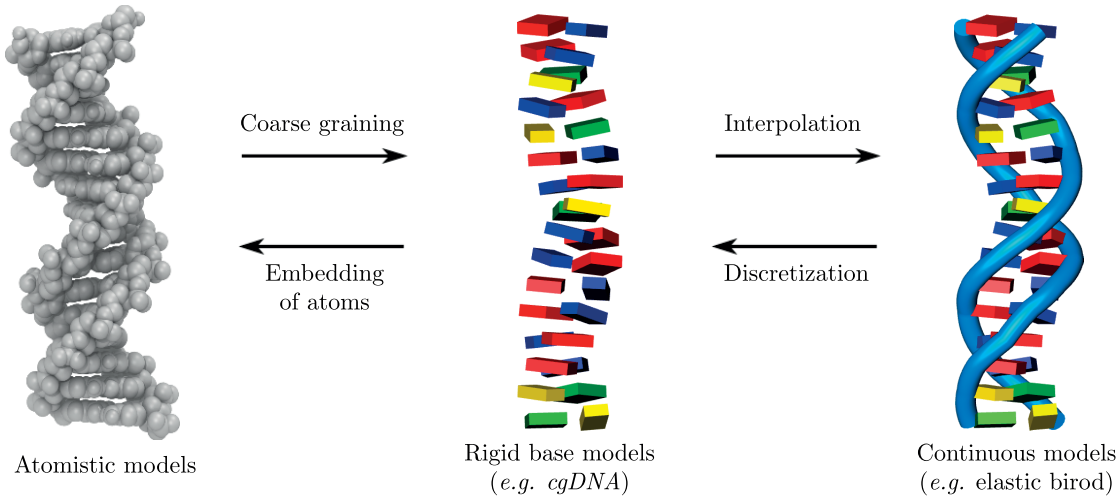


Figure 1. *Hierarchy of DNA coarse graining*

The most detailed information can be obtained from atomistic models [BevBarByu2004; DixBevCas2005; PasMadBev2014], which include every atom of the studied oligomer. Computations in such models are based on classical Molecular Dynamics (MD) where interactions between all pairs of atoms are considered. One issue of this kind of formulation is the difficulty of designing reliable force fields that describe the underlying dynamics. A comparison of the two most widely used force fields used in the context of B-form DNA can be found in [PérLanLuqOro2008]. MD simulations are also extremely computationally intensive, with microsecond simulations for relatively short oligomers (a few tens of base pairs in length) taking days on contemporary supercomputers. Furthermore the analysis of the resulting time series (on the order of TB of data) is rather laborious. Nevertheless MD simulations can provide a source of comprehensive statistics for parametrizing more coarse grain models.

In discrete coarse grain models of DNA certain groups of atoms, such as base pairs (*e.g.* [OlsGorLu1998]) are assumed to be rigid. It has been shown that a common modelling assumption of only nearest neighbour interactions between the rigid bodies is not very accurate for *rigid base pair* models [LanGonHef2009]. Hence, here we consider the discrete *rigid base cgDNA* model [Pet2012; GonPetMad2013; Pet2012]. *cgDNA* can be parametrized from statistics of time series of large scale MD simulations such as [PasMadBev2014]. As a result of coarse graining some of the information is lost, yet *cgDNA* was shown to well reproduce ground state statistics of the training set as well as of independent simulation data [Pet2012; GonPetMad2013].

Continuum formulations can be constructed by interpolating the discrete models. This need not be considered as yet another coarse graining step (because no information is lost in interpolation), unless homogenization (or averaging) techniques are additionally applied to obtain slowly varying continuous constitutive coefficient functions. The great advantage of a continuum treatment is the possibility to decouple the discretization used to numerically solve the problem from the actual physical one. Such a continuous interpolation of the rigid base model has been developed under the name of the *elastic birod model* [MoaMad2005] and has recently been extended with a Hamiltonian formulation and a method of extracting parameter functions from *cgDNA* [Gra2016].

The contributions of this thesis concern the development of methods and tools for analysing DNA oligomers using the *cgDNA* model and the elastic birod model. We have structured the presentation in three parts: first (B) presenting background material, the second (P1) revolving around discrete DNA modelling using *cgDNA*, and the third (P2) dedicated to the continuum birod description.

Chapter B.1 outlines the *cgDNA* model of [Pet2012; GonPetMad2013; Pet2012], where each base is assumed to be a rigid body. The 3D configuration of an oligomer can be described as a position of a reference point of each base and its orientation. Internal coordinates, which eliminate an overall rotation and translation of the entire oligomer, are introduced

to describe the configuration. The energy of an oligomer is assumed to be a shifted quadratic function of the internal coordinate vector $\boldsymbol{w}$, *i.e.* $\mathrm{U}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w} - \widehat{\boldsymbol{w}})^T \boldsymbol{K}(\boldsymbol{w} - \widehat{\boldsymbol{w}})$. It is further assumed that each base interacts only with its 5 nearest neighbours which implies a particular sparsity structure of the stiffness matrix $\boldsymbol{K}$. The (sequence-dependent) stiffness matrix $\boldsymbol{K}$ and ground state configuration vector $\widehat{\boldsymbol{w}}$ for any given oligomer can be reconstructed using a given parameter set. The *cgDNA* parameter sets are extracted from a large data set of molecular dynamics simulation. The parameter estimation procedure has recently been modified to use the maximum absolute entropy fit introduced in Chapter P1.1. All of the original contributions of this thesis in the context of *cgDNA* are gathered in Part P1.

Chapter B.2 describes methods for solving systems of non-linear algebraic equations with a scalar parameter λ. Parameter continuation method starts from a known regular solution and computes the one-dimensional branch of solutions passing through it by changing the parameter λ. A parametrization of the solution branch is required in order to proceed along the branch. The pseudo-arclength parametrization is described, which allows for robust traversal of every solution point during continuation, including *e.g.* fold points, where the branch "turns back" with respect to the parameter λ. A method of branch switching is presented, which can be applied at branch points, where another branch of solutions bifurcates from the one being continued. The parameter continuation method can be used for solving boundary value problems of ordinary differential equations. Such problems can be reduced to algebraic systems through discretization. The *AUTO-07p* implementation [DoeKelKer1991a; DoeKelKer1991b; DoeChaDer2009] of the presented method has been chosen for computations in the birod DNA model, presented in Part P2.

Chapter B.3 is an introduction to elastic rod [CosCos1909] and birod [MoaMad2005; Gra2016] theory. This begins by defining an elastic rod as a long thin object, whose configurations can be described using a curve of reference points $\mathbf{r}(s) \in \mathbb{R}^3$ and orientations of its cross sections $\boldsymbol{R} = \begin{bmatrix}\mathbf{d}_1 & \mathbf{d}_2 & \mathbf{d}_3\end{bmatrix} \in \mathcal{SO}(3)$ (see Figure B.3.1). Balance laws are presented, which describe the equilibrium conditions on the internal stresses acting across the cross section at s. The notion of constitutive relations that describe the connection between strains of the rod and the stresses acting across each cross section is introduced. A variational formulation of the hyperelastic rod problem is described followed by the introduction of the Hamiltonian formulation with the particular choice of unit quaternion representation of the cross section orientation. Two example Boundary Value Problems (BVP) for the elastic rod system are introduced. In the pulling and twisting BVP the rod is fixed at one end and vertical loads are applied at the other end (see Figure B.3.2a). In the closed loop problem the ends of the rod are required to coincide, while the rod cross sections at both ends are required to share a common $\mathbf{d}_3$ director vector (see Figure B.3.2b). An approach to solving these problems using techniques of symmetry breaking [LiMad1996] (see Figure B.3.4) is presented. The same approach is adapted in Part P2 to the DNA birod model.

A birod, as introduced in [MoaMad2005] and studied in [Gra2016], is a system of two rods with a common parametrization interacting elastically (see Figure B.3.5). The configuration of such a system can be described as a rod macrostructure that represents the appropriate average of the two rods, and a microstructure that gives the relative rotation and translation between the two strands. Internal coordinates of the macrostructure are the strains of the average rod. Balance laws of the birod are expressed in terms of the two rod description. Variational principles are known to exist for all the cases of boundary conditions used here [Gra2016]. A method of [Gra2016, sec. 4.2] of extracting birod DNA constitutive coefficients from the *cgDNA* model is outlined. Finally a Hamiltonian formulation in unit quaternions, analogous to the one for rods, is stated.

Part P1 presents original research in the context of the *cgDNA* model.

Chapter P1.1 describes a procedure for maximum entropy fitting for banded covariance matrices (which can be applied in the case of the sparsity pattern of the *cgDNA* model). The existence and uniqueness of maximum entropy fits was first shown in [Dem1972]. A recursive algorithm of completing the covariance to maximize the entropy of the resulting Gaussian model is given. The main contribution is a simple procedure for constructing the inverse covariance (or stiffness) of that Gaussian, which vanishes outside the prescribed sparsity pattern. The presented proof of the result uses only basic concepts of linear algebra. After the results was obtained it was found it can be recovered from prior results of [SpeKii1986], [Lau1996, sec. 5.3] and [JohLun1998] largely couched in significantly different formulations. Maximum entropy fitting improves the parameter extraction procedure of *cgDNA* [GonPetPas], as confirmed by results presented in Chapter P1.4.

In Chapter P1.2 we introduce a method of describing long oligomers of repeating sequence (tandem repeats) within the *cgDNA* model by a particular form of a stiffness matrix $\boldsymbol{K}_p(S)$ and ground state configuration vector $\widehat{\boldsymbol{w}}_p(S)$ of the repeated fragment S. Properties of the new formulation are analysed. The error of approximating the "standard" *cgDNA* energy of a finite tandem repeat using the periodic coefficients is evaluated. A numerical argument is used to show that for any sequence the periodic ground state configuration vector well approximates the standard *cgDNA* ground state configuration vector of the same sequence far ($> 5\,\mathrm{bp}$) from the ends. The periodic coefficients are also shown to be well suited for modelling covalently bonded closed loops of DNA (which is used in Chapter P2.3).

In Chapter P1.3 a method of calculating pitch and radius of superhelices formed by ground state configurations of DNA tandem repeats in the *cgDNA* model with periodic coefficients is shown. The applicability of the procedure is studied. The radius and pitch returned by the method are shown to be invariant under changing the number M of repeats of the basal sequence, under cyclic shifts of the sequence and under Watson-Crick symmetry. An exhaustive study of the superhelical structure of intrinsic shapes of tandem repeats of all possible sequences of length up to $12\,\mathrm{bp}$ is made. Intrinsic shapes of repeats

of fragments of up to 10 base pairs were found to be either very close to straight, or to form left-handed superhelices. For all fragments of 12 bp the superhelices were found to be right-handed. In case of repeats of 11 bp long fragments both left- and right-handed helices were found with an exceptionally wide range of pitches and radii. Two sequences of particular superhelical structure from the study have been chosen for the pulling and twisting numerical experiment in Chapter P2.3.

Chapter P1.4 presents a Monte Carlo approach to calculating DNA persistence lengths in the *cgDNA* model. An efficient Monte Carlo code *cgDNAmc* developed for the purpose is presented as well as results of a number of simulations run with the code.

Part P2 of our considerations is entirely dedicated to the birod model of DNA.

Chapter P2.1 reports two issues encountered while performing parameter continuation in the birod model with the DNA coefficients of [Gra2016, sec. 4.2]. Bifurcation detection was found to fail in *AUTO-07p* for any parameter continuation run in the birod DNA model. The problem is identified as numerical stiffness of the system. An alternative bifurcation detection method for *AUTO-07p* is proposed and implemented. The other problem is discovered to be related to the pronounced discontinuities of the birod DNA coefficients at base pairs. This feature of the coefficients brings about the requirement of excessive discretization that exceeds the number of base pairs of the modelled oligomer several times. Application of an analogue of the coefficient homogenization technique of [Gra2016, sec. 7.3] is proposed, and a positive validation of results obtained using it is presented. All results of birod computations presented herein are performed using homogenized DNA coefficient.

Chapter P2.2 describes the *bBDNA* software for computation in the DNA birod model and visualization of results. Many aspects of the application were modelled on the *VBM* package of [Paf1999a; Paf1999b] as well as on the *PLAUT04* interface of the *AUTO-07p* solver [DoeChaDer2009]. The general structure of the code is outlined together with a justification of the design decisions. Finally a procedure of symmetry breaking from straight, inextensible, unshearable, uniform, transversely isotropic rods directly to full *bBDNA* birods is outlined. This is an application of an approach used previously in case of elastic rods [LiMad1996; DicLiMad1996].

The last Chapter P2.3 is dedicated to example results of solving Boundary Value Problems (BVP) in the birod DNA model using *bBDNA*. Pulling and twisting BVP experiments for two superhelical oligomers of Chapter P1.3, and one intrinsically close to straight sequence of Chapter P1.3 are performed using the interactive computational steering interface of *bBDNA*. Subsequently results in the closed loop BVP, obtained through the automatic symmetry breaking script of Chapter P2.2 are presented.

Background material

B.1 The *cgDNA* model

This chapter reviews necessary background material. While certain original refinements are presented here, the material in essence pre-dates this thesis. *cgDNA* is a sequence-dependent, nearest neighbour rigid base model of the statistical mechanics of B-form DNA in solution [Pet2012; GonPetMad2013; PetPasGonMad2014]. DNA configurations in the *cgDNA* model are expressed in internal coordinates [LanGonHef2009], which eliminate an overall translation and rotation of the DNA fragment. The internal coordinates describe the relative translational and rotational displacement between bases within a base pair and between neighbouring base pairs. For any sequence of DNA a free energy minimizing configuration (called the ground state configuration) together with a stiffness matrix are predicted by the *cgDNA* model. Using this ground state configuration vector and stiffness matrix the free energy can then be evaluated for any given configuration of the DNA oligomer. The sequence dependent parameters for the *cgDNA* model are extracted from a large data base of time series of Molecular Dynamics (MD) simulations [BevBarByu2004; DixBevCas2005; LavZakBev2010]. The MD simulations are constantly being extended [PasMadBev2014], which allows ongoing improvement of the *cgDNA* parameters. The authors of the model have shown that it well reproduces ground state statistics of MD simulations at the length scale of a few tens of base pairs.

B.1.1 Configuration of a DNA oligomer

In the *cgDNA* model each base is modelled as a rigid body. There is no explicit description of the backbone, although its influence is encoded in the parameters of the model. Each base of a DNA molecule is described using the the *Curves+* [LavMoaMad2009] version of the standard Tsukuba frame [OlsBanBur2001]. We denote the ath main strand base frame (position and orientation of the ath base in the main or reading strand) as $(\boldsymbol{D}_a^+, \mathbf{r}_a^+) \in \mathcal{SE}(3)$ (see Appendix (A.1) for the notation for rigid body displacements). Similarly the ath (in the order of the main strand) complementary strand base frame will be written as $(\boldsymbol{D}_a^-, \mathbf{r}_a^-) \in \mathcal{SE}(3)$. Here the positions and orientations are expressed with respect to the chosen laboratory frame.

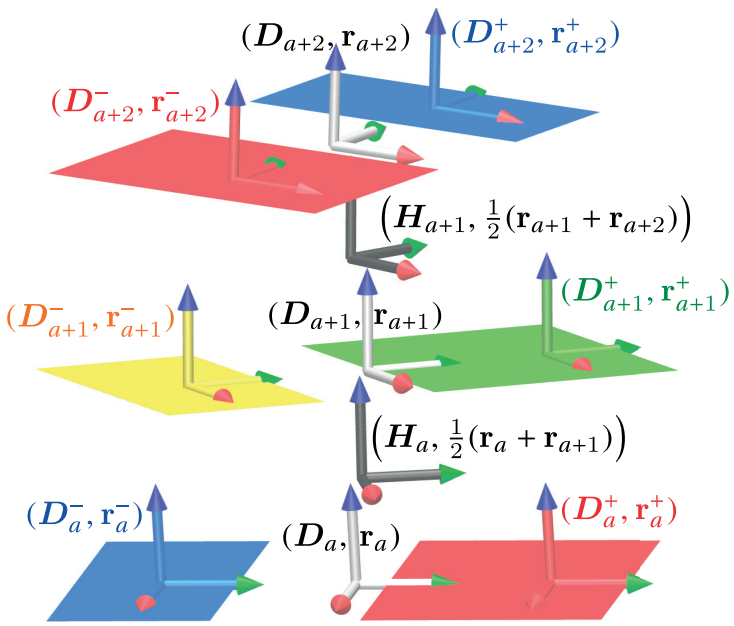


Figure B.1.1. *A schematic view of base, base pair and junction frames in cgDNA. Here the reading (Watson) strand is* `AGT` *(red, green, blue) and so the complementary (Crick) strand is* `TCA`*. The reading strand base frames* $(\boldsymbol{D}_a^+, \mathbf{r}_a^+)$ *and complementary strand base frames* $(\boldsymbol{D}_a^-, \mathbf{r}_a^-)$ *are indicated in colour. Their* $\mathcal{SE}(3)$ *averages, i.e. the base pair frames* $(\boldsymbol{D}_a, \mathbf{r}_a)$ *cgDNA are shown in light grey. The junction frames* $\left(\boldsymbol{H}_a, \frac{1}{2}(\mathbf{r}_a + \mathbf{r}_{a+1})\right)$ *that are the* $\mathcal{SE}(3)$ *averages of the consecutive base pair frames are marked as dark grey. Note that the complementary frames are flipped. The figure also presents the convention for visualizing rigid base representations of DNA oligomers used throughout the thesis. Each base is visualized as a plate which is the bounding rectangle of centres of all atoms of the base. The shown configuration is of no particular significance.*

B.1.2 Internal coordinates

The model uses internal coordinates (*i.e.* shape coordinates that are invariant under an overall translation and rotation of the molecule) [LanGonHef2009; Pet2012].

Define the relative rotation from the complementary to the main strand:

$$\boldsymbol{\Lambda}_a = (\boldsymbol{D}_a^-)^T \boldsymbol{D}_a^+ \quad \in \mathcal{SO}(3) \quad , \tag{B.1.1}$$

the base pair orientation with respect to the laboratory frame:

$$\boldsymbol{D}_a = \boldsymbol{D}_a^- \sqrt{\boldsymbol{\Lambda}_a} \quad \in \mathcal{SO}(3) \quad , \tag{B.1.2}$$

and the relative translation between base pairs expressed with respect to the base pair frame $\boldsymbol{D}_a$:

$$\boldsymbol{\varrho}_a = (\boldsymbol{D}_a)^T (\mathbf{r}_a^+ - \mathbf{r}_a^-) \quad \in \mathbb{R}^3 \quad . \tag{B.1.3}$$

Let $\mathbf{r}_a$ be the base pair position with respect to the laboratory frame:

$$\mathbf{r}_a = \frac{1}{2}(\mathbf{r}_a^+ + \mathbf{r}_a^-) \quad \in \mathbb{R}^3 \quad , \tag{B.1.4}$$

and let $\boldsymbol{L}_a$ be the relative rotation between base pair a and $a+1$:

$$\boldsymbol{L}_a = (\boldsymbol{D}_a)^T \boldsymbol{D}_{a+1} \quad \in \mathcal{SO}(3) \quad , \tag{B.1.5}$$

so that the orientation of the ath junction $\boldsymbol{H}_a$ can be defined as:

$$\boldsymbol{H}_a = \boldsymbol{D}_a \sqrt{\boldsymbol{L}_a} \quad \in \mathcal{SO}(3) \quad . \tag{B.1.6}$$

The base pair step translational coordinates $\boldsymbol{\rho}_a$ with respect to $\boldsymbol{H}_a$ are defined as:

$$\boldsymbol{\rho}_a = (\boldsymbol{H}_a)^T (\mathbf{r}_{(i+1)} - \mathbf{r}_a) \quad \in \mathbb{R}^3 \quad . \tag{B.1.7}$$

In B-form DNA the relative rotations $\boldsymbol{\Lambda}_a$ and $\boldsymbol{L}_a$ in the above definitions are small (specifically the angle of rotation is much smaller than π) and so can be parametrized by Cayley vectors (see Section A.1.1.3):

$$\boldsymbol{\vartheta}_a = \mathrm{cay}^{-1}(\boldsymbol{\Lambda}_a) \tag{B.1.8}$$

$$\boldsymbol{\theta}_a = \mathrm{cay}^{-1}(\boldsymbol{L}_a) \quad . \tag{B.1.9}$$

Configurations with coordinates close to rotations by π are out of the scope of the *cgDNA* model. Hence the singularity of the Cayley vector parametrization at rotation angles close to π (see Section A.1.1.3) is not problematic.

Given the above, the six degrees of freedom between bases of the ath base pair are described by a vector $\boldsymbol{y}_a = \begin{bmatrix}\boldsymbol{\vartheta}_a & \boldsymbol{\varrho}_a\end{bmatrix}^T$ called the *intra base pair* coordinates. Similarly, the six degrees of freedom between bases of the ath junction (*i.e.* of base pair a and base pair $a+1$) are defined by a vector $\boldsymbol{z}_a = \begin{bmatrix}\boldsymbol{\theta}_a & \boldsymbol{\rho}_a\end{bmatrix}^T$ called *inter base pair* coordinates. The conventional name of each coordinate as well as its geometrical interpretation is shown in Figure B.1.2

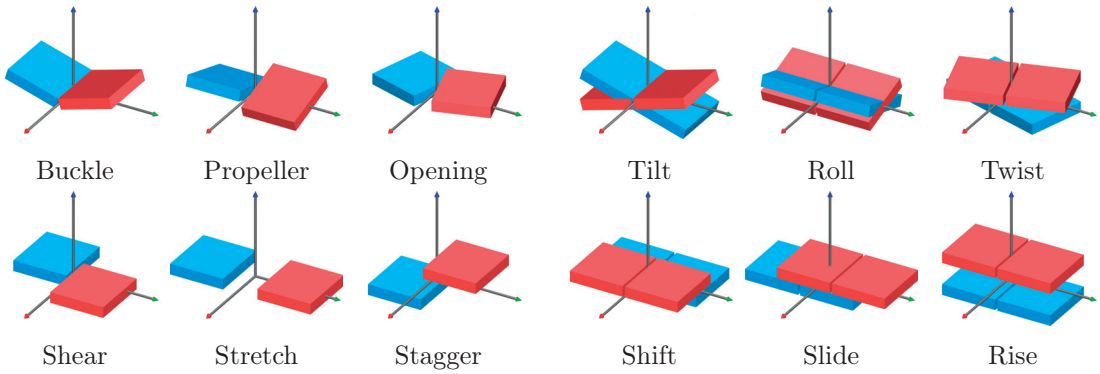


Figure B.1.2. *Intra base pair (left) and inter base pair (right) coordinates of the cgDNA model. Rotational coordinates are in the top row, translational coordinates in the bottom row. Bases are represented as flat rectangle plates. In each case the respective coordinate of value* 3 *is shown. In case of twist, shift and slide additionally the value of* 1 *in rise is added for clarity.*

Finally the configuration of an entire oligomer of N base pairs is described by a vector of coordinates:

$$\boldsymbol{w} = \Big[\underbrace{\boldsymbol{\vartheta}_1\ \boldsymbol{\varrho}_1}_{\boldsymbol{y}_1}\ \underbrace{\boldsymbol{\theta}_1\ \boldsymbol{\rho}_1}_{\boldsymbol{z}_1}\ \underbrace{\boldsymbol{\vartheta}_2\ \boldsymbol{\varrho}_2}_{\boldsymbol{y}_2}\ \underbrace{\boldsymbol{\theta}_2\ \boldsymbol{\rho}_2}_{\boldsymbol{z}_2}\ \ldots\ \underbrace{\boldsymbol{\theta}_{N-1}\ \boldsymbol{\rho}_{N-1}}_{\boldsymbol{z}_{N-1}}\ \underbrace{\boldsymbol{\vartheta}_N\ \boldsymbol{\varrho}_N}_{\boldsymbol{y}_N}\Big]^T \in \mathbb{R}^{12N-6} \tag{B.1.10}$$

called the *configuration vector* or *shape vector*.

For reasons detailed in [GonPetMad2013, sec. II.D] a characteristic scale is introduced: $\ell = 1\text{Å}$ for translations and $g = \frac{1}{5}$ [radians] for rotational variables so that the non-dimensionalized variables used in the *cgDNA* model are defined as:

$$\begin{aligned} \underline{\boldsymbol{\vartheta}}_a &= \frac{1}{g}\boldsymbol{\vartheta}_a = 5\boldsymbol{\vartheta}_a & \underline{\boldsymbol{\varrho}}_a &= \frac{1}{\ell}\boldsymbol{\varrho}_a = \boldsymbol{\varrho}_a \\ \underline{\boldsymbol{\theta}}_a &= \frac{1}{g}\boldsymbol{\theta}_a = 5\boldsymbol{\theta}_a & \underline{\boldsymbol{\rho}}_a &= \frac{1}{\ell}\boldsymbol{\rho}_a = \boldsymbol{\rho}_a \end{aligned} \tag{B.1.11}$$

and the non-dimensionalized intra $\underline{\boldsymbol{y}}_a$ and inter $\underline{\boldsymbol{z}}_a$ vectors as well as non-dimensionalized shape vector $\underline{\boldsymbol{w}}$ can be defined by expressions analogous to (B.1.10).

B.1.3 Reconstruction of 3D configuration

In what follows we use the notation $\mathcal{D}_a^+$, $\mathcal{D}_a^-$ and $\mathcal{D}_a$ to mean the homogeneous coefficients (see Section A.1.2) of the reading strand base frame, complementary base frame, and the base pair frame, respectively, so that:

$$\mathcal{D}_a^+ := \begin{bmatrix} \boldsymbol{D}_a^+ & \mathbf{r}_a^+ \\ \mathbf{0}^T & 1 \end{bmatrix}, \qquad \mathcal{D}_a^- := \begin{bmatrix} \boldsymbol{D}_a^- & \mathbf{r}_a^- \\ \mathbf{0}^T & 1 \end{bmatrix}, \quad \text{and} \quad \mathcal{D}_a := \begin{bmatrix} \boldsymbol{D}_a & \mathbf{r}_a \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{B.1.12}$$

with $\mathbf{0} \in \mathbb{R}^3$. Homogeneous coordinates are introduced as they allow for clean and compact notation. The question of efficient numerical realization of the procedures below is addressed in Section P1.4.2.3.

Given any position and orientation $\mathcal{D}_1$ of the first base pair (usually chosen to be the identity), the entire configuration can be recovered from any given shape vector (B.1.10) as [LanGonHef2009]:

$$\boldsymbol{L}_a = \operatorname{cay}(g\underline{\boldsymbol{\theta}}_a) \tag{B.1.13a}$$

$$\mathcal{L}_a := \begin{bmatrix} \boldsymbol{L}_a & -\ell\sqrt{\boldsymbol{L}_a}\underline{\boldsymbol{\rho}}_a \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{B.1.13b}$$

$$\mathcal{D}_{a+1} = \mathcal{D}_a\mathcal{L}_a \tag{B.1.13c}$$

$$\boldsymbol{\Lambda}_a = \operatorname{cay}(g\underline{\boldsymbol{\vartheta}}_a) \tag{B.1.14a}$$

$$\mathcal{B}_a^+ := \begin{bmatrix} \sqrt{\boldsymbol{\Lambda}_a} & \frac{\ell}{2}\underline{\boldsymbol{\varrho}}_a \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{B.1.14b}$$

$$\mathcal{B}_a^- := \begin{bmatrix} \left(\sqrt{\boldsymbol{\Lambda}_a}\right)^T & -\frac{\ell}{2}\underline{\boldsymbol{\varrho}}_a \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{B.1.14c}$$

$$\mathcal{D}_a^+ = \mathcal{D}_a\mathcal{B}_a^+ \tag{B.1.14d}$$

$$\mathcal{D}_a^- = \mathcal{D}_a\mathcal{B}_a^- \tag{B.1.14e}$$

In the above the notation $\sqrt{\boldsymbol{\Lambda}_a}$ and $\sqrt{\boldsymbol{L}_a}$ mean the principal square roots of the respective rotation matrices, which are interpreted as representing half of the rotation represented by the matrices themselves (see Section A.1.1.4).

In many applications, *e.g.* the Monte Carlo simulations presented in Chapter P1.4, it suffices to reconstruct only the base pair frames (B.1.13) and not the individual base frames (B.1.14), which is possible for the tree-like connectivity expressed in the decoupling of (B.1.13) from (B.1.14).

B.1.4 *cgDNA* energy

In the *cgDNA* model for a given DNA molecule in a heat bath the equilibrium distribution of its configurations $\underline{\boldsymbol{w}}$ is assumed to be given by the density:

$$\rho(\underline{\boldsymbol{w}}) = \frac{e^{-\beta \mathrm{U}(\underline{\boldsymbol{w}})} J(\underline{\boldsymbol{w}})}{\int e^{-\beta \mathrm{U}(\underline{\boldsymbol{w}})} J(\underline{\boldsymbol{w}}) \, \mathrm{d}\underline{\boldsymbol{w}}} \tag{B.1.15}$$

where $\beta = \frac{1}{k_B T}$ with k_B the Boltzmann constant, T the temperature of the bath, and

$$J = \left[\sum_{a=1}^{n-1} \left(1 + \frac{|g\underline{\boldsymbol{\theta}}_a|^2}{4}\right)^{-2}\right] \left[\sum_{a=1}^{n} \left(1 + \frac{|g\underline{\boldsymbol{\vartheta}}_a|^2}{4}\right)^{-2}\right] \tag{B.1.16}$$

is the Jacobian factor associated with the particular "non-flat" nature of the non-dimensionalized Cayley vector rotational coordinates [WalGonMad2010].

The internal energy U is assumed to be a shifted quadratic function of the configuration $\underline{\boldsymbol{w}}$ *i.e.*:

$$\mathrm{U}(\underline{\boldsymbol{w}}) = \frac{1}{2}(\underline{\boldsymbol{w}} - \widehat{\underline{\boldsymbol{w}}})^T \boldsymbol{K} (\underline{\boldsymbol{w}} - \widehat{\underline{\boldsymbol{w}}}) \tag{B.1.17}$$

with the ground state shape vector $\widehat{\underline{\boldsymbol{w}}} \in \mathbb{R}^{12n-6}$ (energy minimizer) and the symmetric positive definite stiffness matrix $\boldsymbol{K} \in \mathbb{R}^{(12n-6)\times(12n-6)}$. The parameters $\widehat{\underline{\boldsymbol{w}}}$ and $\boldsymbol{K}$ depend on the base sequence of the particular oligomer. It is clear from (B.1.15) that the probability density ρ is the same for any constant shift of the energy U. Hence, without changing any statistical properties (B.1.17) assumes the energy $U(\widehat{\underline{\boldsymbol{w}}})$ of the ground state configuration to be zero.

A scale of $k_B T$ is used to non-dimensionalize the energy U:

$$\underline{\mathrm{U}}(\underline{\boldsymbol{w}}) = \frac{1}{2}(\underline{\boldsymbol{w}} - \widehat{\underline{\boldsymbol{w}}})^T \underline{\boldsymbol{K}} (\underline{\boldsymbol{w}} - \widehat{\underline{\boldsymbol{w}}}) \tag{B.1.18}$$

with $\underline{\boldsymbol{K}} = \frac{1}{k_B T} \boldsymbol{K}$. For convenience the $\underline{\cdot}$ notation will be dropped and the variables will be assumed to be non-dimensionalized throughout, unless indicated otherwise.

B.1.5 Nearest neighbour assumption

In addition to the rigid base assumption, the *cgDNA* model also assumes that each base interacts only with its 5 nearest neighbours (the complementary base, the two bases in the base pair upstream and the two in the base pair downstream). This implies a particular 18×18 overlapping block structure with 6×6 overlaps. For an oligomer of sequence $S = \mathtt{X}_1\mathtt{X}_2 \ldots \mathtt{X}_{n-1}\mathtt{X}_n$ (with $\mathtt{X}_i \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$) the stiffness matrix $\boldsymbol{K}(S)$ can be constructed

by summing diagonal coefficient blocks: $\boldsymbol{K}^{\mathtt{X}_a} \in \mathbb{R}^{6\times 6}$ that represent stiffness contribution from intra interactions within ath base pair and $\boldsymbol{K}^{\mathtt{X}_a\mathtt{X}_{a+1}} \in \mathbb{R}^{18\times 18}$ that represent stiffness contribution from inter interactions between all bases in base pairs a and $a+1$. The sequence-dependence of the parameter blocks $\boldsymbol{K}^{\mathtt{X}_a}$ and $\boldsymbol{K}^{\mathtt{X}_a\mathtt{X}_{a+1}}$ is an assumption of the *cgDNA* model independent of the nearest neighbour assumption. An analogous procedure can be used to compute the so called *weighted shape vector* $\boldsymbol{\sigma}(S)$ from intra $\boldsymbol{\sigma}^{\mathtt{X}_a} \in \mathbb{R}^6$ and inter $\boldsymbol{\sigma}^{\mathtt{X}_a\mathtt{X}_{a+1}} \in \mathbb{R}^{18}$ coefficients (see Figure B.1.3). The weighted shape vector satisfies the relation:

$$\boldsymbol{\sigma}(S) = \boldsymbol{K}(S)\widehat{\boldsymbol{w}}(S) \tag{B.1.19}$$

so that the ground state configuration vector $\widehat{\boldsymbol{w}}(S)$ is computed using a linear solve that inverts the relation (B.1.19).

Note that the inverse of the stiffness matrix $\boldsymbol{K}(S)$ is dense, so $\widehat{\boldsymbol{w}}(S)$ has a non-trivial dependence on the entire sequence. Note also that the first and last diagonal 6×6 block of $\boldsymbol{K}(S)$ as well as the first 6 and last 6 entries of $\boldsymbol{\sigma}(S)$ are the sum of only two overlapping blocks/vectors (not three as with the other 6×6 overlaps). As a result the parameters for a given base pair step are different based on whether it appears inside or at an end of a given oligomer. This gives rise to *end effects*.

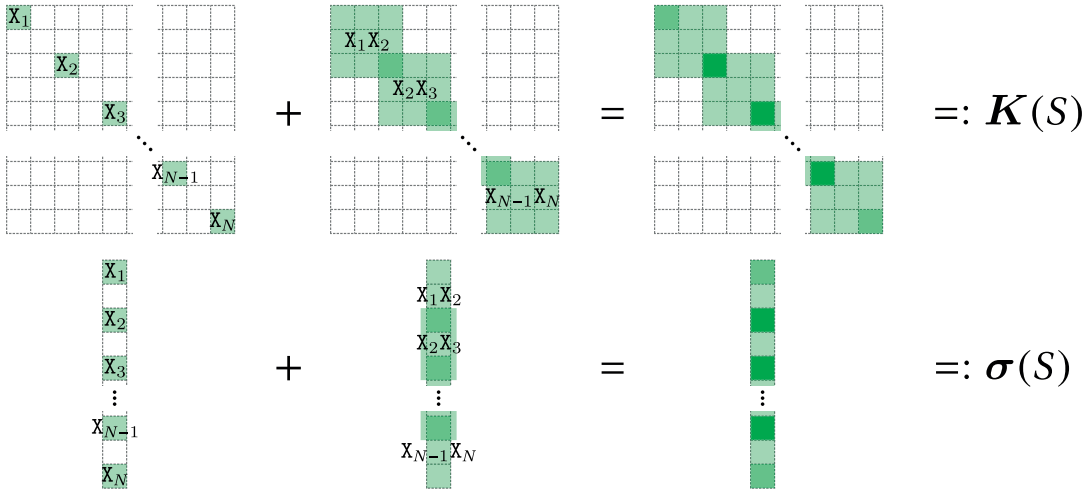


Figure B.1.3. *A sketch showing the procedure of building a cgDNA stiffness matrix* $\boldsymbol{K}(S)$ *with its particular sparsity pattern (top) and the weighted shape vector* $\boldsymbol{\sigma}(S)$ *(bottom) for an oligomer sequence* $S = \mathtt{X}_1\mathtt{X}_2\mathtt{X}_3\ldots\mathtt{X}_{n-1}\mathtt{X}_n$*. Each cell of the matrix is of dimension* 6×6 *and each cell in the vector is of dimension* 6×1*.*

B.1.6 *cgDNA* parameter sets

Computations of sequence dependent energies U in the *cgDNA* model require a set of mononucleotide $\boldsymbol{\sigma}^{\mathtt{X}}$, $\boldsymbol{K}^{\mathtt{X}}$ and dinucleotide $\boldsymbol{\sigma}^{\mathtt{XY}}$, $\boldsymbol{K}^{\mathtt{XY}}$ coefficients for all $\mathtt{X}, \mathtt{Y} \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ called a *parameter set*. The original one called *cgDNAparamset1*, published together with the model, was extracted from a large ensemble of full-atom Molecular Dynamics (MD) simulations data of the ABC collaboration [BevBarByu2004; DixBevCas2005; LavZakBev2010]. The extraction process has been split into three independent steps. First a time series of configurations $\left\{\boldsymbol{w}(S)^{(j)}\right\}$ for each oligomer S is extracted from MD snapshots using the *Curves+* package [LavMoaMad2009], is assumed to follow a Gaussian distribution:

$$\rho_o^S(\boldsymbol{w}) = \frac{1}{Z_o^S} e^{-(\boldsymbol{w}-\widehat{\boldsymbol{w}}_o^S)^T \boldsymbol{K}_o^S (\boldsymbol{w}-\widehat{\boldsymbol{w}}_o^S)} \tag{B.1.20}$$

and to be ergodic. These assumptions allow one to extract an observed ground state shape vector $\widehat{\boldsymbol{w}}_o^S$ and an observed covariance (inverse stiffness) matrix $(\boldsymbol{K}_o^S)^{-1}$ as first and second moments of ρ_o^S approximated in the standard way from the ensemble of snapshots $\left\{\boldsymbol{w}(S)^{(j)}\right\}$.

In the second step, for each simulated oligomer S a stiffness matrix $\boldsymbol{K}_{nn}^S$ with the 18×18 sparsity pattern induced by the nearest neighbour assumption (see Section B.1.5) is computed. This is done by minimizing the Kullback-Leibler divergence [KulLei1951] (or maximizing the relative entropy):

$$D(\rho_{nn}^S,\, \rho_o^S) := \int \rho_{nn}^S(\boldsymbol{w}) \ln\left(\frac{\rho_{nn}^S(\boldsymbol{w})}{\rho_o^S(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w} \tag{B.1.21}$$

between ρ_{nn}^S (whose first and second moment are $\widehat{\boldsymbol{w}}_{nn}^S = \widehat{\boldsymbol{w}}_o^S$ and $(\boldsymbol{K}_{nn}^S)^{-1}$, respectively) and the observed distribution ρ_o.

Finally the parameter set $\mathcal{P}_1^* = \{\boldsymbol{\sigma}^{\mathtt{X}}, \boldsymbol{K}^{\mathtt{X}}, \boldsymbol{\sigma}^{\mathtt{XY}}, \boldsymbol{K}^{\mathtt{XY}}\}$ is computed by minimizing the sum over all oligomers S of Kullback-Leibler divergences

$$\mathcal{P}_1^* = \underset{\mathcal{P}}{\operatorname{argmin}} \sum_S D(\rho_{\mathcal{P}}^S,\, \rho_{nn}^S) \quad . \tag{B.1.22}$$

Each $\rho_{\mathcal{P}}^S$ is defined by the ground state shape vector $\widehat{\boldsymbol{w}}_{\mathcal{P}}(S)$ and stiffness matrix $\boldsymbol{K}_{\mathcal{P}}(S)$ constructed from the parameter set $\mathcal{P}$ for the oligomer S.

Chapter P1.1 presents an alternative approach to step two that involves maximizing the absolute entropy. The resulting parameter set labelled as *cgDNAparamset2* (derived from the same molecular dynamics data) has proven to have certain advantages over the original *cgDNAparamset1* and was used to produce the results of this thesis.

B.2 A parameter continuation method for solving boundary value problems

In this chapter we briefly present elements of the *(single) parameter continuation* method of solving systems non-linear algebraic equations as described *e.g.* in [DoeKelKer1991a] or [Paf1999a]. The method can be used to numerically solve boundary value problems of ordinary differential equations that are reduced to algebraic systems through discretization [DoeKelKer1991b]. This approach implemented in the *AUTO-07p* package, described in Section B.2.4, has been chosen for computations in the birod model of DNA of [Gra2016] (briefly outlined in Section B.3.3). The choice was motivated by the fact that previous versions of *AUTO* proved to be very useful and robust for computations in symmetry breaking within elastic rod models [LiMad1996; DicLiMad1996] as well as in the context of DNA modelling [ManMadKah1996; FurManMad2000].

B.2.1 Continuation of solutions

In simple terms single parameter continuation means exploration of the solution space of a non-linear algebraic system of the form:

$$F(\boldsymbol{v}\,;\,\lambda) = \mathbf{0} \quad , \qquad F : \mathbb{R}^{n+1} \to \mathbb{R}^{n} \quad , \qquad \lambda \in \mathbb{R} \tag{B.2.1}$$

for the state variable $\boldsymbol{v}$ by changing the value of its parameter λ. For simplicity in what follows we introduce the notation $\boldsymbol{x} := \begin{bmatrix}\boldsymbol{v} & \lambda\end{bmatrix}^T$, so the system (B.2.8) can be written as:

$$F(\boldsymbol{x}) = \mathbf{0} \quad . \tag{B.2.2}$$

In case of a *regular solutions* $\boldsymbol{x}_\mathrm{r}$ of (B.2.2) for which the Jacobian $F_{\boldsymbol{x}}(\boldsymbol{x}_\mathrm{r})$ has rank n, there exists a unique one-dimensional *branch* $\boldsymbol{x}(\tau) = \begin{bmatrix}\boldsymbol{v}(\tau) & \lambda(\tau)\end{bmatrix}^T$ of solutions passing through $\boldsymbol{x}_\mathrm{r}$ [DoeKelKer1991a], where τ is a chosen parametrization of the branch. The principle of parameter continuation is to start from a known initial regular solution $\boldsymbol{x}_0$ and compute a nearby solution on the solution branch by a perturbation in λ. An entire set of solutions may be generated by applying this procedure to subsequently computed solution points. Projections of such a solution set into some subset of system variables, such as the one shown schematically in Figure B.2.1, will be referred to as *bifurcation diagrams*.

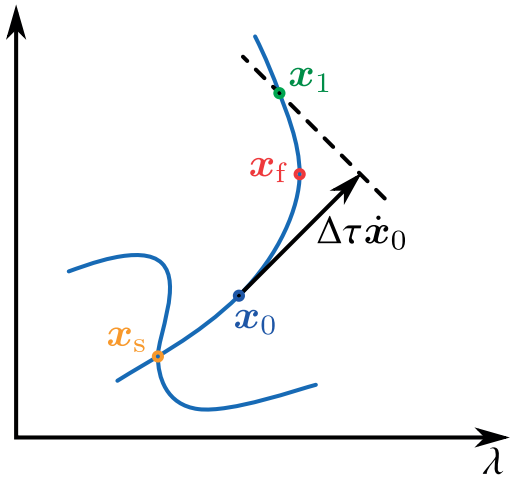


Figure B.2.1. *A schematic picture of pseudo-arclength continuation. Two branches of solutions are indicated as blue lines.* $\boldsymbol{x}_0 = \boldsymbol{x}(\tau_0)$ *is the known starting solution. The black arrow indicates the displacement of* $\Delta\tau$ *in the direction* $\dot{\boldsymbol{x}}_0 = \frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{x}(\tau_0)$. $\boldsymbol{x}_1$ *is the new solution found in the hyperspace (indicated by the dashed line) orthogonal to* $\dot{\boldsymbol{x}}_0$ *at distance* $\Delta\tau$. *Additionally a fold* $\boldsymbol{x}_\mathrm{f}$ *with respect to the parameter* λ *is shown. An example of a simple singular solution* $\boldsymbol{x}_\mathrm{s}$ *that is a branching point, where two solution branches cross is also indicated.*

Let $p(\boldsymbol{x}(\tau)\,;\,\tau) = \mathbf{0}$ be the function that defines the chosen parametrization. The system that needs to be solved to perform continuation can then be written as:

$$\begin{cases} F(\boldsymbol{x}(\tau)) = \mathbf{0} \\ p(\boldsymbol{x}(\tau)\,;\,\tau) = 0 \end{cases} . \tag{B.2.3}$$

For a known solution $\boldsymbol{x}_0 = \boldsymbol{x}(\tau_0)$ and $\dot{\boldsymbol{x}}_0$ the next solution $\boldsymbol{x}_1 = \boldsymbol{x}(\tau_0 + \Delta\tau)$ can be found using Newton iteration:

$$\begin{cases} \boldsymbol{x}_1^0 = \boldsymbol{x}_0 + \Delta\tau\,\dot{\boldsymbol{x}}_0 \\ \begin{bmatrix} F_{\boldsymbol{x}}\left(\boldsymbol{x}_1^i\right) \\ p_{\boldsymbol{x}}\left(\boldsymbol{x}_1^i\,;\,\Delta\tau\right) \end{bmatrix} \Delta\boldsymbol{x}_1^i = -\begin{bmatrix} F\left(\boldsymbol{x}_1^i\right) \\ p\left(\boldsymbol{x}_1^i\,;\,\Delta\tau\right) \end{bmatrix} \\ \boldsymbol{x}_1^{i+1} = \boldsymbol{x}_1^i + \Delta\boldsymbol{x}_1^i \end{cases} . \tag{B.2.4}$$

The next direction vector $\dot{\boldsymbol{x}}_1$ can be found after convergence of the Newton method. The following relation (result of differentiation of (B.2.3) with reference to τ) can be used:

$$\begin{bmatrix} F_{\boldsymbol{x}}(\boldsymbol{x}_1) \\ p_{\boldsymbol{x}}(\boldsymbol{x}_1\,;\,\Delta\tau) \end{bmatrix} \dot{\boldsymbol{x}}_1 = -\begin{bmatrix} \mathbf{0} \\ p_\tau(\boldsymbol{x}_1\,;\,\Delta\tau) \end{bmatrix} \tag{B.2.5}$$

The direction vector should then be normalized so that $|\dot{\boldsymbol{x}}_1| = 1$.

B.2.2 Choice of parametrization

A crucial question of parameter continuation is the choice of the parametrization function $p(\cdot\,;\,\cdot)$. The most straightforward idea might be to use λ, *i.e.* $p(\boldsymbol{v}\,;\,\lambda) = \lambda - \tau$. However this method, called *natural parametrization*, fails at certain points called *simple folds* $\boldsymbol{x}_\mathrm{f}$. These are points where $F_{\boldsymbol{v}}(\boldsymbol{x}_\mathrm{f})$ has rank $n-1$ and $F_\tau(\boldsymbol{x}_\mathrm{f})$ is not in its range. It can be shown that for simple folds $\frac{\mathrm{d}}{\mathrm{d}\tau_\mathrm{a}}\lambda = 0$, where τ_a is the arclength parametrization of the curve [DoeKelKer1991a]. This means that at such points the branch "turns back" with reference to λ (see Figure B.2.1). There is an extensive theory describing the fact that fold points are associated with stability exchange, *e.g.* [Mad1987].

The *AUTO-07p* package, presented in Section B.2.4, uses the so called *pseudo-arclength parametrization*, which is an approximation to the arclength parametrization $p(\boldsymbol{v}\,;\,\lambda) = \left\|\frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{v}\right\| + \left(\frac{\mathrm{d}}{\mathrm{d}\tau}\lambda\right)^2 - 1$ of the branch. Geometrically this method searches for a new solution $\boldsymbol{x}_1$ to (B.2.2) on a hyperplane perpendicular to $\dot{\boldsymbol{x}}_0 := \frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{x}_0$ located at distance $\Delta\tau$ (see Figure B.2.1). This can be written as:

$$\begin{cases} F(\boldsymbol{x}_1) = \mathbf{0} \\ (\boldsymbol{x}_1 - \boldsymbol{x}_0)^T\dot{\boldsymbol{x}}_0 - \Delta\tau = 0 \end{cases} . \tag{B.2.6}$$

It can be shown that pseudo-arclength continuation works for every regular solution point $\boldsymbol{x}_0$ (including folds) provided that $\Delta\tau$ is sufficiently small [DoeKelKer1991a].

B.2.3 Singular points and bifurcation detection

An important type of non-regular solutions are *simple singular points* $\boldsymbol{x}_\mathrm{s}$ where $F_{\boldsymbol{v}}(\boldsymbol{x}_\mathrm{s})$ becomes singular and $F_{\boldsymbol{x}}(\boldsymbol{x}_\mathrm{s})$ has rank $n-1$, so that its null space is two-dimensional. Note that by differentiating (B.2.2) with respect to τ we have $F_{\boldsymbol{x}}(\boldsymbol{x}(\tau))\,\dot{\boldsymbol{x}}(\tau) = \boldsymbol{0}$ for any solution $\boldsymbol{x}(\tau)$ of (B.2.2). Hence the directions $\dot{\boldsymbol{x}}(\tau)$ of branches are in the null space of $F_{\boldsymbol{x}}(\boldsymbol{x}(\tau))$. This further implies that another branch might be bifurcating from the current one (see an example in Figure B.2.1).

When a simple singular point $\boldsymbol{x}_\mathrm{s}$ is reached during continuation the direction of the current branch $\dot{\boldsymbol{x}}_\mathrm{s}$ is already computed as shown before. A method that determines whether $\boldsymbol{x}_\mathrm{s}$ is a branching point (*i.e.* whether a bifurcating branch exists) is known. If such a branch exists the method can also compute its direction $\boldsymbol{x}'_\mathrm{s}$ exactly. This approach, however, is computationally very expensive. Instead in practical branch switching the *orthogonal direction method* is used, where $\boldsymbol{x}'_\mathrm{s}$ is approximated by a null vector of $F_{\boldsymbol{x}}(\boldsymbol{x}(\tau))$ orthogonal to $\dot{\boldsymbol{x}}_\mathrm{s}$. This may fail if the branches are far from being perpendicular, but is successful in most cases and is implemented in *AUTO-07p*.

Let $\boldsymbol{x}_\mathrm{s} = \boldsymbol{x}(\tau_\mathrm{s})$ be a simple singular point of a branch $\boldsymbol{x}(\tau)$ and:

$$G(\boldsymbol{x}\,;\,\tau) := \begin{bmatrix} F(\boldsymbol{x}) \\ (\boldsymbol{x}-\boldsymbol{x}_\mathrm{s})^T\dot{\boldsymbol{x}}_\mathrm{s} - \tau \end{bmatrix} \quad . \tag{B.2.7}$$

It can be proved that if $\boldsymbol{x}(\tau)$ and $G(\boldsymbol{x}\,;\,\tau)$ are sufficiently smooth and $\det G_{\boldsymbol{x}}(\boldsymbol{x}(\tau)\,;\,\tau)$ changes sign at $\tau = \tau_\mathrm{s}$, then $\boldsymbol{x}_\mathrm{s}$ is a branching point. Hence detection of branching points can be performed by monitoring the sign of the determinant of the Jacobian $G_{\boldsymbol{x}}$. This does not, however, ensure that all bifurcation points are found, because bifurcations can occur at singular points with a higher even dimensional nullspace where the appropriate determinant does not change sign.

B.2.4 Solving boundary value problems in *AUTO-07p*

We now briefly present the *AUTO-07p* parameter continuation software [DoeChaDer2009]. As mentioned before previous versions of *AUTO* were successfully used for computations in the elastic rod, and so the latest version has been chosen for the birod computations presented in Chapter P2.3.

Amongst other things *AUTO-07p* allows for solving Boundary Value Problems (BVP) for systems of Ordinary Differential Equations (ODE) of the form:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(t\,;\,\boldsymbol{\rho}) = f\Big(\boldsymbol{u}(t\,;\,\boldsymbol{\rho})\,;\,\boldsymbol{\rho}\Big) \quad , \qquad & \boldsymbol{u}(\cdot\,;\,\cdot)\,,\ f(\cdot\,;\,\cdot) \in \mathbb{R}^{n_\mathrm{d}} \quad , \\ & \boldsymbol{\rho} \in \mathbb{R}^{n_\mathrm{p}} \quad , \\ & t \in [0,1] \quad , \end{aligned} \tag{B.2.8}$$

subject to boundary conditions:

$$b\Big(\boldsymbol{u}(0\,;\,\boldsymbol{\rho})\,,\,\boldsymbol{u}(1\,;\,\boldsymbol{\rho})\,;\,\boldsymbol{\rho}\Big) = 0 \quad , \qquad\qquad b\,(\cdot\,,\,\cdot\,;\,\cdot) \in \mathbb{R}^{n_\mathrm{b}} \quad . \tag{B.2.9}$$

where $\boldsymbol{\rho}$ are the *system parameters*.

To solve such BVPs the *AUTO-07p* solver uses the Gauss collocation method to discretize the system [DoeKelKer1991b]. Branches of continuous piecewise polynomials approximations $\boldsymbol{u}_s(t\,;\,\boldsymbol{\rho})$ of critical solutions of the BVP are computed using the single parameter pseudo-arclength continuation described above, applied to the algebraic system that is the result of discretization.

To solve a BVP in *AUTO-07p* the user has to provide an ANSI C or Fortran source file that we will refer to as the *problem script*. The script defines functions to evaluate the right-hand side $f(\cdot\,;\,\cdot)$ of Equation (B.2.8), the boundary conditions $b\,(\cdot\,,\,\cdot\,;\,\cdot)$ of Equation (B.2.9) and a known starting point $\boldsymbol{u}_0(\cdot\,;\,\cdot)$. Note that each of those functions can depend on a number of system parameters each one of which can be used by *AUTO-07p* for parameter continuation. This way continuation can be performed both in the system itself and in boundary conditions. The *AUTO-07p* package provides shell scripts for compiling problem scripts with pre-compiled *AUTO* solver routines into a single binary.

A continuation run involves execution of such a binary with options provided in a file with the necessary *AUTO* constants. These *AUTO* constants control all variable aspects of a run such as:

- the number n_i of discretization mesh intervals and the number m_c of collocation points per interval,
- the solver accuracy,
- the step size and number of steps along a branch of solutions,
- requests of branch switching at bifurcation points,
- what solution to start from (compute a starting point using the function defined in the *AUTO* script or read a solution from a solution file),
- requests for reporting solutions with a particular value of one of the system parameters (such solutions will be called *user requested points*),
- stopping conditions including stopping at user requested points.

The continuous piecewise polynomial solutions along a branch computed during a run are stored subsequently in a solution file. For each solution the state variables as well as the system parameters of the solution are printed out. The state variables are given as values of the degree m_c polynomial pieces at $m_\mathrm{c}+1$ points in each of the n_i discretization mesh intervals. As a result the values of the state variables of the solution for any value of $t \in [0,1]$ can be recovered *e.g.* using the standard Lagrange interpolation method [PreTeuVetFla2007, sec. 3.2].

If not specified otherwise the *AUTO* constants for a run are read from a file called '`fort.2`', while the default name of the input solution file, from which a pre-computed starting point can be loaded, is '`fort.3`'. The solutions computed during a run are stored by default in a file '`fort.8`'. These files are used by the *bBDNA* software introduced in Chapter P2.2 as an input-output interface to *AUTO-07p*.

B.3 Elements of rod and birod theory

Simple continuum elastic rod models in the form of the worm-like chain have long been used in statistical mechanics of polymers [KraPor1949; BugFuj1969; Yam1976]. More sophisticated rod models have subsequently been developed in the particular context of DNA [Ben1977; Ben1979; MarSig1994; ManMadKah1996; FurManMad2000]. This chapter begins with a brief introduction of the *Cosserat rod theory* [CosCos1909], which generalizes all the mentioned models. An example similar to that of [LiMad1996] of symmetry breaking in the case of closed loops of elastic rods model will be shown. In Section P2.2.2 an analogous method of symmetry breaking has been shown to be useful in the birod model.

Finally, we move on to an extension of rod theory called the elastic birod theory, originally introduced in [MoaMad2005] and further studied in the particular context of DNA modelling in [Gra2016], where a method of extracting sequence-dependent coefficients for the continuous birod system from the discrete *cgDNA* model [Pet2012; GonPetMad2013; PetPasGonMad2014] is presented. The notations introduced here closely follow those of [Gra2016].

B.3.1 Cosserat elastic rod theory

Cosserat rod theory models long thin elastic objects, whose configuration can be described by a curve in the special Euclidean group $\mathcal{G}(s) = \big(\boldsymbol{R}(s), \mathbf{r}(s)\big) \in \mathcal{SE}(3)^{]0,L[}$ (see Section A.1.2). The translational part $\mathbf{r}(s)$ represents the centreline of the rod, while $\boldsymbol{R}(s)$ describes its cross section (see Figure B.3.1). In our treatment the parameter $s \in]0, L[$ will be the arclength in the unstressed shape of the rod. We define also the directors $\{\mathbf{d}_i(s)\}$ as the columns of the matrix representation of $\boldsymbol{R}(s)$ (see Figure B.3.1), so that:

$$\boldsymbol{R}(s) =: \begin{bmatrix} | & | & | \\ \mathbf{d}_1(s) & \mathbf{d}_2(s) & \mathbf{d}_3(s) \\ | & | & | \end{bmatrix} \quad . \tag{B.3.1}$$

As the parameter s varies within $]0, L[$ the rod configuration can be recovered from the relations:

$$\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{R}(s) = \boldsymbol{R}(s)\,[\boldsymbol{U}(s)^{\times}] \tag{B.3.2a}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{r}(s) = \boldsymbol{R}(s)\boldsymbol{V}(s) \tag{B.3.2b}$$

with the notation $[\cdot^{\times}]$ of Equation (A.1.22), for a skew matrix associated with the *rotational strain* $\boldsymbol{U}(s) \in \mathbb{R}^3$, which is sometimes referred to as the *Darboux vector*, while $\boldsymbol{V}(s) \in \mathbb{R}^3$ will be called the *translational strain*.

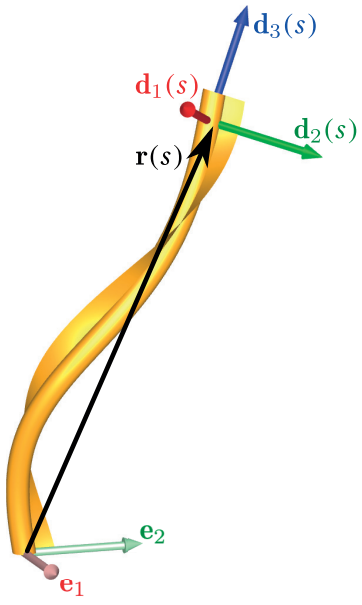


Figure B.3.1. *An example rod configuration. The first cross section orientation* $\boldsymbol{R}(0)$ *is here aligned with the laboratory frame* $\{\mathbf{e}_i\}$*. In our visualization the tube represents the centreline* $\mathbf{r}(s)$ *of the rod, while the ribbon indicates the direction* $\mathbf{d}_2$*. This choice was motivated by the fact, that in the standard Tsukuba framing for DNA this is the stiffer direction (approximately indicating the backbone of a given base pair).*

We also introduce the notation:

$$\boldsymbol{\xi}(s) := \begin{bmatrix} \boldsymbol{U}(s) \\ \boldsymbol{V}(s) \end{bmatrix} \in \mathbb{R}^6 \quad . \tag{B.3.3}$$

We will henceforth refer to the vector $\boldsymbol{\xi}(s)$ as the *rod strains*.

B.3.1.1 Balance laws

A rod is said to be in *equilibrium* if the total moment and total force acting on the cross section at each value of $s \in]0, L[$ vanish. In our treatment we will assume that external loads will only be applied at the ends of the rod. As a result the *balance laws* of the internal moment $\mathbf{m}(s)$ around $\mathbf{r}(s)$ and force $\mathbf{n}(s)$ acting across the cross section at s can be written as:

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big(\mathbf{m}(s) + \mathbf{r}(s) \times \mathbf{n}(s)\Big) = \mathbf{0} \tag{B.3.4a}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{n}(s) = \mathbf{0} \tag{B.3.4b}$$

We also introduce the triples $\boldsymbol{n}(s)$ and $\boldsymbol{m}(s)$ to mean the components of, respectively, the internal force $\mathbf{n}(s)$ and moment $\mathbf{m}(s)$ in the local body frame $\boldsymbol{R}(s)$, *i.e.*:

$$\mathbf{m}(s) = \boldsymbol{R}(s)\boldsymbol{m}(s) \tag{B.3.5a}$$

$$\mathbf{n}(s) = \boldsymbol{R}(s)\boldsymbol{n}(s) \tag{B.3.5b}$$

and the notation:

$$\boldsymbol{\zeta}(s) := \begin{bmatrix} \boldsymbol{m}(s) \\ \boldsymbol{n}(s) \end{bmatrix} \in \mathbb{R}^6 \quad . \tag{B.3.6}$$

B.3.1.2 Constitutive relations

We will now introduce the *constitutive relations* between the rod configuration and stresses, which define the material properties of the rod. In particular we will consider the case of local hyperelastic constitutive relations. Locality means that the stresses $\boldsymbol{\zeta}(s)$ only depend on the local configuration and its derivatives at s. The assumption of hyperelasticity implies that there exists an energy density function $W(\boldsymbol{\xi}(s)\,;\,s) : \mathbb{R}^6 \to \mathbb{R}$ with the following form of the constitutive relations:

$$\boldsymbol{\zeta}(s) = \partial_{\xi} W(\boldsymbol{\xi}(s)\,;\,s) \quad . \tag{B.3.7}$$

B.3.1.3 The rod variational principle

It can be shown that in the particular case of the hyperelastic constitutive relations (B.3.7) a *variational principle* can be applied in all the considered cases of boundary conditions. Denote the potential elastic energy due to the deformation of the rod by the functional:

$$E[\mathcal{G}] = \int_0^L W(\boldsymbol{\xi}(s)\,;\,s)\,\mathrm{d}s \quad . \tag{B.3.8}$$

Then, stationary configurations of the energy functional (B.3.8) that satisfy appropriate boundary conditions also realize the balance laws (B.3.4). In other words the Euler-Poincaré equations of [Gra2016] for (B.3.8) are equivalent to the balance laws (B.3.4).

B.3.1.4 Hamiltonian formulation of the rod governing equations

It can be shown that for any energy density $W(\boldsymbol{\xi}\,;\,s)$ strictly convex in $\boldsymbol{\xi}$ the rod *equilibrium conditions* (B.3.4), (B.3.7) admit a *Hamiltonian structure*. The *Legendre transform* of the energy density W, defined as:

$$H(\boldsymbol{\zeta}\,;\,s) = \max_{\boldsymbol{\zeta}\in\mathbb{R}^6}\{\boldsymbol{\xi}\cdot\boldsymbol{\zeta} - W(\boldsymbol{\xi}\,;\,s)\} \tag{B.3.9}$$

yields the rod Hamiltonian function. Henceforth we limit our considerations to linear hyperelastic rods, with a shifted quadratic form of the energy density, namely:

$$W(\boldsymbol{\xi}(s)\,;\,s) = \frac{1}{2}\left(\boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s)\right)\cdot\boldsymbol{K}(s)\left(\boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s)\right) \tag{B.3.10}$$

with s symmetric positive definite stiffness $\boldsymbol{K}(s)\in\mathbb{R}^{6\times6}$ and intrinsic shape $\widehat{\boldsymbol{\xi}}(s)\in\mathbb{R}^6$. In this case the Hamiltonian function can be written explicitly as:

$$H(\boldsymbol{\zeta}(s)\,;\,s) = \frac{1}{2}\boldsymbol{\zeta}(s)\cdot\boldsymbol{H}(s)\boldsymbol{\zeta}(s) + \widehat{\boldsymbol{\xi}}(s)\cdot\boldsymbol{\zeta}(s) \tag{B.3.11}$$

with the Hamiltonian matrix $\boldsymbol{H}(s) = \boldsymbol{K}(s)^{-1}$.

B.3.1.5 Unit quaternion representation of the cross section orientation

For computational purposes it is convenient to introduce a unit quaternion parametrization $\boldsymbol{q}(s)$ (described in more detail in Section A.1.1.2) of the orientation $\boldsymbol{R}(s)$ of the rod cross section. For $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}^T \in \mathbb{R}^4$ let the matrices $\left[\boldsymbol{x}^{\boldsymbol{B}}\right]$ and $\left[\boldsymbol{x}^{\boldsymbol{F}}\right]$ be defined as:

$$\left[\boldsymbol{x}^{\boldsymbol{B}}\right] := \begin{bmatrix} x_4 & -x_3 & x_2 \\ x_3 & x_4 & -x_1 \\ -x_2 & x_1 & x_4 \\ -x_1 & -x_2 & -x_3 \end{bmatrix} \qquad \left[\boldsymbol{x}^{\boldsymbol{F}}\right] := \begin{bmatrix} x_4 & x_3 & -x_2 \\ -x_3 & x_4 & x_1 \\ x_2 & -x_1 & x_4 \\ -x_1 & -x_2 & -x_3 \end{bmatrix} \quad . \tag{B.3.12}$$

In this setting the Lagrangian of the rod system takes the form:

$$\mathcal{L}(\boldsymbol{q},\, \boldsymbol{q}',\, \boldsymbol{V}\,;\, s) = \int_0^L \left\{ W\left(\frac{2\left[\boldsymbol{q}^{\boldsymbol{B}}\right]^T \boldsymbol{q}'}{|\boldsymbol{q}|^2},\, \boldsymbol{V}\,;\, s \right) + \nu \boldsymbol{q}\boldsymbol{q}' \right\} \mathrm{d}s \tag{B.3.13}$$

with $\boldsymbol{q}' \equiv \frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{q}$ and ν the Lagrange multiplier associated with the pointwise constraint of constant norm of $\boldsymbol{q}(s)$ (see [LiMad1996] for more detail). The variable conjugate to $\boldsymbol{q}(s)$, which we will refer to as the *impetus*, can be expressed as:

$$\boldsymbol{\mu}(s) := \partial_{\boldsymbol{q}'}\mathcal{L} = \frac{2}{|\boldsymbol{q}(s)|^2}\left[\boldsymbol{q}(s)^{\boldsymbol{B}}\right]\partial_{\boldsymbol{U}} W + \nu \boldsymbol{q}(s) \quad . \tag{B.3.14}$$

The moment $\mathbf{m}(s)$ and its coordinates in the body frame $\boldsymbol{m}(s)$ can be recovered from $\boldsymbol{\mu}(s)$ and $\boldsymbol{q}(s)$ as:

$$\mathbf{m}(s) = \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{F}}\right]^T \boldsymbol{\mu} \qquad\qquad \boldsymbol{m}(s) = \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{B}}\right]^T \boldsymbol{\mu} \quad . \tag{B.3.15}$$

The Hamiltonian form of the governing equations in this case is given by:

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{r} = \boldsymbol{R}\boldsymbol{V} \qquad\qquad \frac{\mathrm{d}}{\mathrm{d}s}\mathbf{n} = \mathbf{0} \tag{B.3.16a}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{q} = \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{B}}\right]\boldsymbol{U} \qquad\qquad \frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{\mu} = \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{B}}\right]\boldsymbol{U} - \boldsymbol{D}(\boldsymbol{q}\,;\,\mathbf{n})\,\boldsymbol{V} \tag{B.3.16b}$$

where from B.3.11 the Hamiltonian version of the constitutive relations reads:

$$\begin{bmatrix} \boldsymbol{U}(s) \\ \boldsymbol{V}(s) \end{bmatrix} = \boldsymbol{H}\begin{bmatrix} \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{B}}\right]^T \boldsymbol{\mu}(s) \\ \boldsymbol{R}(\boldsymbol{q})^T \mathbf{n}(s) \end{bmatrix} + \begin{bmatrix} \widehat{\boldsymbol{U}}(s) \\ \widehat{\boldsymbol{V}}(s) \end{bmatrix} \tag{B.3.17}$$

with $\boldsymbol{R}(\boldsymbol{q})$ given in Equation (A.1.16), and

$$\boldsymbol{D}(\boldsymbol{q}\,;\,\mathbf{n}) := \begin{bmatrix} | & | & | \\ \partial_{\boldsymbol{q}}\mathbf{d}_1(\boldsymbol{q})^T\mathbf{n} & \partial_{\boldsymbol{q}}\mathbf{d}_2(\boldsymbol{q})^T\mathbf{n} & \partial_{\boldsymbol{q}}\mathbf{d}_3(\boldsymbol{q})^T\mathbf{n} \\ | & | & | \end{bmatrix} \tag{B.3.18}$$

where the expressions for the directors $\{\mathbf{d}_i\}$ as functions of $\boldsymbol{q}$ can be recovered from Equation (A.1.16), so that each $\partial_{\boldsymbol{q}}\mathbf{d}_i$ is a 3×4 matrix.

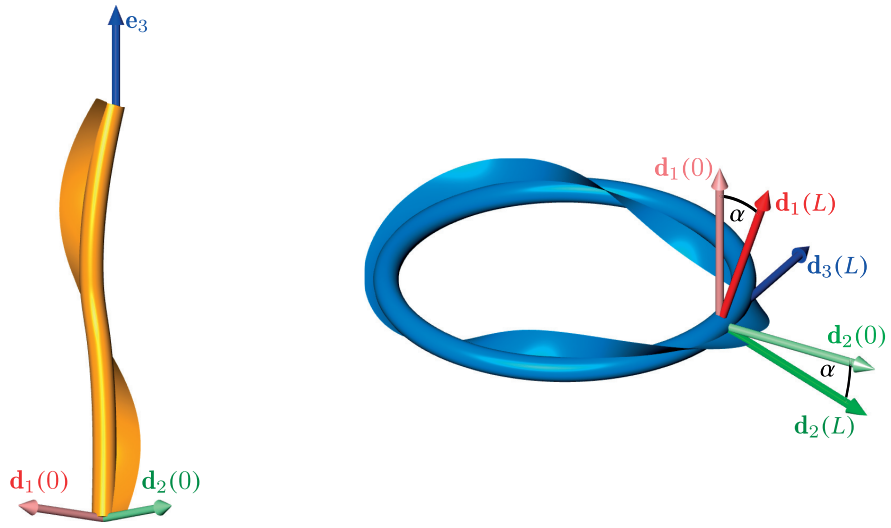
Note that a very desirable feature of the Hamiltonian formulation (in contrast to the Lagrangian description) is that all special cases of rod constitutive relations, including inextensibility and unshearability, are smooth limits where the appropriate entries of the Hamiltonian matrix $\boldsymbol{H}$ tend to zero. For example the inextensible, unshearable, straight, uniform, transversely isotropic rod (as *e.g.* in [LiMad1996]) is given by the Hamiltonian:

$$H(\boldsymbol{\zeta}(s)\,;\,s) = \frac{1}{2}\boldsymbol{\zeta}(s)\cdot\begin{bmatrix} \frac{1}{K_1} & 0 & 0 & \\ 0 & \frac{1}{K_1} & 0 & \mathbf{0}_{3\times3} \\ 0 & 0 & \frac{1}{K_3} & \\ & \mathbf{0}_{3\times3} & & \mathbf{0}_{3\times3} \end{bmatrix}\boldsymbol{\zeta}(s) + \widehat{\boldsymbol{\xi}}\cdot\boldsymbol{\zeta}(s) \tag{B.3.19}$$

with $\mathbf{0}_{3\times3} \in \mathbb{R}^{3\times3}$ a zero block, stiffnesses K_1 and K_3, and $\widehat{\boldsymbol{\xi}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T$.

B.3.2 Examples of rod boundary value problems and symmetry breaking

Here we will introduce certain Boundary Value Problems (BVP) in the rod system and discuss certain aspects of solving them through parameter continuation, as described in Chapter B.2. In particular we will consider two types of BVP that we will call the pulling and twisting problem (Section B.3.2.2, Figure B.3.2a) and the closed loop problem (Section B.3.2.3, Figure B.3.2b). Example solutions of birod versions of both of these problems will be presented in Chapter P2.3.



(a) (b)

Figure B.3.2. *Schematic pictures of the two chosen rod boundary value problems. The pulling and twisting example is shown in panel (a). The boundary conditions at* $s = 0$ *fix the orientation of the cross section* $\boldsymbol{R}(0)$ *(partially indicated in the picture by* $\mathbf{d}_1(0)$ *and* $\mathbf{d}_2(0)$*) to a fixed rotational offset from the laboratory frame* $\{\mathbf{e}_i\}$*. The boundary conditions at* $s = L$ *ask for the force and moment (in the laboratory frame) to be* $\begin{bmatrix} 0 & 0 & \mathrm{n}_3 \end{bmatrix}^T$ *and* $\begin{bmatrix} 0 & 0 & \mathrm{m}_3 \end{bmatrix}^T$*, respectively, i.e. the only applied loads are along the indicated* $\mathbf{e}_3$ *axis. The values of* n_3 *and* m_3 *can be used for parameter continuation. Panel (b) shows an example of the closed loop BVP. Here the directors* $\mathbf{d}_3(0)$ *and* $\mathbf{d}_3(L)$ *are aligned with* $\mathbf{e}_3$ *and only a rotation of* $\boldsymbol{R}(L)$ *around* $\mathbf{e}_3 = \mathbf{d}_3(0) = \mathbf{d}_3(L)$ *is allowed. The rotation angle* α *between* $\mathbf{d}_1(0)$ *and* $\mathbf{d}_1(L)$ *(or in other words between* $\mathbf{d}_2(0)$ *and* $\mathbf{d}_2(L)$*, as indicated) can be used in parameter continuation.*

B.3.2.1 Finding a starting point

First we address the very important issue in solving any BVP using parameter continuation, which is how to get a starting point. A straightforward approach is to start from a known solution of a slightly modified problem and use continuation in the formulation to get

to the problem at hand. As explained *e.g.* in [LiMad1996], in case of rod problems the "simplification" may be reduction of complexity of constitutive relations of the rod. Closed-form solutions are known *e.g.* for inextensible and unshearable, transversely isotropic and uniform rods. As a result computation can be started with some "simplified" uniform constitutive coefficients $\widehat{\boldsymbol{\xi}}_0$ and $\boldsymbol{H}_0$, where closed-form solutions are known. The desired, possibly s-dependent, constitutive coefficients can then be reached through the following continuation in the Hamiltonian coefficients:

$$(1-\sigma)\,\widehat{\boldsymbol{\xi}}_0 + \sigma\,\widehat{\boldsymbol{\xi}}(s) \tag{B.3.20a}$$

$$(1-\sigma)\,\boldsymbol{H}_0 + \sigma\,\boldsymbol{H}(s) \tag{B.3.20b}$$

by increasing the value of the homotopy parameter σ from 0 to 1.

It should also be noted that the symmetries of the simple constitutive relations may introduce non-isolation of solutions to the boundary value problem at fixed parameter values, which can make computation of solution sets and the symmetry breaking procedure delicate. Examples are considered later.

B.3.2.2 Pulling and twisting a rod

The pulling and twisting rod BVP, as presented below, is a simple variation of the classic strut BVP presented in a setting similar to ours *e.g.* in [LiMad1996; MadManPaf1997]. In this case we ask one end of the rod (at $s = 0$) to be clamped with respect to the laboratory frame, *i.e.* the centreline position $\mathbf{r}(0)$ is fixed at $\mathbf{0} \in \mathbb{R}^3$, while the orientation of the cross section is fixed at a constant rotation from the laboratory frame $\{\mathbf{e}_i\}$. The load applied to the rod will be a force and a moment acting along the axis $\mathbf{e}_3$, with transverse loads set to vanish. Figure B.3.2a presents a schematic picture of the problem.

Concretely the boundary conditions can be written as:

$$\begin{cases} \mathbf{r}(0) = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T & \text{(B.3.21a)} \\ \boldsymbol{q}(0) = \boldsymbol{q}_0 & \text{(B.3.21b)} \\ \mu_4(0) = 0 & \text{(B.3.21c)} \\ \\ \mathbf{n}(L) = \begin{bmatrix} 0 & 0 & \mathrm{n}_3 \end{bmatrix}^T & \text{(B.3.21d)} \\ \dfrac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{F}}\right]^T(L)\,\boldsymbol{\mu}(L) = \begin{bmatrix} 0 & 0 & \mathrm{m}_3 \end{bmatrix}^T & \text{(B.3.21e)} \end{cases}$$

where $\boldsymbol{q}_0$ represents a given, fixed rotation. The boundary condition (B.3.21c) removes the gauge freedom $\boldsymbol{\mu} \to \boldsymbol{\mu} + \varepsilon\boldsymbol{q}$ (see [LiMad1996]). The boundary condition (B.3.21e) fixes the value of the moment $\mathbf{m}$ in the laboratory frame (see Equation (B.3.15)) at $s = L$.

For this BVP the problem of finding a starting point is relatively simple. The important observation here is that the unstressed configuration of any rod can be computed as a solution to an appropriate initial value problem. In the particular case of piecewise constant strains, *i.e.*:

$$\widehat{\boldsymbol{\xi}}(s) \equiv \widehat{\boldsymbol{\xi}}_i \in \mathbb{R}^6 \qquad \text{for } s \in (s_i, s_{i+1}], \quad 0 = s_0 < s_i < s_N = L, \tag{B.3.22}$$

$$i \in \{1, \dots, N\} \tag{B.3.23}$$

and any general, s-dependent Hamiltonian matrix $\boldsymbol{H}(s)$ the solution has a piecewise helical centreline and is known analytically to be:

$$\begin{cases} \mathbf{r}(0) = \mathbf{0} & \text{(B.3.24a)} \\ \mathbf{r}(s) = \mathbf{r}(s_i) + \boldsymbol{R}\big(\boldsymbol{q}(s_i)\big)\, \mathbf{r}_h\big(s - s_i, \widehat{\boldsymbol{U}}_i, \widehat{\boldsymbol{V}}_i\big) & \text{for } s \in (s_i, s_{i+1}] \quad \text{(B.3.24b)} \end{cases}$$

$$\begin{cases} \boldsymbol{q}(0) = \boldsymbol{q}_0 & \text{(B.3.24c)} \\ \boldsymbol{q}(s) = \boldsymbol{q}(s_i)\, \boldsymbol{q}_h\big(s - s_i, \widehat{\boldsymbol{U}}_i\big) & \text{for } s \in (s_i, s_{i+1}] \quad \text{(B.3.24d)} \end{cases}$$

$$\mathbf{n}(s) \equiv \mathbf{0} \in \mathbb{R}^3 \tag{B.3.24e}$$

$$\boldsymbol{\mu}(s) \equiv \mathbf{0} \in \mathbb{R}^4 \quad , \tag{B.3.24f}$$

where $\boldsymbol{R}(\boldsymbol{q})$ is the rotation matrix associated with the quaternion $\boldsymbol{q}$ (see Equation A.1.16), and

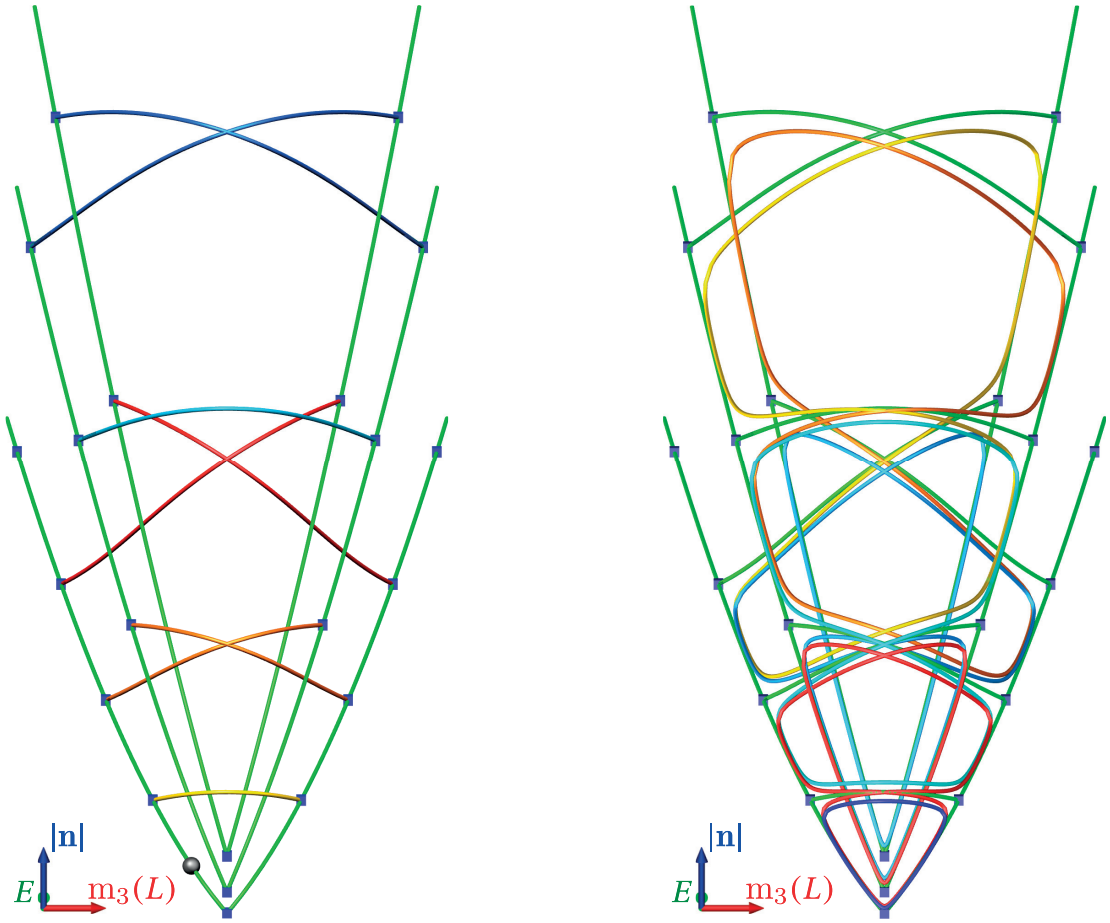
$$\mathbf{r}_h\big(s, \widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}}\big) = \begin{cases} s\widehat{\boldsymbol{V}} & \text{if } \left|\widehat{\boldsymbol{U}}\right| = 0 \\ \left(s\boldsymbol{I} + \frac{1-\cos\left(s\left|\widehat{\boldsymbol{U}}\right|\right)}{\left|\widehat{\boldsymbol{U}}\right|}\left[\boldsymbol{k}_i{}^{\times}\right] + \frac{s-\sin\left(s\left|\widehat{\boldsymbol{U}}\right|\right)}{\left|\widehat{\boldsymbol{U}}\right|}\left[\boldsymbol{k}_i{}^{\times}\right]^2\right)\widehat{\boldsymbol{V}} & \text{if } \left|\widehat{\boldsymbol{U}}\right| \neq 0 \end{cases} \tag{B.3.25a}$$

$$\boldsymbol{q}_h\big(s, \widehat{\boldsymbol{V}}\big) = \begin{cases} \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T & \text{if } \left|\widehat{\boldsymbol{U}}\right| = 0 \\ \begin{bmatrix} \sin\left(\frac{1}{2}s\left|\widehat{\boldsymbol{U}}\right|\right)\boldsymbol{k}_i & \cos\left(\frac{1}{2}s\left|\widehat{\boldsymbol{U}}\right|\right) \end{bmatrix}^T & \text{if } \left|\widehat{\boldsymbol{U}}\right| \neq 0 \end{cases} , \tag{B.3.25b}$$

with $\boldsymbol{k}_i = \frac{\widehat{\boldsymbol{U}}_i}{\left|\widehat{\boldsymbol{U}}_i\right|}$ describe the framed helix defined by $\widehat{\boldsymbol{U}}$ and $\widehat{\boldsymbol{V}}$. The additional rotation of the configuration by $\boldsymbol{q}_0$ is associated with the initial condition (B.3.21b).

This special case is pertinent to the DNA coefficients of the birod model of Section B.3.3.3, where the unstressed shape is exactly piecewise helical.

B.3.2.3 Closed loops of a rod



(a) (b)

Figure B.3.3. *Fragments of solution sets of the closed loop boundary value problems for (a) an ideal and (b) a perturbed rod. The solution sets are presented in the two dimensional projection of twisting moment* $\mathrm{m}_3(L)$ *and a composition of the energy* E *of (B.3.8) and* $|\mathbf{n}|$*. Panel (a) shows the lower energy part of the solution set for an inextensible and unshearable, uniform, transversely isotropic, intrinsically straight, untwisted rod with* $\frac{K_3}{K_1} = 0.78$ *(computed from averaged DNA data for the sequence* $S^{\lambda''}$ *of Section P2.3.2). Each point in the bifurcation diagram is a representative of a family of symmetry-related solutions. The family associated with the solution indicated by the black ball is presented in Figure B.3.4. Bifurcation points are indicated as blue boxes. Any other apparent crossing of branches is an artefact of the projection. The structure of the solution set is described in the main text. In panel (b) the entire set of panel (a) is indicated in green. When a small perturbation in the constitutive relations is made the highly connected set breaks into a number of disconnected components indicated in different colours. The specific perturbation here is the introduction a very localized bend of* $\frac{\pi}{8}$ *in the middle of the rod. Solutions in the disconnected components are isolated. This panel is analogous to Figure 6.1 of [LiMad1996].*

The second rod BVP that we will consider is the closed loop case as described in [LiMad1996] and schematically presented in Figure B.3.2b. This formulation has been shown to be useful in modelling DNA minicircles [ManMadKah1996; FurManMad2000]. In Section P2.3.2 results of applying a similar approach in the birod DNA model will be presented.

As indicated in Figure B.3.2b, in the closed loop problem we ask for the initial $\boldsymbol{R}(0)$ and final $\boldsymbol{R}(L)$ cross section orientations to share a common third director vector, *i.e.* $\mathbf{d}_3(0) = \mathbf{d}_3(L)$. The boundary conditions for this problem read:

$$\begin{cases} \mathbf{r}(0) = \mathbf{0} \in \mathbb{R}^3 & \text{(B.3.26a)} \\ \mathrm{Im}(\boldsymbol{q}(0)) = \mathbf{0} \in \mathbb{R}^3 & \text{(B.3.26b)} \\ \mu_4(0) = 0 & \text{(B.3.26c)} \\ \mathbf{r}(L) = \mathbf{0} \in \mathbb{R}^3 & \text{(B.3.26d)} \\ \boldsymbol{q}(L) = \left[0 \quad 0 \quad -\sin\left(\frac{\alpha}{2}\right) \quad -\cos\left(\frac{\alpha}{2}\right)\right]^T & \text{(B.3.26e)} \end{cases}$$

where the imaginary (or vector) part $\mathrm{Im}(\boldsymbol{q}(0))$ of a quaternion is its first three components (see Equation (A.1.5a)). Again, as in (B.3.21), the boundary condition (B.3.26c) removes the gauge symmetry of $\boldsymbol{\mu}(s)$.

Note that in this setting the loop is closed only in $\mathbf{r}(s)$ but can be seen as open in $\boldsymbol{R}(s)$, as only $\mathbf{d}_3$ is necessarily continuous. In the context of DNA modelling such solutions will be referred to as *partially closed loops*, while the ones where also $\boldsymbol{R}(0) = \boldsymbol{R}(L)$ will be called *fully closed loops*.

For this problem closed-form solutions are known only for certain equilibria with multiply covered circular centrelines in the uniform, straight, transversely isotropic case. Thus parameter continuation has to be started from a symmetric, non-isolated solution. Basic knowledge of the structure of the symmetric solution sets, presented below, makes it easier to understand the symmetry breaking technique. A full discussion of the non-isolation, as well as techniques of computing solution sets of the symmetric problem with modified boundary conditions is presented in [ManMad1999].

A symmetric solution set consists of branches of N-covered ($N \in \{1, 2, \ldots\}$) planar, circular solutions, indicated in green in the bifurcation diagram of Figure B.3.3a, that are pairwise connected with branches of non-planar solutions (indicated in different colours). The details of the connectivity are discussed *e.g.* in [LiMad1996].

As conjectured *e.g.* in [LiMad1996] the solution set in the symmetric case is completely connected (see Figure B.3.3a). This property is lost when the symmetries are broken through the introduction of more complex constitutive coefficients. That is to say that the connected solution set splits into a number of disconnected components (see Figure B.3.3b).

The families of symmetric planar solutions are related by the *register symmetry* associated with rotation of the cross sections locally around $\mathbf{d}_3(s)$, as shown in Figure B.3.4. This symmetry is related to the fact the rod is straight and transverse isotropic. In the examples of this thesis symmetry breaking is started from a chosen representative of such a symmetry family that lies on the plane $(\mathbf{d}_2(0), \mathbf{d}_3(0))$. Closed form equations of such solutions for the Hamiltonian (B.3.19) and the twisting angle α can be found *e.g.* in [LiMad1996; DicLiMad1996; ManMad1999]:

$$\mathbf{r}(s) = \frac{L}{2N\pi}\begin{bmatrix} 0 \\ \cos\left(\frac{2N\pi s}{L}\right) - 1 \\ \sin\left(\frac{2N\pi s}{L}\right) \end{bmatrix} \tag{B.3.27a}$$

$$\boldsymbol{q}(s) = \begin{bmatrix} \sin\left(\frac{N\pi s}{L}\right)\cos\left(\frac{\alpha s}{2L}\right) \\ -\sin\left(\frac{N\pi s}{L}\right)\sin\left(\frac{\alpha s}{2L}\right) \\ \cos\left(\frac{N\pi s}{L}\right)\sin\left(\frac{\alpha s}{2L}\right) \\ \cos\left(\frac{N\pi s}{L}\right)\cos\left(\frac{\alpha s}{2L}\right) \end{bmatrix} \tag{B.3.27b}$$

$$\mathbf{n}(s) = \frac{1}{L}\begin{bmatrix} 2NK_3\alpha\pi \\ 0 \\ 0 \end{bmatrix} \tag{B.3.27c}$$

$$\boldsymbol{\mu} = \begin{bmatrix} 4K_1N\pi q_4(s) + 2K_3\alpha q_2(s) \\ -4K_1N\pi q_3(s) - 2K_3\alpha q_1(s) \\ 4K_1N\pi q_2(s) + 2K_3\alpha q_4(s) \\ -4K_1N\pi q_1(s) - 2K_3\alpha q_3(s) \end{bmatrix} \tag{B.3.27d}$$

A starting point of this kind is indicated in the bifurcation diagram of Figure B.3.3a with a black ball. The exact same solution is indicated with a black ball in the bifurcation diagram of Figure B.3.4i that presents the procedure of symmetry breaking started from that point. In this particular case the perturbation in the constitutive coefficients involves introduction of a bend localized in the middle of the rod. The Hamiltonian stiffness matrix stays unchanged.

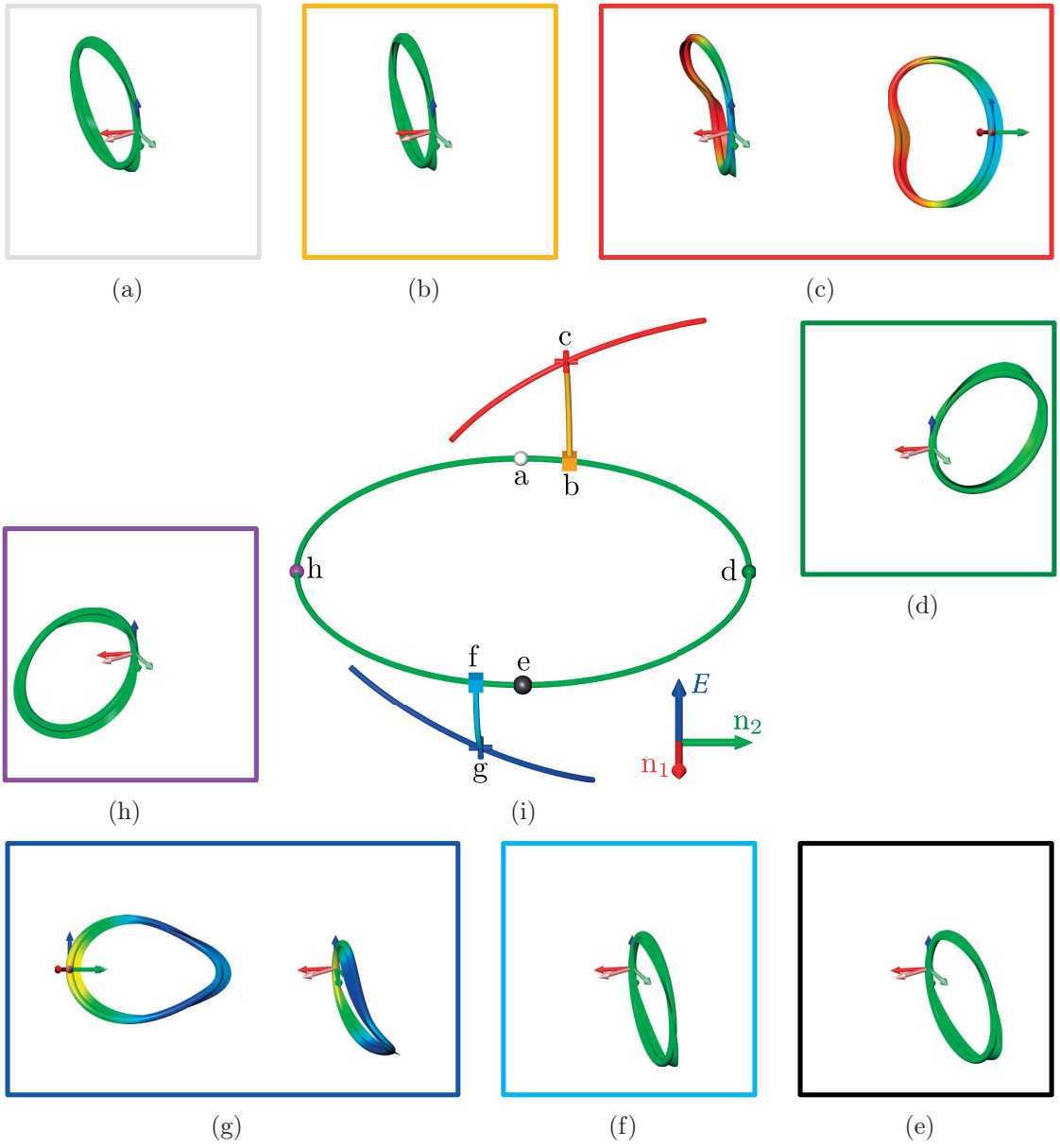


Figure B.3.4. *Symmetry breaking in the closed loop rod boundary value problem. Panel B.3.4i shows a bifurcation diagram of symmetry breaking, while the other panels show tube-ribbon representations of chosen solutions. The tubes/ribbons are coloured using the value of* $|\mathbf{m}|$ *(from blue to red). The colour coding of the frames of the panels matches the one of balls/boxes/crosses in the bifurcation diagram. The symmetry related family that is the result of the first step of symmetry breaking (see main text) is the green branch in the bifurcation diagram. The black ball on that branch represents the starting point of this first computation. This solution was constructed using Equations (B.3.27) so its centreline lies on the plane* $(\mathbf{d}_2(0), \mathbf{d}_3(0))$*, as shown in Panel (e). This is the same solution as the one represented by the black ball in Figure B.3.3a. (Continued on the following page)*

Figure B.3.4. *(Continued from the previous page) The solutions, (f), (h), (a), (b) and (d) are met in this order in the clockwise traversal of the symmetry family starting from (e). The centreline of solution (a) lies in the same plane as the one of (e), while the centrelines of (h) and (d) lie in an orthogonal plane* $(\mathbf{d}_1(0), \mathbf{d}_3(0))$. *The solutions (f) and (b), marked with boxes, are the two bifurcation points in coefficients continuation found on the green branch. The bright blue and orange branches represent the two symmetry breaking runs ended for* $\sigma = 1$ *at solutions (g) (a local minimum) and (c) (a saddle point), marked with the blue and red cross, respectively. Projections of these solutions to the plain of the intrinsic bend are also shown. The red and blue branches are fragments of the red and blue components from the bifurcation diagram of Figure B.3.3b, for the desired constitutive relations.*

The procedure begins with computation of the symmetry-related family of solutions for the starting point. The symmetry family is shown as the green circular branch in Figure B.3.4i. As illustrated schematically in Figure 5 of [FurManMad2000] in the general case close-by perturbed stationary solutions exist for only two points in the entire symmetry-related family. As a result the family can be generated in *AUTO-07p* by requesting continuation in the homotopy parameter σ of the Hamiltonian coefficients (see Equation (B.3.20)), using the boundary conditions (B.3.26). This continuation does not change the value of the homotopy parameter $\sigma = 0$, but computes the symmetry-related family and reports the two points where close-by perturbed solutions exist as bifurcation points (bright blue and orange boxes in Figure B.3.4i).

The following step involves the actual symmetry breaking at the two bifurcation points found previously. Continuation in the homotopy parameter σ is run until the value $\sigma = 1$ is reached, and so the desired constitutive relations are attained. The symmetry breaking branches are the bright blue and orange ones with ends indicated by blue and red cross in Figure B.3.4i.

Note that in the generic case if symmetry breaking is performed on the single covered branch below the first bifurcation point one of the two bifurcation points leads to a local minimum and the other one to a saddle point [FurManMad2000]. This can be seen in Figure B.3.4. For the local minimum presented in panel (g) the intrinsic bend "helps" forming the loop, while in case of the saddle point shown in panel (c) the bend is "inside out".

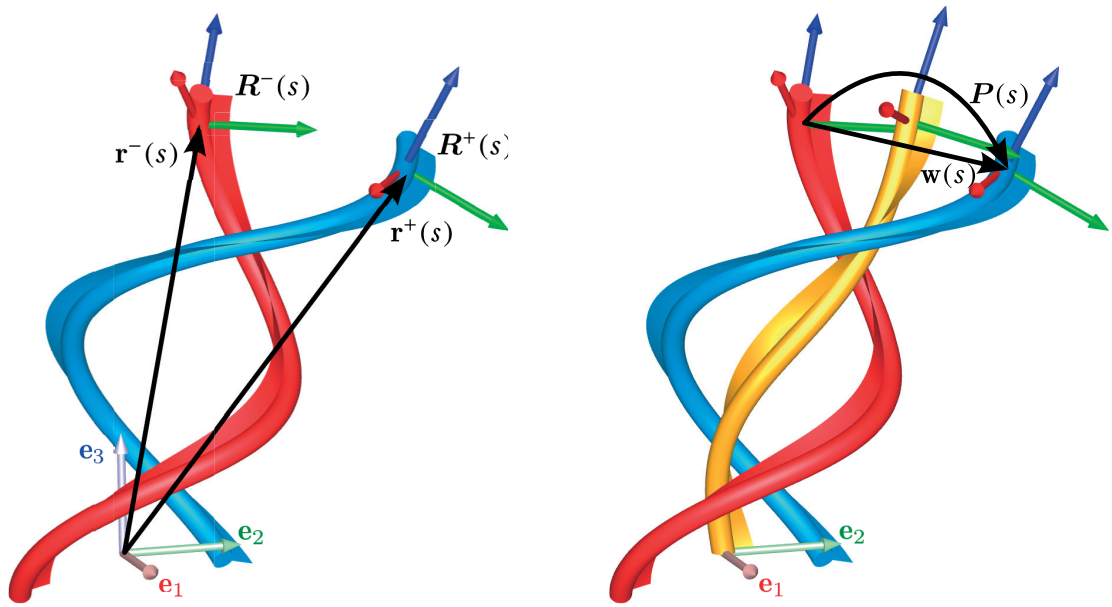
The small portions of the blue and red branches in Figure B.3.4i are fragments of the blue and red disconnected components in Figure B.3.3b shown in a different protection. They were computed by continuation in the twist angle α started at the blue and red cross.

Note, that as shown *e.g.* in [MadManPaf1997; HofManMad2003] in the symmetric problem stable solutions exist only in the part of the solution set corresponding to the blue component of the perturbed system, shown in Figure B.3.3a. For that reason to compute stable stationary points it is probably sufficient to perform symmetry breaking only in that region.

In Section P2.3.2 we will show that the technique presented above can be used to break symmetry from symmetric rods of Equation (B.3.27) directly to birods with sequence dependent DNA constitutive relations, provided that the Hamiltonian form of the equilibrium conditions is adopted.

B.3.3 Elastic birod theory

In this section we briefly outline the birod theory with the particular application to DNA modelling in mind. The model was first introduced in [MoaMad2005], but here we present the modified version described in [Gra2016] where a Hamiltonian form of the Euler-Lagrange equations is introduced. In addition [Gra2016] presents a method of computing continuous birod coefficients starting from the discrete *cgDNA* model [Pet2012; GonPetMad2013; PetPasGonMad2014] described in Chapter B.1.



(a) (b)

Figure B.3.5. *An example birod configuration. The first average frame* $\boldsymbol{R}(0)$ *here is aligned with the laboratory frame* $\{\mathbf{e}_i\}$. *Panel (a) shows the configurations of the two rods,* $\mathcal{G}^+(s) = \left(\boldsymbol{R}^+(s), \mathbf{r}^+(s)\right)$ *and* $\mathcal{G}^-(s) = \left(\boldsymbol{R}^-(s), \mathbf{r}^-(s)\right)$, *that constitute the birod. In panel (b) additionally the average rod is shown as well as the relative rotation* $\boldsymbol{P}(s)$ *and translation* $\mathbf{w}(s) = \boldsymbol{R}(s)\boldsymbol{w}(s)$.

A birod consists of two rods with local elastic interactions (see Figure B.3.5a). The configurations of the two rods, denoted $\mathcal{G}^+(s) = \left(\boldsymbol{R}^+(s), \mathbf{r}^+(s)\right)$ and $\mathcal{G}^-(s) = \left(\boldsymbol{R}^-(s), \mathbf{r}^-(s)\right)$, share a common parametrization $s \in]0, L[$. In the birod setting the configuration is given by a pair $(\mathcal{G}(s), \mathcal{P}(s)) \in \left(\mathcal{SE}(3)^{]0,L[}\right)^2$. $\mathcal{G}(s)$, called the *macrostructure*, represents the average rod, while the *microstructure* $\mathcal{P}(s) = (\boldsymbol{P}(s), \boldsymbol{w}(s)) \in \mathcal{SO}(3) \times \mathbb{R}^3$ characterizes the relative translational and rotational displacement of the two rods (see Figure B.3.5b).

In terms of the two rod configurations $\mathcal{G}^+(s)$ and $\mathcal{G}^-(s)$ the birod macrostructure can be expressed as their $\mathcal{SE}(3)$ average so that:

$$\boldsymbol{R}(s) = \boldsymbol{R}^-(s)\left(\boldsymbol{R}^-(s)^T\boldsymbol{R}^+(s)\right)^{\frac{1}{2}} \tag{B.3.28a}$$

$$\mathbf{r}(s) = \frac{1}{2}\left(\mathbf{r}^+(s) + \mathbf{r}^-(s)\right) \tag{B.3.28b}$$

where $(\,\cdot\,)^{\frac{1}{2}}$ denotes the half rotation as discussed in Section A.1.1.4. The microstructure can be written as:

$$\boldsymbol{P}(s) = \boldsymbol{R}^-(s)^T\boldsymbol{R}^+(s) \tag{B.3.29a}$$

$$\boldsymbol{w}(s) = \boldsymbol{R}(s)^T\left(\mathbf{r}^+(s) - \mathbf{r}^-(s)\right) \tag{B.3.29b}$$

so that $\boldsymbol{P}(s)$ is the relative rotation from $\boldsymbol{R}^-(s)$ to $\boldsymbol{R}^+(s)$, while $\boldsymbol{w}(s)$ is the relative translation between $\mathbf{r}^-(s)$ to $\mathbf{r}^+(s)$ expressed in the average frame $\boldsymbol{R}(s)$.

The inverse relation from $\mathcal{G}(s)$ to $\mathcal{G}^+(s)$ and $\mathcal{G}^-(s)$ reads:

$$\boldsymbol{R}^\pm(s) = \boldsymbol{R}(s)\boldsymbol{P}(s)^{\pm\frac{1}{2}} \tag{B.3.30a}$$

$$\mathbf{r}^\pm(s) = \mathbf{r}(s) \pm \frac{1}{2}\boldsymbol{R}(s)\boldsymbol{w}(s) \tag{B.3.30b}$$

The internal coordinates of the birod macrostructure will be the strains $\boldsymbol{\xi}(s)$ of the average rod defined in equation (B.3.2). For the microstructure we introduce the Cayley vector representation of the rotational part:

$$\boldsymbol{\eta}(s) := \mathrm{cay}(\boldsymbol{P}(s)) \ \in \mathbb{R}^3 \quad . \tag{B.3.31}$$

(where cay$(\cdot)$ is defined in Equation (A.1.35) or equivalently (A.1.38)). The microstructure coordinates are, then:

$$\boldsymbol{y}(s) := \begin{bmatrix} \boldsymbol{\eta}(s) \\ \boldsymbol{w}(s) \end{bmatrix} \in \mathbb{R}^6 \tag{B.3.32}$$

with the derivatives:

$$\boldsymbol{\xi}^{\mathcal{P}}_{\boldsymbol{y}}(s) := \frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{y}(s) = \begin{bmatrix} \boldsymbol{U}_{\boldsymbol{\eta}}(s) \\ \boldsymbol{V}_{\boldsymbol{w}}(s) \end{bmatrix} \in \mathbb{R}^6 \quad . \tag{B.3.33}$$

B.3.3.1 Balance laws

Similarly to the rod case an equilibrium configuration of a birod has the property that total moments and forces vanish at each $s \in]0, L[$. To explain the equilibrium equations of a birod it is useful to treat it as a system of two rods that interact with one another. In this setting the stresses exerted on each strand by its counterpart can be seen as an external field for that strand. Let $\mathbf{m}^{\pm}(s)$ be the total moment around the point $\boldsymbol{R}^{\pm}(s)$ and $\mathbf{n}^{\pm}(s)$ the total force for the strand $\pm$. The balance laws for each strand can be written as:

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big(\mathbf{m}^{\pm}(s) + \mathbf{r}^{\pm}(s) \times \mathbf{n}^{\pm}(s)\Big) = \mathbf{c}^{\pm}(s) + \mathbf{r}^{\pm}(s) \times \mathbf{f}^{\pm}(s) \tag{B.3.34a}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{n}^{\pm}(s) = \mathbf{f}^{\pm}(s) \tag{B.3.34b}$$

where $\mathbf{c}^{\pm}(s) \in \mathbb{R}^3$ is the total external moment density and $\mathbf{f}^{\pm}(s) \in \mathbb{R}^3$ the total external force density at s exerted by the other strand.

B.3.3.2 Variational formulation and constitutive relations for birods

In the context of the double rod description it can be shown [Gra2016] that variational principles exist for all cases of boundary conditions treated in this thesis for a local energy of the form:

$$E^{\mathcal{U}}[\mathcal{G}^+, \mathcal{G}^-] = \int_0^L W^{\mathcal{U}}\Big((\mathcal{G}^-)^{-1}\,\mathcal{G}^+,\, \boldsymbol{\xi}^-,\, \boldsymbol{\xi}^+\,;\, s\Big)\,\mathrm{d}s \tag{B.3.35}$$

where $\boldsymbol{\xi}^-$ and $\boldsymbol{\xi}^+$ are the strains of the two strands $\mathcal{G}^-$ and $\mathcal{G}^+$, respectively, while $(\mathcal{G}^-)^{-1}\,\mathcal{G}^+$ is the rigid body displacements between the strands.

An equivalent formulation in terms of the birod internal coordinates can be written as:

$$E[\mathcal{G}, \mathcal{P}(\boldsymbol{y})] = \int_0^L W\Big(\boldsymbol{y},\, \boldsymbol{\xi}^{\mathcal{P}}_{\boldsymbol{y}}(s),\, \boldsymbol{\xi}\,;\, s\Big)\,\mathrm{d}s \quad . \tag{B.3.36}$$

The variables conjugate to $\boldsymbol{\eta}(s)$, $\boldsymbol{w}(s)$, $\boldsymbol{U}_{\boldsymbol{\eta}}(s)$ and $\boldsymbol{V}_{\boldsymbol{w}}(s)$ will be denoted as $\boldsymbol{m}^{\mathcal{P}}(s)$, $\boldsymbol{n}^{\mathcal{P}}(s)$, $\boldsymbol{c}^{\mathcal{P}}(s)$ and $\boldsymbol{f}^{\mathcal{P}}(s)$, respectively. These conjugate variables are non-trivially related to the stresses $\mathbf{m}^{\pm}(s)$, $\mathbf{n}^{\pm}(s)$ and stress densities $\mathbf{c}^{\pm}(s)$, $\mathbf{f}^{\pm}(s)$, introduced in the previous section.

The stationary conditions for (B.3.36), equivalent to the balance laws (B.3.34), are:

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big(\mathbf{m}(s) + \mathbf{r}(s) \times \mathbf{n}(s)\Big) = \mathbf{0} \qquad \frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{m}^{\mathcal{P}}(s) = \boldsymbol{c}^{\mathcal{P}}(s) \tag{B.3.37a}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{n}(s) = \mathbf{0} \qquad \frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{n}^{\mathcal{P}}(s) = \boldsymbol{f}^{\mathcal{P}}(s) \tag{B.3.37b}$$

with $\mathbf{m}(s)$ the total moment around $\mathbf{r}(s)$, and $\mathbf{n}(s)$ the force acting on the macrostructure.

In analogy to rods we define $\boldsymbol{m}(s)$ and $\boldsymbol{n}(s)$ as the components of the average rod stresses expressed in the frame $\boldsymbol{R}(s)$ (see Equation (B.3.5)). Additionally we introduce the notation:

$$\mathcal{F}^{\mathcal{P}}(s) := \begin{bmatrix} \boldsymbol{c}^{\mathcal{P}}(s) \\ \boldsymbol{f}^{\mathcal{P}}(s) \end{bmatrix} \qquad \boldsymbol{\zeta}^{\mathcal{P}}_{\boldsymbol{y}}(s) := \begin{bmatrix} \boldsymbol{m}^{\mathcal{P}}(s) \\ \boldsymbol{n}^{\mathcal{P}}(s) \end{bmatrix} \qquad \boldsymbol{\zeta}(s) := \begin{bmatrix} \boldsymbol{m}(s) \\ \boldsymbol{n}(s) \end{bmatrix} \quad . \tag{B.3.38}$$

The birod constitutive relations can be written as:

$$\mathcal{F}^{\mathcal{P}}(s) = \partial_{\boldsymbol{y}} W \tag{B.3.39a}$$

$$\boldsymbol{\zeta}^{\mathcal{P}}_{\boldsymbol{y}}(s) = \partial_{\boldsymbol{\xi}^{\mathcal{P}}_{\boldsymbol{y}}} W \tag{B.3.39b}$$

$$\boldsymbol{\zeta}(s) = \partial_{\boldsymbol{\xi}} W \quad . \tag{B.3.39c}$$

B.3.3.3 DNA birod coefficients

In the particular case of the DNA birod model the energy will be assumed to be a shifted quadratic function of the birod internal coordinates and the Lagrangian will be assumed to be of the form:

$$\widehat{E}[\mathcal{G}, \mathcal{P}(\boldsymbol{y})] = \int_0^L \begin{bmatrix} \boldsymbol{y}(s) - \widehat{\boldsymbol{y}}(s) \\ \boldsymbol{\xi}^{\mathcal{P}}_{\boldsymbol{y}}(s) - \widehat{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \\ \boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s) \end{bmatrix} \cdot \boldsymbol{K}(s) \begin{bmatrix} \boldsymbol{y}(s) - \widehat{\boldsymbol{y}}(s) \\ \boldsymbol{\xi}^{\mathcal{P}}_{\boldsymbol{y}}(s) - \widehat{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \\ \boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s) \end{bmatrix} \mathrm{d}s \tag{B.3.40a}$$

$$+ \frac{1}{2}\Big(\boldsymbol{y}(0) - \widehat{\boldsymbol{y}}(0)\Big) \cdot \boldsymbol{K}_0\Big(\boldsymbol{y}(0) - \widehat{\boldsymbol{y}}(0)\Big) \tag{B.3.40b}$$

$$+ \frac{1}{2}\Big(\boldsymbol{y}(L) - \widehat{\boldsymbol{y}}(L)\Big) \cdot \boldsymbol{K}_L\Big(\boldsymbol{y}(L) - \widehat{\boldsymbol{y}}(L)\Big) \quad , \tag{B.3.40c}$$

where $\boldsymbol{K}(s) \in \mathbb{R}^{18}$ is the interior stiffness matrix, $\boldsymbol{K}_0, \boldsymbol{K}_L \in \mathbb{R}^6$ will be called boundary intra stiffness matrices, and $\widehat{\boldsymbol{\xi}}(s), \widehat{\boldsymbol{y}}(s), \widehat{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \in \mathbb{R}^6$ are the internal variables of the ground state configuration. All the mentioned values will be collectively referred to as *birod coefficients*.

For the energy of the form (B.3.40) birod coefficients for a given DNA oligomer can be recovered from a parameter set of the *cgDNA* model [Pet2012; GonPetMad2013; PetPasGonMad2014]. The *cgDNA* parameters were chosen as the starting point for parametrizing the birod model as they have been shown to reproduce well ground state statistics of molecular dynamics simulations at short length scales.

A method of extracting birod coefficients from the *cgDNA* model was presented in [Gra2016, sec. 4.2]. It was shown that the resulting continuum birod energy is consistent with the discrete energy of *cgDNA* up to quadratic terms. Here we present a brief outline of the method.

Let $\widehat{\boldsymbol{w}} = (\widehat{\boldsymbol{y}}_1, \widehat{\boldsymbol{z}}_1, \widehat{\boldsymbol{y}}_2, \widehat{\boldsymbol{z}}_2, \ldots, \widehat{\boldsymbol{y}}_N)$ be the *cgDNA* ground state configuration vector of an oligomer with the sequence $\mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_N$ (see Section B.1.2). Let $\widehat{\mathcal{D}}_1, \ldots, \widehat{\mathcal{D}}_N$ be the base pair frames reconstructed from $\widehat{\boldsymbol{w}}$ using the procedure of Section B.1.3. A piecewise helical interpolation $\widehat{\mathcal{D}}^{(N)}$ of $\widehat{\mathcal{D}}$ can be defined by introducing a local helical interpolation between $\widehat{\mathcal{D}}_n$ and $\widehat{\mathcal{D}}_{n+1}$ for each junction n. More precisely, for each junction n define:

$$h_n^{(N)} \begin{bmatrix} \left[\widehat{\boldsymbol{U}}_n(s)^{\times}\right] & \widehat{\boldsymbol{V}}_n(s) \\ \boldsymbol{0} & 0 \end{bmatrix} := \ln\left\{\widehat{\mathcal{D}}_n^{-1}\widehat{\mathcal{D}}_{n+1}\right\} \tag{B.3.41}$$

with $\left|\widehat{\boldsymbol{V}}_n(s)\right| = 1$, where $\ln\{\cdot\}$ is a matrix logarithm of the 4×4 matrix of homogeneous coordinates (see Section A.1.2) of the rigid body motion that is its argument, and $h_n^{(N)}$ is the length of the local helix of the junction.

Now the nodes $s_n^{(N)}$ of the parametrization, the reference length L and the helical interpolation $\widehat{\mathcal{D}}^{(N)}$ itself can be defined as:

$$s_n^{(N)} = \sum_{k=1}^{N-1} h_k^{(N)} \quad , \tag{B.3.42a}$$

$$L = s_N^{(N)} \quad , \tag{B.3.42b}$$

$$\begin{cases} \widehat{\mathcal{D}}^{(N)}\left(s_n^{(N)}\right) = \widehat{\mathcal{D}}_n \quad , & \text{(B.3.42c)} \\ \dfrac{\mathrm{d}}{\mathrm{d}s}\widehat{\mathcal{D}}^{(N)}(s) = \widehat{\mathcal{D}}^{(N)}(s) \begin{bmatrix} \left[\widehat{\boldsymbol{U}}_n(s)^{\times}\right] & \widehat{\boldsymbol{V}}_n(s) \\ \boldsymbol{0} & 0 \end{bmatrix} \quad \text{for } s \in \left]s_n^{(N)}, s_{n+1}^{(N)}\right[\quad , & \text{(B.3.42d)} \end{cases}$$

In case of the microstructure internal coordinates, piecewise linear interpolation is used, therefore for each junction n we have:

$$\begin{bmatrix} \widehat{\boldsymbol{y}}(s) \\ \widehat{\boldsymbol{\xi}}_{\boldsymbol{y}}^{p}(s) \end{bmatrix} := \begin{bmatrix} \left(1 - \frac{s-s_n^{(N)}}{h_n^{(N)}}\right)\boldsymbol{I}_6 & \frac{s-s_n^{(N)}}{h_n^{(N)}}\boldsymbol{I}_6 \\ -\frac{1}{h_n^{(N)}}\boldsymbol{I}_6 & \frac{1}{h_n^{(N)}}\boldsymbol{I}_6 \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{y}}_n \\ \widehat{\boldsymbol{y}}_{n+1} \end{bmatrix} \tag{B.3.43}$$

where $\boldsymbol{I}_6 \in \mathbb{R}^{6\times 6}$ is the identity matrix.

At this point we merely state the coefficient fitting method without going into details of the justification, which can be found in [Gra2016, sec. 4.2]. For vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ define:

$$\mathbb{P}_1(\boldsymbol{u}) = \frac{1}{1 + \left(\frac{|\boldsymbol{u}|}{2}\right)^2}\left(\boldsymbol{I}_3 + \frac{1}{2}\left[\boldsymbol{u}^{\times}\right]\right) \tag{B.3.44}$$

$$\mathbb{P}_2(\boldsymbol{u}) = \left(\boldsymbol{I}_3 + \boldsymbol{Q}(\boldsymbol{u})^{\frac{1}{2}}\right)^{-1}\mathbb{P}_1(\boldsymbol{u}) \tag{B.3.45}$$

$$\mathbb{L}_{\boldsymbol{u},\boldsymbol{v}}^{(1)} = \begin{bmatrix} \mathbb{P}_1(\boldsymbol{u}) & \boldsymbol{0}_3 \\ \boldsymbol{Q}(\boldsymbol{u})^{\frac{1}{2}}\left[\boldsymbol{v}^{\times}\right]\mathbb{P}_2(\boldsymbol{u}) & \boldsymbol{Q}(\boldsymbol{u})^{\frac{1}{2}} \end{bmatrix} \quad , \tag{B.3.46}$$

where $\boldsymbol{I}_3, \boldsymbol{0}_3 \in \mathbb{R}^{3\times3}$ are the identity and zero matrices and $\boldsymbol{Q}(\boldsymbol{u}) = \mathrm{cay}(\boldsymbol{u})$ (of Equation (A.1.35) or equivalently (A.1.38)).

Let also:

$$\mathrm{Ad}_{\mathcal{Q}}^{-1} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{0}_3 \\ [\boldsymbol{r}^\times]\,\boldsymbol{R} & \boldsymbol{R} \end{bmatrix} \tag{B.3.47}$$

for $\mathcal{Q} = (\boldsymbol{r}, \boldsymbol{R}) \in \mathcal{SE}(3)$ and

$$\boldsymbol{L}_{\boldsymbol{u},\boldsymbol{v}}(s) = \begin{bmatrix} \left(1 - \frac{s-s_n^{(N)}}{h_n^{(N)}}\right)\boldsymbol{I}_6 & \boldsymbol{0}_6 & \frac{s-s_n^{(N)}}{h_n^{(N)}}\boldsymbol{I}_6 \\ -\frac{1}{h_n^{(N)}}\boldsymbol{I}_6 & \boldsymbol{0}_6 & \frac{1}{h_n^{(N)}}\boldsymbol{I}_6 \\ \boldsymbol{0}_6 & \mathrm{Ad}^{-1}_{\widehat{\mathcal{D}}_n^{-1}\,\widehat{\mathcal{D}}^{(N)}(s)}\,\mathbb{L}^{(1)}_{\boldsymbol{u},\boldsymbol{v}} & \boldsymbol{0}_6 \end{bmatrix} . \tag{B.3.48}$$

With all of the above, the sequence dependent DNA birod coefficients can be computed as:

$$\begin{cases} \begin{bmatrix} \widehat{\boldsymbol{y}}(s) \\ \widehat{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \end{bmatrix} := \begin{bmatrix} \left(1 - \frac{s-s_n^{(N)}}{h_n^{(N)}}\right)\boldsymbol{I}_6 & \frac{s-s_n^{(N)}}{h_n^{(N)}}\boldsymbol{I}_6 \\ -\frac{1}{h_n^{(N)}}\boldsymbol{I}_6 & \frac{1}{h_n^{(N)}}\boldsymbol{I}_6 \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{y}}_n \\ \widehat{\boldsymbol{y}}_{n+1} \end{bmatrix} & \text{(B.3.49a)} \\ \widehat{\boldsymbol{\xi}}(s) := \begin{bmatrix} \widehat{\boldsymbol{U}}_n(s) \\ \widehat{\boldsymbol{V}}_n(s) \end{bmatrix} \qquad \text{for } s \in \left]s_n^{(N)}, s_{n+1}^{(N)}\right[& \text{(B.3.49b)} \\ \boldsymbol{K}(s) := \dfrac{1}{h_n^{(N)}} \boldsymbol{L}_{\widehat{\boldsymbol{\theta}}_n,\widehat{\boldsymbol{\rho}}_n}(s)^{-T} \\ \qquad \left(\boldsymbol{K}^{\mathtt{X}_n\mathtt{X}_{n+1}} + \dfrac{1}{2}\begin{bmatrix} \boldsymbol{K}^{\mathtt{X}_n} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{K}^{\mathtt{X}_{n+1}} \end{bmatrix}\right) \boldsymbol{L}_{\widehat{\boldsymbol{\theta}}_n,\widehat{\boldsymbol{\rho}}_n}(s)^{-1} & \text{(B.3.49c)} \end{cases}$$

$$\boldsymbol{K}_0 := \frac{1}{2}\boldsymbol{K}^{\mathtt{X}_1} \tag{B.3.49d}$$

$$\boldsymbol{K}_L := \frac{1}{2}\boldsymbol{K}^{\mathtt{X}_N} \tag{B.3.49e}$$

with the *cgDNA* parameter set blocks $\boldsymbol{K}^{\mathtt{XY}} \in \mathbb{R}^{18\times18}$ and $\boldsymbol{K}^{\mathtt{X}} \in \mathbb{R}^{6\times6}$, with $\widehat{\boldsymbol{U}}_n(s)$ and $\widehat{\boldsymbol{V}}_n(s)$ of Equation (B.3.41), where $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\rho}}_n$ are the *cgDNA* rotational and translational inter base pair coordinates of the junction, respectively (see Section B.1.2). The expressions of Equation (B.3.43) are repeated for completeness.

B.3.3.4 The birod Hamiltonian formulation in unit quaternions

The birod system, just like the rod one, exhibits a Hamiltonian structure. The Hamiltonian function in the internal coordinates, for the general hyperelastic case is [Gra2016, sec. 3.2.8]:

$$H\left(\boldsymbol{y},\, \boldsymbol{\zeta}_{\boldsymbol{y}}^{\mathcal{P}},\, \boldsymbol{\zeta}\,;\, s\right) = \max_{\boldsymbol{\zeta}_{\boldsymbol{y}}^{\mathcal{P}},\boldsymbol{\zeta}\in\mathbb{R}^6} \left\{\boldsymbol{\xi}\cdot\boldsymbol{\zeta} + \boldsymbol{\xi}^{\mathcal{P}}\cdot\boldsymbol{\zeta}_{\boldsymbol{y}}^{\mathcal{P}} - W\left(\boldsymbol{y},\, \boldsymbol{\xi}^{\mathcal{P}},\, \boldsymbol{\xi}\,;\, s\right)\right\} \tag{B.3.50}$$

Here for the first time we introduce the birod Hamiltonian formulation with the unit quaternion representation of the average cross section in complete analogy to what has been previously done for elastic rods [LiMad1996; DicLiMad1996; ManMadKah1996] (see Section B.3.1.5).

In the birod DNA model formulation used in this thesis an alternative, equivalent [Gra2016, app. A.4] formulation of the Lagrangian (B.3.40) will be used. It is obtained using the change of variable:

$$\widetilde{\boldsymbol{y}}(s) = \begin{bmatrix}\widetilde{\boldsymbol{\eta}}(s)\\ \widetilde{\boldsymbol{w}}(s)\end{bmatrix} := \boldsymbol{y}(s) - \widehat{\boldsymbol{y}}(s) \qquad \widetilde{\mathcal{F}^{\mathcal{P}}}(s) = \begin{bmatrix}\widetilde{\boldsymbol{c}}^{\mathcal{P}}(s)\\ \widetilde{\boldsymbol{f}}^{\mathcal{P}}(s)\end{bmatrix} \tag{B.3.51a}$$

$$\widetilde{\boldsymbol{\xi}}_{\boldsymbol{y}}^{\mathcal{P}}(s) = \begin{bmatrix}\widetilde{\boldsymbol{U}}_{\boldsymbol{\eta}}\\ \widetilde{\boldsymbol{V}}_{\boldsymbol{w}}\end{bmatrix} := \boldsymbol{\xi}_{\boldsymbol{y}}^{\mathcal{P}}(s) - \widehat{\boldsymbol{\xi}}_{\boldsymbol{y}}^{\mathcal{P}}(s) \qquad \widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(s) = \begin{bmatrix}\widetilde{\boldsymbol{m}}^{\mathcal{P}}(s)\\ \widetilde{\boldsymbol{n}}^{\mathcal{P}}(s)\end{bmatrix}\quad , \tag{B.3.51b}$$

that eliminates the shifts in the microstructure. The Lagrangian then reads:

$$\begin{aligned}\widehat{E}[\mathcal{G}, \mathcal{P}(\boldsymbol{y})] = &\int_0^L \begin{bmatrix}\widetilde{\boldsymbol{y}}(s)\\ \widetilde{\boldsymbol{\xi}}_{\boldsymbol{y}}^{\mathcal{P}}(s)\\ \boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s)\end{bmatrix}\cdot \boldsymbol{K}(s)\begin{bmatrix}\widetilde{\boldsymbol{y}}(s)\\ \widetilde{\boldsymbol{\xi}}_{\boldsymbol{y}}^{\mathcal{P}}(s)\\ \boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s)\end{bmatrix}\mathrm{d}s\\ &+\frac{1}{2}\widetilde{\boldsymbol{y}}(0)\cdot\boldsymbol{K}_0\widetilde{\boldsymbol{y}}(0) \qquad +\frac{1}{2}\widetilde{\boldsymbol{y}}(L)\cdot\boldsymbol{K}_L\widetilde{\boldsymbol{y}}(L)\quad .\end{aligned} \tag{B.3.52}$$

In this particular case the Hamiltonian takes the explicit form:

$$H\left(\widetilde{\boldsymbol{y}},\, \widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}},\, \boldsymbol{\zeta}\,;\, s\right) = \frac{1}{2}\begin{bmatrix}\widetilde{\boldsymbol{y}}(s)\\ \widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(s)\\ \boldsymbol{\zeta}(s)\end{bmatrix}\cdot\boldsymbol{H}(s)\begin{bmatrix}\widetilde{\boldsymbol{y}}(s)\\ \widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(s)\\ \boldsymbol{\zeta}(s)\end{bmatrix} + \widehat{\boldsymbol{\xi}}(s)\cdot\boldsymbol{\zeta}(s) \tag{B.3.53}$$

with the Hamiltonian matrix $\boldsymbol{H}(s) \in \mathbb{R}^{18\times 18}$:

$$\boldsymbol{H} = \begin{bmatrix}\boldsymbol{K}_2\boldsymbol{K}_3^{-1}\boldsymbol{K}_2^T - \boldsymbol{K}_1 & -\boldsymbol{K}_2\boldsymbol{K}_3^{-1}\\ -\boldsymbol{K}_3^{-1}\boldsymbol{K}_2^T & \boldsymbol{K}_3^{-1}\end{bmatrix} \tag{B.3.54}$$

defined as a function of the stiffness matrix $\boldsymbol{K}(s)$ partitioned as:

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{K}_2 \\ \boldsymbol{K}_2^T & \boldsymbol{K}_3 \end{bmatrix} \tag{B.3.55}$$

with $\boldsymbol{K}_1 \in \mathbb{R}^{6\times6}$, $\boldsymbol{K}_2 \in \mathbb{R}^{6\times12}$ and $\boldsymbol{K}_3 \in \mathbb{R}^{12\times12}$.

Note that by the Sylvester's law of inertia the signature (the number of positive and the number of negative eigenvalues) of a matrix is invariant under change of base. As a result the following change of base of the Hamiltonian matrix:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{I}_6 & \boldsymbol{K}_2 \\ \boldsymbol{0} & \boldsymbol{I}_{12} \end{bmatrix} \begin{bmatrix} \boldsymbol{K}_2\boldsymbol{K}_3^{-1}\boldsymbol{K}_2^T - \boldsymbol{K}_1 & -\boldsymbol{K}_2\boldsymbol{K}_3^{-1} \\ -\boldsymbol{K}_3^{-1}\boldsymbol{K}_2^T & \boldsymbol{K}_3^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_6 & \boldsymbol{0} \\ \boldsymbol{K}_2^T & \boldsymbol{I}_{12} \end{bmatrix} \tag{B.3.56a}$$

$$= \begin{bmatrix} -\boldsymbol{K}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_3^{-1} \end{bmatrix} \tag{B.3.56b}$$

(with $\boldsymbol{I}_6$ and $\boldsymbol{I}_{12}$ the identity matrices of the indicated dimensions) shows that $\boldsymbol{H}$ has six negative and twelve positive eigenvalues provided that $\boldsymbol{K}$ is positive definite.

Exactly as described in Section B.3.1.5 in the context of rods and using the same notation we now introduce the unit quaternion parametrization of the average rod cross section orientation $\boldsymbol{R}(s)$. The birod Hamiltonian system for this choice can be written as (with explicit s dependence skipped for clarity):

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbf{r} = \boldsymbol{R}\boldsymbol{V} \qquad \frac{\mathrm{d}}{\mathrm{d}s}\mathbf{n} = \boldsymbol{0} \tag{B.3.57a}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{q} = \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{B}}\right]\boldsymbol{U} \qquad \frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{\mu} = \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{B}}\right]\boldsymbol{U} - \boldsymbol{D}(\boldsymbol{q}\,;\,\mathbf{n})\,\boldsymbol{V} \tag{B.3.57b}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\widetilde{\boldsymbol{w}} = \widetilde{\boldsymbol{V}}_{\boldsymbol{w}} \qquad \frac{\mathrm{d}}{\mathrm{d}s}\widetilde{\boldsymbol{n}}^{\mathcal{P}} = \widetilde{\boldsymbol{f}}^{\mathcal{P}} \tag{B.3.57c}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\widetilde{\boldsymbol{\eta}} = \widetilde{\boldsymbol{U}}_{\boldsymbol{\eta}} \qquad \frac{\mathrm{d}}{\mathrm{d}s}\widetilde{\boldsymbol{m}}^{\mathcal{P}} = \widetilde{\boldsymbol{c}}^{\mathcal{P}} \tag{B.3.57d}$$

with the Hamiltonian constitutive relations:

$$\begin{bmatrix} -\widetilde{\boldsymbol{c}}^{\mathcal{P}}(s) \\ -\widetilde{\boldsymbol{f}}^{\mathcal{P}}(s) \\ \widetilde{\boldsymbol{U}}_{\boldsymbol{\eta}}(s) \\ \widetilde{\boldsymbol{V}}_{\boldsymbol{w}}(s) \\ \boldsymbol{U}(s) \\ \boldsymbol{V}(s) \end{bmatrix} = \boldsymbol{H}(s) \begin{bmatrix} \widetilde{\boldsymbol{\eta}}(s) \\ \widetilde{\boldsymbol{w}}(s) \\ \widetilde{\boldsymbol{m}}^{\mathcal{P}}(s) \\ \widetilde{\boldsymbol{n}}^{\mathcal{P}}(s) \\ \frac{1}{2}\left[\boldsymbol{q}^{\boldsymbol{B}}\right]^T\boldsymbol{\mu}(s) \\ \boldsymbol{R}(\boldsymbol{q})^T\mathbf{n}(s) \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \widehat{\boldsymbol{U}}(s) \\ \widehat{\boldsymbol{V}}(s) \end{bmatrix} . \tag{B.3.58}$$

The (sequence-dependent) DNA coefficients of this system are the Hamiltonian matrix $\boldsymbol{H}(s)$ computed from (B.3.49c) using Equation (B.3.54) and the strains $\widehat{\boldsymbol{\xi}}(s) = \left[\widehat{\boldsymbol{U}}(s) \quad \widehat{\boldsymbol{V}}(s)\right]^T$ of the intrinsic shape defined in Equation (B.3.49b). The boundary stiffness matrices of Equations (B.3.49d) and (B.3.49e) will be used in the definition of free end microstructure boundary conditions (see Equation (B.3.61)). The intrinsic internal microstructure coordinates $\widehat{\boldsymbol{y}}(s)$ (see Equation (B.3.49a)) will only be used for computed solutions to reconstruct the 3D double rod configuration for visualization. The reconstruction can be done using Equation (B.3.51a), the inverse of (B.3.31) (defined in (A.1.40) or equivalently in (A.1.36)) and Equation (B.3.30).

Note that the macrostructure and microstructure can be decoupled by setting to zero the sub-blocks $\boldsymbol{H}_{(1,12),(13,18)}$ and $\boldsymbol{H}_{(13,18),(1,12)}$ (that correspond to $-\boldsymbol{K}_2\boldsymbol{K}_3^{-1}$ and its transpose in Equation (B.3.54)). The microstructure variables can also be frozen by asking for the leading 12×12 diagonal sub-block $\boldsymbol{H}_{(1,12),(1,12)}$ to vanish. This way rod computations can be performed within the birod system $\boldsymbol{H}_{(13,18),(1,12)}$. In particular the Hamiltonian:

$$H\left(\widetilde{\boldsymbol{y}},\, \widetilde{\boldsymbol{\zeta}}^{\mathcal{P}}_{\boldsymbol{y}},\, \boldsymbol{\zeta}\,;\, s\right) = \frac{1}{2}\begin{bmatrix} \widetilde{\boldsymbol{y}}(s) \\ \widetilde{\boldsymbol{\zeta}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \\ \boldsymbol{\zeta}(s) \end{bmatrix} \cdot \begin{bmatrix} \mathbf{0}_{12\times12} & \mathbf{0}_{12\times6} & \\ & \left(\boldsymbol{K}^d\right)^{-1} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{6\times12} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{y}}(s) \\ \widetilde{\boldsymbol{\zeta}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \\ \boldsymbol{\zeta}(s) \end{bmatrix} + \widehat{\boldsymbol{\xi}} \cdot \boldsymbol{\zeta}(s) \tag{B.3.59}$$

(with $\mathbf{0}_{12\times12}, \mathbf{0}_{6\times12}, \mathbf{0}_{12\times6}, \mathbf{0}_{3\times3}$ zero blocks of indicated dimensions, a diagonal stiffness block $\boldsymbol{K}^d = \begin{bmatrix} K_1 & 0 & 0 \\ 0 & K_1 & 0 \\ 0 & 0 & K_3 \end{bmatrix}$ and $\widehat{\boldsymbol{\xi}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T$) describes an inextensible and unshearable, straight, uniform, transversely isotropic rod. In fact the rod examples of Section B.3.2 were computed using *AUTO-07p* with the same problem script as all of the birod DNA examples of Chapter P2.3, but with different Hamiltonian coefficients. The rod computations agree with previously computed solutions.

B.3.4 Example birod boundary value problems

The Boundary Value Problems (BVP) that will be considered in what follows will simply be natural extensions of the rod BVP presented in Section B.3.2. This means that the boundary conditions on the macrostructure will be either the pulling and twisting conditions of Equation (B.3.21) or the closed loop conditions of Equation (B.3.26). In this section we will address the question of the microstructure part of the system.

B.3.4.1 Starting point

In all cases the birod starting point will consist of the macrostructure part defined for each BVP of interest in Section B.3.4 and the shifted microstructure part will be set to zero *i.e.*:

$$\widetilde{\boldsymbol{y}}(s) \equiv \mathbf{0} \in \mathbb{R}^6 \tag{B.3.60a}$$

$$\widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(s) \equiv \mathbf{0} \in \mathbb{R}^6 \tag{B.3.60b}$$

In the pulling and twisting BVP this definition is possible because the starting point will always be chosen as the unstressed shape. The closure condition introduces non-zero stress in the macrostructure for a closed loop BVP but the starting point will always be an ideal rod where the macrostructure is decoupled from microstructure. The microstructure can, then, be unstressed.

B.3.4.2 Microstructure boundary conditions

We will consider two types of microstructure boundary conditions. The first will be referred to as free-end microstructure boundary conditions, where the microstructure stresses vanish at the ends. This can be achieved by setting:

$$\begin{cases} \widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(0) = \boldsymbol{K}_0 \, \widetilde{\boldsymbol{y}}(0) & \text{(B.3.61a)} \\ -\widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(L) = \boldsymbol{K}_L \, \widetilde{\boldsymbol{y}}(L) & \text{(B.3.61b)} \end{cases}$$

The other set of microstructure boundary conditions, used in Section P2.3.2, will be periodic conditions, *i.e.*:

$$\begin{cases} \widetilde{\boldsymbol{y}}(0) = \widetilde{\boldsymbol{y}}(L) & \text{(B.3.62a)} \\ \widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(0) = \widetilde{\boldsymbol{\zeta}}_{\boldsymbol{y}}^{\mathcal{P}}(L) \quad . & \text{(B.3.62b)} \end{cases}$$

Part 1

Discrete DNA modelling

P1.1 Maximum entropy fitting for covariance matrices with overlapping squares sparsity

In this chapter we present a maximum entropy fitting procedure that is a core element of an improved method of parameter fitting [GonPetPas] for the *cgDNA* model [Pet2012; GonPetMad2013; PetPasGonMad2014] introduced in Section B.1.6. More precisely, consider a covariance matrix $\boldsymbol{C}$ prescribed only within a sparsity pattern of overlapping diagonal squares. The procedure completes $\boldsymbol{C}$ to a dense covariance $\widetilde{\boldsymbol{C}}$ in such a way that the inverse $\widetilde{\boldsymbol{C}}^{-1}$ vanishes outside the pattern. For any covariance $\boldsymbol{C}$ and any overlapping squares sparsity pattern such a completion $\widetilde{\boldsymbol{C}}$ exists and is unique [Dem1972]. $\widetilde{\boldsymbol{C}}$ is also the covariance of the Gaussian model with maximum entropy amongst those whose covariances are equal to $\boldsymbol{C}$ inside the pattern [Dem1972].

Our main result is a direct way of computing the inverse $\widetilde{\boldsymbol{C}}^{-1}$. This involves local inversion of appropriate diagonal sub-blocks of $\boldsymbol{C}$ and is of particular importance in the context of the *cgDNA* model. In the second step of parameter fitting (see Section B.1.6) the simple maximum (*absolute*) entropy fit is meant to replace the numerical optimization procedure necessary in case of the maximum *relative* entropy fit used originally in [Pet2012; GonPetMad2013; PetPasGonMad2014]. As shown in [GonPetPas] the maximum (absolute) entropy is a more natural choice for the *cgDNA* parameter fitting. In Chapter P1.4, the resultant parameter set *cgDNAparamset2* (used throughout this thesis) is also shown to allow for better predictive capabilities of persistence lengths within the *cgDNA* model than the initial *cgDNAparamset1*.

It should be pointed out that after being stated and proved the presented result for the inverse covariance $\widetilde{\boldsymbol{C}}^{-1}$ was found to be a particular case of prior work of [SpeKii1986], [Lau1996, sec. 5.3] and [JohLun1998] that is stated in rather different languages. The proof presented here (similar to that of [Lau1996, sec. 5.3]) is expressed in terms of recursive Schur factorization and (unlike the prior works) also provides a method of constructing $\widetilde{\boldsymbol{C}}$ itself.

P1.1.1 Notation and definitions

Throughout this chapter we use the notation $\boldsymbol{I}$, $\boldsymbol{0}$ for, respectively, identity and null matrices with size set by context, while indices i, $i+1$ *etc.* run over the implied range for the expression at hand between all entries of the associated matrix. All square matrices and matrix partitions are symmetric unless stated otherwise.

Let $\mathcal{N}$ denote an index set, *i.e.* a sub-set of all pairs of indices of an $n \times n$ matrix. An index set with the property that $(i,j) \in \mathcal{N} \iff (j,i) \in \mathcal{N}$ will be called symmetric. We will also denote by $\mathcal{N}^c$, the set of indices complementary to $\mathcal{N}$. For $1 \le p \le n$ denote by $\mathcal{N}_p$ and $\mathcal{N}_p^c$ the corresponding subsets of $\mathcal{N}_p$ and $\mathcal{N}_p^c$ limited to indices of at most p. We also introduce the notation $[[\boldsymbol{A}]]_{\mathcal{N}}$ to mean the subset of elements of the matrix $\boldsymbol{A}$ defined by $\mathcal{N}$. It should be understood that matrix equalities involving $[[\cdot]]_{\mathcal{N}}$, $[[\cdot]]_{\mathcal{N}^c}$, or $[[\cdot]]_{\mathcal{N}_p}$ hold only on matrix entries with indices in the associated index set, so for example relations such as

$$[[\boldsymbol{A}]]_{\mathcal{N}_p} = [[\boldsymbol{B}]]_{\mathcal{N}_p} \;, \qquad [[\boldsymbol{A}]]_{\mathcal{N}_{p+1}} \neq [[\boldsymbol{B}]]_{\mathcal{N}_{p+1}} \;, \qquad [[\boldsymbol{A}]]_{\mathcal{N}^c} \neq [[\boldsymbol{B}]]_{\mathcal{N}^c} \;, \tag{P1.1.1}$$

are all compatible. Figure P1.1.1 schematically illustrates the notions.

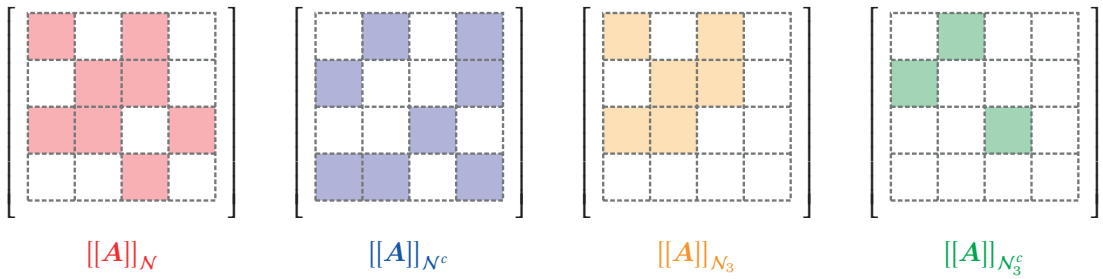


Figure P1.1.1. *An example of a symmetric index set* $\mathcal{N} = \{(1,1),(1,3),(3,1),(2,2),(2,3),(3,2),(3,4),(4,3)\}$ *with the complement* $\mathcal{N}^c = \{(1,2),(2,1),(1,4),(4,1),(2,4),(4,2),(3,3),(4,4)\}$ *and the sub-sets* $\mathcal{N}_3 = \{(1,1),(1,3),(3,1),(2,2),(2,3),(3,2)\}$ *and* $\mathcal{N}_3^c = \{(1,2),(2,1),(3,3)\}$. *Each small square of the grid represents a single element of a matrix* $\boldsymbol{A} \in \mathbb{R}^{4\times 4}$.

For integer numbers $1 \le i \le j \le m$, and $1 \le k \le l \le n$, and a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ we write

$$\boldsymbol{A}_{(i,j),(k,l)} = \begin{bmatrix} A_{i,k} & A_{i,k+1} & \cdots & A_{i,l} \\ A_{i+1,k} & A_{i+1,k+1} & \cdots & A_{i+1,l} \\ \vdots & \vdots & \ddots & \vdots \\ A_{j,k} & A_{j,k+1} & \cdots & A_{j,l} \end{bmatrix} \tag{P1.1.2}$$

to mean a sub-block of $\boldsymbol{A}$ consisting of elements of the intersection of rows i to j and columns k to l.

We will now define the class of overlapping squares index sets. Block tridiagonal is a simple special case of such an index set where for block size l in the tridiagonal structure, the corner set is $\{i_s, j_s\} = \{(s-1)l+1, (s+1)l\}$ which are the top right corners of $2l \times 2l$ diagonal blocks. However, an overlapping block index sets can be more general. Figure P1.1.2 illustrates a particular overlapping squares index set for a matrix $\boldsymbol{A} \in \mathbb{R}^{14\times 14}$. The conditions (P1.1.3) require that the (exterior) corner points $\{(i_s, j_s)\}_{s=1}^{k}$ lie strictly above the diagonal, with the implied interior corner points $\{(i_s, j_{s-1})\}_{s=2}^{k}$ lying on or above the diagonal. In particular the conditions (P1.1.3) exclude cases of trivial decouplings of the index set $\widetilde{\mathcal{N}}$, which must include all diagonal entries.

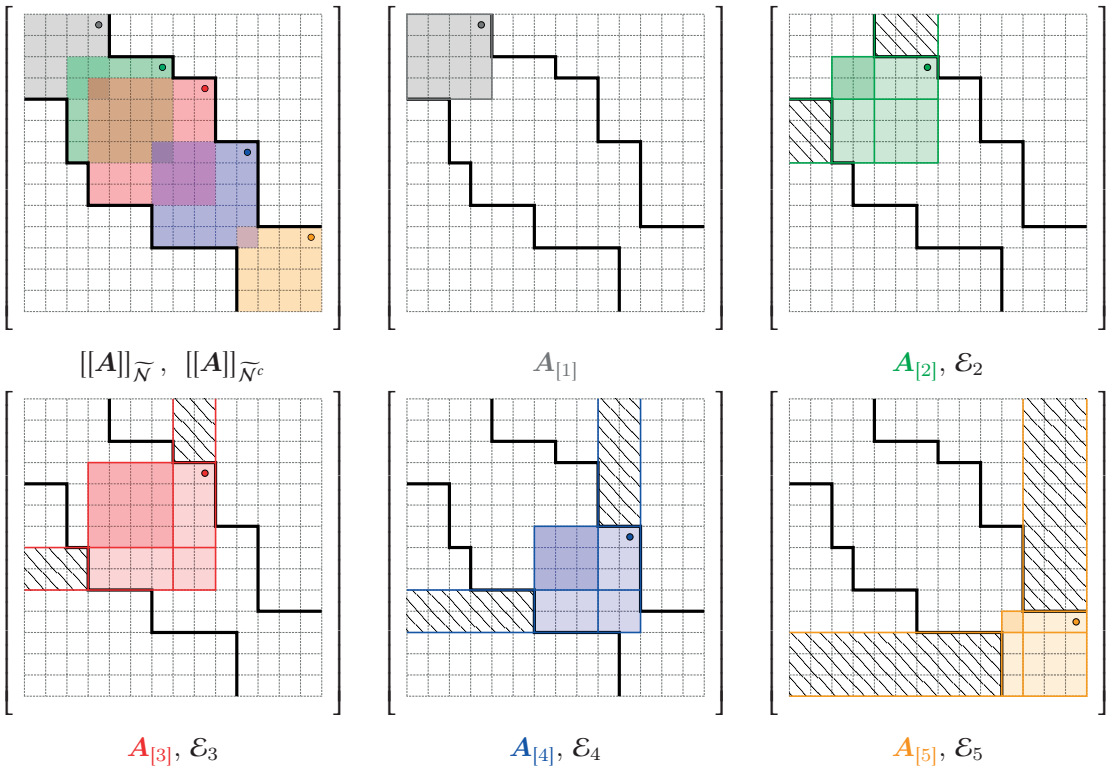


Figure P1.1.2. *An example of an overlapping squares index set and the partitioning of Definition P1.1.2 induced by it for a matrix* $\boldsymbol{A} \in \mathbb{R}^{14\times 14}$. *Each small cell corresponds to a matrix entry. The thick lines on the edges of cells away from the principal diagonal link the* $k = 5$ *corners (indicated by coloured dots) of the corner set* $\{(1,4), (3,7), (4,9), (7,11), (11,14)\}$. *The associated overlapping squares index set* $\widetilde{\mathcal{N}}$ *corresponds to all entries, or cells, lying within the black lines. Each overlapping square is indicated in a different colour in the first panel with overlaps shaded in mixtures of the colours. Note that multiple overlaps, such as that of* $\boldsymbol{A}_{[1]}$, $\boldsymbol{A}_{[2]}$, $\boldsymbol{A}_{[3]}$ *or* $\boldsymbol{A}_{[2]}$ $\boldsymbol{A}_{[3]}$, $\boldsymbol{A}_{[4]}$ *are also allowed. All elements of* $[[\boldsymbol{A}]]_{\widetilde{\mathcal{N}}^c}$ *are those in the striped region. Each one of the subsequent panels illustrates one of the overlapping diagonal square blocks of the sequence* $\left\{\boldsymbol{A}_{[s]}\right\}_{s=1}^{k}$. *The* 2×2 *partitioning of* $\boldsymbol{A}_{[s]}$ *implied by* $\boldsymbol{A}_{[s-1]}$ *is also marked, with the overlap* $\boldsymbol{A}_{[s]_{1,1}}$ *indicated in darker colour. The striped regions are elements with indices in the index sets* $\mathcal{E}_s$. *Note that* $i_5 = j_4$ *so that equality is achieved in conditions (P1.1.3), and, as a consequence, blocks* $\boldsymbol{A}_{[4]}$ *and* $\boldsymbol{A}_{[5]}$ *overlap by only one element.*

More precisely we introduce the following definitions:

Definition P1.1.1. *A **corner set** is a sequence of pairs of indices* $\{(i_s, j_s)\}_{s=1}^{k}$ *satisfying*

$$\begin{aligned} &i_1 = 1, && i_s < i_{s+1} \leq j_s, \\ & && j_s < j_{s+1}, && j_k = n. \end{aligned} \tag{P1.1.3}$$

Definition P1.1.2. *Let* $\{(i_s, j_s)\}_{s=1}^{k}$ *be a corner set and* $\boldsymbol{A} \in \mathbb{R}^{j_k \times j_k}$*. The corner set defines a symmetric **overlapping squares index set** $\widetilde{\mathcal{N}}$ whose indices coincide with all the indices of matrix entries lying within the* k *overlapping diagonal sub-blocks* $\{\boldsymbol{A}_{[s]}\}_{s=1}^{k}$ *with top right-hand corners given by the corner set. This is to say:*

$$\boldsymbol{A}_{[s]} = \boldsymbol{A}_{(i_s, j_s),(i_s, j_s)} \tag{P1.1.4}$$

For $s \geq 2$ *the overlap between* $\boldsymbol{A}_{[s-1]}$ *and* $\boldsymbol{A}_{[s]}$ *implies the* 2×2 *partitioning*

$$\boldsymbol{A}_{[s]} = \begin{bmatrix} \boldsymbol{A}_{[s]_{1,1}} & \boldsymbol{A}_{[s]_{1,2}} \\ \boldsymbol{A}_{[s]_{2,1}} & \boldsymbol{A}_{[s]_{2,2}} \end{bmatrix} := \begin{bmatrix} \boldsymbol{A}_{(i_s, j_{s-1}),(i_s, j_{s-1})} & \boldsymbol{A}_{(i_s, j_{s-1}),(j_{s-1}+1, j_s)} \\ \boldsymbol{A}_{(j_{s-1}+1, j_s),(i_s, j_{s-1})} & \boldsymbol{A}_{(j_{s-1}+1, j_s),(j_{s-1}+1, j_s)} \end{bmatrix} \tag{P1.1.5}$$

so that $\boldsymbol{A}_{[s]_{1,1}}$ *is the overlap.*

For $s \geq 2$ *the corner set* $\{(i_s, j_s)\}_{s=1}^{k}$ *also introduces a partitioning of the complementary set* $\widetilde{\mathcal{N}}^c$ *into a union of disjoint symmetric index sets*

$$\begin{aligned} \mathcal{E}_s = \{(i,j) : \quad & 1 \leq i \leq (i_s - 1) \quad (j_{s-1} + 1) \leq j \leq j_s\} \\ \cup \{(i,j) : \quad & (j_{s-1} + 1) \leq i \leq j_s \quad 1 \leq j \leq (i_s - 1)\} \quad . \end{aligned} \tag{P1.1.6}$$

Remark P1.1.1. *Note that for* $s \geq 2$ *the overlap* $\boldsymbol{A}_{[s]_{1,1}}$ *between* $\boldsymbol{A}_{[s]}$ *and* $\boldsymbol{A}_{[s-1]}$ *from Definition P1.1.2 is a trailing diagonal sub-block of* $[[\boldsymbol{A}]]_{\widetilde{\mathcal{N}}_{j_{s-1}}}$ *(see Figure P1.1.2).*

Remark P1.1.2. *If the overlaps* $\boldsymbol{A}_{[s]_{1,1}}$ *from Definition P1.1.2 are all invertible then each* $\boldsymbol{A}_{[s]}$ *can be written as*

$$\boldsymbol{A}_{[s]} = \begin{bmatrix} \boldsymbol{A}_{[s]_{1,1}} & \boldsymbol{A}_{[s]_{1,2}} \\ \boldsymbol{A}_{[s]_{2,1}} & \boldsymbol{A}_{[s]_{2,2}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{\Psi}_s(\boldsymbol{A}) & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{[s]_{1,1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_s(\boldsymbol{A}) \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{\Omega}_s(\boldsymbol{A}) \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \tag{P1.1.7}$$

with

$$\boldsymbol{\Omega}_s(\boldsymbol{A}) = \left(\boldsymbol{A}_{[s]_{1,1}}\right)^{-1} \boldsymbol{A}_{[s]_{1,2}} \quad , \tag{P1.1.8a}$$

$$\boldsymbol{\Psi}_s(\boldsymbol{A}) = \boldsymbol{A}_{[s]_{2,1}} \left(\boldsymbol{A}_{[s]_{1,1}}\right)^{-1} \tag{P1.1.8b}$$

and

$$\boldsymbol{H}_s(\boldsymbol{A}) = \boldsymbol{A}_{[s]_{2,2}} - \boldsymbol{A}_{[s]_{2,1}} \left(\boldsymbol{A}_{[s]_{1,1}}\right)^{-1} \boldsymbol{A}_{[s]_{1,2}} \quad . \tag{P1.1.9}$$

The factorization (P1.1.7) is one way of showing the well known result that if additionally $\boldsymbol{A}_{[s]}$ *is invertible so is* $\boldsymbol{H}_s(\boldsymbol{A})$*. The main application of the partitioning of definition Definition P1.1.2 will be to symmetric positive definite covariance matrices, for which all the diagonal sub-blocks are invertible.*

Remark P1.1.3. *Following the assumptions and notations of Remark P1.1.2 (in particular invertibility of $\boldsymbol{A}_{[s]}$) Equation (P1.1.7) shows that:*

$$\left(\boldsymbol{A}_{[s]}\right)^{-1} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\Omega}_s \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \left(\boldsymbol{A}_{[s]_{1,1}}\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_s^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{\Psi}_s & \boldsymbol{I} \end{bmatrix} \tag{P1.1.10}$$

$$= \begin{bmatrix} \left(\boldsymbol{A}_{[s]_{1,1}}\right)^{-1} + \boldsymbol{\Omega}_s \boldsymbol{H}_s^{-1} \boldsymbol{\Psi}_s & -\boldsymbol{\Omega}_s \boldsymbol{H}_s^{-1} \\ -\boldsymbol{H}_s^{-1} \boldsymbol{\Psi}_s & \boldsymbol{H}_s^{-1} \end{bmatrix} \tag{P1.1.11}$$

and so

$$\left(\boldsymbol{A}_{[s]}\right)^{-1} - \begin{bmatrix} \left(\boldsymbol{A}_{[s]_{1,1}}\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_s \boldsymbol{H}_s^{-1} \boldsymbol{\Psi}_s & -\boldsymbol{\Omega}_s \boldsymbol{H}_s^{-1} \\ -\boldsymbol{H}_s^{-1} \boldsymbol{\Psi}_s & \boldsymbol{H}_s^{-1} \end{bmatrix} \tag{P1.1.12}$$

where $\boldsymbol{\Psi}_s = \boldsymbol{\Psi}_s(\boldsymbol{A})$, $\boldsymbol{\Omega}_s = \boldsymbol{\Omega}_s(\boldsymbol{A})$ and $\boldsymbol{H}_s = \boldsymbol{H}_s(\boldsymbol{A})$.

Remark P1.1.4. *Equation (P1.1.6) from Definition P1.1.2 implies that:*

$$\widetilde{\mathcal{N}}_{j_s}^c = \bigcup_{i=2}^{s} \mathcal{E}_i \tag{P1.1.13}$$

so that

$$\widetilde{\mathcal{N}}_{j_s}^c = \widetilde{\mathcal{N}}_{j_{s-1}}^c \cup \mathcal{E}_s \tag{P1.1.14}$$

P1.1.2 Existence and uniqueness of sparse maximum entropy fit

Given the notation of the previous section we are now ready to state the following theorem, which is a reformulation of the result of [Dem1972, p. 160].

Theorem P1.1.1. *Let $\boldsymbol{C}$ be a symmetric positive definite matrix and $\mathcal{N}$ a general (not necessarily overlapping squares nor symmetric) index set containing the diagonal. Let also $C(\boldsymbol{C})$ be the set of all symmetric positive matrices equal to $\boldsymbol{C}$ within $\mathcal{N}$, specifically:*

$$C(\boldsymbol{C}) := \{\boldsymbol{Z} : \quad \boldsymbol{Z} = \boldsymbol{Z}^T, \quad \boldsymbol{Z} > 0, \quad [[\boldsymbol{Z}]]_{\mathcal{N}} = [[\boldsymbol{C}]]_{\mathcal{N}}\} \quad .$$

Then there exists a unique matrix $\widetilde{\boldsymbol{C}} \in C(\boldsymbol{C})$ such that its inverse $\widetilde{\boldsymbol{C}}^{-1}$ vanishes outside $\mathcal{N}$:

$$\left[\left[\widetilde{\boldsymbol{C}}^{-1}\right]\right]_{\mathcal{N}^c} = \boldsymbol{0} \quad .$$

Furthermore $\widetilde{\boldsymbol{C}}$ is the matrix with the maximum determinant of all matrices in $C(\boldsymbol{C})$:

$$\det\left(\widetilde{\boldsymbol{C}}\right) = \max\{\det(\boldsymbol{Z}) : \boldsymbol{Z} \in C(\boldsymbol{C})\} \quad .$$

Theorem P1.1.1 arises in maximum entropy parameter estimation. In that context $\boldsymbol{C}$ is a covariance matrix that can be regarded as partially (*i.e.* within $\mathcal{N}$) observed from data, and $\widetilde{\boldsymbol{C}}$ is the covariance matrix of the Gaussian probability distribution that is the maximum entropy fit subject to the constraints of the subset of covariances $[[\boldsymbol{C}]]_{\mathcal{N}}$ being prescribed. In fact the determinant of $\widetilde{\boldsymbol{C}}$ is simply related to the entropy of the associated Gaussian probability distribution [Dem1972].

P1.1.3 Maximum entropy fitting for overlapping squares index sets

We will now show that in the particular case of a covariance matrix $\boldsymbol{C}$ and an overlapping squares index set $\widetilde{\mathcal{N}}$ there is a simple formula for the inverse covariance $\widetilde{\boldsymbol{C}}^{-1}$ of Theorem P1.1.1, involving inverses of only the sub-blocks $\boldsymbol{C}_{[s]}$ and $\boldsymbol{C}_{[s]_{1,1}}$. This is the main result, which can be applied in practical maximum entropy fitting. At the same time we will provide an explicit, recursive algorithm involving only partial Gaussian elimination by blocks of $\boldsymbol{C}$ that yields the covariance matrix $\widetilde{\boldsymbol{C}}$ itself. The procedure for the covariance matrix itself was inspired by the simpler version of tridiagonal matrix completion of [StrNgu2004]. The procedures are schematically presented in Figures P1.1.3 and P1.1.4, respectively.

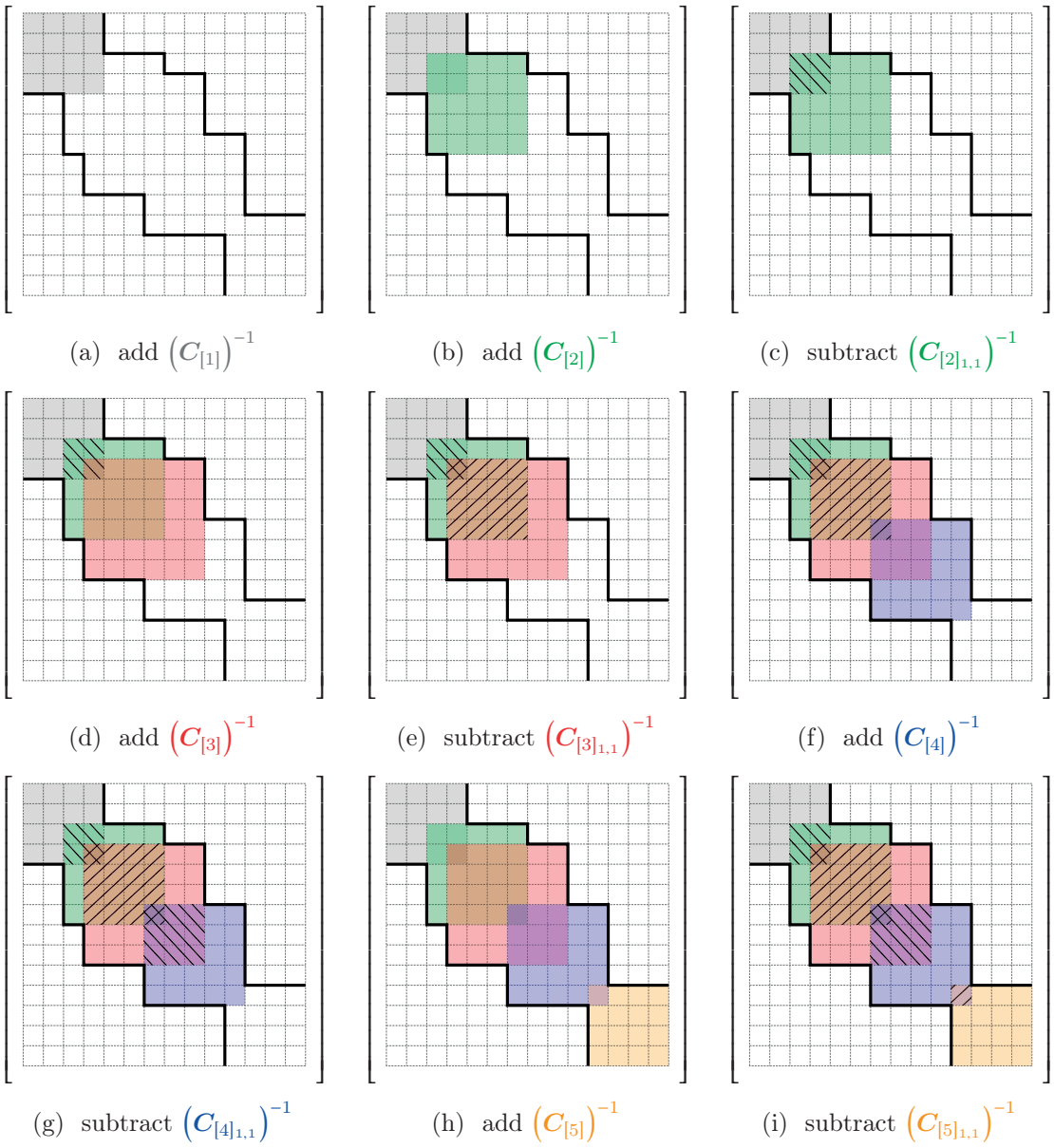
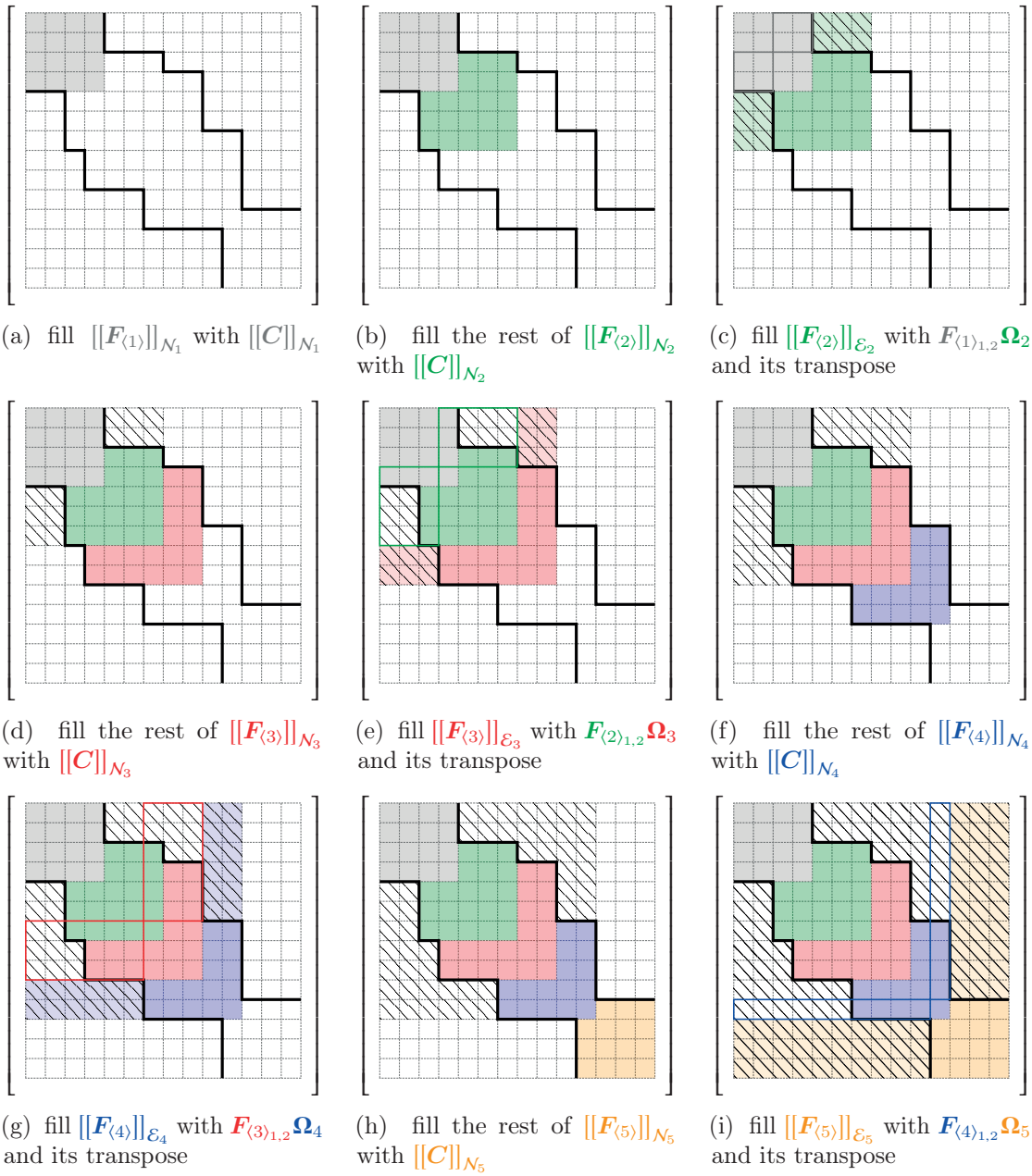


Figure P1.1.3. *An example of the procedure of Corollary P1.1.1a) for computing the inverse* $\widetilde{C}^{-1}$ *of the maximum entropy fit for a covariance matrix* $C \in \mathbb{R}^{14\times14}$ *and the overlapping squares index set of* $\mathcal{N}$ *defined by the corner set* $\{(1,4),(3,7),(4,9),(7,11),(11,14)\}$. *The coloured blocks indicate the inverses* $\left(C_{[s]}\right)^{-1}$ *of the overlapping square blocks, which are added. The striped blocks indicate the inverses* $\left(C_{[s]_{1,1}}\right)^{-1}$ *of the overlaps, which are subtracted.*



(a) fill $[[\boldsymbol{F}_{\langle 1\rangle}]]_{\mathcal{N}_1}$ with $[[\boldsymbol{C}]]_{\mathcal{N}_1}$

(b) fill the rest of $[[\boldsymbol{F}_{\langle 2\rangle}]]_{\mathcal{N}_2}$ with $[[\boldsymbol{C}]]_{\mathcal{N}_2}$

(c) fill $[[\boldsymbol{F}_{\langle 2\rangle}]]_{\mathcal{E}_2}$ with $\boldsymbol{F}_{\langle 1\rangle_{1,2}}\boldsymbol{\Omega}_2$ and its transpose

(d) fill the rest of $[[\boldsymbol{F}_{\langle 3\rangle}]]_{\mathcal{N}_3}$ with $[[\boldsymbol{C}]]_{\mathcal{N}_3}$

(e) fill $[[\boldsymbol{F}_{\langle 3\rangle}]]_{\mathcal{E}_3}$ with $\boldsymbol{F}_{\langle 2\rangle_{1,2}}\boldsymbol{\Omega}_3$ and its transpose

(f) fill the rest of $[[\boldsymbol{F}_{\langle 4\rangle}]]_{\mathcal{N}_4}$ with $[[\boldsymbol{C}]]_{\mathcal{N}_4}$

(g) fill $[[\boldsymbol{F}_{\langle 4\rangle}]]_{\mathcal{E}_4}$ with $\boldsymbol{F}_{\langle 3\rangle_{1,2}}\boldsymbol{\Omega}_4$ and its transpose

(h) fill the rest of $[[\boldsymbol{F}_{\langle 5\rangle}]]_{\mathcal{N}_5}$ with $[[\boldsymbol{C}]]_{\mathcal{N}_5}$

(i) fill $[[\boldsymbol{F}_{\langle 5\rangle}]]_{\mathcal{E}_5}$ with $\boldsymbol{F}_{\langle 4\rangle_{1,2}}\boldsymbol{\Omega}_5$ and its transpose

Figure P1.1.4. *An example of the procedure of Corollary P1.1.1b) for computing the maximum entropy fit* $\widetilde{\boldsymbol{C}} = \boldsymbol{F}_{\langle 5\rangle}$ *for a covariance (and so symmetric positive definite) matrix* $\boldsymbol{C} \in \mathbb{R}^{14\times 14}$ *and the overlapping squares index set of* $\boldsymbol{\mathcal{N}}$ *defined by the corner set* $\{(1,4),(3,7),(4,9),(7,11),(11,14)\}$. *In each step the procedure sets the entries of the coloured, stripped regions* $[[\boldsymbol{F}_{\langle s\rangle}]]_{\mathcal{E}_s}$ *using the symmetric version of Equation (P1.1.27) and (P1.1.28), with* $\boldsymbol{\Omega}_s$ *of Remark P1.1.2 used in Definition P1.1.3. The thick coloured lines in steps (c), (e), (g) and (i) indicate* $\boldsymbol{F}_{\langle s-1\rangle_{1,2}}$ *and* $\boldsymbol{F}_{\langle s-1\rangle_{2,1}}$, *which for our symmetric matrix is equal to* $\left(\boldsymbol{F}_{\langle s-1\rangle_{1,2}}\right)^T$, *used in Equation (P1.1.27). Note that the result of each step may depend on values computed in previous steps.*

To prove the result we introduce the following recursive formulae that are the basic building blocks for our algorithm:

Definition P1.1.3. *Let $\boldsymbol{C}$ be any matrix and $\left\{\boldsymbol{C}_{[s]}\right\}_{s=1}^{k}$ be a sequence of square diagonal sub-blocks of $\boldsymbol{C}$ (of Definition P1.1.2), introduced by an overlapping squares index set $\widetilde{\boldsymbol{\mathcal{N}}}$. Assume also that each $\boldsymbol{C}_{[s]}$ and each overlap $\boldsymbol{C}_{[s]_{1,1}}$ is invertible (which is true if $\boldsymbol{C}$ is a covariance matrix).*
Define $\{\boldsymbol{\Phi}_{\langle s\rangle}\}_{s=1}^{k}$ and $\{\boldsymbol{F}_{\langle s\rangle}\}_{s=1}^{k}$ to be the sequences of matrices satisfying the following recursive relations:

1. *For $s = 1$*

$$\boldsymbol{\Phi}_{\langle 1\rangle} = \left(\boldsymbol{C}_{[1]}\right)^{-1} \tag{P1.1.15}$$

$$\boldsymbol{F}_{\langle 1\rangle} = \boldsymbol{C}_{[1]} \tag{P1.1.16}$$

2. *For $s \geq 2$, with the notations of Remark P1.1.2 (which is valid given the assumptions on $\boldsymbol{C}_{[s]}$ and $\boldsymbol{C}_{[s]_{1,1}}$) the recursive relations can be written as:*

$$\boldsymbol{\Phi}_{\langle s\rangle} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & -\boldsymbol{\Omega}_s \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle} & \boldsymbol{0} & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{H}_s^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{\Psi}_s & \boldsymbol{I} \end{bmatrix} \tag{P1.1.17}$$

$$\boldsymbol{F}_{\langle s\rangle} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi}_s & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_{\langle s-1\rangle} & \boldsymbol{0} & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{H}_s \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{\Omega}_s \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \tag{P1.1.18}$$

where $\boldsymbol{\Psi}_s = \boldsymbol{\Psi}_s(\boldsymbol{C})$, $\boldsymbol{\Omega}_s = \boldsymbol{\Omega}_s(\boldsymbol{C})$ and $\boldsymbol{H}_s = \boldsymbol{H}_s(\boldsymbol{C})$ of Remark P1.1.2.

The following lemma shows certain properties of the sequences $\{\boldsymbol{\Phi}_{\langle s\rangle}\}_{s=1}^{k}$ and $\{\boldsymbol{F}_{\langle s\rangle}\}_{s=1}^{k}$ of Definition P1.1.3 that are crucial in the proof of our result.

Lemma P1.1.1. *Let $\{(i_s, j_s)\}_{s=1}^{k}$ be a corner set and $\widetilde{\boldsymbol{\mathcal{N}}}$ the associated overlapping squares index set. Let $\left\{\boldsymbol{C}_{[s]}\right\}_{s=1}^{k}$ be the sequence of sub-blocks of Definition P1.1.2 of a matrix $\boldsymbol{C} \in \mathbb{R}^{j_k \times j_k}$ with each $\boldsymbol{C}_{[s]}$ and $\boldsymbol{C}_{[s]_{1,1}}$ invertible. Let also $\{\boldsymbol{\Phi}_{\langle s\rangle}\}_{s=1}^{k}$ and $\{\boldsymbol{F}_{\langle s\rangle}\}_{s=1}^{k}$ be the sequences of Definition P1.1.3.*

Then each pair of blocks $\boldsymbol{\Phi}_{\langle s\rangle}$ and $\boldsymbol{F}_{\langle s\rangle}$ has the following properties:

$$\boldsymbol{\Phi}_{\langle s\rangle} = \left(\boldsymbol{F}_{\langle s\rangle}\right)^{-1} \ , \tag{P1.1.19a}$$

$$\left[\left[\boldsymbol{F}_{\langle s\rangle}\right]\right]_{\widetilde{\mathcal{N}}_{j_s}} = \left[\left[\boldsymbol{C}\right]\right]_{\widetilde{\mathcal{N}}_{j_s}} \ , \tag{P1.1.19b}$$

$$\left[\left[\left(\boldsymbol{F}_{\langle s\rangle}\right)^{-1}\right]\right]_{\widetilde{\mathcal{N}}_{j_s}^{c}} = \boldsymbol{0} \quad . \tag{P1.1.19c}$$

Proof. The proof is by induction.

For $s = 1$ all the properties follow straight from the definition of $\boldsymbol{\Phi}_{\langle 1\rangle}$ and $\boldsymbol{F}_{\langle 1\rangle}$, the assumption that $\boldsymbol{C}_{[1]}$ is invertible and the fact that $\widetilde{\mathcal{N}}^c_{j_1}$ is empty.

Suppose now that the properties (P1.1.19) are valid for $s-1$.

Given the assumption that $\boldsymbol{\Phi}_{\langle s-1\rangle} = (\boldsymbol{F}_{\langle s-1\rangle})^{-1}$ property (P1.1.19a) for s can be verified *e.g.* through direct computation of the matrix product $\boldsymbol{\Phi}_{\langle s\rangle}\boldsymbol{F}_{\langle s\rangle}$ using the expressions (P1.1.18) and (P1.1.17).

The assumption that $[[\boldsymbol{F}_{\langle s-1\rangle}]]_{\widetilde{\mathcal{N}}_{j_{s-1}}} = [[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}_{j_{s-1}}}$ together with Remark P1.1.1 implies that $\boldsymbol{F}_{\langle s-1\rangle}$ can be partitioned as:

$$\boldsymbol{F}_{\langle s-1\rangle} = \begin{bmatrix} \boldsymbol{F}_{\langle s-1\rangle_{1,1}} & \boldsymbol{F}_{\langle s-1\rangle_{1,2}} \\ \boldsymbol{F}_{\langle s-1\rangle_{2,1}} & \boldsymbol{C}_{[s]_{1,1}} \end{bmatrix} \quad . \tag{P1.1.20}$$

Equation (P1.1.18) can then be written as:

$$\begin{aligned} \boldsymbol{F}_{\langle s\rangle} &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi}_s^T & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_{\langle s-1\rangle_{1,1}} & \boldsymbol{F}_{\langle s-1\rangle_{1,2}} & \boldsymbol{0} \\ \boldsymbol{F}_{\langle s-1\rangle_{2,1}} & \boldsymbol{C}_{[s]_{1,1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{H}_s \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{\Omega}_s \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{F}_{\langle s-1\rangle_{1,1}} & \boldsymbol{F}_{\langle s-1\rangle_{1,2}} & \boldsymbol{F}_{\langle s-1\rangle_{1,2}}\boldsymbol{\Omega}_s \\ \boldsymbol{F}_{\langle s-1\rangle_{2,1}} & & \boldsymbol{C}_{[s]} \\ \boldsymbol{\Psi}_s\boldsymbol{F}_{\langle s-1\rangle_{2,1}} & & \end{bmatrix} \end{aligned} \tag{P1.1.21}$$

where $\boldsymbol{\Psi}_s = \boldsymbol{\Psi}_s(\boldsymbol{C})$, $\boldsymbol{\Omega}_s = \boldsymbol{\Omega}_s(\boldsymbol{C})$ and $\boldsymbol{H}_s = \boldsymbol{H}_s(\boldsymbol{C})$ of Remark P1.1.2.

Equation (P1.1.21) shows that $\boldsymbol{C}_{[s]}$ is a trailing diagonal sub-block of $\boldsymbol{F}_{\langle s\rangle}$. It follows from Definition P1.1.2 that all elements in $[[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}_{j_s}}$ are only those of $[[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}_{j_{s-1}}} = [[\boldsymbol{F}_{\langle s-1\rangle}]]_{\widetilde{\mathcal{N}}_{j_{s-1}}}$ and of $\boldsymbol{C}_{[s]}$. Hence $[[\boldsymbol{F}_{\langle s\rangle}]]_{\widetilde{\mathcal{N}}_{j_s}} = [[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}_{j_s}}$ and the property (P1.1.19b) is also valid for s.

Finally let:

$$\boldsymbol{\Phi}_{\langle s-1\rangle} = \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle_{1,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{1,2}} \\ \boldsymbol{\Phi}_{\langle s-1\rangle_{2,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{2,2}} \end{bmatrix} \tag{P1.1.22}$$

be the 2×2 partitioning, analogous to (P1.1.20), where $\boldsymbol{\Phi}_{\langle s-1\rangle_{2,2}}$ is of the same dimensions as the overlap $\boldsymbol{C}_{[s]_{1,1}}$. Equation (P1.1.17) can be written as:

$$\begin{aligned} \boldsymbol{\Phi}_{\langle s\rangle} &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & -\boldsymbol{\Omega}_s \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle_{1,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{1,2}} & \boldsymbol{0} \\ \boldsymbol{\Phi}_{\langle s-1\rangle_{2,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{2,2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{H}_s^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{\Psi}_s & \boldsymbol{I} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle_{1,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{1,2}} & \boldsymbol{0} \\ \boldsymbol{\Phi}_{\langle s-1\rangle_{2,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{2,2}} + \boldsymbol{\Omega}_s\boldsymbol{H}_s^{-1}\boldsymbol{\Psi}_s & -\boldsymbol{\Omega}_s\boldsymbol{H}_s^{-1} \\ \boldsymbol{0} & -\boldsymbol{H}_s^{-1}\boldsymbol{\Psi}_s^T & \boldsymbol{H}_s^{-1} \end{bmatrix} \quad . \end{aligned} \tag{P1.1.23}$$

The middle block in the partitioning in equation (P1.1.23) corresponds exactly to the overlap $\boldsymbol{C}_{[s]_{1,1}}$, whose elements all have indices within $\widetilde{\mathcal{N}}$. This together with the assumption that $[[\boldsymbol{\Phi}_{\langle s-1\rangle}]]_{\widetilde{\mathcal{N}}^c_{j_{s-1}}} = \mathbf{0}$ implies that:

$$\left[\left[\begin{array}{cc} \boldsymbol{\Phi}_{\langle s-1\rangle_{1,1}} & \boldsymbol{\Phi}_{\langle s-1\rangle_{1,2}} \\ \left(\boldsymbol{\Phi}_{\langle s-1\rangle_{1,2}}\right)^T & \boldsymbol{\Phi}_{\langle s-1\rangle_{2,2}} + \boldsymbol{\Omega}_s \boldsymbol{H}_s^{-1} \boldsymbol{\Psi}_s \end{array}\right]\right]_{\widetilde{\mathcal{N}}^c_{j_{s-1}}} = \mathbf{0} \quad . \tag{P1.1.24}$$

The above 2×2 partitioning, which defines the 3×3 partitioning of $\boldsymbol{\Phi}_{\langle s\rangle}$ in equation (P1.1.23), implies that the set of pairs of indices of elements of both zero blocks in Equation (P1.1.23) is exactly $\mathcal{E}_s \subset \widetilde{\mathcal{N}}^c$. After Remark P1.1.4 the above reasoning shows that:

$$[[\boldsymbol{\Phi}_{\langle s\rangle}]]_{\widetilde{\mathcal{N}}^c_{j_s}} = \mathbf{0} \tag{P1.1.25}$$

Validity of property (P1.1.19c) for s, then, follows from Equation (P1.1.25) and the previously proven property (P1.1.19a) for s. □

From the proof of Lemma P1.1.1 we can write explicit algorithms for computing $\boldsymbol{\Phi}_{\langle k\rangle}$ and $\boldsymbol{F}_{\langle k\rangle}$ of Definition P1.1.3.

Corollary P1.1.1. *Procedures for computing the matrices $\boldsymbol{F}_{\langle k\rangle}$ and $\boldsymbol{\Phi}_{\langle k\rangle}$ of Definition P1.1.3.*

a) The procedure for computing $\widetilde{\boldsymbol{C}}^{-1} = \boldsymbol{\Phi}_{\langle k\rangle}$ directly is defined by equation (P1.1.15) and the following reformulation of equation (P1.1.23) based on Remark P1.1.3:

$$\boldsymbol{\Phi}_{\langle s\rangle} = \begin{bmatrix} \boldsymbol{\Phi}_{\langle s-1\rangle} & & \mathbf{0} \\ & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{C}_{[s]}\right)^{-1} & \\ \mathbf{0} & & \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{C}_{[s]_{1,1}}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{P1.1.26}$$

where the 3×3 partitioning is the one of equation (P1.1.23). Note that the recursive relation for $\boldsymbol{\Phi}_{\langle s\rangle}$ depends only on the original values of the overlapping block $\boldsymbol{C}_{[s]}$ and the overlap $\boldsymbol{C}_{[s]_{1,1}}$. As a result of this decoupling, the procedure, which involves only inversions of symmetric positive definite matrices, can efficiently be performed in parallel.

A schematic example is shown in Figure P1.1.3.

b) The recursive procedure for constructing $\widetilde{\boldsymbol{C}} = \boldsymbol{F}_{\langle k \rangle}$ is given by equations (P1.1.16) and (P1.1.21). At each step $s > 1$ elements inside the sparsity pattern are simply taken from $\boldsymbol{C}$, while those outside the sparsity pattern, defined by indices in $\mathcal{E}_s$ are replaced by the blocks:

$$\boldsymbol{F}_{\langle s \rangle_{1,2}} \boldsymbol{\Omega}_s \quad \text{and} \quad \boldsymbol{\Psi}_s \boldsymbol{F}_{\langle s \rangle_{2,1}} \tag{P1.1.27}$$

that are transposes of one another in case of symmetric $\boldsymbol{C}$. Note that $\boldsymbol{F}_{\langle s \rangle_{1,2}}$ and $\boldsymbol{F}_{\langle s \rangle_{2,1}}$ are computed in preceding steps and their values depend only on elements of the original matrix that lie within the sparsity pattern. In each of the steps $\boldsymbol{\Omega}_s$ and $\boldsymbol{\Psi}_s$ can be computed as solutions of the linear systems:

$$\boldsymbol{C}_{[s]_{1,1}} \boldsymbol{\Omega}_s = \boldsymbol{C}_{[s]_{1,2}} \quad \text{and} \quad \left(\boldsymbol{C}_{[s]_{1,1}}\right)^T \boldsymbol{\Psi}_s^T = \left(\boldsymbol{C}_{[s]_{2,1}}\right)^T \tag{P1.1.28}$$

(see Equation (P1.1.8)). Again for symmetric $\boldsymbol{C}$ we have $\boldsymbol{\Omega}_s = \boldsymbol{\Psi}_s^T$.

A schematic example is shown in Figure P1.1.4.

Finally we state the entropy maximization result in the form of the following theorem.

Theorem P1.1.2. *Let $\boldsymbol{C}$ be a covariance matrix and $\widetilde{\mathcal{N}}$ an overlapping squares index set.*

Then the matrix $\widetilde{\boldsymbol{C}} = \boldsymbol{F}_{\langle k \rangle}$ of Definition P1.1.3 is the covariance of the maximum entropy Gaussian model with $[[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}}$ prescribed and the inverse covariance $\widetilde{\boldsymbol{C}}^{-1} = \boldsymbol{\Phi}_{\langle k \rangle}$ vanishes outside the pattern.

Proof. The fact that $\boldsymbol{C}$ is a covariance matrix implies it is symmetric positive definite. This, in turn, means that the block $\boldsymbol{C}_{[s]}$ and the overlap $\boldsymbol{C}_{[s]_{1,1}}$ for each $s \in \{1, \ldots, k\}$ are symmetric positive definite (and so invertible). As a result $\boldsymbol{C}$ satisfies the hypothesis of Lemma P1.1.1 and so $\widetilde{\boldsymbol{C}} = \boldsymbol{F}_{\langle k \rangle} = (\boldsymbol{\Phi}_{\langle k \rangle})^{-1}$ satisfies $\left[\left[\widetilde{\boldsymbol{C}}\right]\right]_{\widetilde{\mathcal{N}}} = [[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}}$ and $\left[\left[\widetilde{\boldsymbol{C}}^{-1}\right]\right]_{\widetilde{\mathcal{N}}} = \boldsymbol{0}$. Theorem P1.1.1 implies uniqueness of such a fit $\widetilde{\boldsymbol{C}}$, so that it has to be the symmetric positive definite maximum entropy fit to $[[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}}$. □

As mentioned before the main result of the simple maximum entropy scheme of Corollary P1.1.1a) for the inverse covariance can be recognized as a particular case of analogous results in [SpeKii1986], [Lau1996, sec. 5.3] and [JohLun1998]. To the best of our knowledge the forward procedure of Corollary P1.1.1b) of finding the completion $\widetilde{\boldsymbol{C}}$ for $[[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}}$ was previously known only in the simpler case of tri-diagonal matrices [StrNgu2004].

For completeness we also provide a formula for evaluating the determinant of the maximum entropy fit covariance, which is directly proportional to the value of the entropy of the Gaussian model whose covariance the fit represents.

Corollary P1.1.2. *Let $\boldsymbol{C}$ be a symmetric positive definite matrix, $\widetilde{\mathcal{N}}$ an overlapping squares index set and $\widetilde{\boldsymbol{C}}$ the maximum entropy fit for $[[\boldsymbol{C}]]_{\widetilde{\mathcal{N}}}$ of Theorem P1.1.2.*

Then:

$$\det\left(\widetilde{\boldsymbol{C}}\right) = \frac{\prod_{i=1}^{k} \det\left(\boldsymbol{C}_{[s]}\right)}{\prod_{j=2}^{k} \det\left(\boldsymbol{C}_{[s]_{1,1}}\right)} \tag{P1.1.29}$$

with $\left\{\boldsymbol{C}_{[s]}\right\}_{s=1}^{k}$ (the overlapping blocks) and $\left\{\boldsymbol{C}_{[s]_{1,1}}\right\}_{s=1}^{k}$ (the overlaps) of Definition P1.1.2.

Proof. Note that the factorization (P1.1.18) implies that for $s \geq 2$

$$\det\left(\boldsymbol{F}_{\langle s\rangle}\right) = \det\left(\boldsymbol{F}_{\langle s-1\rangle}\right) \det\left(\boldsymbol{H}_s\right) \quad . \tag{P1.1.30}$$

while equation (P1.1.7) of Remark P1.1.2 shows that

$$\det\left(\boldsymbol{C}_{[s]}\right) = \det\left(\boldsymbol{C}_{[s]_{1,1}}\right) \det\left(\boldsymbol{H}_s\right) \quad . \tag{P1.1.31}$$

The block $\boldsymbol{C}_{[s]_{1,1}}$ is symmetric positive definite as a diagonal sub-block of a symmetric positive definite matrix $\boldsymbol{C}$. Equation (P1.1.30) can then be rewritten as:

$$\det\left(\boldsymbol{F}_{\langle s\rangle}\right) = \det\left(\boldsymbol{F}_{\langle s-1\rangle}\right) \frac{\det\left(\boldsymbol{C}_{[s]}\right)}{\det\left(\boldsymbol{C}_{[s]_{1,1}}\right)} \tag{P1.1.32}$$

The equality (P1.1.29) is simply the expansion of the recursive relation (P1.1.32) for $\boldsymbol{F}_{\langle k\rangle} = \widetilde{\boldsymbol{C}}$ with the step for $s = 1$ given by Equation P1.1.16. □

P1.1.4 Application to parameter extraction for the *cgDNA* model

The scheme of Corollary P1.1.1 can be applied to general (non-symmetric and indefinite) matrices $\boldsymbol{A}$ that satisfy the hypotheses of Lemma P1.1.1. In that case the procedure yields an invertible completion of $[[\boldsymbol{A}]]_{\mathcal{N}}$, and its inverse, which is zero outside the sparsity pattern.

Here, however, we focus on the main application to maximum entropy fitting of inverse covariances, where constraints on the probability distribution can be expressed as an overlapping squares sparsity pattern of the inverse covariance.

In particular this can be used for parameter extraction for the *cgDNA*model [Pet2012; GonPetMad2013; PetPasGonMad2014], summarized briefly in Section B.1.6. The 18×18 overlapping squares sparsity pattern of the *cgDNA* stiffness (inverse covariance) matrix (see Section B.1.5) can be described using the corner set $\left\{\left(12(s-1)+1,\ 12s+6\right)\right\}_{s=1}^{N}$, where N is the number of base pairs of the modelled oligomer.

Chapter P1.1. Maximum entropy fitting for covariance matrices with overlapping squares sparsity

The presented maximum (absolute) entropy fitting procedure of Corollary P1.1.1a) can be used in the second step of the parameter extraction of Section B.1.6, to replace the maximum relative entropy fitting used originally [Pet2012; GonPetMad2013; PetPasGonMad2014]. As argued in [GonPetPas] in this step the entries of the observed covariances close to the diagonal are much better converged than those further out. This is a possible reason why the maximum (absolute) entropy fit, which relies only on entries within the relatively narrow sparsity pattern, is an apparently more natural choice than maximum relative entropy fit, which uses all entries of covariance.

On a more practical note the direct formula for maximum absolute entropy fitting in Corollary P1.1.1a) replaces the more complicated and slower numerical optimization routine of computing the maximum relative entropy fit.

A full discussion of the modified parameter estimation procedure behind *cgDNAparamset2*, which was used to used to generate the results of this thesis, can be found in [GonPetPas].

P1.2 *cgDNA* model coefficients for periodic DNA molecules

(joint work with A. Grandchamp)

Tandem repeats are DNA sequences that consist of multiple (not just two), end-to-end repeats of a shorter fragment that we will refer to as the *basal sequence* of the repeat. Such sequences, are common in nature (*e.g.* they form 3.9% of the human genome [PadZelGas2015]) and while they do not, in most cases, code for genes they play a vital role in the biology of a living cell. They are crucial *e.g.* for chromosome structure and cell division [PadZelGas2015]. Multiple diseases are related to instability (expansion) of such regions in the genome [La Tay2010].

In this chapter we present a new method of modelling such structures within the *cgDNA* model (see Chapter B.1 for a description of the *cgDNA* model as introduced in [Pet2012; GonPetMad2013; PetPasGonMad2014]). In particular we describe a method of constructing what we call a *cgDNA* periodic stiffness matrix and periodic ground state configuration vector for a basal sequence that characterizes an infinite DNA tandem repeat of the basal sequence. In contrast a standard *cgDNA* parameter reconstruction has end effects, as it physically should. But for periodic sequences far from an end it is convenient to appropriately eliminate those end effects. The difference of approximating the standard *cgDNA* shape vector of a finite tandem repeat using the periodic coefficients is also evaluated. We finally show that the periodic coefficients introduced in the context of very long (infinite) linear repeating sequences, can also be used to describe closed loops of DNA.

P1.2.1 Notation and definitions

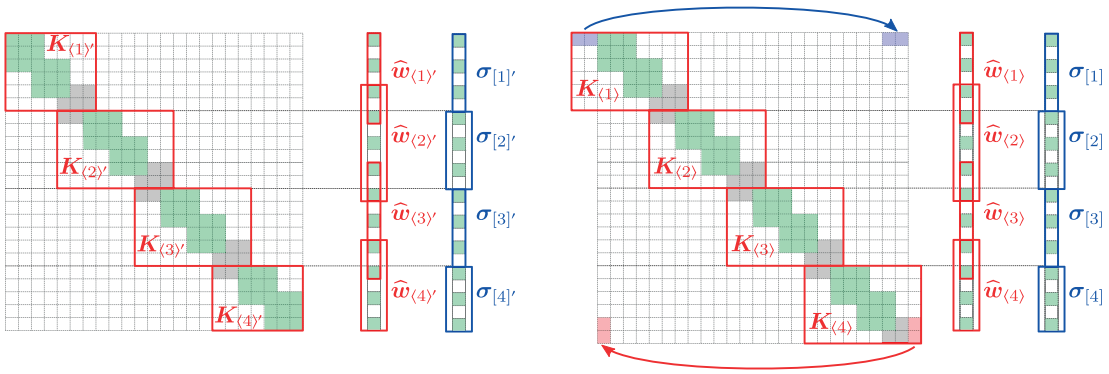
Let $S = \mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_N$ (with $\mathtt{X}_i \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$) be any DNA sequence (the basal sequence) of length N. Let $S_M = \underbrace{SS\ldots S}_{M}$ (the tandem repeat) denote the concatenation of M instances of the basal sequence S. For any such sequence S_M let $\boldsymbol{K}(S_M) \in \mathbb{R}^{(12MN-6)\times(12MN-6)}$ and $\widehat{\boldsymbol{w}}(S_M) \in \mathbb{R}^{12MN-6}$ denote the *cgDNA* stiffness matrix and the ground state configuration vector, respectively, that can be built using the standard *cgDNA* procedure ([PetPasGonMad2014] – see Section B.1.5).

We now introduce notation helpful in stating and proving our result. As previously, by $\boldsymbol{I}_n$ we mean the identity matrix of dimension n, while $\boldsymbol{0}$ will represent a zero matrix (block) of dimension set by context.

For integers $1 \le i \le j \le m$, and $1 \le k \le l \le n$, a vector $\boldsymbol{v} \in \mathbb{R}^m$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ we introduce the notation:

$$\boldsymbol{v}_{(i,j)} = \begin{bmatrix} v_i & v_{i+1} & \ldots & v_j \end{bmatrix}^T \tag{P1.2.1}$$

for a sub-vector of $\boldsymbol{v}$ consisting of elements i to j.



(a) (b)

Figure P1.2.1. *Examples of vectors and blocks of Definition P1.2.1. Panel (a) Equations (P1.2.4) – (P1.2.6) applied to a cgDNA stiffness matrix* $\boldsymbol{K}(S_M) \in \mathbb{R}^{(12NM-6)\times(12NM-6)}$*, a ground state configuration vector* $\boldsymbol{w}(S_M) \in \mathbb{R}^{12NM-6}$ *and a weighted shape vector* $\boldsymbol{\sigma}(S_M) \in \mathbb{R}^{12NM-6}$ *for a sequence* S *of length* $N = 3$ *repeated* $M = 4$ *times. The sub-blocks* $\boldsymbol{K}(S_M)_{\langle s\rangle'}$ *and sub-vectors* $\widehat{\boldsymbol{w}}(S_M)_{\langle s\rangle'}$ *are marked with red. Sub-vectors* $\boldsymbol{\sigma}(S_M)_{[s]'}$ *are marked with blue. Entries of the matrix corresponding to each repeat of* S *are indicated with green. Panel (b) Equations (P1.2.7) – (P1.2.9) applied to a periodic cgDNA stiffness matrix* $\boldsymbol{K}_p(S_M) \in \mathbb{R}^{12NM\times 12NM}$*, a ground state configuration vector* $\boldsymbol{w}_p(S_M) \in \mathbb{R}^{12NM}$ *and a weighted shape vector* $\boldsymbol{\sigma}_p(S_M) \in \mathbb{R}^{12NM}$ *(defined below). The colour scheme is analogous to that of Panel (a). Additionally, the extra "non-local" elements of* $\boldsymbol{K}_p(S_M)_{\langle 1\rangle}$*,* $\boldsymbol{w}_p(S_M)_{\langle 1\rangle}$*,* $\boldsymbol{K}_p(S_M)_{\langle M\rangle}$ *and* $\boldsymbol{w}_p(S_M)_{\langle M\rangle}$ *are shaded in the same way as the elements they are equal to. Here (and in all subsequent figures of this type in this chapter) each cell of the matrix is of dimension* 6×6*. Each cell in the vector is of dimension* 6×1*.*

We also write

$$\boldsymbol{A}_{(i,j),(k,l)} = \begin{bmatrix} A_{i,k} & A_{i,k+1} & \cdots & A_{i,l} \\ A_{i+1,k} & A_{i+1,k+1} & \cdots & A_{i+1,l} \\ \vdots & \vdots & \ddots & \vdots \\ A_{j,k} & A_{j,k+1} & \dots & A_{j,l} \end{bmatrix} \tag{P1.2.2}$$

to mean a sub-block of $\boldsymbol{A}$ consisting of elements of the intersection of rows i to j and columns k to l.

To present our result we also introduce the partitionings of *cgDNA* stiffness matrices and (weighted) shape vectors depicted in Figure P1.2.1. In particular Figure P1.2.1a shows an example of vectors and blocks introduced in Definition P1.2.1a) and b) for a standard *cgDNA* stiffness matrix $\boldsymbol{K}(S_M)$, a weighted shape vector $\boldsymbol{\sigma}(S_M)$ and a ground state configuration vector $\widehat{\boldsymbol{w}}(S_M)$, with $N = 3$ and $M = 4$. These are used to define the periodic *cgDNA* coefficients.

An example of vectors and blocks introduced in Definition P1.2.1c) and d) is shown in Figure P1.2.1b. These will be useful for treatment of the periodic *cgDNA* parameters as defined below.

Definition P1.2.1. *Given two integer numbers $N \geq 1$ and $M \geq 3$ define the sequences of indices $\{i_s\}_{s=1}^M$, $\{j_s\}_{s=1}^M$, $\{j'_s\}_{s=1}^M$ and $\{k_s\}_{s=1}^M$, $\{l_s\}_{s=1}^M$, $\{l'_s\}_{s=1}^M$:*

$$i_s = 12(s-1)N + 1 \qquad \begin{cases} k_1 = 1 \\ k_s = 12(s-1)N - 11 \quad \text{for } 2 \leq s \leq M \end{cases}$$

$$j_s = 12sN \qquad \begin{cases} l_s = 12sN + 6 \quad \textit{for } 1 \leq s \leq M-1 \\ l_M = 12MN \end{cases}$$

$$\begin{cases} j'_s = j_s & \textit{for } 1 \leq s \leq M-1 \\ j'_M = 12MN - 6 \end{cases} \qquad \begin{cases} l'_s = l_s & \textit{for } 1 \leq s \leq M-1 \\ l'_M = 12MN - 6 \end{cases} . \tag{P1.2.3}$$

a) For any vector $\boldsymbol{u} \in \mathbb{R}^{12MN-6}$ define $\left\{\boldsymbol{u}_{[s]'}\right\}_{s=1}^M$ and $\{\boldsymbol{u}_{\langle s\rangle'}\}_{s=1}^M$ to be two sequences of vectors (see Figure P1.2.1a):

$$\boldsymbol{u}_{[s]'} := \boldsymbol{u}_{(i_s, j'_s)} \tag{P1.2.4}$$

$$\boldsymbol{u}_{\langle s\rangle'} := \boldsymbol{u}_{(k_s, l'_s)} \tag{P1.2.5}$$

b) For any matrix $\boldsymbol{A} \in \mathbb{R}^{(12MN-6)\times(12MN-6)}$ define $\{\boldsymbol{A}_{\langle s\rangle'}\}_{s=1}^M$ to be the sequence of blocks defined as (see Figure P1.2.1a):

$$\boldsymbol{A}_{\langle s\rangle'} = \boldsymbol{A}_{(i_s, j'_s),(k_s, l'_s)} \quad . \tag{P1.2.6}$$

c) For any vector $\boldsymbol{v} \in \mathbb{R}^{12MN}$ *define* $\left\{\boldsymbol{v}_{[s]}\right\}_{s=1}^{M}$ *and* $\{\boldsymbol{v}_{\langle s\rangle}\}_{s=1}^{M}$ *to be two sequences of vectors (see Figure P1.2.1b):*

$$\boldsymbol{v}_{[s]} := \boldsymbol{v}_{(i_s,\, j_s)} \tag{P1.2.7}$$

$$\begin{cases} \boldsymbol{v}_{\langle 1\rangle} := \begin{bmatrix} \boldsymbol{v}_{(l_M-12,\, l_M)} \\ \boldsymbol{v}_{(k_1,\, l_1)} \end{bmatrix} & \\ \boldsymbol{v}_{\langle s\rangle} := \boldsymbol{v}_{(k_s,\, l_s)} & \text{for } 2 \geq s \geq M-1 \\ \boldsymbol{v}_{\langle M\rangle} := \begin{bmatrix} \boldsymbol{v}_{(k_M,\, l_M)} \\ \boldsymbol{v}_{(k_1,\, k_1+6)} \end{bmatrix} & \end{cases} \tag{P1.2.8}$$

d) For any matrix $\boldsymbol{B} \in \mathbb{R}^{(12MN)\times(12MN)}$ *define* $\{\boldsymbol{B}_{\langle s\rangle}\}_{s=1}^{M}$ *to be the sequence of blocks defined as (see Figure P1.2.1b):*

$$\begin{cases} \boldsymbol{B}_{\langle 1\rangle} = \begin{bmatrix} \boldsymbol{B}_{(i_1,\, i_1+6),(l_M-12,\, l_M)} & \boldsymbol{B}_{(i_1,\, i_1+6),(k_1,\, l_1)} \\ \boldsymbol{0} & \boldsymbol{B}_{(i_1+6,\, j_1),(k_1,\, l_1)} \end{bmatrix} & \\ \boldsymbol{B}_{\langle s\rangle} = \boldsymbol{B}_{(i_s,\, j_s),(k_s,\, l_s)} & \text{for } 2 \geq s \geq M-1 \\ \boldsymbol{B}_{\langle M\rangle} = \begin{bmatrix} \boldsymbol{B}_{(i_M,\, j_M-12),(k_M,\, l_M)} & \boldsymbol{0} \\ \boldsymbol{B}_{(j_M-12,\, j_M),(k_M,\, l_M)} & \boldsymbol{B}_{(j_M-12,\, j_M),(k_1,\, k_1+6)} \end{bmatrix} & \end{cases} . \tag{P1.2.9}$$

Note that for any $p, q \in \{1, \ldots, M\}$ in Definition P1.2.1 the vectors and blocks:

- $\boldsymbol{u}_{[p]'}$ with $\boldsymbol{u}_{[q]'}$,
- $\boldsymbol{A}_{\langle p\rangle'}$ with $\boldsymbol{A}_{\langle q\rangle'}$,
- $\boldsymbol{v}_{[p]}$ with $\boldsymbol{v}_{[q]}$,
- $\boldsymbol{B}_{\langle p\rangle}$ with $\boldsymbol{B}_{\langle g\rangle}$

share no element. On the other hand for $s \in \{1, \ldots, M-1\}$

- $\boldsymbol{u}_{\langle s\rangle'}$ with $\boldsymbol{u}_{\langle s+1\rangle'}$,
- $\boldsymbol{v}_{\langle s\rangle}$ with $\boldsymbol{v}_{\langle s+1\rangle}$,
- $\boldsymbol{v}_{\langle 1\rangle}$ with $\boldsymbol{v}_{\langle M\rangle}$

do share some of the elements. Note also that the vectors $\boldsymbol{v}_{[s]}$ are all of the same size. Similarly all the blocks $\boldsymbol{B}_{\langle s\rangle}$ are of the same size.

P1.2.1.1 The periodic stiffness matrix and weighted shape vector

Definition P1.2.2. *For $M \geq 3$ let S_M be a tandem repeat of the basal sequence S with the stiffness matrix $\boldsymbol{K}(S_M)$ and weighted shape vector $\boldsymbol{\sigma}(S_M)$. Let $\left\{\boldsymbol{K}(S_M)_{\langle s\rangle'}\right\}_{s=1}^{M}$ and $\left\{\boldsymbol{\sigma}(S_M)_{[s]'}\right\}_{s=1}^{M}$ be the sequences of Definition P1.2.1. Thanks to the locality of cgDNA stiffness matrices and weighted shape vectors, and due to repeats in the sequence for $m, n \in \{2, \ldots, M-1\}$ we have the equality:*

$$\boldsymbol{K}_{\langle m\rangle'}(S_M) = \boldsymbol{K}_{\langle n\rangle'}(S_M) =: \overline{\boldsymbol{K}}_p(S) \in \mathbb{R}^{(12N)\times(12N+18)} \tag{P1.2.10}$$

and

$$\boldsymbol{\sigma}_{[m]'}(S_M) = \boldsymbol{\sigma}_{[n]'}(S_M) =: \boldsymbol{\sigma}_p(S) \in \mathbb{R}^{12N} \tag{P1.2.11}$$

*Note that neither $\overline{\boldsymbol{K}}_p(S)$ nor $\boldsymbol{\sigma}_p(S)$ depends on the number of repeats M, which is indicated by the notation. We call the vector $\boldsymbol{\sigma}_p(S)$ the **periodic weighted shape vector** of the basal sequence S.*

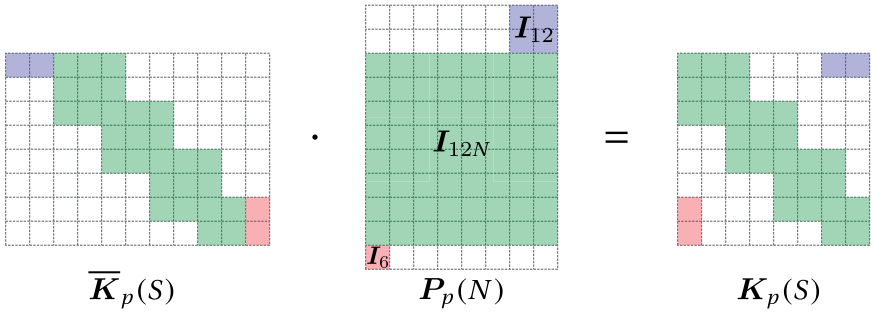


Figure P1.2.2. *Schematic result of multiplying a matrix $\overline{\boldsymbol{K}}_p(S)$ on the right by a matrix $\boldsymbol{P}_p(N)$ (of Definition P1.2.3) for a sequence S of length $N = 4$ In particular, the rearrangement of the sub-blocks coloured with blue and red is indicated.*

Definition P1.2.3. *For any number $N \geq 1$ define the matrix $\boldsymbol{P}_p(N)$ as:*

$$\boldsymbol{P}_p(1) := \begin{bmatrix} \boldsymbol{I}_6 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_6 \\ \boldsymbol{I}_6 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_6 \\ \boldsymbol{I}_6 & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{P}_p(N) := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_6 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{12(N-1)-6} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_6 & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \text{ for } N > 1 \quad . \tag{P1.2.12}$$

*Define the **periodic stiffness matrix** of a basal sequence S of length N as:*

$$\boldsymbol{K}_p(S) := \overline{\boldsymbol{K}}_p(S)\boldsymbol{P}_p(N) \quad \in \mathbb{R}^{12N\times 12N} \tag{P1.2.13}$$

with $\overline{\boldsymbol{K}}_p(S)$ of Definition P1.2.2.

Figure P1.2.2 shows example results of multiplying a stiffness matrix $\overline{\boldsymbol{K}}_p(S)$ on the right by a matrix $\boldsymbol{P}_p(N)$ for a sequence S for $N = 4$.

P1.2.1.2 Constructing a periodic stiffness matrix and a periodic weighted shape vector from a parameter set

For any sequence S of length $N \geq 1$ the periodic stiffness matrix $\boldsymbol{K}_p(S)$ and periodic weighted shape vector $\boldsymbol{\sigma}_p(S)$ can be constructed directly using any given *cgDNA* parameter set. Section B.1.6 provides a general description of *cgDNA* parameter sets. Section P1.1 discusses the *cgDNAparamset2* used to obtain all results in this thesis. The procedure of building $\boldsymbol{K}_p(S)$ and $\boldsymbol{\sigma}_p(S)$ is analogous to the procedure of building the standard *cgDNA* coefficients by adding the overlapping intra and inter contributions (see Section B.1.5). The difference is that the periodic coefficients additionally include an extra set of 6 inter parameters that describe the connection to the next repeat of S. Also extra contributions $\boldsymbol{K}^{\mathtt{X}_N\mathtt{X}_1}$ and $\boldsymbol{\sigma}^{\mathtt{X}_N\mathtt{X}_1}$ that model the coupling to the "upstream" and "downstream" instances of S in the tandem repeat are added.

The procedure of constructing the periodic coefficients is schematically explained in Figure P1.2.3 that indicates in particular how the extra inter block $\boldsymbol{K}^{\mathtt{X}_N\mathtt{X}_1}$ and inter vector $\boldsymbol{\sigma}^{\mathtt{X}_N\mathtt{X}_1}$ are included.

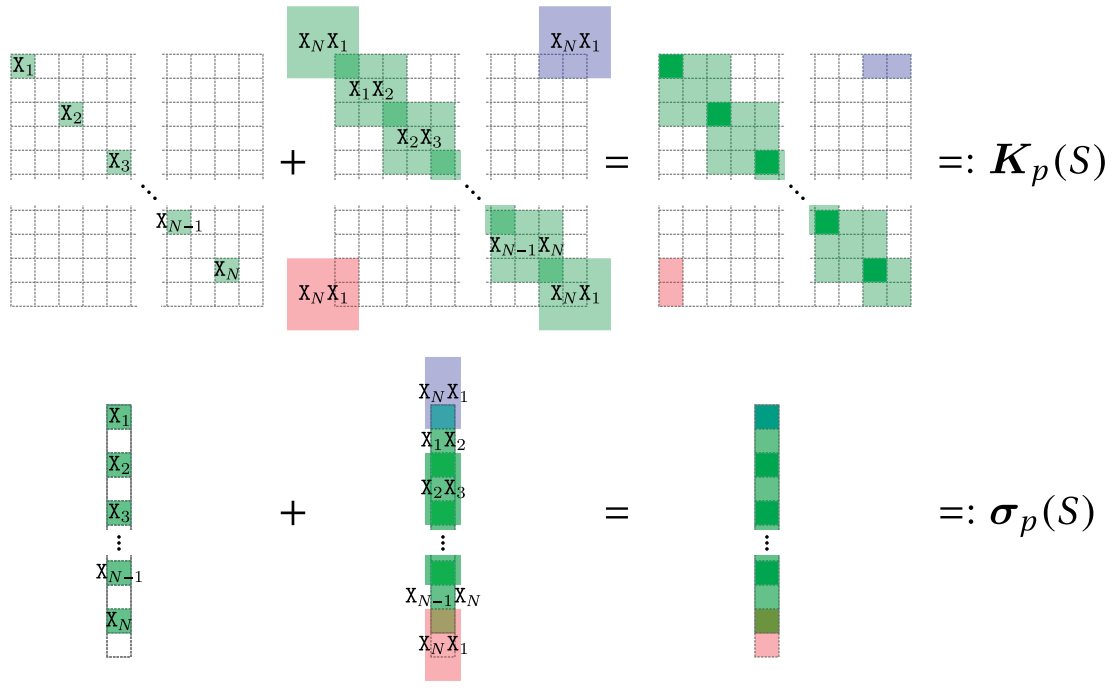


Figure P1.2.3. *Schematic image of how a periodic stiffness matrix* $\boldsymbol{K}_p(S)$ *and a periodic weighted shape vector* $\boldsymbol{\sigma}_p(S)$ *can be constructed for a basal DNA sequence* S *of length* $N \geq 1$. *Note that the two* 6×6 *end blocks that have double overlap in standard cgDNA coefficients, here have additional contributions that give triple overlap.*

We now want to study the definiteness of the periodic stiffness matrices built through the procedure described above using a given *cgDNA* parameter set. It is required from a valid *cgDNA* parameter set that for any sequence S of length $N \geq 2$ the constructed stiffness matrix $\boldsymbol{K}(S)$ (as an inverse of a covariance matrix) is symmetric positive definite [Pet2012; GonPetMad2013]. Lemma P1.2.1 provides sufficient conditions for valid *cgDNA* parameter set so that for any sequence S of length $N \geq 1$ the periodic stiffness matrix $\boldsymbol{K}_p(S)$ is also symmetric positive definite.

Lemma P1.2.1. *Let $\mathcal{P}$ be a valid cgDNA parameter set with the property that for each mononucleotide step $\mathtt{X}_1$ the periodic stiffness matrix $\boldsymbol{K}_p(\mathtt{X}_1)$ is symmetric positive definite and for each dinucleotide step $\mathtt{X}_1\mathtt{X}_2$ the matrix:*

$$\boldsymbol{K}_{\frac{1}{2}}(\mathtt{X}_1\mathtt{X}_2) := \boldsymbol{K}^{\mathtt{X}_1\mathtt{X}_2} + \frac{1}{2}\begin{bmatrix} \boldsymbol{K}^{\mathtt{X}_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{K}^{\mathtt{X}_2} \end{bmatrix} \tag{P1.2.14}$$

is also symmetric positive definite. For any such parameter set the periodic stiffness matrix $\boldsymbol{K}_p(S)$ is symmetric positive definite for any sequence S of length $N \geq 1$. This result can be found in [GonPetPas].

Proof. Let $S = \mathtt{X}_1\mathtt{X}_2 \ldots \mathtt{X}_1\mathtt{X}_N$ be any DNA sequence of length $N \geq 2$ (the case $N = 1$ is covered by the hypothesis). The fact that the periodic stiffness matrix $\boldsymbol{K}_p(S)$ is symmetric follows directly from its definition and the symmetry of all stiffness blocks of a valid parameter set.
Let $\{i_s\}_{s=1}^N$ and $\{j_s\}_{s=1}^N$ be two sequences of indices defined as:

$$i_s = 12(s-1)+1 \qquad \begin{cases} j_s = 12s+6 & \text{for } s \in \{1,\ldots,N\} \\ j_N = 12N \end{cases} \tag{P1.2.15}$$

and let $\boldsymbol{x} \in \mathbb{R}^{12N}$ be an arbitrary non-zero vector. Note that using the overlapping block structure of $\boldsymbol{K}_p(S)$ we can now rewrite the $12N \times 12N$ quadratic form $\boldsymbol{x} \cdot \boldsymbol{K}_p(S)\boldsymbol{x}$ as a sum of small 18×18 quadratic forms as:

$$\begin{aligned} \boldsymbol{x} \cdot \boldsymbol{K}_p(S)\boldsymbol{x} = &\sum_{s=1}^{N-1} \left(\boldsymbol{x}_{(i_s,j_s)} \cdot \boldsymbol{K}_{\frac{1}{2}}(\mathtt{X}_s\mathtt{X}_{s+1})\boldsymbol{x}_{(i_s,j_s)} \right) \\ &+ \begin{bmatrix} \boldsymbol{x}_{(i_N,j_N)} \\ \boldsymbol{x}_{(1,6)} \end{bmatrix} \cdot \boldsymbol{K}_{\frac{1}{2}}(\mathtt{X}_N\mathtt{X}_1) \begin{bmatrix} \boldsymbol{x}_{(i_N,j_N)} \\ \boldsymbol{x}_{(1,6)} \end{bmatrix} \quad . \end{aligned} \tag{P1.2.16}$$

By the hypothesis each such 18×18 quadratic form is positive definite hence $\boldsymbol{x} \cdot \boldsymbol{K}_p(S)\boldsymbol{x} > 0$ for any non-zero $\boldsymbol{x} \in \mathbb{R}^{12N}$ □

In practice the hypotheses of Lemma P1.2.1 were verified numerically for the *cgDNA-paramset2* used to produce all the results of this thesis. The decomposition (P1.2.16) is easy to see in Figure P1.2.3.

P1.2.1.3 Computing the periodic ground state configuration vector

For any *cgDNA* parameter set satisfying Lemma P1.2.1 the periodic ground state configuration vector can be computed by inverting the analogue of the relation B.1.19 of the standard *cgDNA* model.

Definition P1.2.4. *For any DNA sequence S the* ***periodic ground state configuration vector*** *of S is defined as:*

$$\widehat{\boldsymbol{w}}_p(S) := \boldsymbol{K}_p^{-1}(S)\boldsymbol{\sigma}_p(S) \quad , \tag{P1.2.17}$$

where $\boldsymbol{K}_p(S)$ is the periodic stiffness matrix and $\boldsymbol{\sigma}_p(S)$ is the periodic weighted shape vector of S.

We will now prove a feature of the periodic ground state configuration vectors that will be useful for computationally efficient approximation of standard *cgDNA* ground state configurations of tandem repeats using periodic coefficients.

Lemma P1.2.2. *Let S be any sequence of length N and let $\widehat{\boldsymbol{w}}_p(S)$ be its periodic ground state configuration vector. For any number $M \geq 1$ the periodic ground state configuration vector $\widehat{\boldsymbol{w}}_p(S_M)$ of the tandem repeat S_M is equal to:*

$$\widehat{\boldsymbol{w}}_p(S_M) = \Big[\underbrace{\widehat{\boldsymbol{w}}_p(S) \quad \widehat{\boldsymbol{w}}_p(S) \quad \dots \quad \widehat{\boldsymbol{w}}_p(S)}_{M} \Big]^T =: \overline{\boldsymbol{w}}_p(S_M) \quad , \tag{P1.2.18}$$

i.e. $\widehat{\boldsymbol{w}}_p(S_M)$ can be constructed by concatenating M instances of $\widehat{\boldsymbol{w}}_p(S)$.

Proof. Let $\boldsymbol{K}_p(S_M)$, $\boldsymbol{\sigma}_p(S_M)$ and $\widehat{\boldsymbol{w}}_p(S_M)$ be, respectively, the periodic stiffness matrix, periodic weighted shape vector and periodic ground state configuration of the tandem repeat S_M. Let:

$$\boldsymbol{x} := \boldsymbol{K}_p(S_M)\overline{\boldsymbol{w}}_p(S_M) \quad . \tag{P1.2.19}$$

Note that this can be rewritten as M equalities using the vectors and blocks of Definition P1.2.1c) and d):

$$\boldsymbol{x}_{[s]} = \boldsymbol{K}_p(S_M)_{\langle s \rangle}\overline{\boldsymbol{w}}_p(S_M)_{\langle s \rangle} \quad \text{for } 1 \leq s \leq M \quad . \tag{P1.2.20}$$

Furthermore, thanks to the locality in the definition of $\boldsymbol{K}_p(S_M)$ and due to the fact that the sequence is a tandem repeat of the basal sequence S (which defines the partitioning), we have for any $m, n \in \{1, \dots, M\}$:

$$\boldsymbol{K}_p(S_M)_{\langle m \rangle} = \boldsymbol{K}_p(S_M)_{\langle n \rangle} = \overline{\boldsymbol{K}}_p(S) \tag{P1.2.21}$$

(see Definition P1.2.2).

Directly from the definition of $\overline{\boldsymbol{w}}_p(S_M)$ in Equation (P1.2.18), using the $\boldsymbol{P}_p(N)$ matrix of Definition P1.2.3 we have also for any $m, n \in \{1, \ldots, M\}$:

$$\overline{\boldsymbol{w}}_p(S_M)_{\langle m \rangle} = \overline{\boldsymbol{w}}_p(S_M)_{\langle n \rangle} = \boldsymbol{P}_p(N)\widehat{\boldsymbol{w}}_p(S) \quad . \tag{P1.2.22}$$

This, together with Definition P1.2.3 implies that for each $s \in \{1, \ldots, M\}$ Equation (P1.2.20) can be rewritten as:

$$\boldsymbol{x}_{[s]} = \overline{\boldsymbol{K}}_p(S)\boldsymbol{P}_p(N)\widehat{\boldsymbol{w}}_p(S) = \boldsymbol{K}_p(S)\widehat{\boldsymbol{w}}_p(S) = \boldsymbol{\sigma}_p(S) \quad . \tag{P1.2.23}$$

From the local structure of the periodic weighted shape vector of the tandem repeat S_M we therefore have:

$$\boldsymbol{x} = \boldsymbol{\sigma}_p(S_M) \tag{P1.2.24}$$

and so by definition P1.2.4:

$$\boldsymbol{K}_p(S_M)\widehat{\boldsymbol{w}}(S_M) = \boldsymbol{\sigma}_p(S_M) = \boldsymbol{K}_p(S_M)\overline{\boldsymbol{w}}_p(S_M) \quad . \tag{P1.2.25}$$

Finally from the fact that $\boldsymbol{K}_p(S_M)$ is symmetric positive definite, and so non-singular, we have the equality:

$$\widehat{\boldsymbol{w}}(S_M) = \overline{\boldsymbol{w}}_p(S_M) \quad . \tag{P1.2.26}$$

□

The above lemma shows that the periodic ground state configuration vector of a tandem repeat can be constructed by concatenating the periodic ground state configuration vector of the basal sequence. Of course if the number M of repeats of the basal sequence S in the tandem repeat S_M is large the periodic coefficients can potentially provide a computationally efficient method of characterizing the tandem repeat. This will be further investigated in the next section.

P1.2.2 Coefficients of a linear repeating DNA fragment

As mentioned before the the periodic *cgDNA* coefficients can be seen as characterizing an infinite tandem repeat. Here we want to estimate how good an approximation they are for finite tandem repeats.

Consider a finite tandem repeat S_M ($M > 3$) of the basal sequence S with a *cgDNA* stiffness matrix $\boldsymbol{K}$, ground state configuration vector $\widehat{\boldsymbol{w}}$ and a weighted shape vector $\boldsymbol{\sigma}$. The explicit dependence on the sequence will be dropped for the sake of clarity, wherever it refers to the entire tandem repeat S_M (*e.g.* we will use $\boldsymbol{K}$ instead of $\boldsymbol{K}(S_M)$, but keep writing $\boldsymbol{K}_p(S)$).

Note that the *cgDNA* internal energy (B.1.17) of a linear DNA fragment of sequence S_M, for any configuration $\boldsymbol{w}$ can be written as:

$$\begin{aligned}
\mathrm{U}(\boldsymbol{w}; S_M) &= \frac{1}{2}\,(\boldsymbol{w}-\widehat{\boldsymbol{w}})\cdot\boldsymbol{K}\,(\boldsymbol{w}-\widehat{\boldsymbol{w}}) && \text{(P1.2.27)}\\
&= \frac{1}{2}\boldsymbol{w}\cdot\boldsymbol{K}\boldsymbol{w} - \boldsymbol{w}\cdot\boldsymbol{K}\widehat{\boldsymbol{w}} + \frac{1}{2}\widehat{\boldsymbol{w}}\cdot\boldsymbol{K}\widehat{\boldsymbol{w}} && \text{(P1.2.28)}\\
&= \frac{1}{2}\boldsymbol{w}\cdot\boldsymbol{K}\boldsymbol{w} - \boldsymbol{w}\cdot\boldsymbol{\sigma} + \frac{1}{2}\widehat{\boldsymbol{w}}\cdot\boldsymbol{\sigma} \quad . && \text{(P1.2.29)}
\end{aligned}$$

Using the notations of Definition P1.2.1 the energy can further be written as:

$$\mathrm{U}(\boldsymbol{w}; S_M) = \frac{1}{2}\sum_{s=1}^{M}\boldsymbol{w}_{[s]'}\cdot\boldsymbol{K}_{\langle s\rangle'}\boldsymbol{w}_{\langle s\rangle'} \;-\; \sum_{s=1}^{M}\boldsymbol{w}_{[s]}\cdot\boldsymbol{\sigma}_{[s]} \;+\; \frac{1}{2}\sum_{s=1}^{M}\widehat{\boldsymbol{w}}_{[s]}\cdot\boldsymbol{\sigma}_{[s]} \quad . \qquad \text{(P1.2.30)}$$

The local structure of the weighted shape vector gives:

$$\boldsymbol{\sigma}_{[i]} = \boldsymbol{\sigma}_p(S) \quad \text{for } n \in \{2,\ldots,M-1\} \quad . \qquad \text{(P1.2.31)}$$

From now on in this section we will only consider periodic deformations in the sense that:

$$\boldsymbol{w}_{[m]} = \boldsymbol{w}_{[n]} := \boldsymbol{w}_p \quad \text{for } m, n \in \{1,\ldots,M\} \quad . \qquad \text{(P1.2.32)}$$

which, for $s \in \{2,\ldots M-1\}$, after Definition P1.2.2 and P1.2.3, gives:

$$\boldsymbol{K}_{\langle s\rangle'}\boldsymbol{w}_{\langle s\rangle'} = \boldsymbol{K}_p(S)\boldsymbol{w}_p \qquad \text{(P1.2.33)}$$

and so:

$$\begin{aligned}
\mathrm{U}(\boldsymbol{w}; S_M) = \frac{1}{2}\boldsymbol{w}_{[1]'}\cdot\boldsymbol{K}_{\langle 1\rangle'}\boldsymbol{w}_{\langle 1\rangle'} &+ \frac{M-2}{2}\boldsymbol{w}_p\cdot\boldsymbol{K}_p(S)\boldsymbol{w}_p &&+ \frac{1}{2}\boldsymbol{w}_{[M]'}\cdot\boldsymbol{K}_{\langle M\rangle'}\boldsymbol{w}_{\langle M\rangle'}\\
- \boldsymbol{w}_{[1]'}\cdot\boldsymbol{\sigma}_{[1]'} &- (M-2)\boldsymbol{w}_p\cdot\boldsymbol{\sigma}_p(S) &&- \boldsymbol{w}_{[M]'}\cdot\boldsymbol{\sigma}_{[M]'}\\
+ \frac{1}{2}\widehat{\boldsymbol{w}}_{[1]'}\cdot\boldsymbol{\sigma}_{[1]'} &+ \frac{1}{2}\sum_{s=2}^{M-1}\widehat{\boldsymbol{w}}_{[s]'}\cdot\boldsymbol{\sigma}_p(S) &&+ \frac{1}{2}\widehat{\boldsymbol{w}}_{[M]'}\cdot\boldsymbol{\sigma}_{[M]'} \quad .
\end{aligned} \qquad \text{(P1.2.34)}$$

Let us now define what we will call the average energy of a repeat as:

$$\overline{\mathrm{U}}(\boldsymbol{w}; S_M) := \frac{1}{M}\mathrm{U}(\boldsymbol{w}; S_M) \tag{P1.2.35}$$

$$= \frac{M-2}{2M}\left(\boldsymbol{w}_p \cdot \boldsymbol{K}_p(S)\boldsymbol{w}_p \;-\; 2\boldsymbol{w}_p(S)\cdot\boldsymbol{\sigma}_p(S)\right) + \frac{1}{2M}\sum_{s=2}^{M-1}\widehat{\boldsymbol{w}}_{[s]'}\cdot\boldsymbol{\sigma}_p \;+\; O\left(\frac{1}{M}\right) \tag{P1.2.36}$$

$$= \frac{M-2}{2M}\left(\boldsymbol{w}_p \cdot \boldsymbol{K}_p(S)\boldsymbol{w}_p \;-\; 2\boldsymbol{w}_p\cdot\boldsymbol{\sigma}_p(S) \;+\; \widehat{\boldsymbol{w}}_p\cdot\boldsymbol{\sigma}_p(S)\right) + \frac{1}{2M}\sum_{s=2}^{M-1}\widehat{\boldsymbol{w}}_{[s]'}\cdot\boldsymbol{\sigma}_p(S) \;-\; \frac{M-2}{2M}\widehat{\boldsymbol{w}}_p\cdot\boldsymbol{\sigma}_p(S) \;+\; O\left(\frac{1}{M}\right) \quad . \tag{P1.2.37}$$

After Definition P1.2.4 this can further be rewritten as:

$$\overline{\mathrm{U}}(\boldsymbol{w}; S_M) = \frac{M-2}{2M}\left(\boldsymbol{w}_p - \widehat{\boldsymbol{w}}_p(S)\right)\cdot\boldsymbol{K}_p(S)\left(\boldsymbol{w}_p - \widehat{\boldsymbol{w}}_p(S)\right) + \frac{1}{2M}\sum_{s=2}^{M-1}\left(\widehat{\boldsymbol{w}}_{[s]'} - \widehat{\boldsymbol{w}}_p(S)\right)\cdot\boldsymbol{\sigma}_p(S) \;+\; O\left(\frac{1}{M}\right) \quad . \tag{P1.2.38}$$

Note that P1.2.38 shows that the accuracy of the periodic approximation:

$$\overline{\mathrm{U}}_p(\boldsymbol{w}_p; S) := \frac{1}{2}\left(\boldsymbol{w}_p - \widehat{\boldsymbol{w}}_p(S)\right)\cdot\boldsymbol{K}_p(S)\left(\boldsymbol{w}_p - \widehat{\boldsymbol{w}}_p(S)\right) \tag{P1.2.39}$$

differs from the average energy of a repeat (P1.2.35) only by a constant term, which depends on how well $\widehat{\boldsymbol{w}}_p$ approximates $\widehat{\boldsymbol{w}}_{[s]'}$ for large M. Unlike the stiffness matrix, which is localized by construction the standard shape vector has non-local sequence dependence, which gives rise to end effects (see Section B.1.5). Hence we ask the question:

P1.2.2.1 How important are end effects?

In this sub-section we study how far from the ends of a DNA molecule the end effects are actually significant. To answer this question we study the error of using the periodic ground state configuration vector $\widehat{\boldsymbol{w}}_p(S)$ of the basal sequence S to approximate the standard *cgDNA* ground state configuration vector $\widehat{\boldsymbol{w}}(S_M)$ of a tandem repeat of S. To evaluate the error we create an ensemble of multiple basal sequences S of lengths $N \in \{1, 2, 3, 4, 5, 10, 30, 60, 120\}$. For lengths $N \in \{1, 2, 3, 4, 5\}$ all possible sequences are generated. For $N \in \{10, 30, 60, 120\}$ we generate only 1000 random sequences (with equal probability of each type of base). For each basal sequence in the ensemble we repeat it M times, so that the resulting tandem repeat S_M is of length 120 bp. For each such tandem repeat we generate the ground state configuration vector $\widehat{\boldsymbol{w}}(S_M) \in \mathbb{R}^{(12MN-6)}$ ($12MN - 6 = 1434$ in each case). For each S_M we also create a vector $\widehat{\boldsymbol{w}}_a(S_M) := \widehat{\boldsymbol{w}}_p(S_M)_{(1,\,12MN-6)} \in \mathbb{R}^{(12MN-6)}$. Here $\widehat{\boldsymbol{w}}_p(S_M)$ is the periodic ground state configuration of the entire tandem repeat constructed by concatenating M times the periodic ground state configuration $\widehat{\boldsymbol{w}}(S)$ of the basal sequence.

The semi-logarithmic plots of Figure P1.2.4 show the absolute value of each component of the difference $\widehat{\boldsymbol{w}}(S_M) - \widehat{\boldsymbol{w}}_a(S_M)$ for all sequences in our ensemble on a semi-log plot. We note that the magnitude of the error does not depend on N nor on M, but rather on the distance from either end of the molecule. For all entries further than five base pairs from either end of the molecule the approximation is accurate up to around 10^{-2}. This suggests that the end effects in the *cgDNA* model are significant up to five base pairs in from either end of the molecule. The independence of M also means that the periodic coefficients are as a good an approximation for non-repeating sequences as they are for tandem repeats far from the ends.

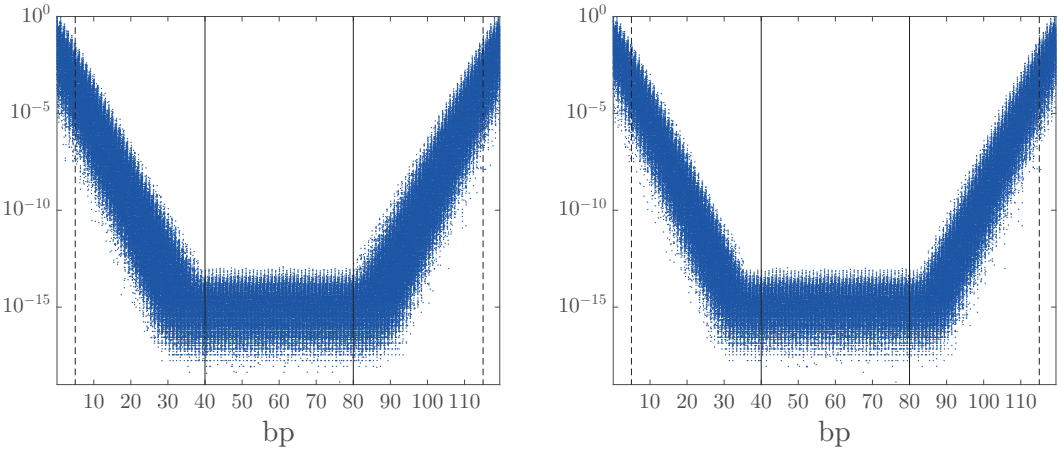


Figure P1.2.4. *The absolute value of the difference between each element of the standard cgDNA ground state configuration vector* $\widehat{\mathbf{w}}(S_M)$ *and its approximation* $\widehat{\mathbf{w}}_a(S_M)$ *constructed by repeating instances of* $\widehat{\mathbf{w}}_p(S)$*. Panel (a) shows results for all sequences* S_M *in our ensemble. Panel (b) shows results only for the* 1000 *random sequences with* $N = 120$ *and* $M = 1$ *. The dashed line indicates coefficients five base pairs from either end, where the error is around* 10^{-2}*. The solid line indicates coefficients* 40 *base pairs from either end, which is where the error is below machine double precision (hence the flattening of the plot). The accuracy does not depend on* N *nor on* M*, but on the distance from ether end.*

P1.2.3 Coefficients of a closed loop of DNA

Although also important in the case of modelling of long (infinite) linear fragments of DNA one particular property of the periodic coefficients is easier to explain in the context of closed loops that, as shown below, can also be modelled using this formalism.

Consider a DNA molecule where the two ends are covalently bonded. In the *cgDNA* model the internal energy of such a closed configuration, as compared to that of an open one, requires an additional term that represents the nearest neighbour interactions of the last base pair with the first. Note that this is exactly reflected in the periodic coefficients, shown in Figure P1.2.3.

We will also show that the energy definition through the periodic coefficients does not depend on where the closed DNA molecule is opened to give a linear sequence. In order to explain what we mean by that let us first introduce the notation.

Definition P1.2.5. *For a given integer number $N \geq 1$ define the **cyclic right shift permutation matrix**:*

$$\boldsymbol{P}_{\circlearrowright}(N) := \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{12(N-1)} & \mathbf{0} \end{bmatrix} \tag{P1.2.40}$$

*and the **cyclic left shift permutation matrix**:*

$$\boldsymbol{P}_{\circlearrowleft}(N) := \boldsymbol{P}_{\circlearrowright}^{-1}(N) = \boldsymbol{P}_{\circlearrowright}^{T}(N) = \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_{12(N-1)} \\ \boldsymbol{I}_{12} & \mathbf{0} \end{bmatrix} \tag{P1.2.41}$$

In the remaining part of this section we will write $\boldsymbol{P}_{\circlearrowright}$ and $\boldsymbol{P}_{\circlearrowleft}$ to mean $\boldsymbol{P}_{\circlearrowright}(N)$ and $\boldsymbol{P}_{\circlearrowleft}(N)$.

Now consider the following lemma:

Lemma P1.2.3. *Let $S = \mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_N$ (with $\mathtt{X}_a \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$) be a given DNA sequence and $S' = \mathtt{X}_N\mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_{N-1}$ be the DNA sequence that is a cyclic shift of the sequence S to the right. Let $\boldsymbol{w}$ be a given periodic cgDNA configuration vector of a closed DNA molecule of sequence S. Finally let $\boldsymbol{w}' := \boldsymbol{P}_{\circlearrowright}\boldsymbol{w}$. The periodic cgDNA energy of $\boldsymbol{w}$ for the sequence S is the same as the energy of $\boldsymbol{w}'$ for the sequence S', or:*

$$\mathrm{U}_p(\boldsymbol{w}; S) = \mathrm{U}_p(\boldsymbol{w}'; S') \quad . \tag{P1.2.42}$$

Proof. To begin with note that using Definition P1.2.5, the local structure of the periodic coefficients and simple algebra it can be shown that:

$$\begin{cases} \boldsymbol{K}_p(S') = \boldsymbol{P}_{\circlearrowright}\boldsymbol{K}_p(S)\boldsymbol{P}_{\circlearrowleft} \\ \boldsymbol{\sigma}_p(S') = \boldsymbol{P}_{\circlearrowright}\boldsymbol{\sigma}_p(S') \end{cases} \quad \Rightarrow \quad \begin{aligned} \widehat{\boldsymbol{w}}_p(S') &= \boldsymbol{K}_p^{-1}(S')\boldsymbol{\sigma}_p(S') \\ &= \boldsymbol{P}_{\circlearrowright}\widehat{\boldsymbol{w}}_p(S) \end{aligned} \tag{P1.2.43}$$

as a result the energy of $\boldsymbol{w}'$ for the sequence S' can be written as:

$$\begin{aligned} \mathrm{U}_p(\boldsymbol{w}'; S') &= \frac{1}{2}\left(\boldsymbol{w}' - \widehat{\boldsymbol{w}}_p(S')\right) \cdot \boldsymbol{K}_p(S') \left(\boldsymbol{w}' - \widehat{\boldsymbol{w}}_p(S')\right) && \text{(P1.2.44)} \\ &= \frac{1}{2}\left(\boldsymbol{P}_{\circlearrowright}\boldsymbol{w} - \boldsymbol{P}_{\circlearrowright}\widehat{\boldsymbol{w}}_p(S)\right) \cdot \boldsymbol{P}_{\circlearrowright}\boldsymbol{K}_p(S)\boldsymbol{P}_{\circlearrowleft} \left(\boldsymbol{P}_{\circlearrowright}\boldsymbol{w} - \boldsymbol{P}_{\circlearrowright}\widehat{\boldsymbol{w}}_p(S)\right) \\ &= \frac{1}{2}\left(\boldsymbol{w} - \widehat{\boldsymbol{w}}_p(S)\right)^T \boldsymbol{P}_{\circlearrowleft}^T\boldsymbol{P}_{\circlearrowright}\boldsymbol{K}_p(S) \left(\boldsymbol{w} - \widehat{\boldsymbol{w}}_p(S)\right) \\ &= \frac{1}{2}\left(\boldsymbol{w} - \widehat{\boldsymbol{w}}_p(S)\right) \cdot \boldsymbol{K}_p(S) \left(\boldsymbol{w} - \widehat{\boldsymbol{w}}_p(S)\right) && \text{(P1.2.45)} \\ &= \mathrm{U}_p(\boldsymbol{w}; S) && \text{(P1.2.46)} \end{aligned}$$

□

Note that the relation $\boldsymbol{w}' = \boldsymbol{P}_{\circlearrowleft}\boldsymbol{w}$ of Lemma P1.2.3 means that the last set of intra and inter coefficients of $\boldsymbol{w}$ are moved to the "front" in $\boldsymbol{w}'$. As a result both vectors can be seen as describing the same configuration of a given closed DNA molecule: $\boldsymbol{w}$ with respect to S, while $\boldsymbol{w}'$ with respect to S' (a different place of opening of the closed molecule). Lemma P1.2.3 shows that the periodic *cgDNA* energy of the molecule is the same using both descriptions (and, as a consequence, any other cyclic shift of S). In other words we have shown that the periodic coefficients provide a consistent description of closed molecules independent of where the molecule is cut to give its linear sequence.

Note that while the above identities are valid for all shifts of periodic configuration vectors, not all periodic internal coordinate vectors correspond to closed loops as there are six non-linear, non-local conditions on $\boldsymbol{w}$ guaranteeing that $(\boldsymbol{D}_i, \mathbf{r}_i)$ are appropriately periodic.

P1.2.4 The structure of periodic *cgDNA* covariance matrices

A natural question that might arise at this point is whether a maximum entropy fitting procedure analogous to the one presented in Chapter P1.1 can be constructed for the sparsity pattern of the periodic *cgDNA* coefficients. As indicated in Figure P1.2.5 a simple extensions of the algorithm of Chapter P1.1 involving local inversions of sub-blocks of the covariance matrix is not a solution.

To analyse the question in a slightly more general context we need to briefly introduce certain concepts of graph theory. A graph G is called chordal if any cycle of four or more vertices in G is reducible, *i.e.* there exist an edge (called a chord) that connects two vertices of the cycle but is not part of the cycle (the cycle can be split into two shorter cycles). To our knowledge, to date the most general form of sparsity known to have a local inversion procedure for maximum entropy fitting has to include the diagonal and be given by an adjacency matrices of a chordal graphs [SpeKii1986; Lau1996; JohLun1998].

However, as schematically shown in Figure P1.2.5c, graphs associated with the sparsity pattern of periodic stiffness matrices (for sequences of length $N \geq 4$) are not chordal. Irreducible cycles of length N can be constructed as:

$$\begin{array}{ccc} \overset{V_1}{\cup} & & \overset{V_2}{\cup} \\ u_1 & \leftrightarrow & u_2 \\ \updownarrow & & \updownarrow \\ u_N & \leftrightarrow & \dots \\ \overset{\cap}{V_N} & & \end{array} \tag{P1.2.47}$$

with the sub-sets of vertices $V_i = \{v_{6i-5}, \dots, v_{6i}\}$ indicated in Figure P1.2.5.

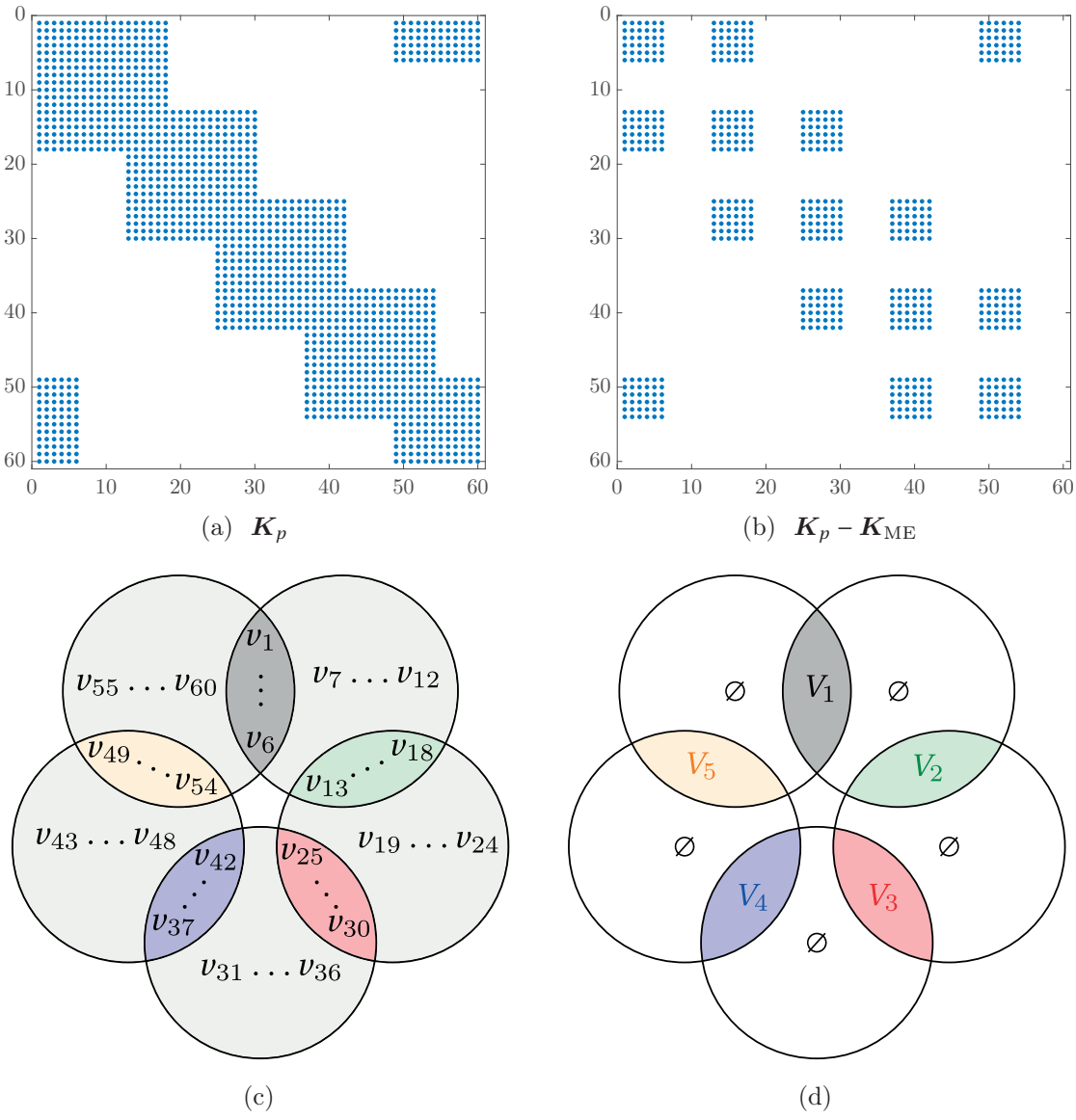


Figure P1.2.5. *The effect of applying an analogue of the maximum entropy procedure in case of periodic covariance (inverse stiffness) matrix. Panel (a) shows the sparsity of periodic cgDNA stiffness matrices* $\boldsymbol{K}_p$ *for sequences of length 5. Panel (b) shows the sparsity of the difference between* $\boldsymbol{K}_p$ *and a matrix* $\boldsymbol{K}_{\mathrm{ME}}$ *computed numerically by applying an analogue of the maximum entropy procedure of Chapter P1.1 to the covariance* $\boldsymbol{K}_p^{-1}$*. Panel (c) shows a schematic representation of the graph* G *whose adjacency matrix is defined by the sparsity pattern of Panel P1.2.5a. Vertex* v_i *in* G *corresponds to row* i *and column* i *of* $\boldsymbol{K}_p$ *with non-zero elements representing edges in* G*. The circles and their intersections represent cliques, i.e. fully connected sub-graphs. Panel (d) indicates the sub-graph* $G' \subset G$ *whose vertices give rise to irreducible cycles of the kind indicated in Equation (P1.2.47). The adjacency matrix of* G' *is given by the sparsity of* $\boldsymbol{K}_p - \boldsymbol{K}_{\mathrm{ME}}$ *shown in Panel (b).*

In the light of the above the existence of a simple local formula for computing the maximum entropy fit periodic stiffness matrix for a given covariance matrix remains an open question, although a positive answer seems to be unlikely.

P1.3 Superhelical structure of DNA tandem repeats

In general, periodic stacking of (close to) identical elements leads to a (close to) helical structure of the resulting construct [ChoGorMad2006], hence *e.g.* the (close to) double helical structure of DNA. Similarly the concatenation of multiple instances of a DNA sequence (that we will call the *basal sequence*) leads to a superhelical shape of the centreline of the ground state of the resulting DNA *tandem repeat*.

For that reason multiple consecutive repeats of relatively short fragments ($\sim 10\,\mathrm{bp}$) have *e.g.* been used in *in vitro* experiments, as such an approach allows for engineering of fragments of a particular intrinsic shape: *e.g.* left- or right-handed superhelices [DubBedFur1994] or straight oligomers [BedFurKat1995; GegVol2010], that are also easy to synthesize [CalDreLuiTra2004, ch. 5].

In this chapter we present a method of analysing the superhelical structure of the ground state configurations of DNA tandem repeats using the *cgDNA* model. Given a basal sequence S of a tandem repeat S_M the presented procedure yields parameters of the superhelix traced by the centreline of the repeat such as a pitch and a radius, which are the same for any number of repeats M.

Finally we present a brief analysis of the results of an exhaustive study of the superhelices traced by tandem repeats with basal sequence of up to 12 base pairs.

P1.3.1 The method

In our treatment we will use the periodic *cgDNA* coefficients of Chapter P1.2. These coefficients model infinitely long repeating sequences. We have shown that they also provide a very good approximation of the standard *cgDNA* ground state configurations of finite DNA fragments far from the ends. This choice is motivated by some properties of the periodic coefficients that will be significant in what follows. The procedure is summarized in Figure P1.3.1.

We recall that by a tandem repeat we mean a DNA sequence $S_M = \overbrace{SS\ldots S}^{M}$ that is a result of concatenating $M \geq 1$ instances of any basal DNA sequence $S = \mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_N$ (with $\mathtt{X}_i \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$) of length $N \geq 1$. Let $\widehat{\boldsymbol{w}}_p(S)$ be the periodic ground state configuration vector of the basal sequence S. Let $\widehat{\mathcal{L}}_a \in \mathcal{SE}(3)$, $a \in \{1,\ldots N-1\}$ be the homogeneous coordinates (see Section A.1.2) of the step between base pair a and $(a+1)$ computed from $\widehat{\boldsymbol{w}}_p(S)$ through the reconstruction procedure described in Section B.1.3. Note that in the case of periodic coefficients the extra set of inter coefficients gives rise to the extra base pair step $\widehat{\mathcal{L}}_N$ that should be interpreted as the step between the last base pair of S and the first base pair of the subsequent instance of S in the tandem repeat.

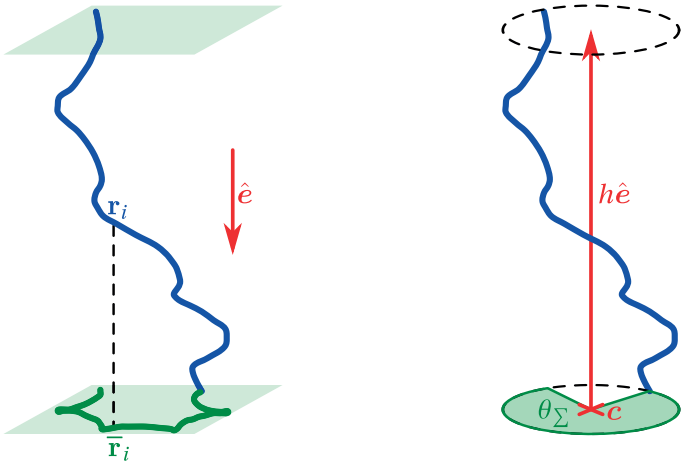


Figure P1.3.1. *A schematic summary of the procedure of computing superhelical pitch and radius. The blue curves represent the base pair positions* $\mathbf{r}_i$ *reconstructed from cgDNA periodic parameters for* 5 *repeats of a sequence of length* N *(see Equation (P1.3.1)).* $\hat{\boldsymbol{e}}$ *is the axis of the rotational displacement* $\boldsymbol{D} = \boldsymbol{D}_{N+1}$ *of a single repeat of Equation (P1.3.2). The green curve of panel (a) indicates the orthogonal projection* $\overline{\mathbf{r}}_i$ *of* $\mathbf{r}_i$ *to the plane perpendicular to* $\hat{\boldsymbol{e}}$ *(see Equation (P1.3.4b)). Note that in this particular example the elevation* $h = \mathbf{r}_{N+1} \cdot \hat{\boldsymbol{e}}$ *(see Equation (P1.3.4a)) indicated in panel (b) is negative, hence* $\hat{\boldsymbol{e}}$ *points down. Panel (b) indicates also the centre* $\boldsymbol{c}$ *of Equation (P1.3.9) as a red cross and the total angle* $\theta_\sum$ *of Equation (P1.3.14). The pitch* p *can be computed from* h *and* $\theta_\sum$ *using Equation (P1.3.15), while the radii as lengths of the displacements* $\widetilde{\mathbf{r}}_i = \overline{\mathbf{r}}_i - \boldsymbol{c}$ *(see Equations (P1.3.8), (P1.3.9) and (P1.3.12)).*

As shown by Lemma P1.2.2, the periodic ground state configuration vector $\widehat{\boldsymbol{w}}_p(S_M)$ of a tandem repeat can be constructed by simple concatenation of M instances of the ground state configuration vectors $\widehat{\boldsymbol{w}}_p(S)$ of the basal sequence. As a result the homogeneous coordinates of the rigid body displacement from base pair 1 of a tandem repeat S_M to base pair $m + (n-1)M$, *i.e.* base pair m in the nth instance of S ($n \in \{1, \dots M\}$, $m \in \{1, \dots N\}$), can be written as:

$$\begin{aligned}\mathcal{D}_{m+(n-1)M} = \mathcal{D}_{m:n} = \begin{bmatrix} \boldsymbol{D}_{m:n} & \mathbf{r}_{m:n} \\ \mathbf{0}^T & 1 \end{bmatrix} &:= \left(\prod_{k=1}^{N} \widehat{\mathcal{L}}_k\right)^{n-1} \prod_{k=1}^{m-1} \widehat{\mathcal{L}}_k \\ &= (\mathcal{D}_{1:2})^{n-1} \prod_{k=1}^{m-1} \widehat{\mathcal{L}}_k \end{aligned} \tag{P1.3.1}$$

with $\mathbf{0} \in \mathbb{R}^3$, $\boldsymbol{D}_{m:n}$ – the relative rotation and $\mathbf{r}_{m:n}$ – the relative translation. For brevity we also introduce the notation:

$$\mathcal{D} = \begin{bmatrix} \boldsymbol{D} & \mathbf{r} \\ \mathbf{0}^T & 1 \end{bmatrix} := \mathcal{D}_{N+1} = \mathcal{D}_{1:2} = \prod_{k=1}^{N} \widehat{\mathcal{L}}_k \quad . \tag{P1.3.2}$$

so that

$$\mathcal{D}_{1+nN} = \mathcal{D}_{1:n+1} = \mathcal{D}^n \quad . \tag{P1.3.3}$$

Most of the time we will use the double index notation $\mathcal{D}_{m:n}$ reverting to the single index notation $\mathcal{D}_k$ where it better illustrates the formula at hand.

In what follows we will only consider cases where the resulting superhelices have non-zero curvature. The other possibilities (*i.e.* $\boldsymbol{D} = \boldsymbol{I}$, $\mathbf{r} = \mathbf{0}$ and $\mathbf{r} \neq \mathbf{0}$ parallel to the axis of rotation of $\boldsymbol{D} \neq \boldsymbol{I}$) are discussed in Section P1.3.2.1. We also note that for some (but not all – see Figure P1.3.4m and P1.3.4n) superhelices very close to straight in the study of Section P1.3.3 the problem of distinguishing between the superhelix and the primary DNA double helix is ill-posed. In such cases, called *atypical*, the method reports pitch and radius of the primary doublehelix.

Let $\theta \in (0, \pi]$ denote the (right-handed) rotation angle of $\boldsymbol{D}$ and let $\hat{\boldsymbol{e}}$ denote the axis of the rotation with the direction implied by θ (see introduction to Section A.1.1). The angle is positive and the axis is well defined under our assumptions. The axis also remains unchanged under the rotation, as discussed in Section A.1.1.1.

Define

$$h_{m:n} := \mathbf{r}_{m:n} \cdot \hat{\boldsymbol{e}} \qquad h := h_{1:2} \tag{P1.3.4a}$$

$$\overline{\mathbf{r}}_{m:n} := \mathbf{r}_{m:n} - h_{m:n}\hat{\boldsymbol{e}} \qquad \overline{\mathbf{r}} := \overline{\mathbf{r}}_{1:2} \quad , \tag{P1.3.4b}$$

so that $\overline{\mathbf{r}}_{m:n}$ is the projection of $\mathbf{r}_{m:n}$ to a plane perpendicular to $\hat{\boldsymbol{e}}$ (see Figure P1.3.3). We will refer to the value of h as the *elevation* of the fragment. Note that the elevation can be negative.

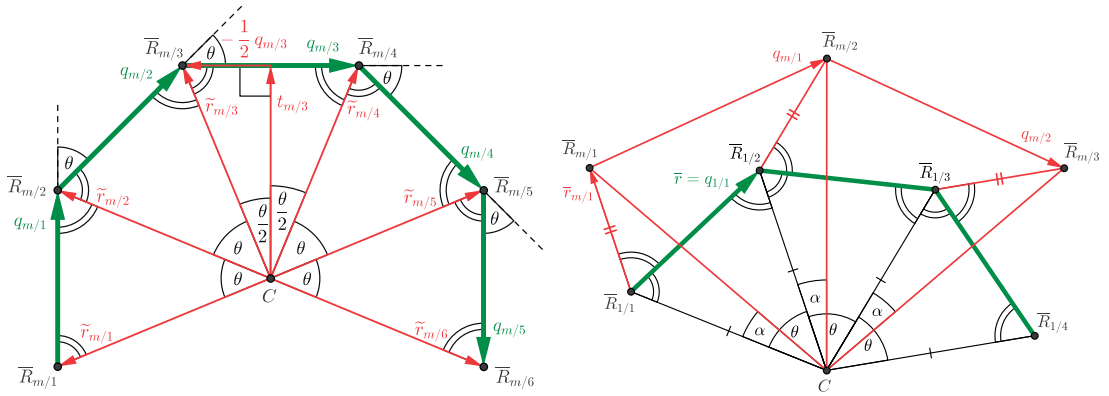


Figure P1.3.2. *In both panels* $\overline{R}_{m:n}$ *are points with coordinates given by the vectors* $\overline{\mathbf{r}}_{m:n}$ *respectively. Both panels show the plane perpendicular to the superhelical axis* $\hat{\boldsymbol{e}}$ *with the axis pointing out of the page; Panel (a) is an example illustrating the relations between the projections* $\overline{\mathbf{r}}_{m:n}$ *and the displacements* $\widetilde{\mathbf{r}}_{m:n}$ *in their common plane.* $\overline{\mathbf{r}}_{1:n}$ *are shown as a sum of* $\boldsymbol{q}_{1:\cdot}$ *(see Equation (P1.3.6)). The point* C *is constructed in such a way that the triangle* $\overline{R}_{m:1}C\overline{R}_{m:2}$ *is isosceles with* $\left|\angle\,\overline{R}_{m:1}C\overline{R}_{m:2}\right| = \theta$*. As a result* $\left|\angle C\overline{R}_{m:2}\overline{R}_{m:1}\right| = \frac{1}{2}(\pi-\theta)$*. The angle between* $\boldsymbol{q}_{m:1}$ *and* $\boldsymbol{q}_{m:2}$ *is* θ *(after Equation (P1.3.6)) and so* $\left|\angle C\overline{R}_{m:2}\overline{R}_{m:3}\right| = \frac{1}{2}(\pi-\theta)$*. This and the fact that* $|\boldsymbol{q}_{1:1}| = |\boldsymbol{q}_{1:2}|$ *(again from Equation (P1.3.6)) and* $\left|\overline{R}_{m:1}C\right| = \left|\overline{R}_{m:2}C\right|$ *(by construction) implies that the triangles:* $\overline{R}_{m:1}C\overline{R}_{m:2}$ *and* $\overline{R}_{m:2}C\overline{R}_{m:3}$ *are congruent. Iteration of the above reasoning shows that all* $\overline{R}_{m:n}$ *(for a fixed* m*) lie on a circle centred at* C*. The figure also explains the meaning of the vectors* $\boldsymbol{t}_{m:n}$ *of Equation (P1.3.8) on an example with* $n = 3$*. Panel (b) illustrates the fact that all helices traced by* $\mathbf{r}_{m:\cdot}$*,* $m \in \{1,\ldots,M\}$ *are coaxial. Let* C *be the centre of the circle through the projections* $\overline{R}_{1:\cdot}$*. Note that* $\left|\angle\overline{R}_{m:1}\overline{R}_{1:1}\overline{R}_{1:2}\right| = \left|\angle\overline{R}_{m:2}\overline{R}_{1:2}\overline{R}_{1:3}\right|$ *and* $|\overline{R}_{1:1}\overline{R}_{m:1}| = |\overline{R}_{1:2}\overline{R}_{m:2}|$ *(which follows from Equation (P1.3.5). By construction also* $\left|\angle C\overline{R}_{1:1}\overline{R}_{1:2}\right| = \left|\angle C\overline{R}_{1:2}\overline{R}_{1:3}\right|$ *and* $\left|C\overline{R}_{1:1}\right| = \left|C\overline{R}_{1:2}\right|$*. As a result the triangles:* $\overline{R}_{m:1}C\overline{R}_{m:1}$ *and* $\overline{R}_{m:2}C\overline{R}_{m:2}$ *are congruent and so the triangle* $\overline{R}_{1:1}C\overline{R}_{m:1}$ *is isosceles with* $|\angle\overline{R}_{1:2}C\overline{R}_{m:2}| = \theta$*. This implies that the points* $\overline{R}_{m:1}$ *and* $\overline{R}_{m:2}$ *lie on a circle centred at* C*. Again, by repeating the same argument it can be shown that all the projections* $\overline{R}_{m:\cdot}$ *lie on that circle. Thanks to Przemysław Głowacki for creating the figures.*

A direct computation using (P1.3.1), (P1.3.2) and (P1.3.3) and the above decomposition of $\mathbf{r}_{m:n}$ leads to:

$$\begin{aligned}\mathbf{r}_{m:n} &:= \left(\sum_{k=0}^{n-2}\boldsymbol{D}^k\mathbf{r}\right) && + \boldsymbol{D}^{n-1}\mathbf{r}_{m:1}\\ &= \left((n-1)h\hat{\boldsymbol{e}} + \sum_{k=0}^{n-2}\boldsymbol{D}^k\overline{\mathbf{r}}\right) && + h_{m:1}\,\hat{\boldsymbol{e}} + \boldsymbol{D}^{n-1}\overline{\mathbf{r}}_{m:1}\end{aligned} \qquad \text{(P1.3.5)}$$

and so the step from the mth base pair of the nth instance of S in S_M to the mth base

pair of the $(n+1)$st instance of S can be written as:

$$\begin{aligned}\mathbf{r}_{m:n+1} - \mathbf{r}_{m:n} &= h\hat{\boldsymbol{e}} \;+\; \boldsymbol{D}^{n-1}\overline{\mathbf{r}} \;+\; \boldsymbol{D}^{n}\overline{\mathbf{r}}_{m:1} \;-\; \boldsymbol{D}^{n-1}\overline{\mathbf{r}}_{m:1} \\ &= h\hat{\boldsymbol{e}} \;+\; \boldsymbol{D}^{n-1}\left(\overline{\mathbf{r}} \;+\; \boldsymbol{D}\overline{\mathbf{r}}_{m:1} \;-\; \overline{\mathbf{r}}_{m:1}\right) \\ &=: h\hat{\boldsymbol{e}} \;+\; \boldsymbol{q}_{m:n} \\ &= h\hat{\boldsymbol{e}} \;+\; \boldsymbol{D}^{n-1}\boldsymbol{q}_{m:1} \end{aligned} \tag{P1.3.6}$$

and for the first base pair:

$$\boldsymbol{q}_{1:1} = \overline{\mathbf{r}} \quad . \tag{P1.3.7}$$

Note that because each $\overline{\mathbf{r}}_{m:n}$ lies in the plane perpendicular to the rotation axis $\hat{\boldsymbol{e}}$ of $\boldsymbol{D}$, so does $\boldsymbol{D}\overline{\mathbf{r}}_{m:n}$. As a result also each $\boldsymbol{q}_{m:n}$ is perpendicular to $\hat{\boldsymbol{e}}$.

This means all the projections $\overline{\mathbf{r}}_{m:\cdot}$ of base pair positions $\mathbf{r}_{m:\cdot}$ can be constructed from $\overline{\mathbf{r}}_{m:1}$ by subsequently adding the offset $\boldsymbol{q}_{m:1}$ rotated in the plane by an extra θ in each step. As a result all those base pair positions lie on a circle (see Figure P1.3.2a). Additionally, a step along $\hat{\boldsymbol{e}}$ between base pair position $\mathbf{r}_{m:n}$ and $\mathbf{r}_{m:n+1}$ is the same for any $m \in \{1,\ldots,N\}$ an $n \in \{1,\ldots,M-1\}$ and is equal to h. All the above shows that indeed for any m all the mth base pairs of all the instances of S in S_N lie on a helix whose axis is parallel to $\hat{\boldsymbol{e}}$. Furthermore all of those helices have the same pitch. Figure P1.3.2b shows also that all the helices are coaxial and indicates that the displacement $\widetilde{\mathbf{r}}_{m:n}$ of base pair position $\mathbf{r}_{m:n}$ from the superhelical axis can be written as:

$$\begin{aligned}\widetilde{\mathbf{r}}_{m:n} &:= -\frac{1}{2}\boldsymbol{q}_{m:n} \;-\; \frac{1}{2}\cot\left(\tfrac{\theta}{2}\right)\hat{\boldsymbol{e}}\times\boldsymbol{q}_{m:n} \\ &=: -\frac{1}{2}\boldsymbol{q}_{m:n} \;+\; \boldsymbol{t}_{m:n} \quad , \end{aligned} \tag{P1.3.8}$$

so that the common centre $\boldsymbol{c}$ of all the circles that are projections of the helices on a plane perpendicular to $\hat{\boldsymbol{e}}$ can be found as:

$$\begin{aligned}\boldsymbol{c} = \overline{\mathbf{r}}_{m:n} - \widetilde{\mathbf{r}}_{m:n} &= \overline{\mathbf{r}}_{m:n} + \frac{1}{2}\left(\boldsymbol{q}_{m:n} \;-\; \boldsymbol{t}_{m:n}\right) \\ = \qquad -\widetilde{\mathbf{r}}_{1:1} \quad &= \qquad \frac{1}{2}\left(\overline{\mathbf{r}} \;+\; \cot\left(\tfrac{\theta}{2}\right)\hat{\boldsymbol{e}}\times\overline{\mathbf{r}}\right) \end{aligned} \tag{P1.3.9}$$

with the last equality coming from Equation (P1.3.7) and the fact that $\mathbf{r}_{1:1} = \overline{\mathbf{r}}_{1:1} = \mathbf{0}$ (as translations from the first base pair to itself). Note that the above provides an alternative formula to compute $\widetilde{\mathbf{r}}_{m:n}$:

$$\widetilde{\mathbf{r}}_{m:n} = \overline{\mathbf{r}}_{m:n} + \widetilde{\mathbf{r}}_{1:1} \quad . \tag{P1.3.10}$$

Note also that:

$$\left|\widetilde{\mathbf{r}}_{m:n}\right| = \left|\widetilde{\mathbf{r}}_{m:n'}\right| \tag{P1.3.11}$$

for any $m \in \{1,\ldots N\}$ and $n, n' \in \{1,\ldots M\}$.

In particular the values:

$$r_{\max} := \max_{m\in\{1,\dots N\}} \{\, |\widetilde{\mathbf{r}}_{m:\cdot}| \,\} \qquad \text{and} \qquad r_{\min} := \min_{m\in\{1,\dots N\}} \{\, |\widetilde{\mathbf{r}}_{m:\cdot}| \,\} \tag{P1.3.12}$$

are the radii of the two cylinders (along $\hat{\boldsymbol{e}}$, centred at $\boldsymbol{c}$) encapsulating all the base pair positions between them.

Let $\theta_m \in (-\pi, \pi)$ be the angle of the right-handed (with respect to $\hat{\boldsymbol{e}}$) rotation between $\widetilde{\mathbf{r}}_{m:n}$ and $\widetilde{\mathbf{r}}_{m+1:n}$ ($m \in \{1, \dots N-1\}$) that can be calculated as:

$$\theta_m := \operatorname{sgn}\left((\widetilde{\mathbf{r}}_{m:n} \times \widetilde{\mathbf{r}}_{m+1:n}) \cdot \hat{\boldsymbol{e}}\right) \frac{\operatorname{acos}\left(\widetilde{\mathbf{r}}_{m:n} \cdot \widetilde{\mathbf{r}}_{m+1:n}\right)}{|\widetilde{\mathbf{r}}_{m:n}|\,|\widetilde{\mathbf{r}}_{m+1:n}|} \tag{P1.3.13}$$

(the angle is obviously the same for any choice of $n \in \{1, \dots M\}$). Note that here we explicitly exclude the case of $\theta_m = \pm\pi$, where the above formula cannot be used (see Section P1.3.2.2). The sign of the projection of the cross product on the axis $\hat{\boldsymbol{e}}$ indicates whether the rotation is right-handed (+) or left handed (−) with respect to the axis. The angle:

$$\theta_\sum := \sum_{k=1}^{M-1} \theta_m \tag{P1.3.14}$$

is, then, the total signed angle the base pairs of S trace around $\hat{\boldsymbol{e}}$.

Note that the value of h (of Equation (P1.3.6)) is the total elevation along the superhelical axis $\hat{\boldsymbol{e}}$ of a single instance of S, while $\theta_\sum$ (as defined above) is the total angle, both with sign. As a result the common pitch of each of the N coaxial helices traced by repeats of each of the N base pairs of S can be calculated as:

$$p := \frac{2\pi}{\theta_\sum} h \quad . \tag{P1.3.15}$$

The sign of the pitch as defined above depends on the sign of both h and $\theta_\sum$ in such a way that the pitch is positive for right-handed helices and negative for left-handed helices (see the discussion of Section P1.3.2.2).

Note that the values of the radii $r_{\min}$, $r_{\max}$ and p depend only in the basal sequence S and not on the number M of repeats in the tandem repeat S_M (see Section P1.3.2.3). The values are also the same for the Watson-Crick complement $\overline{S}$ of S as well as any cyclic shift of S (see Section P1.3.2.4).

For completeness we also state here the formulae to compute the curvature κ and torsion τ for a helix with the pitch p and radius r and *vice versa* (see *e.g.* [ChoMad2004]):

$$\kappa = \frac{r}{r^2 + \frac{p^2}{4\pi^2}} \qquad\qquad \tau = \frac{p}{2\pi r^2 + \frac{p^2}{2\pi}} \tag{P1.3.16a}$$

$$r = \frac{\kappa}{\kappa^2 + \tau^2} \qquad\qquad p = \frac{2\pi\tau}{\kappa^2 + \tau^2} \quad . \tag{P1.3.16b}$$

Another quantity that will be useful in the discussion of the study of Section P1.3.3 is the minimum number of repeats M_t of the basal sequence S for which the superhelix of the tandem repeat S_{M_t} has at least one full turn. This can be computed as:

$$M_t := \left\lceil \left| \frac{p}{h} \right| \right\rceil = \left\lceil \left| \frac{2\pi}{\theta_\Sigma} \right| \right\rceil \quad . \tag{P1.3.17}$$

Finally we denote the length (in number of base pairs) of the tandem repeat S_{M_t} by:

$$T = NM_t \quad , \tag{P1.3.18}$$

where N is the length of the basal sequence S.

P1.3.2 Discussion of the method

In this section we will outline certain features of the presented method of calculating pitch and radius (radii) of the periodic *cgDNA* ground state configuration vector of a DNA tandem repeat S_M.

P1.3.2.1 Degenerate helices

The outline of the method was based on the assumption that $\boldsymbol{D} \neq \boldsymbol{I}$ and $\mathbf{r} \neq \mathbf{0}$ and $\mathbf{r}$ not parallel to the axis $\hat{\boldsymbol{e}}$. Here we briefly discuss the degenerate cases.

First of all note that for $\boldsymbol{D} \neq \boldsymbol{I}$ and $\mathbf{r} \neq \mathbf{0}$ and $\mathbf{r} \parallel \hat{\boldsymbol{e}}$ the resulting shape is exactly straight and twisted.

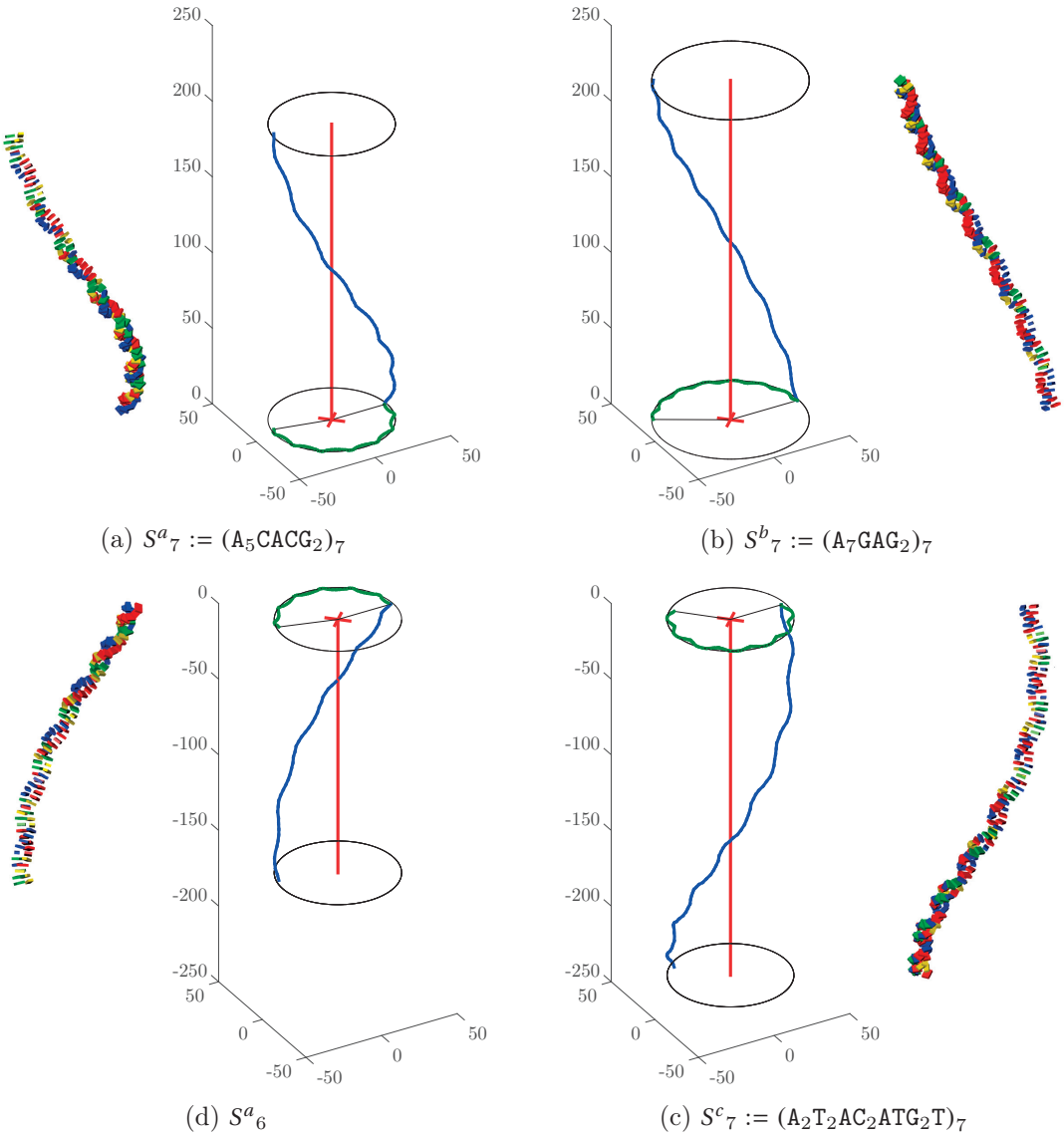
If $\boldsymbol{D} = \boldsymbol{I}$ the rotation axis is undefined (see Section A.1.1.1). In such a case the DNA superhelix degenerates to a straight and untwisted configuration, with the direction of the centerline given by $\mathbf{r}$.

On the other hand if $\mathbf{r} = \mathbf{0}$ the intrinsic shape of the molecule forms a closed loop (with possible rotational misalignment of the first and last base pair). From Equation (P1.3.5) it is clear that subsequent instances of the basal sequence in the tandem repeat will only be rotated and not translated and form petals of a flower-like structure.

In the most degenerate case where both $\boldsymbol{D} = \boldsymbol{I}$ and $\mathbf{r} = \mathbf{0}$ subsequent instances of the basal sequence in the tandem repeat lie on top of one another.

All of the mentioned cases are extremely unlikely for relatively short basal sequences (< 100 bp) and would require special analysis. None such case has been found in the comprehensive study of oligomers up to 12 bp long presented in Section P1.3.3.

P1.3.2.2 Considerations on the angle $\theta_{\sum}$ and chirality of the superhelix



(a) $S^a{}_7 := (\mathtt{A_5CACG_2})_7$ (b) $S^b{}_7 := (\mathtt{A_7GAG_2})_7$

(d) $S^a{}_6$ (c) $S^c{}_7 := (\mathtt{A_2T_2AC_2ATG_2T})_7$

Figure P1.3.3. *Examples of sequences for which the superhelices are left-handed (left column) and right-handed (right column), where the elevation* $\boldsymbol{h}$ *is either positive (top row) or negative (bottom row). In all the plots the centre* $\boldsymbol{c}$*, indicated by the red cross, is chosen to be the centre of the coordinate system.* x*-axis is aligned with the displacement* $\widetilde{\mathbf{r}}_{1:1}$*, the* z*-axis is chosen to be the rotation axis* $\hat{\boldsymbol{e}}$ *and* y*-axis completes the other two. The base pair positions* $\mathbf{r}_{m:n}$ *are indicated in blue. Their projections* $\overline{\mathbf{r}}_{m:n}$ *to the plane perpendicular to* $\hat{\boldsymbol{e}}$ *are indicated in green. The vector* $\boldsymbol{h}\hat{\boldsymbol{e}}$ *is indicated in red. The black circles, shown for clarity, are of radius* $r_{\max}$*. For better perception of length scale a box visualization of bases has been shown. Each box is the bounding box of centres of all atoms of the respective base and has a height of* 1.5 Å*.* *(Continued on the following page)*

Figure P1.3.3. *(Continued from the previous page) All sequences have been chosen as the ones with with* $35\,\text{Å} < r_{\max} < 55\,\text{Å}$ *and the smallest (in absolute value) pitch among all decanucleotides (*S^a*), undecanucleotides (*S^b*) and dodecanucleotides (*S^c*). Note, in particular the difference between Panel (a) and (d). As explained in the text the values of the radii* $r_{\min}$*,* $r_{\min}$ *and the pitch* p *are the same for any number of repeats of the basal sequence, but the sign of* $h(S^a{}_N)$ *(and of* $\theta_\sum(S^a{}_N)$ *accordingly) may be different for different number of repeats* M *(see Equations (P1.3.21) and (P1.3.22)). The numerical values characterizing the presented superhelices are given in Table P1.3.1*

Sequence	p	$r_{\min}$	$r_{\max}$	h	$\theta_\sum$	handedness
S^a				-27.91	0.48	
$S^a{}_6$	-367.02	34.09	36.47	-167.48	2.87	left
$S^a{}_7$				195.39	-3.35	
S^b	539.75	42.94	45.09	32.04	0.37	right
$S^b{}_7$				224.27	2.61	
S^c	374.38	33.04	36.30	33.59	0.56	right
$S^c{}_7$				-235.16	-3.95	

Table P1.3.1. *The numerical values characterizing the sequences of Figure P1.3.3. Note in particular that the absolute value of the total angle* $|\theta_\sum(S^a{}_6)|$ *is just below* π *and so it is the case of (P1.3.22a). On the other hand* $|\theta_\sum(S^a{}_6)|$ *is just above* π*, which makes it (P1.3.22b). The values of* $r_{\max}$*,* $r_{\min}$ *and* p *are the same for any number of repeats of the basal sequence.*

As already mentioned previously the angle $\theta_\sum$ represents the total angle the base pair positions trace around the superhelical axis, including the direction. The total angle is signed to be positive under right-handed with respect to the direction of $\hat{\boldsymbol{e}}$ and, unlike the angle θ of the relative rotation $\boldsymbol{D}$, is not limited to $[0, \pi]$, but can possibly take any real value. Together with the sign of h (the component of the total elevation introduced in Equation (P1.3.6)) the sign of $\theta_\sum$ is used to recover the handedness of the resulting superhelix that is encoded in the sign of the pitch p, as discussed before. Figure P1.3.3 and Table P1.3.1 show examples of all 4 possible cases of pairs of $\text{sgn}\,(\theta_\sum)$ and $\text{sgn}(h)$.

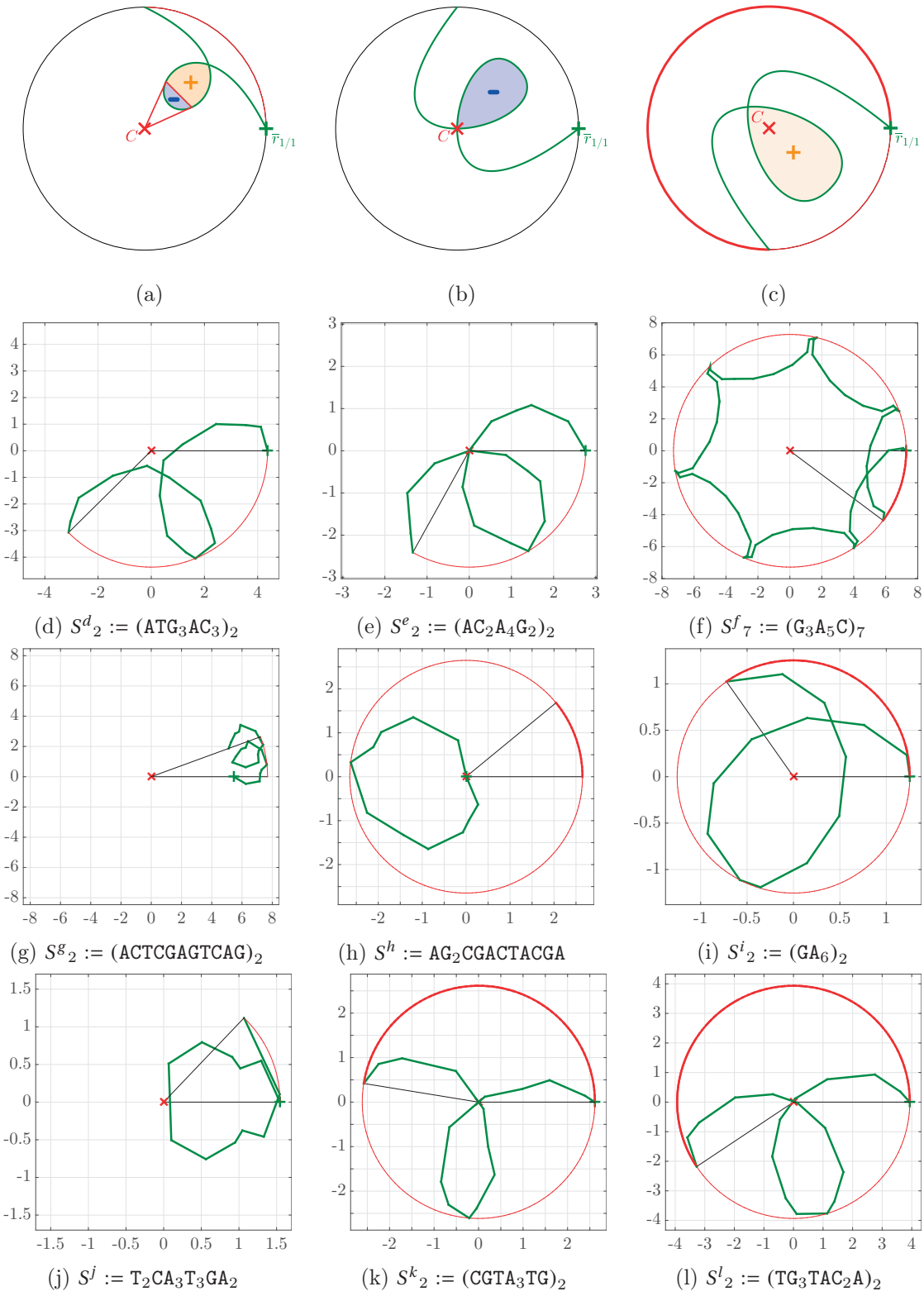
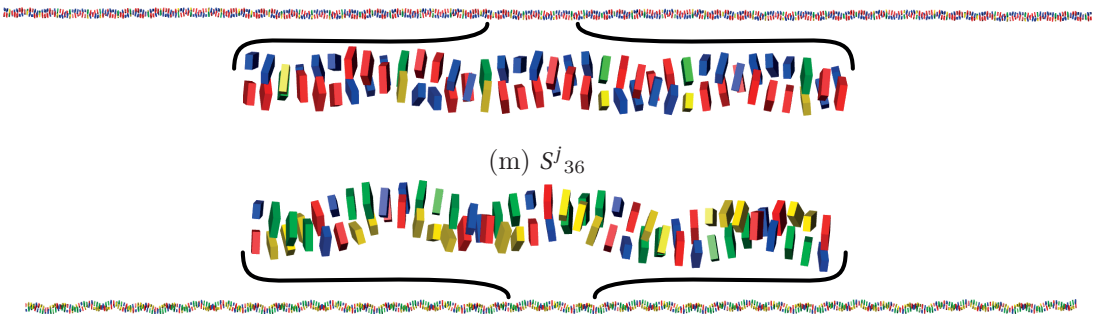


Figure P1.3.4. *See caption on the following page*



(m) $S^{j}{}_{36}$

(n) $S^{l}{}_{48}$

Figure P1.3.4. *Schematic pictures of the three categories of loops introduced in the text and example sequences forming them: category I – left column, category II – middle column, category III – right column. For each category example sequences from the ensemble of Section P1.3.3 are given, forming a left-handed (second row) and right-handed (third row) superhelix. All numbers are in* Å*. As noted in the text no example for category III has been found in the ensemble. Three closest examples are shown instead (see details below). Each picture shows the plane perpendicular to the superhelical axis* $\hat{\mathbf{e}}$ *with the axis pointing out of the page and the centre* $\boldsymbol{c}$ *of the superhelix indicated by a red* ×*. The projections* $\overline{\mathbf{r}}_{:}$. *of base pair positions are indicated in green with* + *marking* $\overline{\mathbf{r}}_{1:1}$*. The angle* $\theta_{\sum}$ *traced by the projections is indicated in red with thicker line indicating double covering. In panel (a) the two fragments of the loop that contribute the same angle with opposite sign are indicated by shading. In panel (b) the loop contributes a positive angle* $< \pi$*, while the contribution from the loop in panel (c) is* $+2\pi$*. The sequence* S^{d} *has been chosen as the one with maximum value of* $(r_{\max} - r_{\min})$ *of all typical (see text) nonanucleotides. The sequence* S^{g} *is the undecanucleotide with the maximal (in absolute value) elevation* h*. The sequence* S^{j} *has the minimum radius* $r_{\max}$ *of all typical superhelices in the ensemble. The sequence happens to be its own complement. The sequence* S^{e} *has the smallest value of* $r_{\min} = 0.012$ Å *among all left-handed superhelices. In fact the loop belongs to category I which can be deduced from the angle it traces. The sequence* S^{h} *has the smallest value of* $r_{\min} = 0.0037$ Å *among all studied superhelices. Again the angle shows the loop falls into category III. The sequence* S^{k} *is an example close to the special case of category II where the centre of the superhelix lies on the line segment connecting* $\overline{\mathbf{r}}_{k}$ *and* $\overline{\mathbf{r}}_{k}$ *(see text). The sequence has been chosen as the one with the largest angle* $\max_{m\in\{1,\ldots N\}}\{|\theta_m|\} = 3.1413$ *(for an oligomer of length* N*) in the entire ensemble. The actual category of the loop is III. The sequence* S^{f} *has the largest radius* $r_{\max}$ *among all nonanucleotides. This particular example shows that for any sequence repeated at least* M_t *times (see Equation (P1.3.17)) a loop of category III is formed (apart from possibly any others).* S^{i} *is the heptanucleotide with the highest (in magnitude) value of the total angle* $\theta_{\sum}$*. The sequence* S^{l} *has the largest radius of all atypical superhelices. Panels P1.3.4m and P1.3.4n show a comparison of tandem repeats of* S^{j} *(with smallest* $r_{\max} = 1.55$ Å *of all typical superhelices) and* S^{l} *(with largest* $r_{\max} = 3.93$ Å *of all atypical superhelices), respectively. The coloured boxes as in Figure P1.3.3. The point here is to show that there is no drastic difference between atypical superhelices and typical ones with small radius (see text). The numerical values characterizing the superhelices traced by ground state configurations of the presented sequences are given in Table P1.3.2. Thanks to Przemysław Głowacki for creating the figures of panels (a), (b) and (c).*

Note, however, that due to the double-helical secondary structure of B-DNA the projections $\overline{\mathbf{r}}_{.:.}$ of base pair positions (indicated in green in Figure P1.3.3) onto the plane perpendicular to the axis $\hat{\boldsymbol{e}}$ of the superhelix may form loops in that plane. The position of those loops with respect to the centre $\boldsymbol{c}$ (see Equation (P1.3.9)), as shown in Figure P1.3.4 is of particular importance. Specifically we separate all the possibilities into three categories:

I $\boldsymbol{c}$ lying outside of the loop (left column of Figure P1.3.4). As shown in Figure P1.3.4a, the total contribution of such a loop to the total angle $\theta_\sum$ is 0, as the indicated two parts of the loop contribute an angle of the same magnitude but opposite handedness (sign).

II $\boldsymbol{c}$ lying on the loop (Figure P1.3.4b). This can be seen as a singular border case of the other two as the minimum radius $r_{\min}$ goes to 0. The singularity is easy to see as the contributions to the total angle $\theta_\sum$ coming from I loops are 0 and those coming from III loops are $\pm 2\pi$ independent of how small the $r_{\min} = \varepsilon > 0$ is. For loops of category II their contribution to $\theta_\sum$ can (in principle) be anywhere from 0 to 2π. The fact that $r_{\min} = |\widetilde{\mathbf{r}}_k| = 0$ for a given oligomer means that the total angle $\theta_\sum$ cannot be computed using formulae (P1.3.13) and (P1.3.14).

It should be noted that this kind of loop is much less likely to be found for relatively short sequences than I or III. There is no such case among oligomers of length up to 12. Figures P1.3.4e and P1.3.4h present examples of superhelices with smallest values of $r_{\min}$ in the ensemble of Section P1.3.3.

A particular type of loops that could also be included in category II are loops where the centre $\boldsymbol{c}$ lies exactly on the line segment connecting projections $\overline{\mathbf{r}}_k$ and $\overline{\mathbf{r}}_{k+1}$ of subsequent base pairs but is not one of the end points. In such a case the displacement vectors $\widetilde{\mathbf{r}}_k$ and $\widetilde{\mathbf{r}}_{k+1}$ are collinear and of opposite direction and so the angle θ_k between them is $\pm\pi$. On the other hand in such case the formula (P1.3.13) for θ_m evaluates to 0, so the presented method cannot be used. Among all oligomers of length at most 12 there is no such case. Figure P1.3.4k shows an example of a superhelix with $\max_{m\in\{1,\ldots N\}} \{|\theta_m|\} = 3.1413$ (for an oligomer of length N) which is the closest to π in the studied ensemble.

III $\boldsymbol{c}$ lying inside the loop (right column of Figure P1.3.4). In such a case the loop adds an entire 2π (or -2π, depending on the handedness) to the total angle $\theta_\sum$.

Note that, as shown in Figure P1.3.4f, for any oligomer repeated at least M_t number of times (see Equation (P1.3.17)) a loop of that type (apart from possibly any others) is formed by the tandem repeat. This is directly related to the discussion of Section P1.3.2.3, in particular the part of Equations (P1.3.22) relating to $\theta_\sum$ and the invariance of pitch p .

Another very important remark is that for some sequences a loop of this type is formed over a fragment of as few as 11 bp, *i.e.* around a single repeat of the double

helix. In such cases the superhelix is so close to straight that the presented method simply picks up the *primary helix* (of the DNA secondary structure) instead. In such cases the problem of distingushing between the primary helix and the superhelix is ill-posed. Due to the singularity that separates category III from I (discussed above) the method introduces a sharp distinction of such oligomers. In the discussion of Section P1.3.3 we will refer to such basal sequences as *atypical*, while the others will be called *typical*. As shown in Figures P1.3.4m and P1.3.4n there is no clear difference in the geometry of atypical superhelices (which all have $r_{\max} \leq 3.93\,\text{Å}$) and typical ones comparably close to straight. In fact, all superhelices with relatively small $r_{\max}$ (comparable to the size of a base, *i.e.* $\sim 5\,\text{Å}$) can be seen as straight, despite the fact that the method is able to distinguish the superhelix from the primary helix. See Section P1.3.3 for further discussion.

Sequence	p	$r_{\min}$	$r_{\max}$	h	θ_Σ	handedness
S^d	-153.43	0.59	4.37	-28.84	1.18	left
S^e	-177.67	0.012	2.76	-29.37	1.04	left
S^f $S^f{}_7$	-182.78	4.91	7.28	-28.78 -201.44	0.99 6.92	left
S^g	1304.08	5.45	7.69	36.35	0.18	right
S^h	35.34	0.0037	2.65	39.20	6.97	right
S^i	34.01	0.60	1.25	-22.92	-4.23	right
S^j	308.05	0.52	1.55	39.65	0.81	right
S^k	35.20	0.18	2.61	-25.95	-4.63	right
S^l	36.05	0.13	3.93	-28.72	-5.01	right

Table P1.3.2. *The numerical values characterizing the sequences of Figure P1.3.4.*

P1.3.2.3 Invariance of pitch and radius for different number of repeats

Let $\boldsymbol{D}(S_M)$ be the relative rotation of Equation (P1.3.2) with the axis $\hat{\boldsymbol{e}}(S_M)$ and angle $\theta(S_M)$, $\theta_\sum(S_M)$ the total angle, $h(S_M)$ the total elevation and $\widetilde{\mathbf{r}}_{m:n}(S_M)$ the displacement from the superhelical axis of the position of the nth instance of base pair m, all defined for the tandem repeat S_M $(M \geq 1)$ so that:

$$\boldsymbol{D}(S_M) = (\boldsymbol{D}(S_1))^M = (\boldsymbol{D}(S))^M \quad . \tag{P1.3.19}$$

Note that it is clear from the statement of the method that $\widetilde{\mathbf{r}}_{m:n}(S_M) = \widetilde{\mathbf{r}}_{m:1}(S)$ for any number of repeats M and so:

$$\begin{aligned} r_{\min}(S_M) &= r_{\min}(S) \\ r_{\max}(S_M) &= r_{\max}(S) \quad . \end{aligned} \tag{P1.3.20}$$

On the other hand for different values of N the axes $\hat{\boldsymbol{e}}(S_M)$ are necessarily parallel (as eigenvectors – see Section A.1.1.1) but may be of opposite sign. This is due to the fact that (to avoid ambiguity, as discussed in Chapter A.1) the rotation angle $\theta(S_M)$ is defined only in $[0, \pi]$. As a result:

$$\begin{cases} \hat{\boldsymbol{e}}(S_M) = \hat{\boldsymbol{e}}(S) \\ \theta(S_M) = \Big((M \cdot \theta(S)) \bmod 2\pi\Big) \end{cases} \quad \text{if} \quad (M \cdot \theta(S)) \bmod 2\pi \leq \pi \tag{P1.3.21a}$$

$$\begin{cases} \hat{\boldsymbol{e}}(S_M) = -\hat{\boldsymbol{e}}(S) \\ \theta(S_M) = 2\pi - \Big((M \cdot \theta(S)) \bmod 2\pi\Big) \end{cases} \quad \text{if} \quad (M \cdot \theta(S)) \bmod 2\pi > \pi \tag{P1.3.21b}$$

The direction of the axis affects the sign of each partial angle $\theta_m(S_M)$ of Equation (P1.3.13) and so the sign of the total angle $\theta_\sum(S_M)$. Exactly in the same way it affects the sign of the elevation $h(S_M)$, so that:

$$\begin{cases} \theta_\sum(S_M) &= M \cdot \theta_\sum(S) \\ h(S_M) &= M \cdot h(S) \end{cases} \quad \text{if} \quad \hat{\boldsymbol{e}}(S_M) \cdot \hat{\boldsymbol{e}}(S) = 1 \tag{P1.3.22a}$$

$$\begin{cases} \theta_\sum(S_M) &= -M \cdot \theta_\sum(S) \\ h(S_M) &= -M \cdot h(S) \end{cases} \quad \text{if} \quad \hat{\boldsymbol{e}}(S_M) \cdot \hat{\boldsymbol{e}}(S) = -1 \quad . \tag{P1.3.22b}$$

This finally leads to:

$$p(S_M) = \frac{2\pi}{M \cdot \theta_\sum(S)} M \cdot h(S) = p(S) \tag{P1.3.23}$$

Figure P1.3.3 and Table P1.3.1 show an example for the sequence $S^a = \mathtt{A_5CACG_2}$. For ${S^a}_6$ of Figure P1.3.3d the absolute value of the total angle is $|\theta_\sum({S^a}_6)|$ is just below π and so we have case (P1.3.22a). For ${S^a}_7$ of Figure P1.3.3a, however, $|\theta_\sum({S^a}_7)|$ is just above π which gives (P1.3.22b). The pitch in both cases is the same and equal to $p(S^a)$.

The properties (P1.3.20) and (P1.3.23) are desired and important features of the presented method that show its consistency.

P1.3.2.4 Invariance of pitch and radius under Watson-Crick symmetry and cyclic shifts of sequence

The Watson-Crick symmetry of the *cgDNA* model [Pet2012; GonPetMad2013] assures that the same values of the radii $r_{\min}$, $r_{\max}$ and the pitch p are found for the Watson-Crick complement of S, namely $\overline{S} = \overline{\mathtt{X}}_N\overline{\mathtt{X}}_{N-1}\ldots\overline{\mathtt{X}}_2\overline{\mathtt{X}}_1$.

What is maybe less evident is that the same three values characterize any cyclic shift of the basal sequence S. Consider the sequence $S' = \mathtt{X}_N\mathtt{X}_1\mathtt{X}_2\ldots\mathtt{X}_{N-1}$ (the cyclic shift of S to the right by one base pair). As indicated by Lemma P1.2.3 the periodic ground state configuration vector for the sequence S' can be obtained by simply "moving" the appropriate set of inter and intra coefficient from the back to the front of the periodic ground state configuration vector of S. As a result the homogeneous coordinates $\mathcal{D}'_k$ of the rigid body displacement from base pair 1 to base pair k for the sequence S' can be written as:

$$\mathcal{D}'_k =: \begin{bmatrix} \boldsymbol{D}'_k & \mathbf{r}'_k \\ \mathbf{0}^T & 1 \end{bmatrix} = \widehat{\mathcal{L}}_N \mathcal{D}_{k-1} \tag{P1.3.24}$$

with $\mathbf{0} \in \mathbb{R}^3$ $\widehat{\mathcal{L}}_N$ as introduced in the previous section and $\mathcal{D}_{k-1}$ describing the rigid body motion from base pair 1 and $k-1$ for the sequence S (see Equation (P1.3.1)). Therefore the rotation $\boldsymbol{D}' := \boldsymbol{D}'_{N+1}$ can be written with respect to $\boldsymbol{D} := \boldsymbol{D}_{N+1}$ as:

$$\boldsymbol{D}' = \widehat{\boldsymbol{L}}_N \boldsymbol{D} \widehat{\boldsymbol{L}}_N^T \quad , \tag{P1.3.25}$$

with $\widehat{\boldsymbol{L}}_N$ the rotational part of $\widehat{\boldsymbol{L}}_N$. The axis of rotation of $\boldsymbol{D}'$ (and so the superhelical axis) is simply $\hat{\boldsymbol{e}}' = \widehat{\boldsymbol{L}}_N\hat{\boldsymbol{e}}$, which can be verified by simple direct calculation.

Let $\widehat{\boldsymbol{L}}_N$ be the rotation matrix associated with $\widehat{\mathcal{L}}_N$ so that from Equations (P1.3.25), (P1.3.5) and (P1.3.6) we have:

$$\mathbf{r}'_{m:n+1} - \mathbf{r}'_{m:n} = \widehat{\boldsymbol{L}}_N \left(\mathbf{r}'_{m:n+1} - \mathbf{r}'_{m:n}\right) \tag{P1.3.26}$$

$$= h\widehat{\boldsymbol{L}}_N\hat{\boldsymbol{e}} + \widehat{\boldsymbol{L}}_N\boldsymbol{q}_{m:n} \tag{P1.3.27}$$

$$:= h\hat{\boldsymbol{e}}' + \boldsymbol{q}'_{m:n} \quad . \tag{P1.3.28}$$

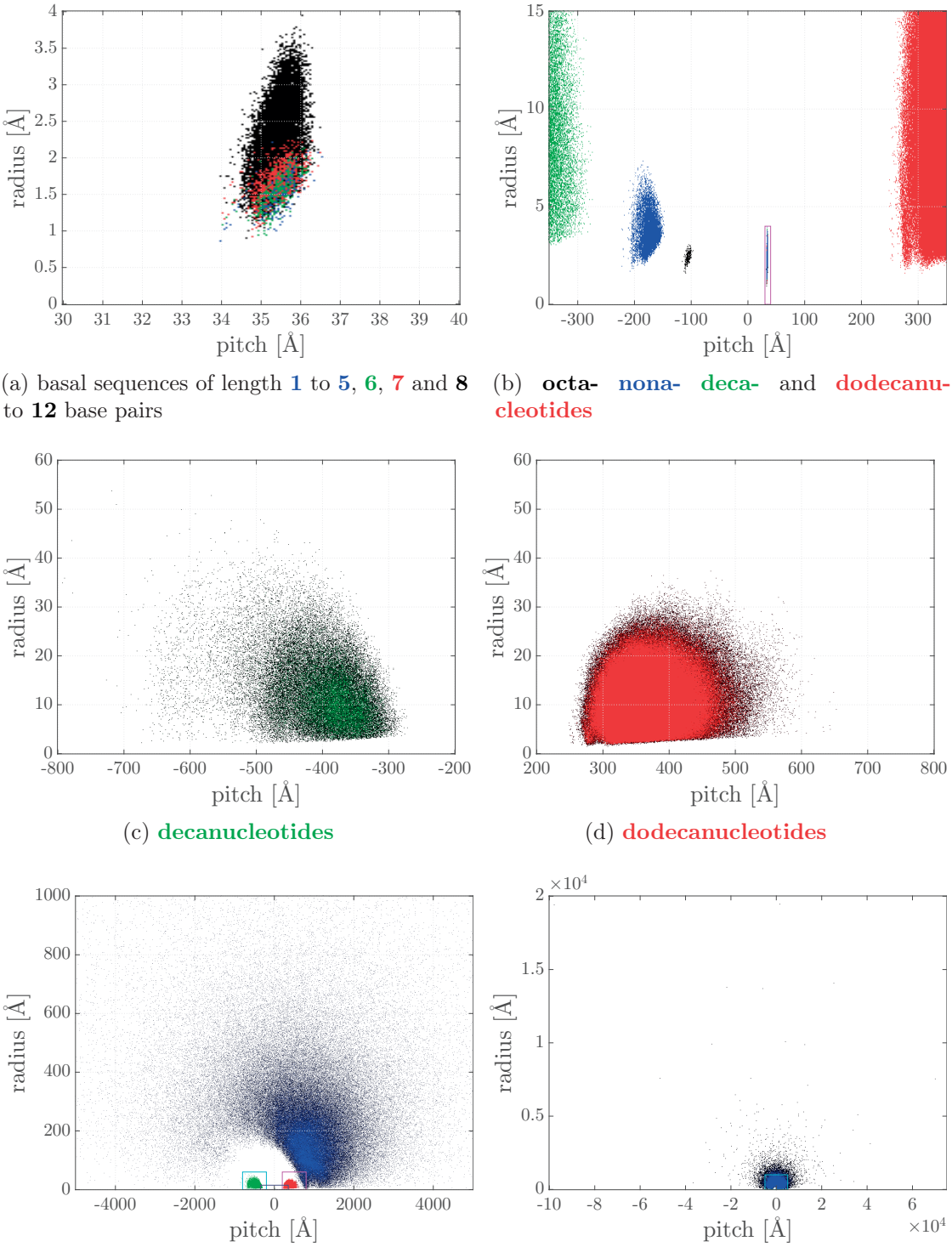
This finally shows that for S' the displacement along the superhelical axis between subsequent instances of the same base pair is the same as for S. The helical axis $\hat{\boldsymbol{e}}'$ as well as the shift vectors $\boldsymbol{q}'_{m:n}$ for S' are all simply rotated by $\widehat{\boldsymbol{L}}_N$ as compared to their counterparts $\hat{\boldsymbol{e}}$ and $\boldsymbol{q}'_{m:n}$ (respectively) for S. Consequently, after Equation (P1.3.8):

$$\widetilde{\mathbf{r}}'_{m:n} = \widehat{\boldsymbol{L}}_N\widetilde{\mathbf{r}}_{m:n} \tag{P1.3.29}$$

so that the formulae (P1.3.13), (P1.3.14) and (P1.3.15) yield the same values for S' and S.

The above observations show that in deed the same values of the radii $r_{\min}$, $r_{\max}$ and pitch p characterize both S and S' and, as a result, any cyclic shift of those. This conclusion will facilitate the exhaustive studies of oligomers of given length, as the one of Section P1.3.3, by substantially limiting the number of cases to be processed.

P1.3.3 An exhaustive study of relatively short oligomers



(a) basal sequences of length **1** to **5**, **6**, **7** and **8** to **12** base pairs

(b) **octa-** **nona-** **deca-** and **dodecanucleotides**

(c) **decanucleotides**

(d) **dodecanucleotides**

(e) **deca-**, **undeca-** and **dodecanucleotides**

(f) **undecanucleotides**

Figure P1.3.5. *See caption on the following page*

Figure P1.3.5. *Scatter plots of pitch p vs. radius $r_{\max}$ of ground state configuration superhelices for basal sequences of up to 12 bp in length. In each panel a coloured pixel indicates at least one sequence within the range of pitch and radius of that pixel. In panels (c), (d), (e) and (f) the intensity of the colour of a pixel grows from dark to bright with the growing number of sequences represented by the pixel. The table below gives the number of sequences per pixel that gives the maximum colour brightness. Note that the resultant intensity of the colour strongly depends also on the resolution (i.e. size of the area represented by a single pixel). The sizes of pixels have been chosen for each panel separately with only panel (c) and (d) having the same pixel size. The exact values are given in the table below. Panel (a) shows all cases where the superhelices are classified as atypical (see text) in the entire ensemble. Note that indicated in black are superhelices formed by fragments of length $8 \leq N12$. Panel (b) shows all data of octa- and nonanucleotides in context of subsets of deca- and dodecanucleotides. The purple region is enlarged in panel (a). Panels (c) and (d) show data for all decanucleotides and dodecanucleotides, respectively. Both have the same scale and pixel size and are directly comparable. Panel (e) shows a subset of undecanucleotides and indicates the much smaller ranges of pitch and radius of deca- and dodecanucleotides. The cyan box indicates the area of panel (c), the purple box the area of panel (d) and the dark grey box – the area of panel (b). Panel (f) shows the entire range of pitch and radius for undecanucleotides. The cyan box indicates the area of panel (e).*

Panel	*(a)*	*(b)*	*(c) and (d)*	*(e)*	*(f)*
# of seq. for full colour	≥ 1	≥ 1	≥ 5	≥ 5	≥ 10
pixel size $(p \times r_{\max})$ [Å]	0.05×0.02	1×0.05	1×0.10	10×1	250×40

length [bp]	8	9	10	11	12
#	4140	14 560	52 632	190 650	699 875
# typical	289	9604	51 740	190 646	690 796
% typical	7%	66%	98%	99.998%	99%
$\min\{p\}$	−113.3	−219.8	−840.0	−97 342.0	250.0
$\max\{p\}$	−94.5	−145.7	−271.5	70 329.3	652.9
median p	−103.5	−172.3	−391.1	860.0	364.1
$\min\{r_{\max}\}$	1.58	1.65	2.20	2.30	1.55
$\max\{r_{\max}\}$	3.03	7.28	53.67	19 341.56	37.26
median $r_{\max}$	2.39	3.64	11.65	195.75	10.58
$\min\left\{\frac{\lvert p\rvert}{r_{\max}}\right\}$	33.75	24.96	9.71	0.000 025	10.31
$\max\left\{\frac{\lvert p\rvert}{r_{\max}}\right\}$	68.21	125.10	292.81	1056.64	199.37
median $\frac{\lvert p\rvert}{r_{\max}}$	43.18	46.77	33.79	5.17	34.63

Table P1.3.3. *Statistics of pitch and radius of all superhelices classified as typical (see text) formed by tandem repeats of sequences of up to 12 bp in length. Pitches and radii are reported in* Å.

This section presents an exhaustive study of the superhelical structure of ground state configurations of DNA tandem repeats for basal sequences up to 12 bp long. The superhelices are characterized by the radii $r_{\min}$ and $r_{\max}$ and pitch p computed using the presented method and the periodic *cgDNA* parameters of Chapter P1.2 with *cgDNAparamset2* introduced in Chapter P1.1.

It should be pointed out that our exhaustive study was greatly facilitated by the fact that $r_{\min}$, $r_{\max}$ and p are the same for Watson-Crick complements and for any cyclic shifts of a given sequence, as discussed in Section P1.3.2.3. For example in the case of dodecanucleotides instead of the complete set of $4^{12} \simeq 17M$ sequences only 699 875 needed to be processed to get complete statistics.

Figure P1.3.5 presents a global picture of the study as scatter plots of $r_{\max}$ vs p for the entire ensemble, while Table P1.3.3 gives global statistics. It can be observed that that the presented method divides all the superhelices in 4 groups:

1. Atypical helices (see discussion of Section P1.3.2.2 for the definition and Figure P1.3.5a for a scatter plot of all such superhelices); basal sequences of length under 7 bp all form atypical superhelices. It is not surprising that sequences considerably shorter than a single repeat of the double helix give rise to helices very close to straight. However, as could be inferred from Figure P1.3.5a, the majority of atypical superhelices in the ensemble (in fact over 90%) are formed by basal sequences of 8 – 12 bp in length. Still, Table P1.3.3 shows that as the number of base pairs of basal sequences increases the ratio of typical to atypical superhelices increases as well. In the particular case of 11 bp there are only 4 atypical sequences, namely: $\mathtt{A}_3\mathtt{GAGA}_2\mathtt{GTC}$, $\mathtt{A}_2\mathtt{CTCT}_2\mathtt{CTCG}$, $\mathtt{A}_2\mathtt{G}_2\mathtt{AGACTAG}$ and $\mathtt{ACGTC}_2\mathtt{TCGCG}$. It should be pointed out here that the ground state configurations of atypical sequences are, in general, not significantly closer to being straight in any sense than some typical superhelices. This has been shown already in the comparison of Figures P1.3.4m and P1.3.4n. More close to straight examples are given in Figure P1.3.6 and are discussed further on. Note that in Figure P1.3.6 only the dinucleotides, shown for reference, are atypical.

   For this group of oligomers the reported radii $r_{\max}$ lie between 33.96 Å and 36.53 Å, while the pitches p are between 0.85 Å and 3.93 Å.

   An interesting subset of this group are all possible dinucleotides. In particular Table P1.3.4 presents those sequences in the descending order of the pitch p, which agrees with ascending order of the total angle θ_Σ. This shows that poly-`A` and poly-`AG` are the two most tightly coiled superhelices (those with highest $\frac{\theta_\Sigma}{N}$ with N the number of base pairs) of the six, while poly-`AT` is the most loosely coiled. In fact poly-`A` is the most tightly coiled superhelix in the whole ensemble and very likely the most tightly coiled of all possible superhelices.

   Note that there is correlation between these observations about poly-dinucleotides and the findings of Chapter P1.4. In particular poly-`A` and poly-`AG` were found

to have very high values of persistence length with that of poly-A being by far the highest of all studied sequences. On the other hand poly-AT has the lowest persistence length of all the dinucleotides. Furthermore the differences in values of pitch p can be observed as differences in the period of oscillations of the tangent-tangent correlation plots of all dinucleotides in Figure P1.4.2a. Also the amplitude of the oscillations is proportional to the value of pitch.

2. Typical oligomers of length $8-10$ bp; (see Figure P1.3.5b and P1.3.5c); all such sequences give rise to left handed superhelices. Note that all typical octanucleotides have very small radii (under 3.03 Å) and so they are all very close to straight. Figure P1.3.6β shows the ground state configuration of the octanucleotide with the maximum radius ($r_{\max} = 3.03$ Å) for which the superhelix is barely visible. For nonanucleotides more cases with larger radius and clearer superhelical structure can be found, *e.g.* those of S^{γ} and S^{δ}. Among decanucleotides the majority (just under 74%) superhelices have $r_{\min} > 5$ Å, while 43% have $r_{\min} > 10$ Å.

3. Typical dodecanucleotides (see Figure P1.3.5d); in this group all the superhelices are right-handed; similarly to decanucleotides the majority (72%) have $r_{\min} > 5$ Å and 34% have $r_{\min} > 10$ Å. However, despite the fact that there is 13 times more dodecanucleotides than decanucleotides the range of radius is considerable smaller for the latter group. This can be seen by comparing Figure P1.3.5d with Figure P1.3.5c or Figure P1.3.6ζ with Figure P1.3.6θ and in Table P1.3.3.

4. Typical undecanucleotides (see Figure P1.3.5f and Figure P1.3.5e); This is the most varied of the four groups. 15% of superhelices in that group are left-handed, while the others are right-handed (see Figure P1.3.5e). Also the range of pitch is over 110 times greater than that of all the other oligomers in the ensemble, while the range of radii is over 370 times (!) greater (see Table P1.3.3). For that reason the ground state configurations of the undecanucleotides with extreme pitch and radius are not included in the comparison of Figure P1.3.6. To give another perspective, consider that the decanucleotide S^{ε} has the highest value of the length $T = 270$ bp (see Equation (P1.3.18)) amongst all sequences of all the other groups. That is the length scale chosen for Figure P1.3.6. On the other hand the maximum T amongst undecanucleotides is 47 927 bp achieved by the sequence $\mathtt{A_2C_2AT_2ACGC}$ with highest pitch in the entire ensemble. For the oligomer ACACTGCACGT with the maximum radius the number is 37 587 bp.

	AT	AC	CC	CG	AG	AA
p	35.82	35.72	35.26	34.99	34.22	33.96
$r_{\max}$	1.99	1.88	2.21	1.14	1.09	0.85
$\theta_{\sum}$	1.112	1.126	1.135	1.173	1.206	1.212

Table P1.3.4. *The ordering of all poly-dinucleotides induced by the descending order of the pitch p. Pitches and radii are reported in* Å*, angles in radians. The ordering agrees with the ascending ordering with respect to the total angle $\theta_{\sum}$. It is similar to the ordering by the persistence lengths found for these sequences in Chapter P1.4 (see Table P1.4.2).*

Following the argument of [DubBedFur1994] about handedness of tandem repeats of basal sequences of length close to the DNA helical repeat we conjecture that for *cgDNAparamset2* the helical repeat is within the accepted range of 10 – 11 bp. For that reason all typical superhelices for basal sequences of under 10 bp are left-handed and all those above 11 bp are right-handed. The exceptional nature of the undecanucleotides can be explained *e.g.* by the supposition that the effective helical repeat of DNA for *cgDNAparamset2* is closer to 11 bp than to 10 bp. This conjecture is supported by the data of Table P1.3.3 on the value of the ratio $\frac{|p|}{r_{\max}}$. The extreme minimum at 11 bp indicates that certain undecanucleotides are very close to forming exact circles (see Figure P1.3.7).

Finally Figure P1.3.8 shows a scatter plot of pitch and radius of superhelices formed by all palindromic (self-complementary) sequences of length 8, 10 and 12. While among all 39 octanucleotide palindromes only 6 are typical, in case of both decanucleotides and dodecanucleotides over 90% of the palindromes are typical. These preliminary results seem to suggest that superhelices formed by palindromic sequences do not stand out from those formed by all the other sequences (*e.g.* in being particularly close to straight), but the question requires further investigation.

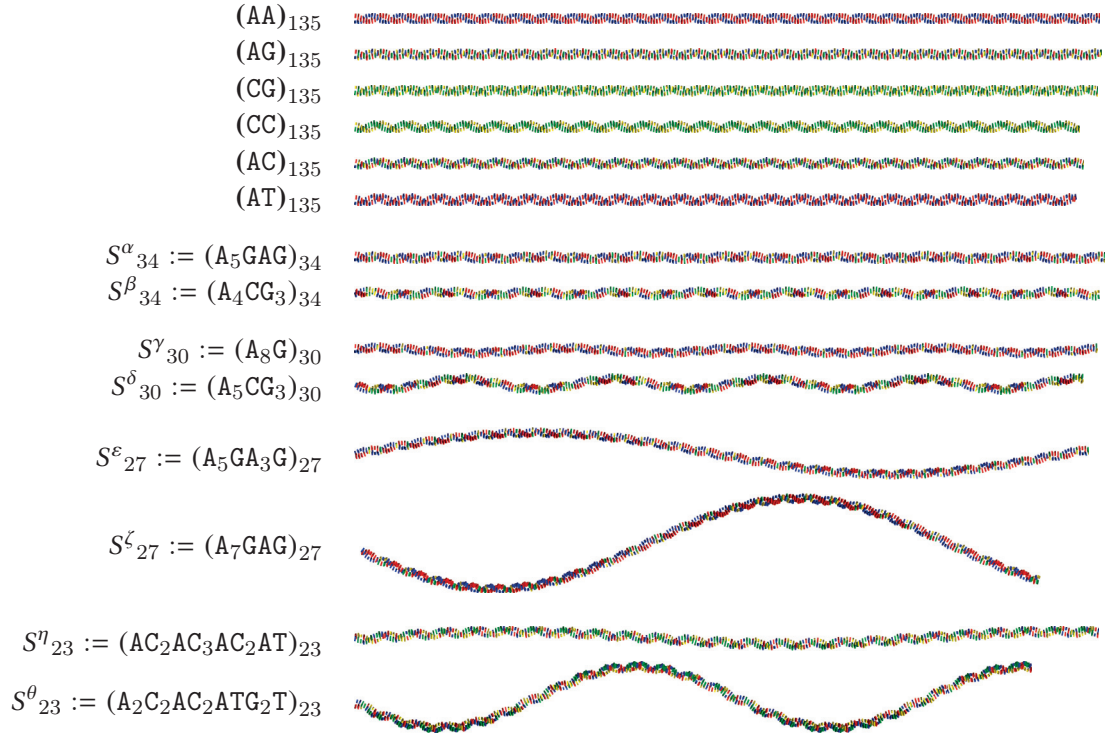


Figure P1.3.6. *Ground state configurations of superhelices with maximum (in absolute value) pitches and radii among octa- (S^{α}, S^{β}) nona- (S^{γ}, S^{δ}) deca- (S^{ε}, S^{ζ}) and dodecanucleotides (S^{η}, S^{θ}). For reference, ground state configurations of all dinucleotides (including poly-*A *and poly-*C*) are shown as the most uniform of the analysed sequences. The coloured boxes as in Figure P1.3.3. The ordering of the dinucleotides is that of Table P1.3.4. The pitches and radii for all other sequences can be found in Table P1.3.3*

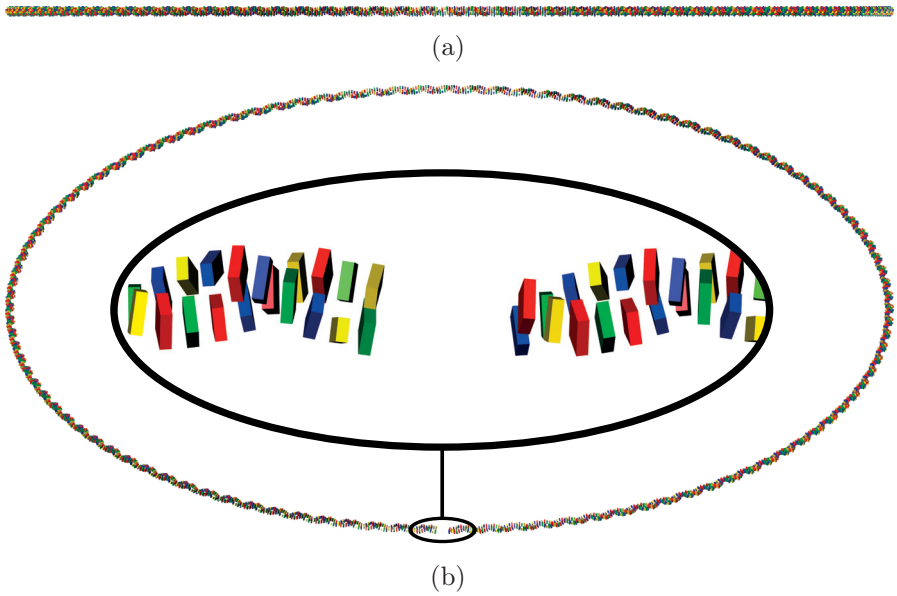


Figure P1.3.7. *Ground state configurations of the undecanucleotide* $S^{\iota}{}_{107}$ = ACAGATACAGC$_{107}$ *with the smallest ratio* $\frac{|p|}{r_{\max}} = 0.000\,025$ *of all sequences in the ensemble (the superhelix closest to a circle). The coloured boxes as in Figure P1.3.3. Panel (a) show a "side" view (with the view direction parallel to the plane perpendicular to the helical axis; helical axis pointing up). Panel (b) shows the view of panel (a) tilted by 30 degrees.*

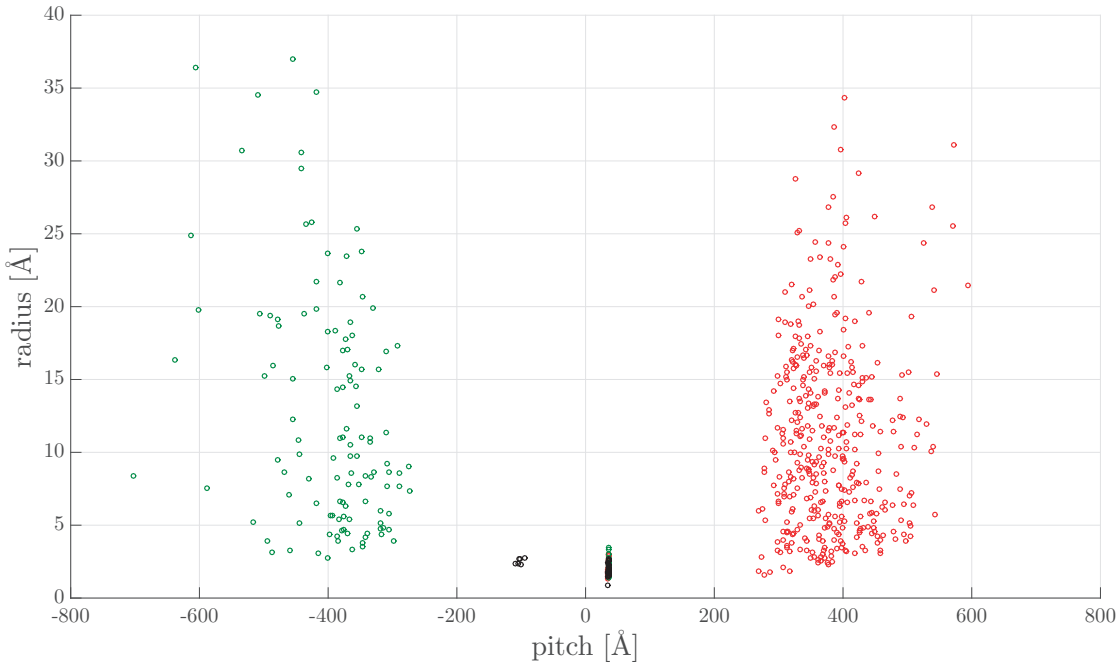


Figure P1.3.8. *A scatter plot of pitch* p *vs. radius* $r_{\max}$ *of ground state configuration superhelices for all palindromic basal sequences of up to* **8**, **10** *and* **12** *bp in length.*

P1.4 Sequence-dependent persistence lengths of DNA

(joint work with Jonathan S. Mitchell, Alexandre Grandchamp and Robert S. Manning)

DNA 'rigidity' is often expressed as a sequence-averaged, single parameter, namely the persistence length, which is sometimes informally described as a measure of the length scale over which correlation between the tangents along a polymer centreline is lost [Hag1988; RitGilKoo2009]. Frequently, the persistence length is extracted by interpreting experimental data with the Kratky-Porod worm like chain (or WLC) model [KraPor1949; PetMah2010] where the persistence length is one of only two free parameters. There is a consensus in the literature that the persistence length of DNA is approximately 150 bp, or 50 nm. This value is estimated using diverse experimental techniques, each with their own assumptions necessary to interpret the data, and often at quite different length scales [Hag1988; RitGilKoo2009]. Consequently the estimate can be regarded as robust, but not necessarily very precise. When sequence-dependence of a DNA fragment is of interest, then a description solely in terms of sequence-averaged persistence lengths is too imprecise [Flo1969; MarOls1988; ThéCouLe Rév1988; SchHar1995; Yam1997].

Consequently, while the WLC has proven extremely successful in interpreting diverse experimental results for DNA, its application to biological problems that depend significantly on sequence is precluded by its simplicity. Accordingly there have been many efforts at developing more detailed, but still coarse grained models. Some such models incorporate an overall fit to a sequence-averaged persistence length [OlsGorLu1998; SulRomOul2012; HinFreWhidPab2013]. There are also sequence-dependent, coarse grain models that *predict* sequence-averaged persistence lengths, for example 15.2 nm [MacSpaLiwSch2014], 20 nm [KnoRatSchdPab2007], 96 bp [SayAvsKab2010], and 75 nm [SavPap2010]. Similarly, estimates of persistence length have been made directly from atomistic MD simulations of (necessarily) relatively short fragments at the scale 20–50 bp, *e.g.* 80 nm for poly(`AT`) and poly(`GC`) [Maz2006], and 43 nm for a mixed sequence fragment [NoyGol2012].

We here assess the ability of the *cgDNA* model [Pet2012; GonPetMad2013; PetPasGonMad2014], as summarized in Chapter B.1, to reproduce the sequence-dependent statistical mechanics properties of B-form double helical DNA, by developing appropriate Monte Carlo (or MC) sampling methods in order to generate associated ensembles of configurations. The MC code developed here allows sampling of *cgDNA* Boltzmann distributions at the scales of tens to thousands of bp. Simulations of sequence-averaged persistence length yield the estimates of 53.5 nm in the sense of Flory (from simulations at the scale of 1 Kbp), and 160 bp in the sense of apparent tangent-tangent correlation decay (from simulations at the scale of 200 bp). These estimates have a standard error of ±0.1 nm/1 bp in the sense of multiple estimates from multiple MC simulations. Error associated with underlying imprecision in *cgDNA* parameters is harder to assess, but is likely to be significantly larger. The tangent-tangent persistence length also has a mild dependence on choices in coarse graining tangents and arc-length.

P1.4.1 Theory

P1.4.1.1 The statistical mechanics of persistence lengths

We will consider two of the classic expectations of polymer physics, see for example [KraPor1949; Flo1973; Sch1974; DoiEdw1986; Yam1997], that depend on a sequence of frames $(\mathbf{r}_n, \boldsymbol{D}_n)$, namely:

$$\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle \tag{P1.4.1a}$$

and

$$\langle \boldsymbol{D}_0^T (\mathbf{r}_i - \mathbf{r}_0) \rangle \quad , \tag{P1.4.1b}$$

where $\langle \cdot \rangle$ denotes the ensemble average, *i.e.* the expectation of the argument with respect to an underlying equilibrium measure, $\mathbf{t}_0$ is a unit vector associated with a specific base pair labelled with index 0 (usually taken to be away from the physical end of the polymer to avoid any possible end effect), and $\mathbf{t}_i$ is the analogous unit vector at the ith base pair along the polymer. Usually $\mathbf{t}_i$ is to be interpreted as some approximation to a unit tangent to the polymer, so that (P1.4.1a) is often described as a tangent-tangent correlation function. Similarly $\boldsymbol{D}_0^T(\mathbf{r}_i - \mathbf{r}_0)$ are the components of the chord vector between the 0th and ith base-pair origins expressed in the chosen reference frame $\boldsymbol{D}_0$. We will call the expectations (P1.4.1b) Flory persistence vectors, as they were apparently first introduced in [Flo1973], see also [MarOls1988; SchHar1995]. Accordingly for each choice of reference frame $\boldsymbol{D}_0$, the expectations (P1.4.1a) and (P1.4.1b) are respectively scalar and vector functions of the index $i \geq 1$.

One of the simplest model ensembles in which to compute the expectations (P1.4.1) is a discrete version of the Kratky-Porod WLC [KraPor1949; Sch1974]. In this model the polymer is assumed to be a chain of rigid links all of length b, so that any configuration is described by unit chord vectors $\mathbf{t}_i := \frac{1}{b}(\mathbf{r}_{i+1} - \mathbf{r}_i)$, and the equilibrium measure is assumed to be Boltzmann with inverse temperature scale $\beta = 1/k_B T$ and free energy (or Hamiltonian) $E = \frac{B}{b}\sum_{i=1}^{N}(1 - \mathbf{t}_i \cdot \mathbf{t}_{i+1})$ with B a (constant) bending rigidity parameter. In particular the minimum energy, or ground, state of the WLC is intrinsically straight with all tangent vectors parallel. Then provided that the non-dimensional parameter $\ell_p := \beta B/b$ is large (compared to 1), it can be calculated analytically that the correlations (P1.4.1a) are well approximated by the formula

$$\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle_{\mathrm{WLC}} = e^{-i/\ell_p}. \tag{P1.4.2}$$

Then the exponential decay scale ℓ_p is the persistence length expressed in bp, while $b\,\ell_p = \beta B$ is the dimensional persistence length expressed in the (arc-)length units of b. Similarly within the discrete WLC model the expectations (P1.4.1b) can be computed to be

$$\langle \boldsymbol{D}_0^T(\mathbf{r}_i - \mathbf{r}_0)\rangle_{\mathrm{WLC}} = b\,\ell_p \begin{bmatrix} 0 & 0 & \left(1 - e^{-i/\ell_p}\right) \end{bmatrix}^T \tag{P1.4.3}$$

provided only that the third column of $\boldsymbol{D}_0$ is chosen to coincide with $\mathbf{t}_0$. In fact the specific functional forms of expressions (P1.4.2) and (P1.4.3) are only exact in the limit of the continuous WLC, in which the dimensional persistence length $\beta B = b\ell_p$ stays constant, while $b \to 0$, $N \to \infty$, $Nb \to L$, and $i\,b \to s \in [0, L]$. Nevertheless, the simple approximations (P1.4.2) and (P1.4.3) suffice for our purposes.

For a DNA fragment with sequence S, and motivated by the WLC formula (P1.4.3), we introduce a Flory persistence length $\ell_F(S)$ as the limiting value of the magnitude (in the usual Euclidean distance $\|\cdot\|$) of the Flory persistence vector (P1.4.1b) as $i \to \infty$, along with its sequence-averaged version $\overline{\ell}_F$:

$$\ell_F(S) = \lim_{i\to\infty} \left\| \langle \boldsymbol{D}_0^T(\mathbf{r}_i - \mathbf{r}_0)\rangle \right\| \quad , \tag{P1.4.4a}$$

$$\overline{\ell}_F = \lim_{i\to\infty} \left\| \{\langle \boldsymbol{D}_0^T(\mathbf{r}_i - \mathbf{r}_0)\rangle\} \right\| \quad . \tag{P1.4.4b}$$

Here the brackets $\{\cdot\}$ in the second expression denote an additional average over an ensemble of sequences S_j of the average $\langle\cdot\rangle$ over an ensemble of configurations of a fragment with fixed sequence. The persistence lengths defined in (P1.4.4) indeed have the dimension of length, which we will report in nanometers (or nm). For the (sequence-independent) WLC, $\ell_F = b\,\ell_p$.

Similarly the WLC chain formula (P1.4.2) motivates the definition of a sequence-dependent tangent-tangent correlation length $\ell_p(S)$ and its sequence-averaged version $\overline{\ell}_p$:

$$e^{-i/\ell_p(S)} \approx \langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle \quad , \tag{P1.4.5a}$$

$$e^{-i/\overline{\ell}_p} \approx \{\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle\} \quad , \tag{P1.4.5b}$$

where the symbol $\approx$ signifies that, for a given sequence S, $\ell_p(S)$ is computed as the number of base pairs equal to the (negative reciprocal) of the slope of the straight line through the origin that is the least squares fit to the plot of $\ln\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle$ vs. i. Similarly $\overline{\ell}_p$ is computed via the analogous semi-log plot of the sequence averaged data $\{\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle\}$ vs. i.

We note that for more realistic DNA free energies than the WLC, there is no *a priori* reason to believe that the dimensionless tangent-tangent persistence lengths $\ell_p(S)$ can be simply related to the Flory persistence lengths $\ell_F(S)$ via the introduction of a single length scale. Furthermore it is well understood that there are some sequences with high intrinsic curvature, for example those containing phased A-tracts *e.g.* [MarOls1988; SchHar1995], for which the exponential fit in (P1.4.5) to obtain $\ell_p(S)$ is an extremely poor approximation at scales of one or two persistence lengths or shorter (indeed for some exceptional sequences of moderate length $\langle \mathbf{t}_i \cdot \mathbf{t}_0 \rangle$ can even become negative so that the semi-log plot fit yielding $\ell_p(S)$ has no sense, in contrast to the more robust definition of $\ell_F(S)$). However for 'reasonable' (*i.e.* non-exceptional) sequence ensembles $\{\cdot\}$ it is believed that the sequence-averaged exponential fit to obtain $\overline{\ell}_p$ is a rather good approximation. Our simulations will confirm these behaviours within the *cgDNA* model.

P1.4.1.2 The choice of Monte Carlo observables

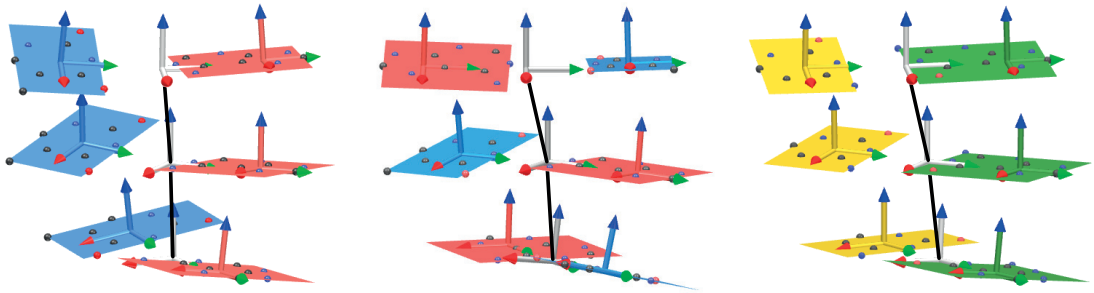
We will use Monte Carlo simulations applied to the *cgDNA* model of the free energies of a number of different sequences in order to generate ensembles that yield numerical estimates of the expectation functions (P1.4.1). We can then obtain estimates of the four related notions of persistence length from the ansatzen (P1.4.4, P1.4.5), along with assessments of the convergence of the norms of the Flory vectors (P1.4.4), and the quality of the fits in the tangent-tangent cases (P1.4.5).

For the Flory vector ensemble the most natural choice is to take the $\mathbf{r}_i$ and $\boldsymbol{D}_i$ to be the *cgDNA* base-pair location and orientation (as defined in Equation (B.1.4) and (B.1.2) of Section B.1.2), after which the simulations are completely specified.

However, for the tangent-tangent persistence length it remains to make precise the choice for the unit vectors $\mathbf{t}_i$. In contrast to the WLC, because the *cgDNA* model encompasses fluctuations in the junction translations of shift, slide and rise (see Figure B.1.2), there are at least two natural choices for $\mathbf{t}_i$. One possibility is the base-pair normal, *i.e.* the third column of each $\boldsymbol{D}_i$, or equivalently the frame vector most closely aligned with the helical

axis (see Figure P1.4.1), which, as a matter of convention, will be denoted $\mathbf{t}_i^{[0]}$. Another natural possibility is the unit tangent to the junction chord between two consecutive base pair origins $\mathbf{t}_i^{[1]} = (\mathbf{r}_{i+1} - \mathbf{r}_i)/\|\mathbf{r}_{i+1} - \mathbf{r}_i\|$ (shown in black in Figure P1.4.1).

As a matter of convention $\ell_p^{[k]}(S)$ will be taken to mean a persistence length of the sequence S evaluated using the formula (P1.4.5), with $k = 0$ for the choice of base pair normals $\mathbf{t}_i^{[0]}$ and $k = 1$ for the junction unit vectors $\mathbf{t}_i^{[1]}$. A comparison between expectations $\ell_p^{[0]}$ and the $\ell_p^{[1]}$ has previously been considered in [FatEslEjt2012] and is briefly addressed in Section P1.4.3.1.



(a) (b) (c)

Figure P1.4.1. *A schematic visualization of three central base pairs in the cgDNA ground state configuration of three icosanucleotides: panel (a)* poly(A), *panel (b)* poly(TA), *and panel (c)* poly(G). *Each nucleotide is represented as a rigid body fit to base atoms that is visualized as a coloured plate (*A *red,* T *blue,* G *green,* C *yellow) along with a base normal. The position and orientation of each base pair frame (light grey) is an appropriate average of the two associated base frames (for visual clarity each base frame is offset by 0.35 nm toward its backbone from the standard Curves+ definition). The junction chords between the origins of adjacent base-pair frames are shown in black. Note that the* poly(A) *sequence has exceptionally high (propeller) intra base pair rotations, and the junction chords are closely aligned with the base pair normal, while for both* poly(TA) *and* poly(G) *there is a significant angle between the junction chords and associated base-pair normals. The angle between the chord and the base pair normal is closely connected to the radius and pitch of the superhelix formed by repeats of a dimer (see Section P1.3.3).*

P1.4.1.3 Direct Monte Carlo sampling

The ensemble expectation $\langle f \rangle$ of any function $f(\boldsymbol{w})$ of the *cgDNA* internal variables can be approximated as the simple average $\frac{1}{M}\sum_{j=1}^{M} f(\boldsymbol{w}_j)$ over a sequence of configurations $\boldsymbol{w}_j$ that is generated by a Monte Carlo method which appropriately samples the associated equilibrium distribution $p(\boldsymbol{w})\,d\boldsymbol{w}$. We will consider two specific cases of the probability density function, a pure Gaussian, or multivariate normal, and a perturbed Gaussian:

$$p(\boldsymbol{w}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{w})} \quad , \tag{P1.4.6a}$$

$$\widetilde{p}(\boldsymbol{w}) = \frac{1}{\widetilde{Z}} J(\boldsymbol{w}) e^{-\beta E(\boldsymbol{w})} \quad , \tag{P1.4.6b}$$

where $E(\boldsymbol{w})$ is the shifted quadratic *cgDNA* energy (B.1.17), $\beta = 1/(k_B T)$ is the inverse temperature scale, Z is the (explicitly known) normalization constant (or partition function), and $J(\boldsymbol{w}) > 0$ is an explicitly known function of $\boldsymbol{w}$, but now the value of the associated normalizing constant $\widetilde{Z}$ is in general not known. There are several possible motivations for the generalization (P1.4.6b), for example modelling contributions to the *cgDNA* free energy from end-loading terms as in single molecule tweezer experiments, or modelling multi-well DNA backbone states as described in [PasMadBev2014]. However we focus here on a third motivation in which $J(\boldsymbol{w})$ is a Jacobian factor required [BecEve2007; LanGonHef2009; WalGonMad2010] by the non-Cartesian nature of any rotational coordinates for the relative rotations between the base pair frames $\boldsymbol{D}_i$ (and the base frames $\boldsymbol{D}_i^{\pm}$). In the scaled Cayley vector rotational coordinates adopted within the *cgDNA* model it can be computed explicitly that the appropriate configuration space equilibrium distribution is of the form (P1.4.6b) with the explicit correction term (B.1.16) (and scales of the rotational variables of (B.1.11)). Essentially we here wish to be able to assess when the differences between the two pdfs in (P1.4.6) (with the same sequence-dependent free energy $E(\boldsymbol{w})$) are sufficiently small that attention can be restricted to the simpler case (P1.4.6a).

One approach to Monte Carlo simulation of multivariate normals such as (P1.4.6a) involves the Cholesky decomposition of the covariance matrix (*e.g.* [**Gentle2003**]). We adapt this approach to take advantage of the sparsity structure of the stiffness matrix $\boldsymbol{K}$, performing the Cholesky decomposition on $\boldsymbol{K}$ itself:

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^T, \tag{P1.4.7}$$

where $\boldsymbol{L}$ is a lower triangular matrix. We can then use the Cholesky factorization to rewrite the *cgDNA* energy (B.1.18) as $E(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}$ where:

$$\boldsymbol{y} = \boldsymbol{L}^T(\boldsymbol{w} - \hat{\boldsymbol{w}}) \quad . \tag{P1.4.8}$$

This distribution can be sampled directly as the product of uncoupled univariate normal distributions:

$$p_{\mathbf{y}}(\boldsymbol{y}) = \prod_{i=1}^{12n-6} \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\beta}{2}y_i^2}. \tag{P1.4.9}$$

For each sample the configuration in the original variables $\boldsymbol{w}$ must first be reconstructed from the configuration $\boldsymbol{y}$ using (P1.4.8). Then the observables $\mathbf{r}_n$ and $\boldsymbol{D}_n$ must also be computed from $\boldsymbol{w}$ as described in Section B.1.3. As both of these computations occur at every draw they should be done efficiently. As described in more detail in Sections P1.4.2.1 and P1.4.2.3, we make full use of the sparsity in the problem, as well as of quaternion multiplication in the many rotation matrix products.

P1.4.1.4 Metropolis Monte Carlo sampling

To sample the perturbed Gaussian distribution (P1.4.6b) we adopt the following simple Metropolis algorithm (see [MetRosRos1953] for a similar treatment). For non-Gaussian distributions of the form (P1.4.6b), a natural implementation of the Metropolis method involves generating candidate configurations by direct sampling of the Gaussian part of the distribution and then adding an acceptance/rejection criterion based on the correction term $J(\boldsymbol{w})$. Specifically, given a prior configuration with internal-variable vector $\boldsymbol{w}$, we follow the direct Monte Carlo procedure from the previous section to generate a new draw of the configuration vector $\boldsymbol{w}^*$ and accept or reject it as follows:

- if $J(\boldsymbol{w}^*) \geq J(\boldsymbol{w})$, we accept $\boldsymbol{w}^*$
- if $J(\boldsymbol{w}^*) < J(\boldsymbol{w})$ we accept $\boldsymbol{w}^*$ with probability $\frac{J(\boldsymbol{w}^*)}{J(\boldsymbol{w})}$ and otherwise reject it (in which case we append a new copy of $\boldsymbol{w}$ to our ensemble).

This acceptance criterion is one way of ensuring the crucial property of *detailed balance*, which requires that

$$\alpha(\boldsymbol{w} \to \boldsymbol{w}^*)\, P(\boldsymbol{w} \to \boldsymbol{w}^*)\, \widetilde{p}_{\mathbf{w}}(\boldsymbol{w}) \;=\; \alpha(\boldsymbol{w}^* \to \boldsymbol{w})\, P(\boldsymbol{w}^* \to \boldsymbol{w})\, \widetilde{p}_{\mathbf{w}}(\boldsymbol{w}^*), \tag{P1.4.10}$$

where $\widetilde{p}_{\mathbf{w}}$ is the probability density function (P1.4.6b), $\alpha(\boldsymbol{y} \to \boldsymbol{z})$ is the conditional probability density in our Metropolis algorithm for choosing state $\boldsymbol{z}$ given prior state $\boldsymbol{y}$ (which in our scheme is independent of $\boldsymbol{y}$ and equals $p_{\mathbf{w}}(\boldsymbol{z})$ from (P1.4.6a)), and $P(\boldsymbol{y} \to \boldsymbol{z})$ is the probability in our Metropolis algorithm of accepting the new state $\boldsymbol{z}$ given a prior state $\boldsymbol{y}$ (which in our scheme is 1 if $J(\mathbf{z}) \geq J(\boldsymbol{y})$ and $\frac{J(\boldsymbol{z})}{J(\boldsymbol{y})}$ otherwise).

P1.4.2 Details regarding the *cgDNAmc* code

This section describes our Monte Carlo implementation in detail. The simulations described here are not particularly intensive, nevertheless we have taken some efforts to make *cgDNAmc*code efficient. Benchmark results presented below were obtained on a mid-range laptop computer.

P1.4.2.1 Monte Carlo sampling

As described in Section P1.4.1.3, each step of our procedure of direct Monte Carlo sampling begins with a draw draw $\boldsymbol{y}$ from the distribution (P1.4.9). To make the draw each component y_i is taken as a random number from the normal distribution with mean 0 and standard deviation $\beta^{-\frac{1}{2}}$. Note that units of the stiffness matrix $\boldsymbol{K}$ in the cgDNA model are such that $\beta = 1$. For the sake of efficiency uniform deviates are generated using the `xorshift1024*` variant[1] of the `xorshift` algorithm [Mar2003] and are subsequently converted to normal deviates using the `ZIGNOR` implementation[2] of the Ziggurat algorithm [MarTsa2000].

The draw of the internal coordinates $\boldsymbol{w}$ corresponding to $\boldsymbol{y}$ is obtained from Equation (P1.4.8) by solving:

$$\boldsymbol{L}^T\boldsymbol{z} = \boldsymbol{y} \tag{P1.4.11a}$$

for $\boldsymbol{z}$ and then setting:

$$\boldsymbol{w} = \boldsymbol{z} + \hat{\boldsymbol{w}} \quad . \tag{P1.4.11b}$$

For efficient sampling, the key property of the Cholesky factorization (P1.4.7) is that if $\boldsymbol{K}$ has bandwidth m (meaning that all non-zero entries are within m rows of the diagonal, so for us $m = 17$), then $\boldsymbol{L}^T$ also has bandwidth m [GolVan1996, p. 154]. As a result the linear solve (P1.4.11a) can be performed very efficiently by an appropriate solver from LAPACK [AndBaiBis1999] specialized for banded triangular matrices.

One alternative approach to obtain direct sampling would involve a spectral decomposition of $\boldsymbol{K}$ in place of the Cholesky factorisation, *i.e.*

$$\boldsymbol{K} = \mathbf{P}\mathbf{D}\mathbf{P}^T, \tag{P1.4.12}$$

with $\mathbf{P}$ orthogonal and $\mathbf{D}$ diagonal. Here a similar change of variable $\boldsymbol{y} = \mathbf{D}^{\frac{1}{2}}\mathbf{P}^T(\boldsymbol{w} + \hat{\boldsymbol{w}})$ can be used so that $\boldsymbol{w} = \mathbf{P}\mathbf{D}^{-\frac{1}{2}}\boldsymbol{y} + \hat{\boldsymbol{w}}$. This has been successfully exploited by Czapla et al. [CzaSwiOls2006] for the case where $\boldsymbol{K}$ is block diagonal. However in our setting with a (potentially large) banded $\boldsymbol{K}$ that approach is significantly less efficient, since the

[1] http://arxiv.org/abs/1404.0390
[2] http://www.doornik.com/research/ziggurat.pdf

matrix $\mathbf{PD}^{\frac{1}{2}}$ would not be sparse, and a dense matrix-vector multiply must be carried out in the construction of each draw. To give an example a simulation calculating $\langle \mathbf{t}_i^{[0]} \cdot \mathbf{t}_0^{[0]} \rangle$ for 1 million configurations of the S^λ sequence (see Section A.2.2.1) of length 300 bp using Cholesky decomposition takes just above 3 minutes on a contemporary laptop, while using spectral decomposition the running time is around 2 hours.

The Metropolis procedure is computationally much more intensive than the direct sampling possible in the pure Gaussian case. In particular the efficiency of any Metropolis method depends strongly on the acceptance rate for the given move set, which can be punitively small. In the particular case of the pdf (P1.4.6b) with the explicit choice (B.1.16) for J, and the cgDNA energy (B.1.17), the observed acceptance rates depend on the length of the simulated oligomers. For oligomers of 300 bp the acceptance rate is approximately 37%, which is perfectly acceptable and 10^6 accepted moves can be generated in 11 minutes on a mid-range laptop computer. For oligomers 5 times as long (1500 bp – as used for computing the Flory persistence vectors) the acceptance rate drops to just under 5%, with a corresponding increase in the number of draws required to obtain convergence. The acceptance criterion involves only the internal coordinates $\boldsymbol{w}$ and the rejected moves absorb comparatively little computational time because the corresponding $(\mathbf{r}_n, \boldsymbol{D}_n)$ configurations need not be reconstructed. Nevertheless for a 1500 bp fragment the achieved performance was 10^6 accepted moves in 6 hours.

P1.4.2.2 Rigid base pair marginals

We remark that many expectations of interest involve only the *inter* part of the configuration variable $\boldsymbol{w}$ so that the number of degrees of freedom can be reduced by a half by computing the marginal distribution for the inter variables. As the original distribution is Gaussian its marginals are also Gaussian, but the resulting marginal stiffness matrix is now dense, so that sparse computations can no longer be used. As a consequence a calculation of $\langle \mathbf{t}_0^{[0]} \cdot \mathbf{t}_i^{[0]} \rangle$ for 1 million configurations of the S^λ fragment (see Section A.2.2.1) using the marginal distribution takes around 23 minutes, which is nearly 7 times slower than generating ensembles in the full $\boldsymbol{w}$ space and discarding all the *intra* variables.

P1.4.2.3 Reconstruction of 3D shapes

The first step in calculating our observables for a given configuration vector is reconstructing a 3D shape of a molecule from a given internal coordinate vector $\boldsymbol{w}$ [LanGonHef2009] as detailed in Section B.1.3. As mentioned in the previous section, the calculation of tangent-tangent correlations, arclengths and Flory vectors require only the *inter* part of $\boldsymbol{w}$. As a result we only reconstruct base pair positions $\mathbf{r}_i$ and orientations $\boldsymbol{D}_i$ (Equation (B.1.13)), which takes only half the time of reconstructing a full 3D configuration of rigid bases (also Equation (B.1.14)). The mentioned reconstruction procedure, implemented by the *cgDNArecon* library, involves evaluating half rotations, composing rotations, applying rotations to vectors and adding vectors.

A careful numerical study of efficiency of different parametrizations of rotations (namely Cayley vectors, unit quaternions and rotation matrices) implemented as a library called *algebra3d* has been performed. The explicit half-rotation formula (A.1.47) for unit quaternions proved to be 60% faster than a similar formula (A.1.51) for Cayley vectors. For rotation matrices, the analogous calculation would require, *e.g.*, an iterative algorithm of computing the principal square root and so was not considered. As expected, for composition of rotations, quaternion multiplication (A.1.9) was faster than matrix multiplication, with our observed difference being 30%. On the other hand, in the case of applying a rotation to a vector, the standard matrix-vector product was 5 times faster than a specialized quaternion rotation operator (A.1.20). In fact the fastest way to apply a rotation given as unit quaternion to a vector was to convert the quaternion to a rotation matrix first using (A.1.16) (this takes only twice the time of the matrix-vector product). Efficiency of converting between all three parametrizations was also analysed. This suggested, for example, that the formula (A.1.35) for computing a rotation matrix for a given Cayley vector is two times slower than conversion of a Cayley vector to quaternion using (A.1.41) and subsequent conversion of the quaternion to a rotation matrix though (A.1.16).

Considerations similar to the above suggested two approaches to the reconstruction procedure. The first one uses directly the Cayley vectors of the configuration variable $\boldsymbol{w}$ to calculate half rotations and converts to rotation matrices for all subsequent calculations. The other one, that finally proved to be 30% faster, begins with converting the Cayley vectors to quaternions, then computes half rotations using quaternions, and finally converts quaternions to matrices when rotations need to be applied to vectors.

P1.4.2.4 Remarks on parallelization

We first note that in *cgDNAmc* pseudo-random numbers have to be generated sequentially to assure correctness and reproducibility of results. Also the reconstruction procedure is inherently sequential. The conversion of the decoupled normal deviates $\boldsymbol{y}$ to a configuration vector $\boldsymbol{w}$, as introduced in Section P1.4.1.3, depends on the underlying LAPACK routine, which might already be optimized to use available multiple cores, but the *cgDNAmc* code has no other explicit parallelization. In part this is because each configuration can be generated and analysed independently of all others, so that the suggested solution for generating large ensembles is to run multiple independent simulations at the same time, with a different seed for the pseudo-random number generator in each instance. By linearity, expectations from multiple runs can be aggregated as a weighted average with weights proportional to the number of configurations generated in each independent run. As an example we achieved a 2.4 speed up in this way by running four independent simulation on a laptop with a dual-core, hyper-threaded CPU.

P1.4.2.5 Run-times of key steps of the algorithm

A simple profile of run times for the key steps of a simulation that calculates three expectations using 1 million configurations of the 300 bp long S^{λ} oligomer (see Section A.2.2.1) is presented in Table P1.4.1. The presented results show that the time of generation of uniform deviates and their tranformation to *cgDNA* configurations takes 67% of all processing time. Given that some of the fastest available codes are used in those steps the efficiency of the other computations (*i.e.* those implemented in *cgDNAmc*) are satisfactory.

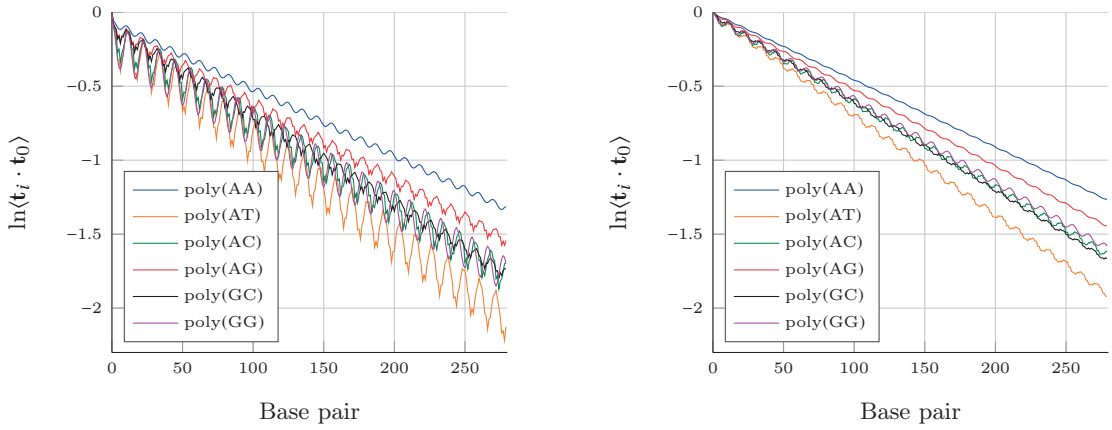
Operation	Run time [s]	% of simulation
Generation of normal deviates $\boldsymbol{y}$	59.79	29.4%
Transformation (P1.4.11) to $\boldsymbol{w}$	75.77	37.3%
Shape reconstruction	43.08	21.2%
Calculating $\langle \mathbf{t}_0^{[0]} \cdot \mathbf{t}_i^{[0]} \rangle$	6.79	3.3%
Calculating $\langle \mathbf{t}_0^{[1]} \cdot \mathbf{t}_i^{[1]} \rangle$	10.36	5.1%
Calculating Flory vectors	6.61	3.3%
Other	0.94	0.5%
Entire simulation	203.3	100.00%

Table P1.4.1. *A run-time profile of a simulation that calculates expectations using 1 million configurations of the 300 bp S^{λ} oligomer (see Section A.2.2.1). The time necessary to evaluate each expectation is a negligible fraction of the total. The simulation was run on a contemporary laptop.*

P1.4.3 Results of simulations

P1.4.3.1 The choice of tangent

In the presentation of results we will limit our consideration to the definition $\mathbf{t}^{[0]}$ of the tangent as the base pair normal. This choice is motivated by the fact that in the case of the unit chord vector $\mathbf{t}^{[1]}$ oscillations of a considerable amplitude due to the intrinsic shape of the oligomer can be observed in the tangent-tangent correlation plots (see Figure P1.4.2a). This can substantially impact the linear fit used to estimate the persistence length leading to low estimates for ℓ_p (see Table P1.4.2). In case of $\mathbf{t}^{[0]}$ (Figure P1.4.2b) the amplitude of the oscillations is much smaller.



(a) $\mathbf{t}_i = \mathbf{t}_i^{[1]}$ (b) $\mathbf{t}_i = \mathbf{t}_i^{[0]}$

Figure P1.4.2. *Comparison of tangent-tangent correlation plots for the two choices of tangent: the base pair normal* $\mathbf{t}^{[0]}$ *and the unit chord vector* $\mathbf{t}^{[1]}$*. Note the much higher amplitude of oscillations introduced by the intrinsic shape in case of* $\mathbf{t}^{[1]}$*.*

	AT	GC	AC	GG	AG	AA	S^λ
ℓ_F [nm]	47.2	54.9	55.5	56.2	64.0	72.7	58.3
$\ell_p^{[0]}$ [bp]	146	166	169	173	192	219	162
$\ell_p^{[1]}$ [bp]	129	153	151	149	175	205	155
$\Delta\ell_p$	11.3 %	7.6 %	10.6 %	14.0 %	9.0 %	6.7 %	4.6 %

Table P1.4.2. *Numerical values of the Flory persistence length* ℓ_F *and the tangent-tangent correlation persistence lengths* $\ell_p^{[0]}$ *and* $\ell_p^{[1]}$ *for all poly-dinucleotides. As reference, data for the sequence* S^λ *is shown (see Section A.2.2.1). The sequence has been chosen as the one with the median value of* $\ell_p^{[0]}$ *over all* 161 *consecutive fragments of* 300 *bp in the genome. For that reason it has been used as a representative average biological sequence in many of the presented examples.*

P1.4.3.2 Sequence is significant

The four panels of Figure P1.4.3 provide normalized histograms of the values of the individual Flory $\ell_F(S)$ (P1.4.4a) and tangent-tangent correlation $\ell_p^{[0]}(S)$ (P1.4.5a) persistence lengths obtained from direct MC simulations of the Gaussian distribution (P1.4.6a) for two ensembles of sequences. One of the ensembles is 1000 random sequences of length 220 bp with equal probabilities for each of the four possible bases at each index i. The other consists of two hundred and twenty 220 bp fragments of the λ-phage genome sequence [SanCouHon1982]. (We remark in passing that there are approximately 10^{120} possible 200-nucleotide sequences.) For each of the selected sequences, the origin base pair index 0 was chosen to be the 11th actual base pair from one end in order to avoid any initial end effects, and similarly statistics were not taken from within 10 bp of the distal end. For the simulations of the Flory persistence length $\ell_F(S)$ in order to obtain good $i \to \infty$ convergence in the base pair index i, (see Equation (P1.4.4)), sequence fragments of approximately 1.5 Kbp are needed, so each sequence was repeated 7 times.

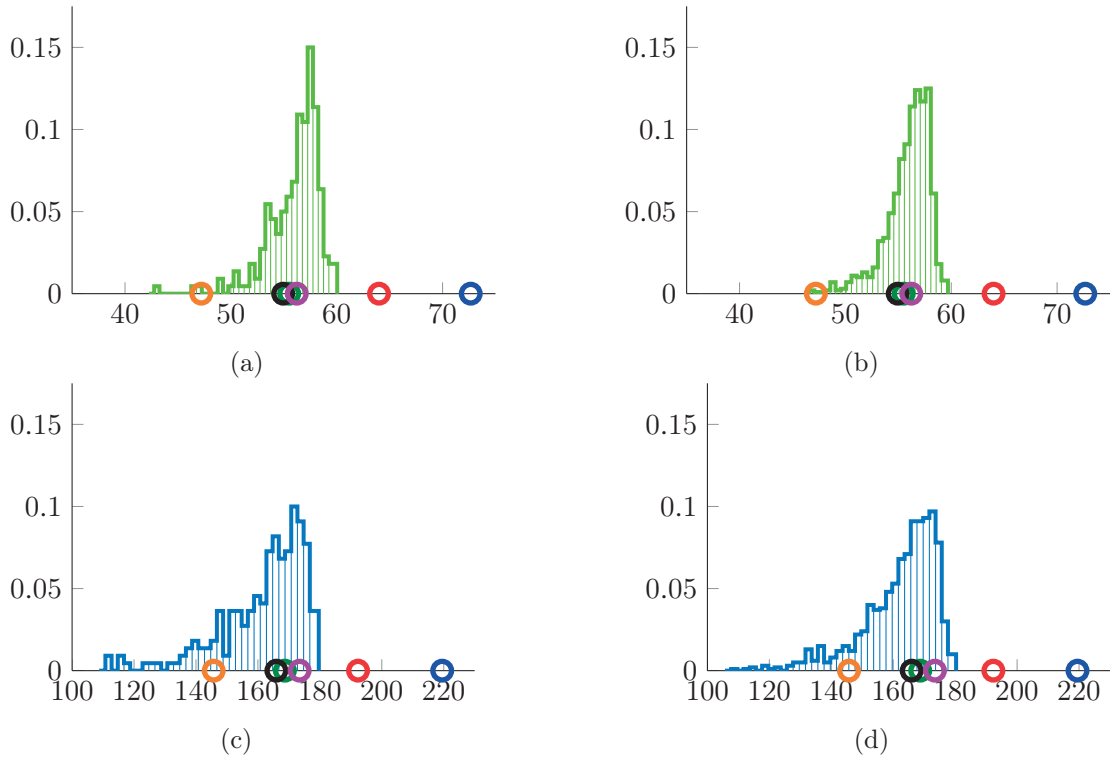


Figure P1.4.3. *Normalized histograms of persistence lengths, $\ell_F(S_j)$, $\ell_p^{[0]}(S_j)$, for* 220 *bp fragments from λ-phage genome [SanCouHon1982] (left) and with random sequence (right). The histograms of $\ell_F(S_j)$ are generated with bin size* 0.5 *nm, while those for $\ell_p^{[0]}(S_j)$ use bins of* 2 *bp. In addition in each panel the associated persistence lengths for the six distinct poly-dinucleotide sequences (see Table P1.4.2) are marked with coloured circles (the colour coding same as in Figure P1.4.2). The harmonic means of $\ell_F(S_j)$ for the λ and random ensembles are, respectively,* 55.7 *nm and* 55.6 *nm and of $\ell_p^{[0]}(S_j)$* 159 *bp and* 160 *bp.*

The histograms indicate that there is strong sequence dependence of both $\ell_F(S)$ and $\ell_p^{[0]}(S)$ with at most small differences between the random and λ-phage ensembles, with perhaps a somewhat more prominent left tail (with many fewer samples) for λ. Both Flory distributions are quite broad and asymmetric, as are both $\ell_p^{[0]}(S)$ histograms which have a notable and abrupt effective maximum close to 180 bp.

In each panel of Figure P1.4.3 the values of the associated persistence lengths for the six distinct poly-dinucleotide sequences are also shown as circles (see Table P1.4.2). It is evident that for these particular sequences there is particularly strong sequence dependence of both persistence lengths.

We take this opportunity to use a similar approach to compare the *cgDNA* parameter set *cgDNAparamset2* extracted using a new parameter fitting procedure of [GonPetPas] that involves maximum (absolute) entropy fitting of Chapter P1.1 with the original *cgDNAparamset1* of [Pet2012; GonPetMad2013; PetPasGonMad2014]. This is to say that we study the difference of predictions of persistent lengths for the *cgDNA* model equipped with either of the two parameter sets using our Monte Carlo simulations. Figure P1.4.4 shows normalized histograms of persistence length $\ell_p^{[0]}(S)$ computed using both parameter sets (the blue histogram in Figure P1.4.4 is the same as the blue one of Figure P1.4.3d). It can be seen that in case of *cgDNAparamset1* the histogram is shifted towards much higher values with the harmonic average equal to 187 bp. This is one of the justifications that *cgDNAparamset2*, used throughout this thesis, has been put forward as the currently preferred parameter set for the *cgDNA* model [GonPetPas].

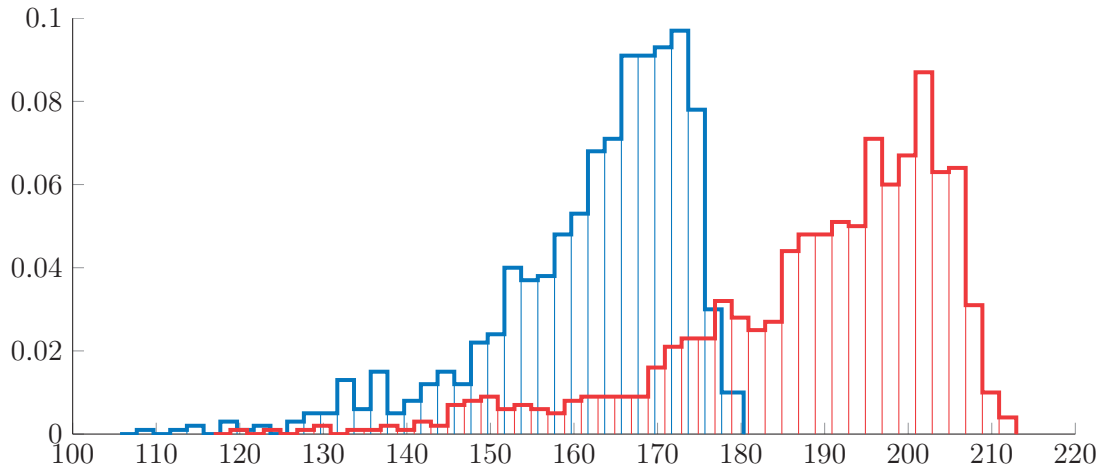


Figure P1.4.4. *Comparison of the normalized histogram of persistence lengths* $\ell_p^{[0]}(S)$ *for the random ensemble of Figure P1.4.3d, obtained for cgDNAparamset2, with analogous data computed in the cgDNA model based on cgDNAparamset1. We recall that for cgDNAparamset2 the harmonic average of* $\ell_p^{[0]}(S_j)$ *is* 160 *bp. For cgDNAparamset1 the harmonic average of* $\ell_p^{[0]}(S_j)$ *is* 187 *bp.*

P1.4.3.3 Sensitivity to the Jacobian perturbation

Figure P1.4.5 provides two examples showing differences between tangent-tangent correlation data for the two ensembles that are small, but perceptible, and accumulating with base pair index. We accordingly conclude that while the effect of the Jacobian perturbation to the equilibrium ensemble deserves further investigation in the case of long segments of DNA, it has a negligible influence on the computation of tangent-tangent persistence lengths at the scale of 200 bp and the simulations yielding the presented results do not include it.

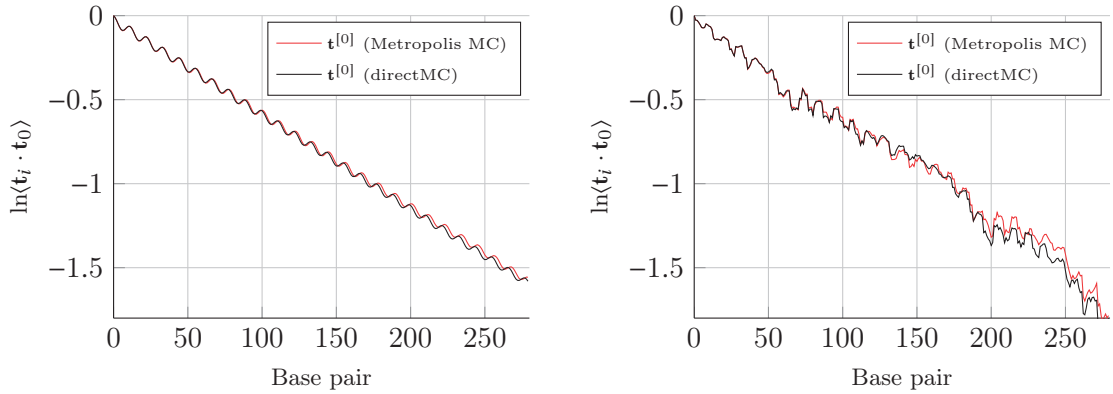


Figure P1.4.5. *Sensitivity of tangent-tangent correlation data to inclusion of the cgDNA Jacobian factor. Direct Monte Carlo simulation (which does not use the Jacobian) in black, Metropolis Monte Carlo (which incorporates Jacobian) in red. Panel (a) is for* 300 *bp* poly(G) *fragment with* $\ell_p^{[0]}(\mathtt{G}) = 173$ *bp for direct and* 175 *bp for Metropolis Monte Carlo. Panel (b) is for* S^λ *with* $\ell_p^{[0]}(S^\lambda) = 162$ *bp for direct Monte Carlo and* 167 *bp for Metropolis. In each case* 10 *bp were excluded from either end.*

P1.4.3.4 Convergence of MC simulations

Irrespective of sequence, estimates of the Flory persistence vectors appeared to be converged to a standard error of less than 0.5 nm for fragments of 1500 bp for multiple estimates each with 10^5 direct Monte Carlo samples. Similarly, in the computation of tangent-tangent correlations of 300 bp fragments, 10^5 direct Monte Carlo samples give a standard error of less than 1 bp in $\ell_p^{[0]}$. For Metropolis MC simulations, longer runs are required for the same level of accuracy: 10^6 accepted samples for ℓ_p and $3 \cdot 10^6$ for ℓ_F. All our reported values for single-molecule ℓ_p and ℓ_F come from samples that meet or exceed these requirements. Figure P1.4.6 shows example convergence plots.

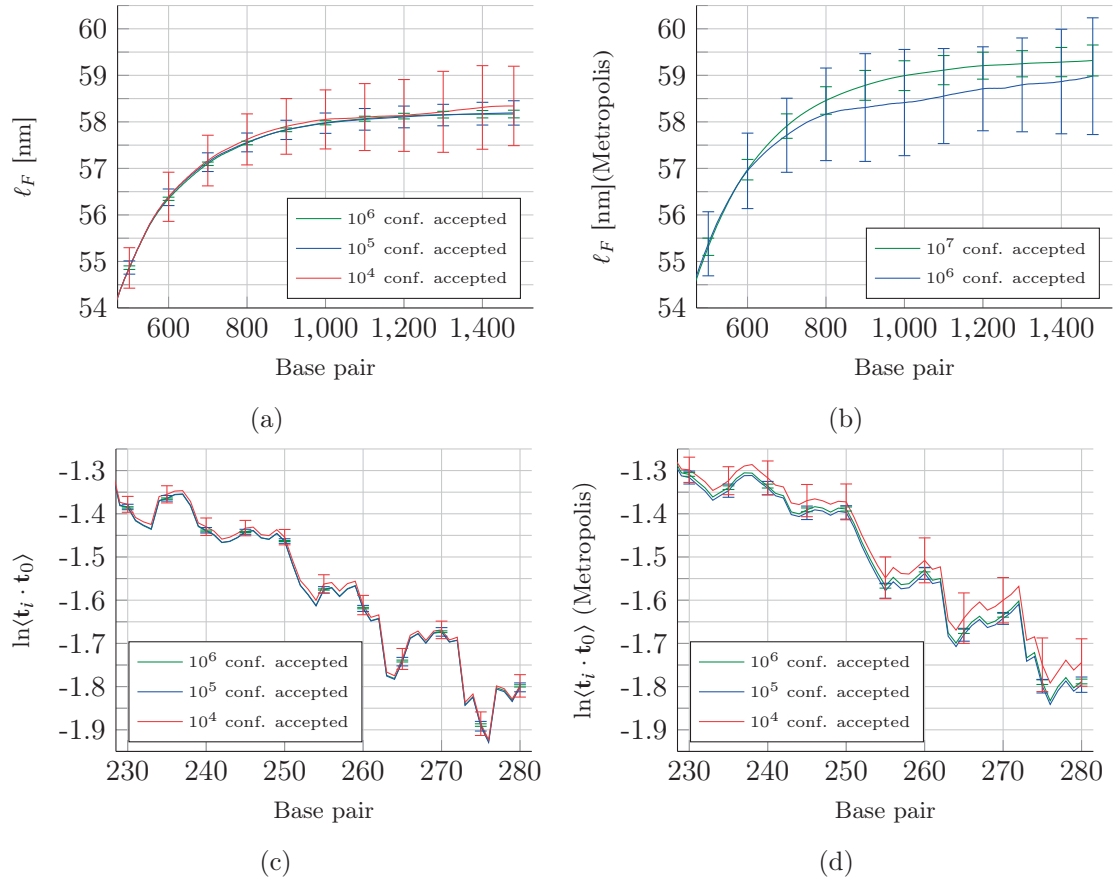


Figure P1.4.6. *Example convergence plots of direct Monte Carlo sampling (left column) and Metropolis Monte Carlo (right column) for the S^λ sequence. In panels (a) and (b) the curves show the norm of the Flory vector (averaged over MC samples of different sizes) plotted against base-pair number. The error bars give the standard error obtained for ten independent MC runs and are plotted every 100 bp. Panels (c) and (d) show the last 50 bp of the tangent-tangent correlation plot relevant for computing $\ell_p^{[0]}$ of a single repeat of λ_3 (averaged over MC samples of different sizes). The error bars, (plotted every 5 bp) give the standard error obtained for ten independent MC runs. The acceptance rate for the Metropolis Monte Carlo was 4% in case (b) and 37% in case (d).*

P1.4.3.5 Sequence-averaged persistence lengths

Table P1.4.3 provides estimates for $\overline{\ell}_F$ and $\overline{\ell}_p^{[0]}$ from evaluation of the sequence ensemble formulae in (P1.4.4b, P1.4.5b) for both of our examples (namely 1000 random 220 bp fragments, and the collection of two hundred and twenty fragments of the λ-phage genome of length 220 bp, as described in Section P1.4.3.2). For the random ensemble, we sample sufficient sequences and MC configurations for each sequence, to produce standard errors below 1 bp/0.1 nm for $\overline{\ell}_p^{[0]}$ and $\overline{\ell}_F$ (by sampling 10^5 configurations for each of the 1000 random sequences). In contrast, λ-phage is a fixed sequence, which we have chosen to divide into consecutive 220 bp fragments; we draw sufficient MC samples to produce the same small standard error for the average over that particular set of fragments, but this does not guarantee the same small variation over different choices of λ-fragments.

	$\overline{\ell}_F$	$\overline{\ell}_p^{[0]}$
random ensemble	53.5 nm	160 bp
λ ensemble	53.4 nm	160 bp

Table P1.4.3. *Sequence averaged persistence lengths for the random and λ sequence ensembles.*

These single ensemble estimates of $\overline{\ell}_p^{[0]}$ and $\overline{\ell}_F$ are very close to the appropriate averages of the histograms of the sequence-dependent quantities illustrated in Figure P1.4.3.

Part 2

Continuum DNA modelling

P2.1 Numerical issues with birod DNA coefficients

In this short chapter we address practical issues arising in numerical computations on the birod system (B.3.57) using the Hamiltonian version (B.3.54), (B.3.58) of the DNA coefficients of [Gra2016] (described in Section B.3.3.3). Identification of those issues, as well as verification of the proposed results was highly facilitated by use of the *bBDNA* software presented in Chapter P2.2.

The first problem is related to the numerical stiffness of the birod system with DNA coefficients. What we mean here is that standard numerical solvers of Initial Value Problems (IVP) of ordinary differential equations cannot reconstruct birod DNA solutions from initial values even with a very small step size This not only makes it practically impossible to reconstruct birod DNA solutions using an IVP solver but also hinders computations in the *AUTO-07p* continuation package (see Section B.2.4) by affecting the provided bifurcation detection procedure. We accordingly present a modification of the original bifurcation detection function in *AUTO-07p* that overcomes this problem.

As indicated by Equation (B.3.49) the DNA coefficients are piecewise continuous within each DNA junction. The rapid variation of coefficients within junctions and discontinuities at base pair positions are, however, very pronounced and pose yet another pragmatic problem. We have observed that computations with *AUTO-07p* fail to converge unless the ends of the discretization mesh intervals are chosen to exactly match the base pairs. This means that the total number of collocation points required to perform continuation in the continuum DNA model needs to be ~ 4 times higher than the number of base pairs. Given the necessity of such a fine discretization the use of a continuum model might seem questionable, even though to date it provides the only way of performing analysis of the kind presented in Chapter P2.3, and the efficiency of such computations is acceptable. We address this issue by applying of a coefficient homogenization technique based on the one presented in [Gra2016, sec. 7.3] in the context of rod models.

P2.1.1 Bifurcation detection in *AUTO-07p*

Difficulties with numerical treatment of the birod system with the DNA parameters described in Section B.3.3.3 manifested themselves when parameter continuation computations using the *AUTO-07p* package were first performed in the model. During each continuation run nearly every computed point was reported to be a bifurcation point. Bifurcation detection in *AUTO-07p* is done by tracking the sign of the determinant of the Jacobian matrix $G_{\boldsymbol{x}}$ (see Equation (B.2.7)) used for computing the solutions (see Section B.2.3), so the supposed rapid changes in the sign seemed to suggest that the Jacobian was close to singular. It has been observed, however, that the matrix remained invertible during the continuation. The conclusion was that the issue had to be related to the method of evaluating the so called *bifurcation function* that represents the sign of the determinant of the Jacobian in *AUTO-07p*.

It should be pointed out that the same deficient behaviour was observed in the case of uniform coefficients obtained through averaging of the sequence dependent ones. Hence, the problem could not be attributed to the aforementioned discontinuities of the coefficient functions at the values of the birod independent parameter $s = s_n^{(N)}$ corresponding to base pairs. Moreover, parameter continuation could be performed in the birod system without trouble for certain randomly generated non-physical coefficients, although no physically sensible scaling introduced in the DNA case was found to eliminate the failures in bifurcation detection.

The conclusion that the DNA birod system is a stiff one came when all the solvers provided by *MATLAB*®, including ones designed for stiff problems, were failing to solve the birod IVPs with initial values obtained from a known solution, computed in *AUTO-07p*. During numerical integration after reaching a value of the birod independent variable s corresponding to around 30 base pairs the birod started to differ significantly from the respective *AUTO-07p* solutions. Figure P2.1.1 presents example results of analysis of linear numerical stiffness of the IVP for the minimum energy minicircle solution of sequence S^{γ} of the following section.

To explain how the fact that our problem is numerically stiff affects bifurcation detection in *AUTO-07p* we need to briefly outline the original implementation of the detection procedure. The process of parameter continuation involves solving the Jacobian system presented in Equation (B.2.4) using Newton's method. As detailed in [DoeKelKer1991b], the linear solve implemented in *AUTO-07p* (needed in every Newton iteration) uses the sparse structure of the Jacobian of the boundary value problem (different from the IVP Jacobian considered above). Using Gauss elimination with full pivoting the system is separated into a large triangular part and a small, dense part. Two sub-blocks of the small square system are related to an approximation of the linearized mapping from the state variables at $s = 0$ to those at $s = L$. Inspection of the source code revealed that the bifurcation function was computed using the determinant of one of those blocks,

which for our system happens to be numerically badly conditioned. We associate the ill-conditioning of the linearized end-to-end mapping with the stiffness of the DNA birod system.

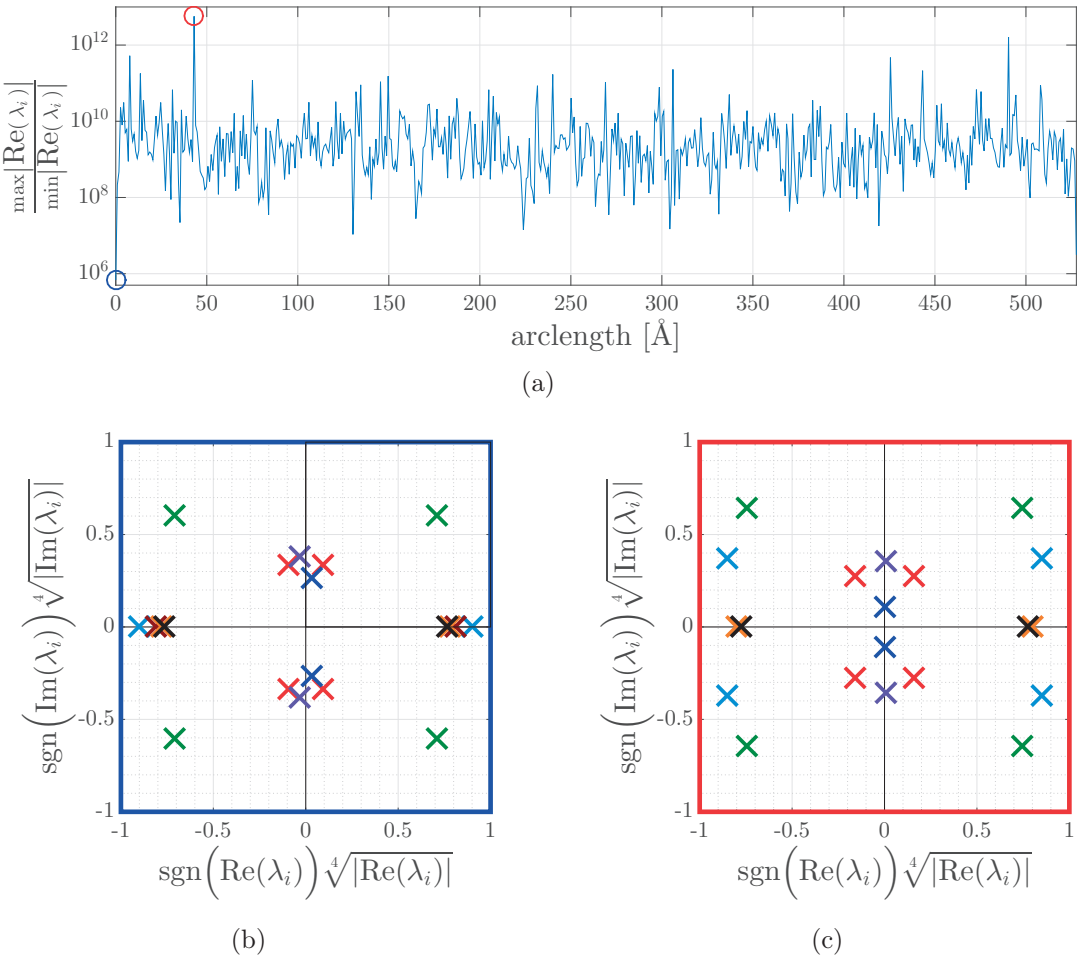


Figure P2.1.1. *Example analysis of numerical stiffness of a birod DNA Initial Value Problem (IVP). Panel (a) shows the condition number of the numerical approximation of the Jacobian of the linearized IVP evaluated in MATLAB® for the particular case of the minimum energy fully closed solution of sequence S^γ presented in the following section. The average value of the condition number of 2.0×10^{10} indicates the system is highly stiff. Panels (b) and (c) present all the (non-zero) eigenvalues λ_i of the Jacobian with the lowest and highest condition number, respectively, indicated by circles in Panel (a). The zero eigenvalue with algebraic multiplicity three and geometric multiplicity six is associated with the fact that the right hand side of the birod system (B.3.16) does not depend on the average rod position $\mathbf{r}(s)$. The real and imaginary parts of the eigenvalues have been transformed through a quartic root to better differentiate the small values. Note that one conjugate pair of eigenvalues is always extremely close to being pure imaginary – its real part is below 9.4×10^{-7} for all values of the independent variable.*

Our proposed solution is very straightforward. The original bifurcation function is replaced with another one that is also, but more directly, related to the sign of the determinant of the Jacobian matrix. As the Gauss elimination of the system is performed the new procedure tracks the sign of each diagonal element $j_{k,k}$ of the resultant triangular system as well as the number n_e of all row and column exchanges due to pivoting. The new choice of the bifurcation function is:

$$f_B(G_{\boldsymbol{x}}) := \min_{k\in\{1,\dots N_G\}} \{|j_{k,k}|\} \cdot (-1)^{n_e} \prod_{k=1}^{N_G} \operatorname{sgn}(j_{k,k}) \tag{P2.1.1}$$

were N_G is the dimension of the Jacobian matrix $G_{\boldsymbol{x}}$. Note that the definition (P2.1.1) is closely related to the standard expression $(-1)^{n_e}\prod_{k=1}^{N_G} j_{k,k}$ for the determinant of a triangular system that is *e.g.* implemented by the `det` function in *MATLAB*®. The dimension N_G of the system is so big that the numerical value of the determinant exceeds the capacity of a double precision floating point number representation. Hence only signs of the diagonal elements are multiplied through in the definition (P2.1.1). The multiplication of the expression for the sign by the minimum absolute value amongst the pivots $j_{k,k}$ makes the bifurcation function continuous, which is a requirement of *AUTO-07p*. This choice is supported by the fact that it is the smallest diagonal entry of the reduced system that vanishes when the system becomes singular.

The *AUTO-07p* package with the proposed modification has been used with success to produce all rod and birod examples of Chapter B.3, all the results of Chapter P2.3, all the test cases of Section P2.1.2, as well as all DNA birod model examples of [Gra2016]. It also passed all the standard tests provided with *AUTO-07p*. We note here that the modified code might be faster or slower than the original, depending on the dimensions of the state variable and adopted discretization. In any case evaluation of the bifurcation function constitutes a very small fraction of all calculations required to compute each solution. For the DNA birod model the difference in time efficiency is unnoticeable.

P2.1.2 Homogenization of the DNA birod coefficients

(joint work with A. Grandchamp)

Parameter continuation in *AUTO-07p* in a birod system (B.3.57) with the original version of the DNA coefficients requires a very fine spacial discretization to be used. For example with coefficients as computed using the procedure (B.3.49), (B.3.54) for an oligomer of 158 base pairs 633 collocation points are necessary for convergence of the numerical method. Nevertheless the computations can be performed rather efficiently. Symmetry breaking resulting in generation of a bifurcation diagram of the kind presented in Section P2.3.2, containing around 500 solutions, takes around 15 minutes.

As argued in [Gra2016, sec. 6.3] solutions computed through such continuation in the birod DNA model, when discretized, provide very good starting points for a discrete

solver that given an initial approximation of a *cgDNA* shape finds the critical point of the *cgDNA* energy. For reference we only mention here that the current version of the discrete solver takes approximately 5 minutes to converge for a single solution.

We recall here that the *cgDNA* model was shown to reproduce well the ground state statistics of molecular dynamics simulations [Pet2012; GonPetMad2013] and for that reason was chosen as reference of the birod DNA model. So far the only way to find stationary solutions for *cgDNA*, is exactly to run the discrete solver on guesses constructed from continuous solutions computed in the DNA birod model. In this context non-local and non-linear constraints in *cgDNA* variables are replaced by simple boundary conditions in the birod formulation. This technique will also be used to evaluate the appropriateness of the procedure for homogenization of DNA coefficients that is presented below.

The procedure of homogenization of the DNA birod coefficients presented here is a direct application of the method presented in [Gra2016, sec. 7.3] in the context of computing certain expectations in the rod model, where a theoretical justification of the process, based on an argument of scale separation can be found. Here we only state the method and evaluate its efficacy using the mentioned discrete solver for the *cgDNA* model.

The procedure comprises three steps. At first the piecewise helix $\widehat{\mathcal{D}}^{(N)}(s)$ defined by $\widehat{\boldsymbol{\xi}}(s)$ (see Equation (B.3.42)) is factored out:

$$\widehat{\boldsymbol{\xi}}_1(s) :\equiv \mathbf{0} \in \mathbb{R}^6 \tag{P2.1.2a}$$

$$\boldsymbol{H}_1(s) := \begin{bmatrix} \boldsymbol{I}_{12\times 12} & \boldsymbol{I}_{12\times 6} \\ \boldsymbol{I}_{6\times 12} & \mathrm{Ad}_{\widehat{\mathcal{D}}^{(N)}(s)} \end{bmatrix} \boldsymbol{H}(s) \begin{bmatrix} \boldsymbol{I}_{12\times 12} & \boldsymbol{I}_{12\times 6} \\ \boldsymbol{I}_{6\times 12} & \mathrm{Ad}^T_{\widehat{\mathcal{D}}^{(N)}(s)} \end{bmatrix} \quad , \tag{P2.1.2b}$$

where $\mathrm{Ad}_{\widehat{\mathcal{D}}^{(N)}(s)}$ is defined in Equation (B.3.47).

Next the actual homogenization of the Hamiltonian matrix $\boldsymbol{H}(s)$ is performed by window averaging for a chosen half-window size Δs as:

$$\boldsymbol{H}_2(s) := \frac{1}{2\Delta s} \int_{s-\Delta s}^{s+\Delta s} \boldsymbol{H}_1(s)\, \mathrm{d}s \quad . \tag{P2.1.3}$$

Finally a constant helix that matches the final base pair frame, defined as:

$$\begin{bmatrix} \left[\overline{\boldsymbol{U}}^{\times}\right] & \overline{\boldsymbol{V}} \\ \mathbf{0} & 0 \end{bmatrix} := \frac{1}{L} \ln\left\{\widehat{\mathcal{D}}^{(N)}(L)\right\} \tag{P2.1.4}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\overline{\mathcal{D}}(s) := \overline{\mathcal{D}} \begin{bmatrix} \left[\overline{\boldsymbol{U}}^{\times}\right] & \overline{\boldsymbol{V}} \\ \mathbf{0} & 0 \end{bmatrix} \quad , \tag{P2.1.5}$$

(with $\ln\{\cdot\}$ as in Equation (B.3.41)), is factored in through a procedure inverse to (P2.1.2).

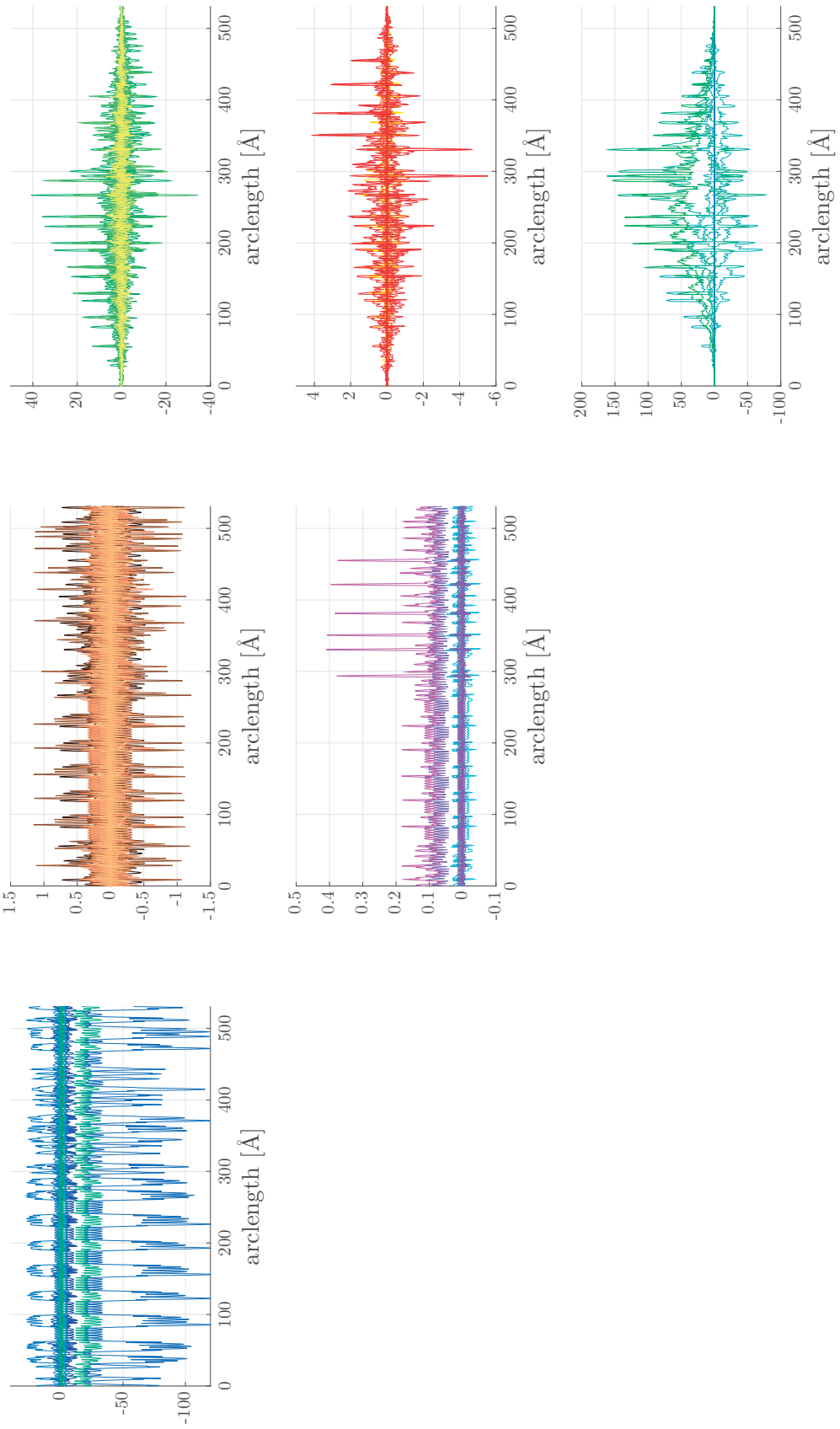


Figure P2.1.2. *Hamiltonian coefficients of the oligomer S^Y without homogenization. Each panel presents coefficients of a 6×6 sub-block of the Hamiltonian matrix $\boldsymbol{H}(s)$ plotted against the arclength s of the unstressed configuration. The intrinsic shape has been factored out and the constant helix has been factored in to make the plots comparable with the homogenized coefficients shown on the following page.*

Figure P2.1.3. Hamiltonian coefficients of the oligomer S^{γ} after homogenization.

The last step, which can be written as:

$$\widehat{\boldsymbol{\xi}}_3(s) :\equiv \left[\begin{array}{c} \boldsymbol{U} \\ \hline \boldsymbol{V} \end{array}\right] \tag{P2.1.6a}$$

$$\boldsymbol{H}_3(s) := \begin{bmatrix} \boldsymbol{I}_{12\times 12} & \boldsymbol{I}_{12\times 6} \\ \boldsymbol{I}_{6\times 12} & \mathrm{Ad}^{-1}_{\mathcal{D}(s)} \end{bmatrix} \boldsymbol{H}(s) \begin{bmatrix} \boldsymbol{I}_{12\times 12} & \boldsymbol{I}_{12\times 6} \\ \boldsymbol{I}_{6\times 12} & \mathrm{Ad}^{-T}_{\mathcal{D}(s)} \end{bmatrix} \tag{P2.1.6b}$$

ensures that the last frame $(\mathbf{r}(L), \boldsymbol{R}(L))$ is the same for the original coefficients and the homogenized ones. This is necessary for consistency in the definition of boundary conditions at $s = L$. Simpler versions of this factorization for rod models have previously been published in [KehMad2000; ReyMad2000].

For practical reasons coefficients have to be given to the birod *AUTO* problem script as piecewise polynomials (see Section P2.2.1). As a result the implementation of the homogenization described above after the first step (P2.1.2) divides the entire sequence into subregions of requested number of base pairs N_i, possibly different for different intervals. The actual homogenization step (P2.1.3) is performed in each interval for possibly different values of the half-window Δs_i. Finally after the last step (P2.1.6) a 4th degree polynomial is fit to $\widehat{\boldsymbol{\xi}}_3(s)$ and $\boldsymbol{H}_3(s)$ in each interval and the piecewise polynomial is returned.

This approach allows for selection of regions of interest in the sequence where the coefficients are only weakly homogenized and others where the homogenization smoothing is stronger. A comprehensive study of effects of such homogenization remains to be performed. Below we will show an evaluation of the (uniform) choice of sub-region length $N_i = 10$ base pairs and the half-window size $\boldsymbol{H}_3(s) = 3\,\text{Å}$ (*i.e.* around 1 base pair) made for the purposes of this thesis. This kind of homogenization reduces the required density of the discretizations ten times so that, roughly speaking, one collocation point can be used every 2.5 base pairs. A comparison of non-homogenized coefficients for the sequence S^γ of [KahCro1992] with the result of such homogenization is presented in Figures P2.1.2 and P2.1.3, respectively. An almost tenfold speed-up in "homogenized" vs "non-homogenized" computations can be observed, *e.g.* the 849 solutions constituting the bifurcation diagram of Figure P2.3.4 were computed in under 3 minutes.

The boundary value problem chosen as a test case are the closed loop boundary conditions in the average rod part (see Equation (B.3.26)), and free end conditions in the microstructure (see Equation (B.3.61)). This choice was motivated by the fact that such boundary conditions are consistent with the constraints of the mentioned *cgDNA* discrete solver of [Gra2016, sec. 6.3]. The solver was used to assess how well the birod solutions of the homogenized system approximate stationary solution of the *cgDNA* model.

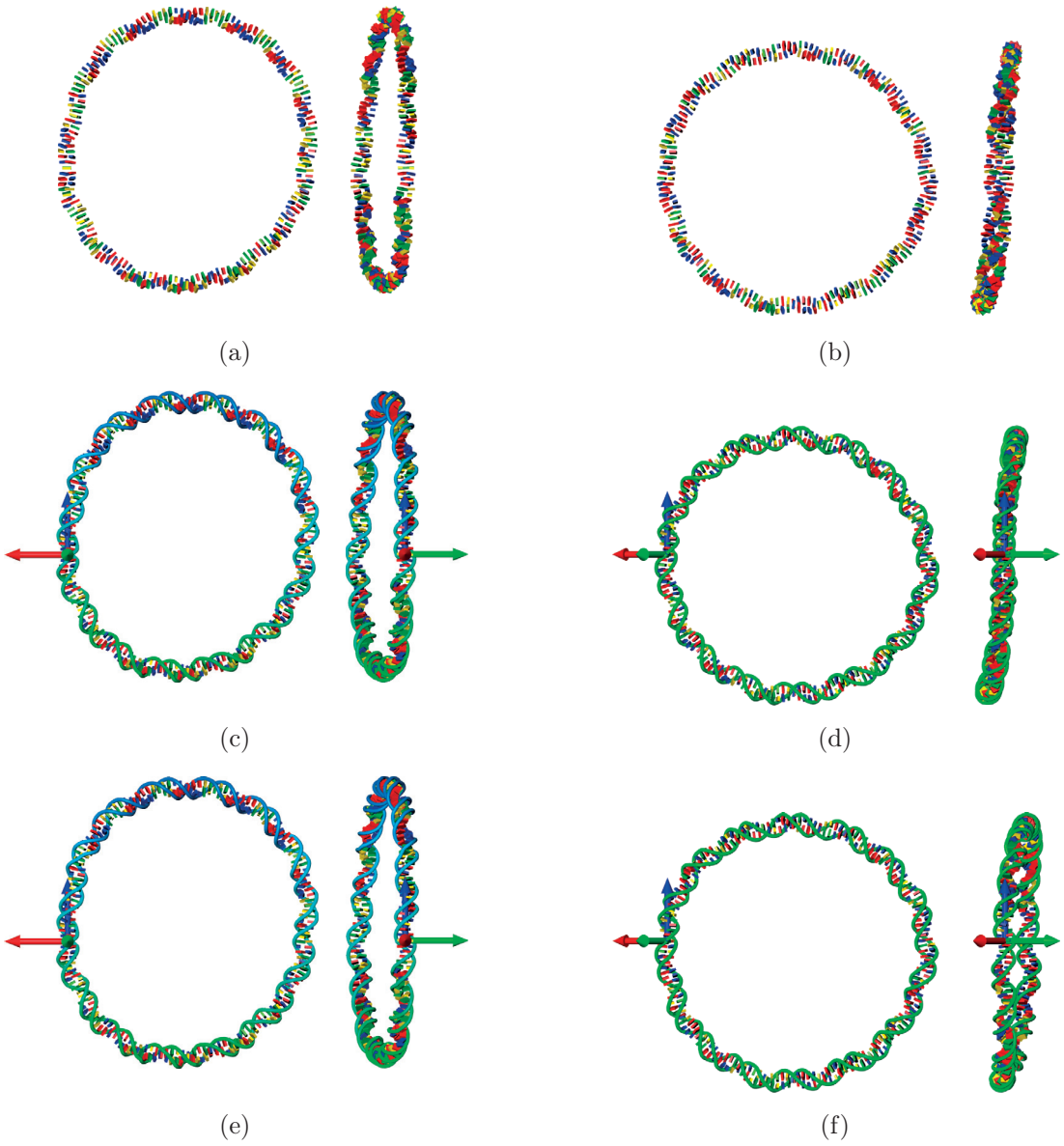


Figure P2.1.4. *Comparison of 3D shapes of fully closed DNA birod loops computed with and without homogenization of coefficients. Left column presents the minimum energy fully closed loop of sequence* S^{γ}*, with* $Lk = 15$*. The right column shows the second lowest energy* $Lk = 15$ *fully closed loop of sequence* $S^{\lambda''}$*. Configurations in the top row are results of the discrete solver, while the other two were computed in bBDNA. The middle row was computed with non-homogenized coefficients, while the computations for the bottom row used homogenized coefficients. In case of the* S^{γ} *sequence differences between all three solutions are almost impossible to see. For the* $S^{\lambda''}$ *sequence small differences in register can be observed (particularly for the homogenized solution (f)), which is manifested as higher differences reported for that sequence in Tables P2.1.1 and P2.1.2.*

Two DNA sequences were considered in our test, namely a highly bent one S^{γ} [KahCro1992] and a relatively straight one $S^{\lambda''}$ (the same as the sequences used in Section P2.3.2). For each sequence four fully closed loops (for which both $\mathbf{r}(0) = \mathbf{r}(L)$ and $\boldsymbol{R}(0) = \boldsymbol{R}(L)$) will be considered, namely the lowest energy (conjectured to be stable, labelled as E_{low}) and second lowest energy (most likely unstable, labelled as E_{hi}) solutions of linking number $Lk = 14$ and $Lk = 15$. Note that in Figure P2.1.4 the compared solutions are aligned by the base pair, where the closure conditions were requested in *bBDNA*. On the other hand the data in Tables P2.1.1 and P2.1.2 was computed after aligning the solutions by the rigid body displacement that minimizes the least squares distance between respective base pair positions.

	S^{γ}				$S^{\lambda''}$			
	$Lk = 14$		$Lk = 15$		$Lk = 14$		$Lk = 15$	
	E_{low}	E_{hi}	E_{low}	E_{hi}	E_{low}	E_{hi}	E_{low}	E_{hi}
ΔE	1.2 %	1.4 %	0.3 %	1.7 %	1.5 %	1.1 %	1.3 %	1.9 %
$\Delta\mathbf{r}_{\max}\left(s_n^{(N)}\right)$ [Å]	0.7	0.9	0.3	1.4	1.0	1.6	0.4	2.0
$\Delta\boldsymbol{R}_{\max}\left(s_n^{(N)}\right)$	3.4°	3.7°	3.7°	3.8°	6.7°	6.2°	4.6°	7.1°

Table P2.1.1. *Comparison of base pair positions and orientations between fully closed loops computed with and without homogenization and sampled at base pairs. The first row presents relative differences ΔE of the Lagrangian energy, as computed using bBDNA – see Chapter P2.2. The second row contains maximum distance $\Delta\mathbf{r}_{\max}\left(s_n^{(N)}\right)$ between positions of respective base pairs. In the last row relative rotations between respective base pair orientations are indicate using the angle of rotation.*

		S^{γ}				$S^{\lambda''}$			
		$Lk = 14$		$Lk = 15$		$Lk = 14$		$Lk = 15$	
		E_{low}	E_{hi}	E_{low}	E_{hi}	E_{low}	E_{hi}	E_{low}	E_{hi}
ΔE	nh	7.3 %	6.6 %	8.2 %	9.7 %	7.7 %	7.3 %	9.2 %	9.4 %
	h	8.4 %	8.0 %	8.4 %	11.2 %	9.1 %	8.4 %	10.4 %	11.1 %
$\Delta\mathbf{r}_{\max}\left(s_n^{(N)}\right)$ [Å]	nh	0.14	0.50	0.03	0.15	0.23	0.25	0.22	0.23
	h	0.84	0.53	0.40	1.68	0.67	1.33	1.13	2.36
$\Delta\boldsymbol{R}_{\max}\left(s_n^{(N)}\right)$	nh	1.7°	3.6°	1.7°	3.3°	2.7°	3.1°	2.1°	3.3°
	h	3.7°	5.7°	3.0°	6.2°	5.6°	8.8°	5.8°	9.8°

Table P2.1.2. *Comparison of base pair positions and orientations between fully closed loops computed with bBDNA and sampled at base pairs with the respective stationary solutions of the cgDNA model. The rows of the table are analogous to those of Table P2.1.1. Here, however each of the three rows compares the cgDNA stationary solution to its two continuous approximations computed with non-homogenized (nh) and homogenized (h) birod coefficients.*

Table P2.1.1 presents a comparison of solutions computed with and without homogenization. The errors due to homogenization are presented as differences of the Lagrangian energy, maximum displacement between positions of respective base pairs and angles of relative rotations between respective base pair orientations. Table P2.1.2 presents a comparison of the continuous birod DNA solutions (both homogenized and not) with the respective stationary solutions of the *cgDNA* model. The *cgDNA* solutions were obtained using the discrete solver of [Gra2016, sec. 6.3]. We note that in all eight test cases the discrete *cgDNA* solver converged to the same discrete solution when initialized from discretized non-homogenized and homogenized birod solutions. Finally, Figure P2.1.4 presents the 3D configurations of both continuous solutions (homogenized and non-homogenized) and the discrete one for two test cases. One of the test cases is the higher energy $Lk = 15$ loop for $S^{\lambda\prime\prime}$, which shows the biggest discrepancy in all three presented numerical values of Table P2.1.1. The other selected test case is the lower energy, $Lk = 15$ solution for S^{γ}, which happens to be the minimum energy fully closed loop for that sequence. This minimum energy loop of S^{γ} shows the smallest difference of continuum energies.

The most striking observation based on the presented results is the big ($\sim 10\,\%$) discrepancy between the discrete and continuum energies demonstrated in Table P2.1.2. This error is of the same order for solutions computed using homogenized and non-homogenized coefficients. This difference is conjectured to be due to the use of piecewise polynomial instead of piecewise helical interpolation for computing continuum solutions, although this question is still being investigated.

The errors in the 3D configuration are also generally greater for the higher energy solution for the same sequence and the same linking number. This might be associated with the conjectured stability of the lower energy solutions and instability of the higher energy ones (these conjectures are further discussed in Section P2.3.2).

The generally higher discrepancies in the case of sequence $S^{\lambda\prime\prime}$ as compared to S^{γ} should most certainly be attributed to the intrinsic shape of those sequences. In case of S^{γ} the plane in which the loops are formed is strongly determined by the high intrinsic bend of that sequence. On the other hand in the case of the intrinsically straight $S^{\lambda\prime\prime}$ the energy landscape around the chosen fully closed loops is much more flat. As a result a small difference in register (*i.e.* a rotation of each cross section locally around $\mathbf{d}_3$ – see Section B.3.2.3) occurred, which is visible between Figures P2.1.4d and P2.1.4f. The intrinsic shapes of both sequences are presented in Section P2.3.2.

Other than that, in accord with [Gra2016, sec. 6.3], we find good agreement between the solutions of birod DNA computations and their *cgDNA* counterparts. The proposed homogenized birod DNA model provide satisfactory approximation to the non-homogenized one. The homogenization introduces only small errors in the resulting solutions, which pose no difficulty to the discrete *cgDNA* solver when discretizations of those continuous

solutions are used as initial guesses. Nevertheless the subject of smoothing the DNA birod coefficients requires further study. For example the effect of different parameters of the described homogenization remain to be investigated.

In short, *AUTO-07p* computations on the birod DNA model with homogenized coefficients are notably faster than those with non-homogenized coefficients. No significant additional error is introduced, as running the discrete solver of [Gra2016, sec. 6.3] on sampled solutions for homogenized coefficients converged to the same discrete solution as when running the solver from initial guesses sampled from solutions for non-homogenizes coefficients.

P2.2 The *bBDNA* software for interactive parameter continuation and visualization of birod DNA

This chapter introduces the *bBDNA* software which is an interactive parameter continuation and visualization tool for the continuum birod model of DNA [Gra2016] as described in Section B.3.3. The design of core aspects of the user interface of the application was modelled on an older package, called Visualization for Bifurcation Manifolds (*VBM*) of [Paf1999a; Paf1999b]. Certain ideas were also drawn from the graphical user interface to *AUTO-07p* called *PLAUT04* [DoeChaDer2009]. That is to say that *bBDNA* provides a graphical user interface that allows for visualizing families of solutions to the birod system as one-dimensional bifurcation diagrams with special solutions, such as branch points, marked accordingly. It also allows inspection of the data of each solution in the diagram through probes. For a given solution a probe allows reconstruction of the 3D configurations of the DNA molecule represented by the solution at different levels of details. It also provides 2D plots of the dependent variables of the birod system and predefined functions of these variables as well as access to raw numerical data. A key feature of *bBDNA* is the ability to extend bifurcation diagrams by interactively starting computations from a solution selected using a probe and to include the outcome in the diagram.

However, it should be pointed out that *bBDNA* is not exactly a successor of *VBM*. The latter provided a general framework for solving various parameter continuation problems using different solvers. From the very beginning *bBDNA* was meant to be specialized for solving boundary value problems in the continuum birod DNA model of [Gra2016], using the *AUTO-07p* solver [DoeChaDer2009]. The goal was to build a tool that would allow the study of DNA molecules using an approach that has been successfully applied to an elastic rod model of DNA [ManMadKah1996; FurManMad2000]. The birod model is a natural extension of the rod model but, unlike for rods (as shown in [LanGonHef2009]), for birods there exist apparently accurate sequence-dependent DNA coefficients [Gra2016].

We also point out that *bBDNA* was developed alongside the DNA birod model of Grandchamp. Computations performed using the application influenced the final formulation of the birod system, briefly described in Section B.3.3. They also helped in identifying practical issues (addressed in Chapter P2.1) concerning the sequence-dependent DNA birod coefficients as extracted from the *cgDNA* model [Gra2016, sec. 4.2] (see Section B.3.3.3).

All birod DNA results of this thesis were computed within *bBDNA* and all the related figures were made using it. In fact all the figures of the background Chapter B.3 also originated from *bBDNA*.

P2.2.1 The design of the software

We remark here that what we describe in this section is the entire software pipeline developed to prepare and run parameter continuation in the birod DNA model. This includes *MATLAB*® scripts for preparing input coefficients files for the *AUTO* birod DNA problem script, the *AUTO* script itself and finally the *bBDNA* GUI that constitutes the largest part of the mentioned pipeline. This way *bBDNA* can be explained in a wider context and certain design decisions become clear.

P2.2.1.1 Scripts for computing birod DNA coefficients

Sequence-dependent DNA birod coefficients are a prerequisite for any computation in the birod model of DNA. As discussed in Section B.3.3.3, the Lagrangian coefficients $\big($*i.e.* the stiffness matrix $\boldsymbol{K}(s)$, the boundary intra stiffness matrices $\boldsymbol{K}_0$, $\boldsymbol{K}_L$, and the intrinsic shape internal parameters $\widehat{\boldsymbol{y}}(s)$, $\widehat{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s)$ and $\widehat{\boldsymbol{\xi}}(s)\big)$ can be computed using a parameter set of the *cgDNA* model [Gra2016, sec. 4.2]. In all our cases this is done using the latest available *cgDNAparamset2* introduced in Chapter P1.1, but any other *cgDNA* parameter set could be used.

As subsequently explained in Section B.3.3.4, in the final formulation of the DNA birod Hamiltonian system the coefficients are only the intrinsic macrostructure strains $\widehat{\boldsymbol{\xi}}(s)$ and the Hamiltonian matrix $\boldsymbol{H}(s)$, which can be computed from the stiffness $\boldsymbol{K}(s)$ using the Legendre transform (B.3.54). This is because the microstructure variables $\widetilde{\boldsymbol{y}}(s)$, $\widetilde{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s)$ describe perturbations from the respective intrinsic values $\widehat{\boldsymbol{y}}(s)$, $\widehat{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s)$ and the intrinsic values do not appear in the Hamiltonian formulation (B.3.57), (B.3.58). The boundary intra stiffness matrices $\boldsymbol{K}_0$, $\boldsymbol{K}_L$ are necessary for the free end microstructure boundary conditions (B.3.61).

We recall that the coefficients mentioned so far are piecewise continuous with discontinuities at base pairs. As pointed out in Section P2.1.2 the discontinuities were found to be very pronounced and as a result excessively fine discretization mesh was required in order

for the *AUTO-07p* solver to converge. For that reason we introduced the procedure described in Section P2.1.2 that yields homogenized version of the Hamiltonian coefficients (see Equation (P2.1.6)), where the macrostructure strains $\widehat{\boldsymbol{\xi}}_3$ are constant along s.

To include both the original piecewise discontinuous coefficients and their homogenized version, as well as possible future modifications of the birod parametrization procedure a piecewise polynomial representation of $\boldsymbol{H}(s)$ and $\widehat{\boldsymbol{\xi}}(s)$ has been proposed for the birod DNA coefficient file. The division of the interval $]0, L[$ into polynomial pieces can be arbitrary. So can the degree of the polynomial, although it must be the same for all pieces. Such an approach allows for a complete decoupling of the implementation of the birod DNA system from the actual functional form of $\boldsymbol{H}(s)$ and $\widehat{\boldsymbol{\xi}}(s)$, which can be highly complicated. It also introduces an approximation error, which can however be minimized by careful choice of the polynomial degree and intervals.

Note also that in order to reconstruct the 3D configuration together with other quantities, such as the (dimensional) stresses $\mathbf{m}(s)$ and $\mathbf{n}(s)$ additional information is required. For that reason the *bBDNA* supplementary data including the sequence of the oligomer, the values $s_n^{(N)}$ of the independent birod parameters relating to base pairs, the values of the (piecewise linear) microstructure intrinsic coordinates $\widehat{\boldsymbol{y}}(s)$, as well as the value of the energy scale k_BT are all included in coefficient files. Keeping all data in a single file reduces the probability of error, while the storage cost of including the extra information is small with respect to the size of the entire file.

The procedures of extracting the Lagrangian coefficients from *cgDNA* (see Section B.3.3.3), transformation to Hamiltonian coefficients (see Section B.3.3.4), homogenization (see Section P2.1.2) and piecewise polynomial fitting were all implemented in *MATLAB*®. The resultant sequence-dependent DNA coefficients together with uniform coefficients for starting points (obtained by averaging) and the supplementary data are stored in a text file that serves as input for the *AUTO* birod DNA problem script. The data is also used by *bBDNA* for the reconstructions of computed solutions.

P2.2.1.2 The *AUTO* birod DNA problem script

As mentioned before the tool chosen for performing computations in the birod DNA model is the *AUTO-07p* parameter continuation package [DoeKelKer1991a; DoeKelKer1991b; DoeChaDer2009] (see Section B.2.4). In *AUTO-07p* a Boundary Value Problem (BVP) is defined in a problem script. The solver requires that the problem scripts are implemented either in ANSI C or Fortran. For the birod DNA model script ANSI C language was chosen. The choice was motivated by the higher flexibility of ANSI C over Fortran.

Additionally, as described in the following section, the code can be directly reused in *bBDNA* (which is written in C++03). The birod DNA *AUTO* problem script defines four required functions:

- `func` evaluates the right-hand side of the birod Hamiltonian system (B.3.57) for a given value of the state variables; helper functions are provided to evaluate the (piecewise polynomial) Hamiltonian coefficients $\boldsymbol{H}(s)$ and $\widehat{\boldsymbol{\xi}}(s)$, and the constitutive relations (B.3.58); an extra state variable has to be included for the birod independent variable s because *AUTO-07p* requires the defined system to be autonomous; yet another state variable is added, whose derivative is the shifted quadratic energy density of the Lagrangian (B.3.40) with an initial value given by the boundary terms (B.3.40b, B.3.40c); the energy density can easily be computed during evaluation of the constitutive relations; as a result, at virtually no extra cost, the Lagrangian energy of the oligomer is provided as the value at $s = L$ of the second extra state variable; this is what is described as the value of energy in all our examples.

- `stpnt` depending on the chosen options evaluates either the pulling and twisting starting point (B.3.24) or the closed loop starting point (B.3.27) for the birod macrostructure; as explained in Section B.3.4.1 the microstructure variables of starting points are always set to be identically equal to zero;

- `bcnd` evaluates the boundary condition functions, defined as in Equation (B.2.9), depending on the selected options; the macrostructure and microstructure boundary conditions can be selected independently.

- `pvls` defines solution measures through the *AUTO-07p* mechanism called *parameter overspecification* [DoeChaDer2009, sec. 10.7.10]; the measures defined for the DNA birod system include the value of the Lagrangian energy of the oligomer (as the value of the extra energy state variable described above at $s = L$) and the third component $q_3(p)$ of the average rod quaternion at $s = L$, which can be used for detecting fully closed solutions (see Section P2.2.2).

A data structure `BirodData` to store the DNA birod coefficients together with the supplementary data has been implemented. Functions to initialize and finalize instances of `BirodData`, as well as a function to read the data from a given coefficients file (of the form described in the previous section) have been provided.

In addition to the birod coefficients and supplementary data the `BirodData` structure includes member fields related to the options specifying the problem to be solved. These options include a path to the coefficients file to be used, the type of the BVP (pulling and twisting/closed loop), parameters of the starting point and versions of boundary conditions. For the pulling and twisting problem a starting point option defines the rotation $\boldsymbol{R}_0$ of the first base pair (see Equation (B.3.24)). For the closed loop problem, starting point options select a planar solution branch and a starting point within it (see

the description of the ideal rod solution set of Figure B.3.3a Section B.3.2.3). Options should be provided in a simple key-value '`options.ini`' file.

For each *AUTO* parameter continuation run the options file and the selected coefficient file are read once only, at the very beginning during initialization. The data is stored in a global instance of `BirodData`, which is then used during the run. This global instance is scheduled to be deleted at program termination.

Necessary linear algebra operations are performed using LAPACK [AndBaiBis1999] through a simplified interface based on introduced `Matrix` and `Vector` structures equipped with a set of helper functions.

The birod DNA *AUTO* problem script does not depend on *bBDNA* in any way and can be run in *AUTO-07p* separately. On the other hand *bBDNA* includes the script to reuse some of the functions implemented in it.

P2.2.1.3 The functional layer of *bBDNA*

The *bBDNA* software itself should be seen as consisting of two layers: the functional layer and the user interface layer. We first describe the design of the functional layer that is responsible for managing output data files generated by *AUTO-07p*, organizing and interpreting the data and running *AUTO-07p* itself. It was implemented entirely in C++03 and uses the established *boost 1.58* [1] library for handling XML files and file system paths. It also makes use of the *algebra3d* library introduced in Chapter P1.4. We point out that the functional layer can be used independently of the GUI layer. For example the automatic symmetry breaking script of Section P2.2.2 is a command line tool implemented using only the functional layer of *bBDNA*.

Interaction with *AUTO-07p* is based on two main types of files described briefly in Section B.2.4. All aspects of a continuation run, such as details of discretization, solver accuracy, step size along the solution branch or stopping conditions are controlled through *AUTO* constants files. A `Constants` class is used to store data of *AUTO* constants files and provides methods of reading and writing of such files. Solutions generated during a run are stored in another type of file. A `Solution` class provides functions of reading and writing of solutions in the format used by *AUTO-07p*. In this format the state variables of an *AUTO* solution are stored as values of the continuous piecewise polynomial at points uniformly distributed within each polynomial piece.

A `Probe` class provides a base for classes that give problem-specific interpretation of the raw data stored in an underlying `Solution`. The superclass defines only a very general interface that allows for a problem-specific method of sampling the piecewise polynomial data defined by a `Solution`. The default sampling $t_i \in [0, 1]$ defined by `Probe` matches

[1] http://www.boost.org/

exactly that of the underlying `Solution`, with the possibility of denser sampling through Lagrange interpolation [PreTeuVetFla2007, sec. 3.2]. The data that can be obtained from a `Probe` can be categorized into two main types:

- a pointwise data field is given by a scalar-valued function of the *AUTO* independent variable t and describes a quantity defined along the solution; the function is returned as its values at points of $t_i \in [0, 1]$ defined by the sampling introduced above; in case of `Probe` the only pointwise data fields are the state variables
- solution measure data that provides a single number characterizing a given solution; this can for example be the value of a continuation parameter or the value of a given pointwise data field at a boundary (*i.e.* at $t = 0$ or $t = 1$ of the *AUTO* independent variable).

`Probe` implements a lazy computation approach in the sense that data provided by the underlying `Solution` is (re)sampled only once on first access to the particular data field.

Instances of `Probe` can also be equipped with a pointer to an instance of the class `SolutionCoefficients`. `SolutionCoefficients` is an abstract base class for classes that store problem-specific data, such as system coefficients.

bBDNA has two specializations of the `Probe` class. The `RodProbe` interprets data generated by the Fortran*AUTO* problem script for closed loops of elastic rods used in *VBM* [Paf1999a]. In addition to the state variables `RodProbe` defines other pointwise data fields such as the norms of the stresses $|\mathbf{n}|$, $|\mathbf{m}(s)|$ or the components of moment in the laboratory frame $\mathrm{m}_i(s)$ and the body frame $m_i(s)$ (see Section B.3.1.1). This type of probe was added to allow the use of the *VBM* [Paf1999a] rod *AUTO* script in *bBDNA* for verification purposes.

The other specialization of `Probe` is the `BirodProbe` meant for the birod DNA *AUTO* problem script. For full functionality the class should be equipped with a pointer to an instance of the `BirodSolutionCoefficients` (which is a subclass of `SolutionCoefficients`) that can be seen as a wrapper of the `BirodData` structure of the birod DNA *AUTO* problem script described in the previous section. It also serves as an interface to the functions of the birod *AUTO* script. Thanks to `BirodSolutionCoefficients` a `BirodProbe` can use a sampling of the pointwise data fields based on base pair locations $s_n^{(N)}$ with the possibility of denser sampling. If homogenized coefficients are used data in `BirodSolutionCoefficients` also allows for reconstruction of the true 3D configuration and stresses of the oligomer by factoring back in the intrinsic shape removed during homogenization (see Section P2.1.2).

`BirodProbe` defines additional pointwise data fields including those defined by `RodProbe` with the addition of the norms of microstructure internal parameters $|\widetilde{\boldsymbol{\eta}}(s)|$, $|\widetilde{\boldsymbol{w}}(s)|$ and of the microstructure stress variables $\left|\widetilde{\boldsymbol{m}}^{\mathcal{P}}(s)\right|$, $\left|\widetilde{\boldsymbol{n}}^{\mathcal{P}}(s)\right|$ (see Section B.3.3.4).

`Probe`s in *bBDNA* are wrapped in `DiagramNode`s. The `DiagramNode` class allows for defining connectivity between probes, which is based on the continuation parameter used to generate a subsequent `Solution` from a previous one during continuation. Using this connectivity information `DiagramNode`s are organized in a `BifurcationDiagram` class into a tree structure. `BifurcationDiagram` provides a method of loading a portion of a branch of solutions from an *AUTO* output solution file that includes the loaded solutions into the diagram structure.

The description of the set-up of a problem to be solved using *bBDNA* can be defined in a computation configuration XML file that is handled by a class called `Configuration`. Configurations files define such aspects of a computation as:

- the problem type
- the working directory,
- the *AUTO* problem script to be used,
- an *AUTO* constants file with initial values of *AUTO* constants
- a coefficients file (necessary in case of computations in the birod DNA model)
- additional compilation flags required to build the *AUTO* binary using the chosen script
- any precomputed data to be loaded.

The problem type specifies what kind of `Probe` should be used. Currently three types are available: a generic problem that adds no interpretation to the data, for which `Probe`s are used, a rod problem that assumes a *VBM* rod script is provided (`Probe`s are used), and a birod problem that requires a birod DNA script and birod DNA coefficients file (`BirodProbe`s are used). The `DiagramNode` and `BifurcationDiagram` are agnostic of the particular implementation of the `Probe` that they handle. Creation of `Probe`s based on the chosen problem type is delegated to a `ProbeFactory` class.

Global management of data as well as execution of *AUTO-07p* continuation runs is performed by a `ComputeEngine` class. Given an instance of `Configuration` a `ComputeEngine` sets up a working directory for the continuation with all the necessary files. It provides methods for building a binary for the chosen *AUTO-07p* problem script as well as for running continuations described by an internal instance of the `Constants` class. The *AUTO* input files '`fort.2`' and '`fort.3`' (see Section B.2.4) are created accordingly. Child process spawning as well as interprocess communication is realized using the standard UNIX interfaces `execlp()`/`waitpid()` and `dup2()`/`pipe()`, respectively.

A `Result` class was introduced for reporting the outcome of different operations from loading of *AUTO* solution files to execution of continuation runs. This slightly simplistic solution avoids exception handling and replaces a possibly large hierarchy of exceptions.

P2.2.1.4 The graphical user interface of *bBDNA*

The graphical user interface layer of *bBDNA* was built in C++03 using the *Qt 4.8* [2] windowing library. 3D visualization is realized with the *Coin3D 3.1.3* [3] implementation of the *Open Inventor 2.1* API which is a high level, retained-mode 3D graphics toolkit. The interface between *Qt 4.8* and *Coin3D 3.1.3* is realized by the *SoQt 1.5* [4] library. Two-dimensional plotting uses the *Qwt 6.0* [5] library.

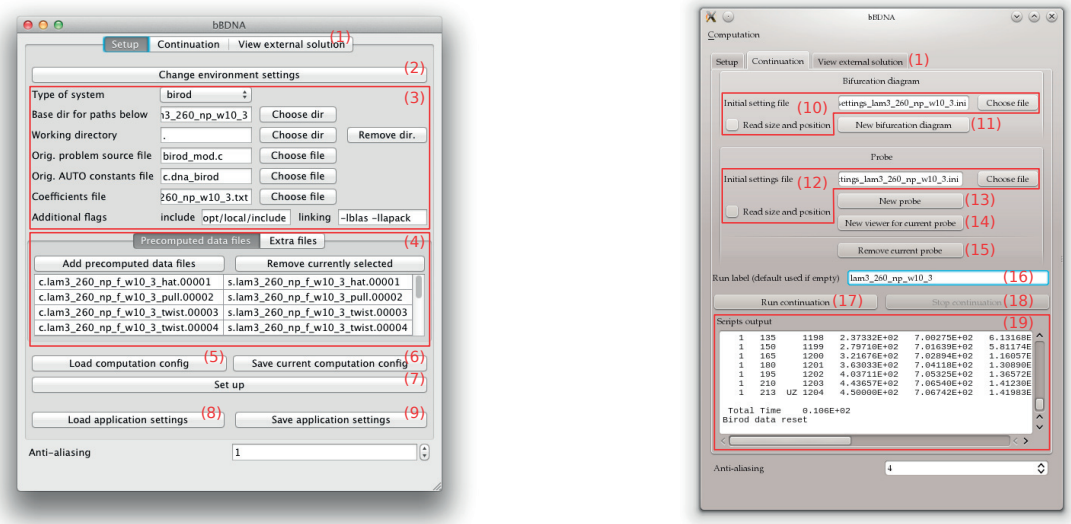
Certain extensions to the *SoQt 1.5* library have been implemented and gathered in a library called *SoQtExtensions*, used by *bBDNA*. These extensions, which can be used in other contexts than the *bBDNA* user interface, include the user interface of the 3D scene viewer of *SoQt 1.5* (*e.g.* additional mouse and keyboard interaction methods, GLSL shaders suited for use with orthographic projection, a widget for setting scales of the 3D scene) and definition of certain 3D objects (*e.g.* arrows, tube representation of curves, box representation of DNA oligomers).

The *bBDNA* application can be compiled for both Linux and Mac OS X. For Mac OS X a self-contained DMG installation file has also been prepared, built on version 10.7.5 of the operating system.

The class hierarchy of the graphical user interface part of *bBDNA* closely reflects the structure of the functional layer, described in Section P2.2.1.3. In fact two groups of classes can be identified: one related to 3D representations of bifurcation diagrams, and DNA birod solutions themselves, and another one related directly to the GUI *i.e.* the main window of the application and other dialog windows. A short description of the graphical user interface follows.

When *bBDNA* is started the user is presented with the main window of the application (see Figure P2.2.1b). From here all the necessary settings can be adjusted and continuation runs can be set up (see Figure P2.2.1a). It also allows for continuation to be run in *AUTO-07p* and for dialog windows to be opened for viewing the bifurcation diagram and for particular solutions (`ProbeDialog`). A configuration of a run can be saved in or loaded from an XML configuration file suitable for the `Configuration` class (see Section P2.2.1.3). The settings of the application (including all chosen options) are saved between subsequent runs, but can also be stored in an INI file. The INI files provide a way of preparing demonstrations based on precomputed data.

[2] http://doc.qt.io/qt-4.8/
[3] https://bitbucket.org/Coin3D/
[4] https://bitbucket.org/Coin3D/soqt
[5] http://qwt.sourceforge.net/



(a) (b)

Figure P2.2.1. *Screenshots of the main window of bBDNA. The different tabs of the main window can be selected using* (1). *The* ***Settings*** *and* ***Continuation*** *tabs are presented in the figure. The* ***View external solution*** *tab is meant for quick previews of a single AUTO birod solution file without the need to set up a continuation run. Panel (a) shows the settings tab of the main window as it appears on Mac OS X. Global preferences (path where AUTO-07p is installed, compilers to be used) can be changed using* (2). *Configuration of a run (see description of the* `Configuration` *class in the previous section) can be defined using* (3) *including a choice of precomputed data to be loaded* (4). *A configuration of a run can be loaded from* (5) *or saved into* (6) *an XML file appropriate for the* `Configuration` *class. The* ***Set up*** *button* (7) *prepares a continuation run by creating a working directory (if it doesn't already exist) and copying all the necessary files into it. All the options chosen in all widgets of the main window can be loaded from* (8) *or stored into an INI file* (9). *Panel (b) shows the continuation tab as it appears on Linux in the KDE 4.14 environment. From here new bifurcation diagram viewers can be open with* (11) *possibly with all the settings loaded from an INI file* (10) *(see Figure P2.2.2). Probe viewers (see Figure P2.2.3) can be opened by adding a new probe in the bifurcation diagram* (13) *or by opening a viewer for the probe that is currently selected in the diagram* (14) *(see Figure P2.2.2). All settings of the newly opened probe viewer can be loaded from an INI file* (12). *The currently selected probe can also be removed* (15), *together with all the attached viewers. Continuation runs in AUTO-07p can be started using* (17), *in which case bBDNA asks for the AUTO constants (see Section B.2.4) and birod boundary condition options to use. By default runs are started from the solution pointed to by the currently selected probe. If necessary the birod DNA AUTO script is compiled. The output from AUTO-07p is reported in* (19). *Runs can be stopped on demand using* (18).

Computed solution sets can be visualized in three dimensions in bifurcation diagram dialog windows (see Figure P2.2.2). Bifurcation diagram viewers allow three solution measures to be chosen (as defined in the description of the `Probe` class in Section P2.2.1.3) and used as the spatial coordinates. Another two solution measures can be indicated as colour of the line/tube representation of the diagram and colour of solution point indicators (as in Figure P2.2.2b). Solutions are divided into thee categories: regular solutions (indicated with balls), bifurcation points (indicated with boxes) and user requested points reported by *AUTO-07p* (indicated with crosses – see description of *AUTO-07p* constants in Section B.2.4 for the definition of user points).

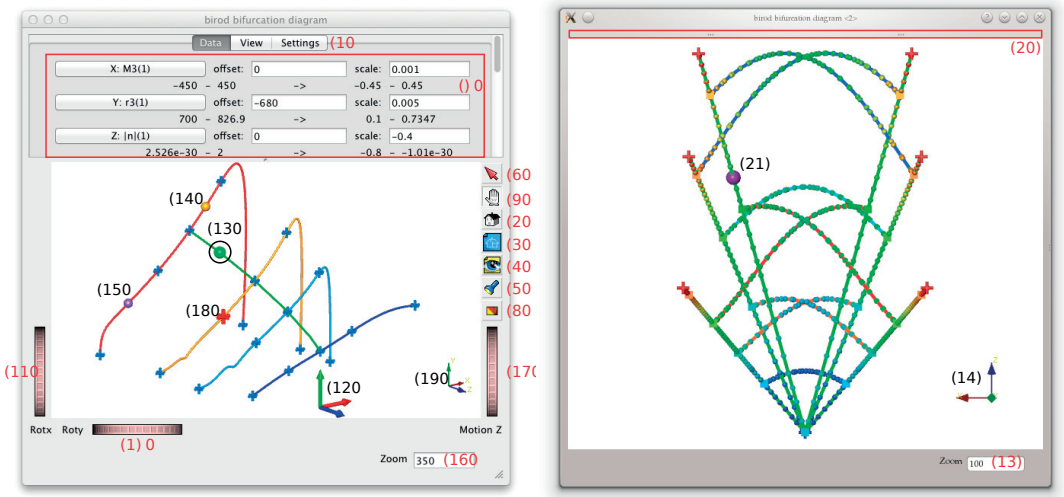
Markers representing solution probes are also shown in the bifurcation diagram viewers with colour matching the colour of the probe dialog window. These markers can be moved around the diagram using the keyboard arrows or by picking a point in the diagram using the mouse. This not only allows the user to select a solution to visualize, but also provides a way of selecting points to start new continuation runs from. Probe marker indicators are balls, boxes or crosses (depending on the type of solution they point to) that are bigger than analogous solution indicators mentioned above (see Figure P2.2.2b).

Important elements of the GUI of all 3D viewers in *bBDNA* are labelled in red in both panels of Figure P2.2.2. We describe these widgets in the context of the bifurcation diagram viewer using these labels. A settings panel can be opened or closed by dragging the handle (20) with a mouse. Different settings of the presented view can be set here. In case of bifurcation diagram viewers these settings include *e.g.* the choice of spatial coordinates (2), visibility of the different types of solution indicators, choice of line/tube representation of the diagram. 3D viewer settings can be saved to or loaded from an INI file. The settings panel is divided into tabs, which can be selected using (1).

The right-hand side toolbox provides the following buttons:

- (3) for turning on the mouse picking mode; in this mode left mouse button clicks select solutions or probe markers in the diagram,
- (4) for turning on the mouse view rotation mode; the view can be panned with Shift button pressed,
- (5) for going back to a saved view of the scene,
- (6) for saving the current view of the scene to be loaded using (5),
- (7) for resetting the view so that all elements of the 3D scene are included,
- (8) for turning on scene lighting options mode; this button is inactive; instead a lighting options dialog window can be shown by pressing the L button on the keyboard)
- (9) for switching between orthographic and perspective view.

The dials allow for zooming (10) and rotating the scene (11), (12). The toolbox (3–9) together with the dials (10–12) can be shown or hidden by pressing Enter on the keyboard. The zoom value widget (13) sets a prescribed camera zoom. Through this option different viewers can use the same scale to allow direct comparison.



(a) (b)

Figure P2.2.2. *Screenshots of bifurcation diagram viewers of bBDNA on Mac OS X (left) and on Linux (right). Panel (a) shows a different projection of the bifurcation diagram of Figure P2.3.3a. Branches of the bifurcation diagram are shown as tubes coloured by a user defined solution measure meant to visually separate different parts of the solution set. The blue crosses are indicators of user requested points reported by AUTO-07p (see description of AUTO-07p constants in Section B.2.4). The red cross* (19) *is a marker of a red probe that points to a user requested solution. Note that the marker cross indicator is bigger than the other crosses. Altogether four probes are marked in the diagram:* (16), (17), (18) *and* (19). *The one currently selected is* (16), *which is indicated by the bigger size. The selected probe can be moved around the diagram using keyboard arrows or by picking a point in the bifurcation diagram with the right mouse button. Continuation can be started from a point selected in this way. The middle of the coordinate system can be marked using a frame* (15). *An additional, constantly visible, orientation indicator is provided by the viewer in the bottom right hand corner of the window* (14). *Panel (b) shows a different projection* ($\mathrm{m}_3(L)$ *vs.* $|\mathbf{n}|$) *of the bifurcation diagram of Figure B.3.3a. Branches are indicated with lines coloured as in Figure B.3.3a. Additionally indicators of all regular points and bifurcation points are shown and coloured by the energy E. A single probe marker* (21) *is also visible. Different elements of the GUI, labelled in red, which are common for all 3D viewers in bBDNA are described in main text. In panel (b) (almost) all of those GUI elements are hidden to extend the 3D viewer.*

Particular solutions can be studied in probe viewers (see Figure P2.2.3). The 3D configuration of the oligomer can be presented at different levels of detail. Tube and ribbon representation of the average rod (as in Figure P2.2.3b) and of the two interacting rods can be shown. Due to the particular embedding of the base frame used by the *cgDNA* model (see Section B.1.1), the distance between the two rods is very small with respect to the length of the oligomer. In fact the two-rod visualization is turned on in Figure P2.2.3b, but the rods are indistinguishable by eye from the average rod. For that reason an offset within the base pair plane can be added between the two rods. By default this offset is set to the average position of the C_1' atom of the deoxyribose that the nucleobases are covalently bonded to. This default offset is used in all birod examples in the thesis.

The probe viewer also allows for visualization of each base of the oligomer. This is done by evaluating the piecewise polynomial solution of *AUTO-07p* at values of the independent birod parameter $s = s_n^{(N)}$ corresponding to base pairs. The values $s_n^{(N)}$ are given in the DNA coefficients file, as explained in Section P2.2.1.1. Bases can be visualized as boxes coloured by the type (by default: `A` is red, `T` is blue, `G` is green and `C` is yellow). Another possibility is to show all atoms of idealized bases as in Figure P2.2.3a (by default carbon is black, nitrogen is blue and oxygen is red). The ideal base data is provided with *cgDNA*.

Another way to look at the DNA birod data in a probe viewer is by plotting selected state variables (or predefined function of those) against the birod independent variable s of the base pair index. An example of such a plot is presented in Figure P2.2.3d). Numerical values of the data can also be inspected (as shown in Figure P2.2.3c).

The layout of the user interface of probe dialog windows is analogous to that of the bifurcation diagram dialogs. We describe the elements specific to probes using the red labels of Figure P2.2.3 as references. All settings can be adjusted in the settings panel, which can be opened exactly as in the case of bifurcation diagrams. The way the piecewise polynomial solution computed by *AUTO-07p* is sampled can be defined using (1–3). The sampling can be based on the values $s_n^{(N)}$ of the birod independent variable corresponding to base pairs ((2) checked) or on the discretization mesh used in *AUTO-07p* ((2) unchecked). The 3D configuration and 2D plots can be shown for data before intrinsic shape refactoring (3). As is clear from the comparison of Figures P2.2.3a and P2.2.3b, the refactoring is required to allow interpretation of the data if the homogenized coefficients of Section P2.1.2 are used.

Different viewer panes can be turned on by selecting the corresponding settings tabs (4). The 3D viewer is selected together with tabs ***3D Mesh***, ***3D Colours*** and ***3D View***, 2D plotting is turned on by the ***2D Plot*** tab and numerical data is shown with ***Data*** tab. The settings of the 3D view consist of parameters of the tubes, ribbons and the discrete representations of bases including the possibility to hide some of them. In the ***2D Plot*** tab the variables to be plotted can be selected (14) and plot styles can be altered. 2D plots allow for zooming into a selected fragment (15) and provide a way to check the numerical values of the point under the mouse cursor (16).

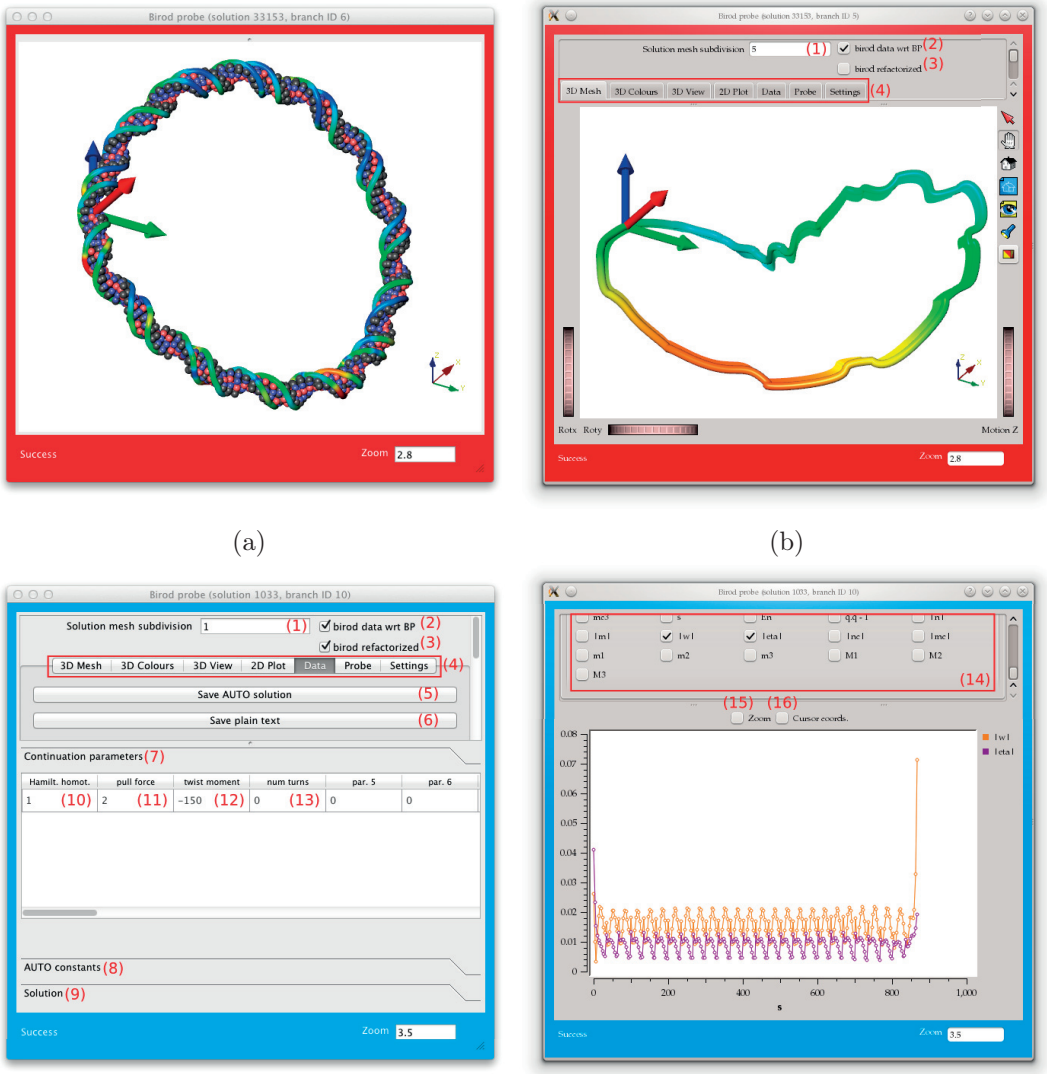


Figure P2.2.3. *Screenshots of probe viewers of bBDNA on Mac OS X (left) and on Linux (right). Panels (a) (b) show two different visualizations of the lowest energy minicircle of linking number* $Lk = 14$ *of the* S^{γ} *sequence (see Figure P2.3.4d). In panel (a) the 3D configuration of the solution is shown with all base atoms indicated. The two tubes are coloured with* $|\boldsymbol{\eta}(s)|$ *and* $|\boldsymbol{w}(s)|$*. Panel (b) shows the same solution as computed in AUTO-07p using the homogenized coefficients of Section P2.1.2 (without the intrinsic shape refactoring* (3)*). The tube and ribbon represent the average rod configuration, coloured by the norm of the (non-refactored) moment* $|\mathbf{m}(s)|$*. Panels (c) and (d) show data of the pulled and under-twisted configuration of the* S^{a} *oligomer of Figure P2.3.3c. Panel P2.2.3c shows the numerical data inspection pane, with continuation parameters section* (7) *open. The pulling force* (11) *is* 2 pN*, while the twisting moment* (12) *is* −150 pNÅ*. The value of the Hamiltonian homotopy parameter of* 1 *in* (10) *means that full DNA sequence-dependent coefficients were used. Panel (d) shows the 2D plotting pane with plots of the norms of microstructure internal parameters* $|\boldsymbol{\eta}(s)|$ *and* $|\boldsymbol{w}(s)|$*. A description of the GUI elements labelled in red can be found in the main text.*

The numerical data pane allows for inspecting the values of continuation parameters (7) as well as the *AUTO* constants (8) and solution (9) exactly as read from *AUTO-07p* output files. Figure P2.2.3c presents all four continuation parameters that can be used in case of the DNA birod system. These are the Hamiltonian homotopy parameter (10) that allows continuation from uniform to sequence-dependent coefficients, the pulling force n_3 (11) and twisting moment m_3 (12) used in the pulling and twisting boundary value problem (see Section B.3.2.2) and the angle α of rotation between the first and the last cross section (13) of the closed loop boundary value problem (see Section B.3.2.3).

P2.2.2 Automatic symmetry breaking

As mentioned before, in our efforts to provide computational tools for the birod model of DNA we have assessed the applicability of the symmetry breaking technique for the closed loop boundary value problem used in the Hamiltonian formulation of rods (*e.g.* [LiMad1996; DicLiMad1996; MadManPaf1997] – see Section B.3.2.3) in case of the birod DNA Hamiltonian of [Gra2016] (see Section B.3.3.4). We have found that indeed the method can easily be extended to the birod case. Symmetry breaking from an inextensible and unshearable, straight, uniform, transversely isotropic rod can be performed directly to a fully DNA sequence-dependent birod.

In particular for a set of target sequence-dependent Hamiltonian coefficients $\boldsymbol{H}(s)$ and $\widehat{\boldsymbol{\xi}}(s)$ the initial uniform birod Hamiltonian coefficients $\boldsymbol{H}_{\mathrm{u}}$ and $\widehat{\boldsymbol{\xi}}_{\mathrm{u}}$ of an ideal rod are defined as:

$$\boldsymbol{H}_{\mathrm{u}} := \begin{bmatrix} \boldsymbol{H}^{\mathcal{P}} & \mathbf{0}_{12\times 6} \\ & \boldsymbol{H}^{\boldsymbol{U}} \quad \mathbf{0}_{3\times 3} \\ \mathbf{0}_{6\times 12} & \mathbf{0}_{3\times 3} \quad \mathbf{0}_{3\times 3} \end{bmatrix} \tag{P2.2.1a}$$

$$\widehat{\boldsymbol{\xi}}_{\mathrm{u}} := \begin{bmatrix} 0 & 0 & \overline{U}_3 & 0 & 0 & 1 \end{bmatrix}^T \quad , \tag{P2.2.1b}$$

with $\mathbf{0}_{12\times 6}, \mathbf{0}_{6\times 12}, \mathbf{0}_{12\times 6}, \mathbf{0}_{3\times 3}$ – zero blocks of indicated dimensions. The block $\boldsymbol{H}^{\mathcal{P}} \in \mathbb{R}^{12\times 12}$ is an average over s of the respective elements of $\boldsymbol{H}(s)$. In the diagonal block

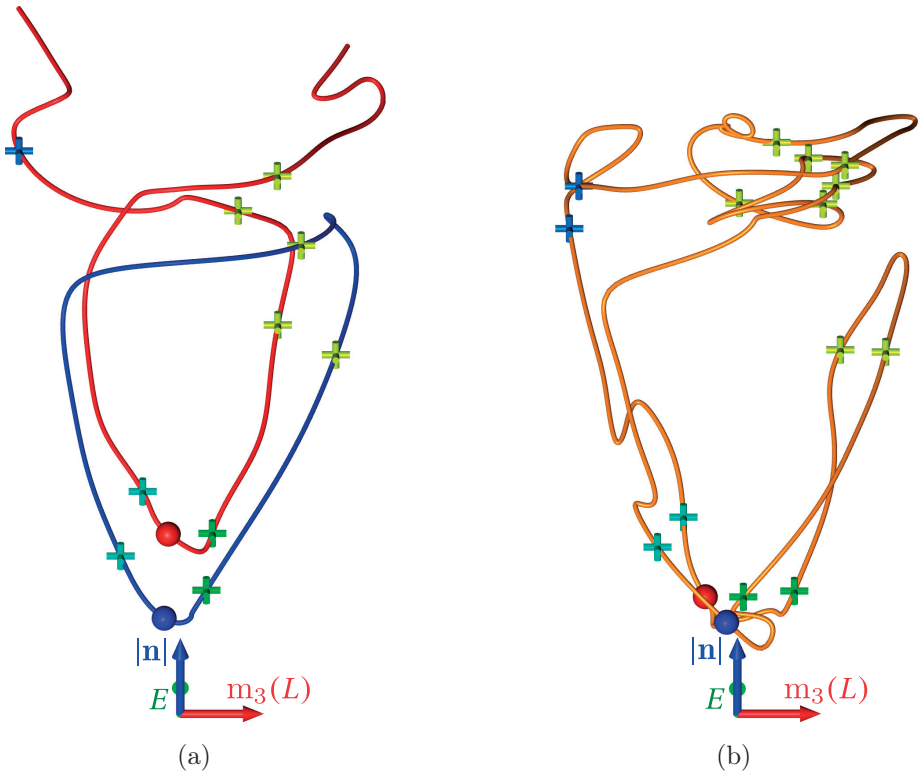
$$\boldsymbol{H}^{\boldsymbol{U}} = \begin{bmatrix} \frac{1}{2}(\overline{H}_1 + \overline{H}_2) & 0 & 0 \\ 0 & \frac{1}{2}(\overline{H}_1 + \overline{H}_2) & 0 \\ 0 & 0 & \overline{H}_3 \end{bmatrix} \tag{P2.2.2}$$

the values $\overline{H}_1$, $\overline{H}_2$ and $\overline{H}_3$ are averages over s of the elements $H_{13}(s)$, $H_{14}(s)$ and $H_{15}(s)$ of $\boldsymbol{H}(s)$, respectively. Finally $\overline{U}_3$ is the average over s of the third component of $\widehat{\boldsymbol{\xi}}(s)$.

These coefficients were chosen to make the initial uniform parameters close to the target sequence-dependent ones. It is easy to see that as long as the microstructure variables

are decoupled from the macrostructure (the zero blocks in $\boldsymbol{H}_{\mathrm{u}}$) the value of the leading diagonal block $\boldsymbol{H}^{\mathcal{P}}$ of $\boldsymbol{H}_{\mathrm{u}}$ is immaterial for the boundary value problem with boundary conditions on the microstructure that allow them to stay zero (*e.g.* periodic or zero Dirichlet boundary conditions).

In this setting the symmetry breaking procedure in exactly the form previously used for elastic rods [LiMad1996; MadManPaf1997; ManMad1999] (see Section B.3.2.3) can be performed automatically. The results of such a computation are indicated in Figures P2.2.4a and P2.2.4b as blue and red balls that are the fully sequence dependent DNA birod solutions that originated from a single ideal rod starting point.



(a) (b)

Figure P2.2.4. *Example results of the automatic symmetry breaking procedure of bBDNA. The two presented bifurcation diagrams are different projections of those shown in Figures P2.3.4 (for the highly bent sequence S^{γ} of [KahCro1992]) and Figures P2.3.5 (for the close to straight sequence $S^{\lambda''}$). The blue and red balls mark the ends of the symmetry breaking that were the starting points for the fully sequence-dependent computations. The crosses mark the closed solutions of linking number* 13 *(blue),* 14 *(cyan),* 15 *(green),* 16 *(yellow). Fully closed solutions are used by the automatic procedure to detect closure of the components of the solution set (see text). The blue branch of panel (a) as well as the orange branch of panel (b) is closed. The red branch of the bifurcation diagram in panel (a) has not been continued past the two end points as the requested maximum energy of* $100\,k_BT$ *was reached. Note the different connectivity of the components of the solution sets in the two cases. Note also that in both cases other components of the solution set exist in the shown subspace.*

Subsequently our automated procedure starts continuation runs in the angle α between the director vectors $\mathbf{d}_1(0)$ and $\mathbf{d}_1(s)$ (or equivalently between $\mathbf{d}_2(0)$ and $\mathbf{d}_2(s)$). Continuation is done in both the positive and negative directions of α in steps that are terminated when a fully closed solution is found. The detection of full closure can be done by tracking the value of the third component $q_3(L)$ which vanishes in such a case $\left(i.e.\ \boldsymbol{q}(L) = \begin{bmatrix}0 & 0 & 0 & 1\end{bmatrix}^T \text{ or } \boldsymbol{q}(L) = \begin{bmatrix}0 & 0 & 0 & -1\end{bmatrix}^T\right)$. After each such step the initial values of the two fully closed solutions that are the ends of the branch being extended are compared. If the initial values all agree up to the accuracy requested from *AUTO-07p* a branch is deemed closed. For example the blue branch of Figure P2.2.4a was closed after 5 such steps, while the orange branch of Figure P2.2.4b was closed after 15 steps. For the red branch computations where finished at the two visible ends due to another termination condition which was the value of the Lagrangian energy reaching $100\,k_BT$. The maximum number of required steps as well as the maximum value of energy are parameters of the automatic procedure.

Note that the connectivity of the perturbed solution set depends on the sequence, which is indicated in Figure P2.2.4. The connectivity of the solution set for the highly bent sequence S^{γ} of [KahCro1992] (Figure P2.2.4a) resembles closely the one of the lowest part of the rod perturbed diagram of Figure B.3.3b (the corresponding blue and red components are marked with the same colour in both figures). The connectivity of Figure B.3.3b seems to be a common feature in cases where the problem is perturbed primarily in shape (*i.e.* when the shape seems to predominate the response of the system). This is the case for most diagrams presented *e.g.* in [LiMad1996; ManMadKah1996; MadManPaf1997] where the perturbation was in shape only.

For the intrinsically close to straight sequence $S^{\lambda''}$ (Figure P2.2.4a) the entire lower energy part of the solution set is a single connected component. In fact two continuations started from the blue and red solutions in Figure P2.2.4a yielded the same component.

P2.3 Examples of birod DNA computations

This chapter presents example results of computations using the birod DNA model of [Gra2016], described in Section B.3.3. All the computations were performed using the *bBDNA* software described in Chapter P2.2. The results of Section P2.3.1 were obtained through interactive computational steering performed using the graphical user interface of *bBDNA*. The solution sets of Section P2.3.2 were generated by the automatic symmetry breaking script of Section P2.2.2.

In our examples we make use of some of the results presented in the other chapters. First of all we point out that the DNA coefficients used throughout this chapter are the homogenized ones of Section P2.1.2. The pulling and twisting numerical experiment uses two of the repeating sequences with pronounced superhelical structure introduced in Chapter P1.3, as well as an, on average, straight oligomer from Chapter P1.4. These sequences are shown to exhibit considerably different responses to over- and under-twisting. We also show how the periodic *cgDNA* parameters of Chapter P1.2 can be used in modelling of short, covalently bonded loops of DNA. A comparison of properties of such loops for two sequences: an intrinsically bent one and an intrinsically straight one (both analysed in Chapter P1.4) are presented in this context.

P2.3.1 Pulling and twisting of DNA in the birod model

We present here the pulling and twisting boundary value problem first introduced in Section B.3.2.2 in the simpler context of elastic rods. As mentioned in Section B.3.4 for the birod system in this problem we ask the average rod to be fixed in the laboratory frame at one end, while force and moment along the vertical axis are applied at the other end. This is expressed more precisely in Equation (B.3.21). The microstructure of the birod is left free to equilibrate by application of the free end boundary conditions (B.3.61).

As our DNA fragments we have chosen tandem (*i.e.* consecutive) repeats of two of the example sequences of Chapter P1.2 shown to have well pronounced superhelical structure. Specifically, the selected sequences are: ${S^a}_{26}$ (26 repeats of the decanucleotide $S^a = \mathtt{A_5CACG_2}$) and ${S^c}_{22}$ (the 264 base pair long fragment constructed as 22 repeats of the dodecanucleotide $S^a = \mathtt{A_5CACG_2}$).

Intrinsic shapes of both of these oligomers are of very similar length and are superhelices with very close values of radius and pitch, but of opposite chirality (see Figure P2.3.1). The length was chosen in such a way that both superhelices have nearly exactly two helical repeats. As a result they were expected to exhibit very different responses when over- and under-twisted. All of these choices were highly facilitated by the results of Chapter P1.3.

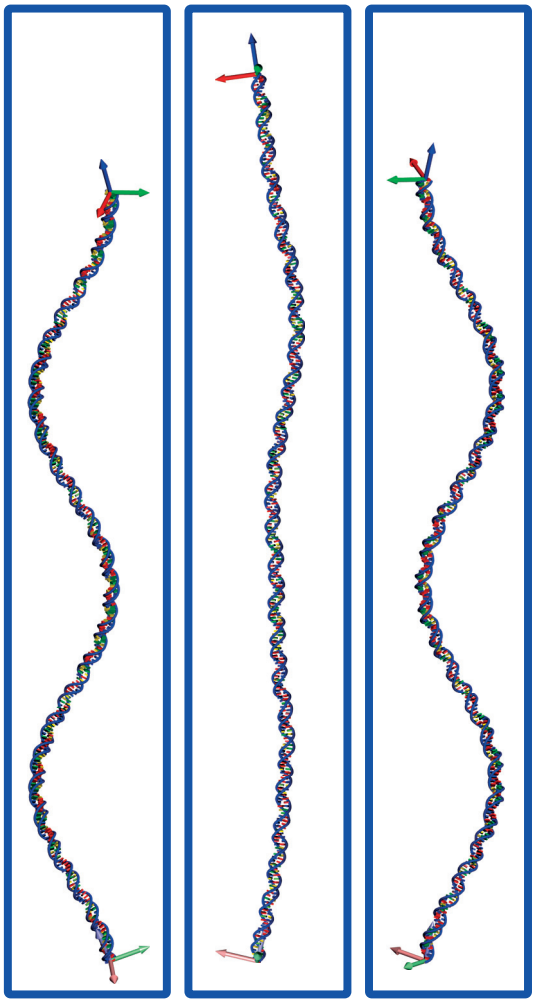
For reference a third sequence with no particular motif in the unstressed configuration was subjected to the same loading conditions. This fragment is made of base pairs 36901–37160 of the genome of λ-phage [SanCouHon1982]. In fact these are the first 260 base pairs of the S^λ fragment presented in Chapter P1.4 as the one with mean persistence length among all consecutive 300 bp long fragments of the genome (see Section A.2.2.1).

The exact boundary conditions imposed on each of the sequences involve pulling and twisting in the direction of the respective end-to-end vector $\mathbf{r}(L) - \mathbf{r}(0)$ of the intrinsic configuration. This could be achieved in boundary conditions (B.3.21) by asking for an appropriate rotation $\boldsymbol{R}_0$ to be applied to the first average rod cross section orientation.

Figures P2.3.2a, P2.3.3a and P2.3.3b show bifurcation diagrams of twist versus extension of the numerical experiments for $S^{\lambda\prime}$, S^a, S^c, respectively. Blue crosses indicate the unstressed shapes, obtained through homotopy continuation in the Hamiltonian coefficients (see Equation B.3.31) from starting points computed using Equations (B.3.24). The coefficients $\boldsymbol{H}_\mathrm{u}$ and $\widehat{\boldsymbol{\xi}}_\mathrm{u}$ of the starting points were computed as averages over s of the sequence-dependent ones for each sequence separately. An analogous approach for rods was discussed in Section B.3.2.2.

Subsequently, the oligomers were pulled up to the value of the vertical force of $n_3 = 2\,\mathrm{pN}$. This continuation is represented by the green branch that ends with a red cross. The 3D configurations corresponding to the resulting stretched solutions are shown in the red panels P2.3.3d for S^a, P2.3.3g for S^a and P2.3.2c for $S^{\lambda\prime}$.

The other branches of the bifurcation diagrams correspond to continuations in the imposed torque m_3. The over- and under-twisting was performed at 4 different values of the pulling force: 0 (dark blue branch), 0.5 (bright blue branch), 1 (orange branch) and 2 (red branch), all units pN. One under-twisted and one over-twisted solution on the red branch is selected for each sequence. The value of the twisting moment for all of them is ±150 pNÅ. These solutions are marked with bright blue (under-twist) and orange crosses (over-twist) and their 3D configurations are shown with coloured frames indicating the correspondence.



(a) $S^a{}_{26}$ (b) $S^{\lambda\prime}$ (c) $S^c{}_{22}$

Figure P2.3.1. *The unstressed shapes of the sequences used for the pulling and twisting example. Note that* $S^a{}_{26}$ *and* $S^c{}_{22}$ *form superhelices of comparable pitch and radius but opposite handedness. All three configurations are oriented so that the respective end-to-end vectors* $\mathbf{r}(L) - \mathbf{r}(0)$ *are aligned with the vertical axis* $\mathbf{e}_3$*. The presented solutions (a), (b) and (c) correspond to the dark blue crosses in the bifurcation diagrams of Figures P2.3.3a, P2.3.2a and P2.3.3b, respectively. The tubes in all visualizations of the 3D shapes in this section are coloured using the norm of the average rod moment* $|\mathbf{m}(s)|$ *with blue representing* $|\mathbf{m}(s)| = 0\,pN\text{Å}$ *and red representing* $|\mathbf{m}(s)| = 175\,pN\text{Å}$*.*

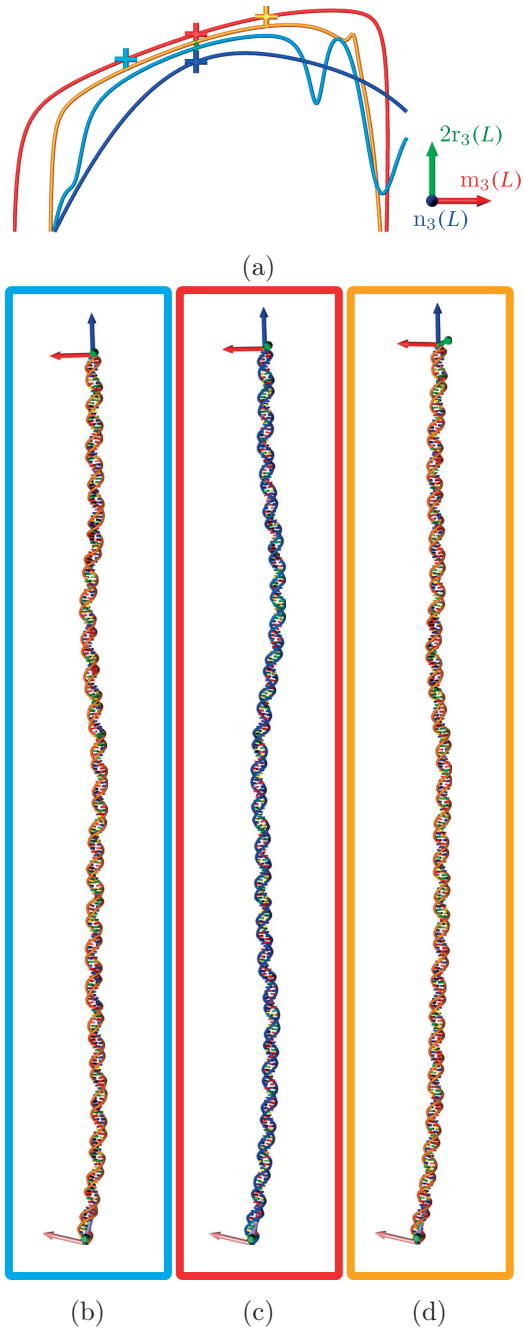


Figure P2.3.2. *Results of the pulling and twisting numerical experiment for a straight oligomer. Panel (a) shows a load-extension bifurcation diagram of the experiment with* $\mathbf{r}_3(L) \in [800\,\text{Å}, 847.1\,\text{Å}]$ *and* $m_3 \in [-450\,\text{pNÅ}, 450\,\text{pNÅ}]$. *The scale of the* $\mathbf{r}_3(L)$ *axis of the bifurcation diagram is twice the one used in Figure P2.3.2, which is indicated by the label of the respective axis. Panels (b), (c) and (d) correspond to the bright blue, red and orange crosses in the bifurcation diagram, respectively.*

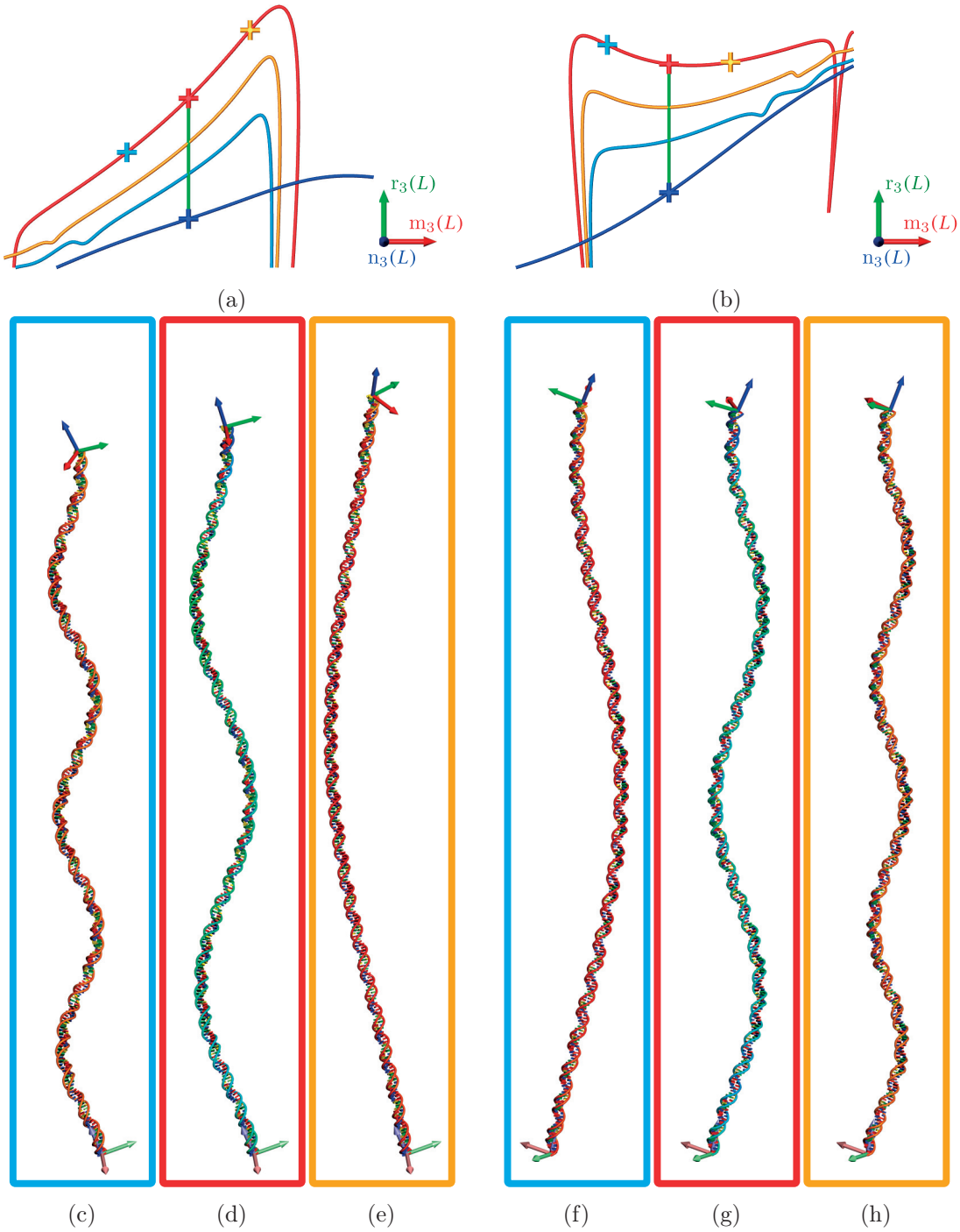


Figure P2.3.3. *Results of the pulling and twisting numerical experiment for left- and right-handed DNA superhelices. The conventions used here are the same as in Figure P2.3.2. The left set of panels corresponds to the sequence* S^a*, the right set to* S^c*. The two bifurcation diagrams are shown in the same scale, with* $\mathbf{r}_3(L) \in [700\,\text{Å}, 826.9\,\text{Å}]$ *and* $\mathrm{m}_3 \in [-450\,\text{pNÅ}, 450\,\text{pNÅ}]$.

We conjecture that all the visualized stationary points are stable, as they were all achieved by continuation without fold points.

As expected all three solutions reacted very differently to the applied stresses. In case of $S^{\lambda\prime}$ barely any change in shape can be seen. In fact even the bifurcation diagram had to be plotted with twice the scaling of the $\mathrm{r}_3(L)$ component used in case of the other sequences, to visually separate the different twisting branches. Nevertheless a little shoulder can be observed suggesting that for small values of the twisting moment, over-twisting induces extension and under-twisting induces shortening for all the studied pulling forces.

Seemingly analogous behaviour can be observed in the bifurcation diagram for S^a, but the shoulder is much more pronounced. Specifically the extension between the red and orange solution is 9 times bigger (33.1 Å vs 3.7 Å) than the analogous value for $S^{\lambda\prime}$. As is clear from the reconstructed shapes. The reason for this is that right-handed twist unravels the left-handed superhelical structure lengthening the configuration, while left-handed twist tightens it, leading to shortening.

The most interesting response to loading is demonstrated by the S^a fragment. For very small pulling forces the tightening of the superhelix through right-handed twist causes its extension, while unravelling under left-handed twist shortens it. As the pulling force is increased such response is weaker and weaker until for the pulling force of 2 pN it is exactly opposite: the superhelix is shortening as it is tightened and *vice versa*.

Our findings confirm the argument of [ĐurGorMad2013] that the answer to the question whether a polymer such as DNA extends or contracts under over- and under-twisting depends on the particular constitutive relations of the polymer. In particular, we have shown that both responses to twist loading may arise at different force loadings.

P2.3.2 Computing equilibria of DNA minicircles using birods

The other example of a computation within the birod DNA model is solving the closed loop boundary value problem, outlined for the elastic rod system in Section B.3.2.3. In this problem the ends of the elastic rod are required to close. Additionally the cross sections at both ends are required to share a common director vector $\mathbf{d}_3$. As a result the only freedom left is a rotation of the cross section orientation $\boldsymbol{R}(L)$ with respect to $\boldsymbol{R}(0)$ around $\mathbf{d}_3(L)$. In the computations the angle α between $\mathbf{d}_1(L)$ and $\mathbf{d}_1(L)$ (or equivalently $\mathbf{d}_2(L)$ and $\mathbf{d}_2(L)$) is used as a continuation parameter.

This formulation was previously used in the elastic rod model to model so-called *minicircle*, *i.e.* rigs of relatively short fragments of DNA (< 1000 bp) both backbones are covalently bonded (see *e.g.* [FurManMad2000]). In this approach solutions for different values of α are computed through parameter continuation. The solutions of the angle reaching a multiple of 2π represent minicircles in the rod model.

In the birod DNA model to assure full closure (*i.e.* exact matching of both backbones) periodicity in both macro- and microstructure is required. Consequently in our approach we will use periodic boundary conditions on the microstructure, as in Equation (B.3.62). As was done with rods we will treat all other $\alpha \neq 0$ solutions only as intermediate steps of the chosen numerical method.

Another requirement of the proposed modelling technique is that the base pairs at $s = 0$ and at $s = L$ should be identified. To achieve that we will use a similar concept to the one proposed in the definition of the periodic *cgDNA* coefficients. Specifically, for a sequence of length N, we will introduce an extra Nth junction modelling the interactions between base pair N and 1. This new junction replaces the end terms in the Lagrangian (B.3.52), which in our periodic case will take the form:

$$\widehat{E}_p[\mathcal{G}, \mathcal{P}(\boldsymbol{y})] = \int_0^L \begin{bmatrix} \widetilde{\boldsymbol{y}}(s) \\ \widetilde{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \\ \boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s) \end{bmatrix} \cdot \boldsymbol{K}(s) \begin{bmatrix} \widetilde{\boldsymbol{y}}(s) \\ \widetilde{\boldsymbol{\xi}}^{\mathcal{P}}_{\boldsymbol{y}}(s) \\ \boldsymbol{\xi}(s) - \widehat{\boldsymbol{\xi}}(s) \end{bmatrix} \mathrm{d}s \quad . \tag{P2.3.1}$$

All in all, the periodic birod DNA coefficients are constructed exactly the way described in Section B.3.3.3, but an extra junction is introduced instead of the end stiffness matrices $\boldsymbol{K}_0$, $\boldsymbol{K}_L$. The periodic *cgDNA* shape $\widehat{\boldsymbol{w}}_p$ is used in place of the standard one $\widehat{\boldsymbol{w}}$ and the periodic version of the 3D configuration reconstruction procedure is used to define the interpolation.

We note that in light of the above discussion the formulation of the birod closed loop BVP of Section P2.1.2 that used the original procedure of constructing birod parameters as well as natural boundary conditions on the microstructure served only for the verification purposes of that section. These solutions should not be treated as representing real, fully closed loops of DNA. They could however be important for computing DNA J-factors, *i.e.* the probability of fully closed loops forming.

Figure P2.3.4. An example of the lowest energy portion of the closed loop bifurcation diagram of the sequence S^{γ}. The crosses indicate the minimum and second minimum energy fully closed solutions of linking number 14 (red and orange) and 15 (bright blue and dark blue). The 3D configurations of these solutions are presented in the four panels with colours of frames corresponding to colours of crosses. The backbone tubes of the 3D configurations are coloured by the norm of a total moment $|\mathbf{m}|$. Note the large energy difference between the corresponding solutions.

Figure P2.3.5. An example of the lowest energy portion of the closed loop bifurcation diagram for the sequence $S^{\lambda''}$. The conventions here are the same as those of Figure P2.3.5. However, as discussed in Section P2.2.2, in this case the lowest energy part of the solution set forms a single connected component contrary to the two components for sequence S^{γ} (see Figure P2.3.4a).

Symmetry breaking computations have been performed using the automatic symmetry breaking script of Section P2.2.2 for the present periodic birod coefficients. Two sequences of length 158 bp have been used. The first one, labelled S^{γ}, was engineered to have a big intrinsic bend (see Section A.3.1.1) and was used in *in vitro* cyclization experiments [KahCro1992]. The other one, $S^{\lambda\prime\prime}$, is taken as base pairs 36901–37058 of the λ-phage genome [SanCouHon1982] and is intrinsically close to straight. The latter sequence is a fragment of $S^{\lambda\prime}$ used in the pulling and twisting experiment, which in turn is a fragment of S^{λ} introduced in Chapter P1.4. The results of symmetry breaking computations for these oligomers are presented in Figures P2.3.4 and P2.3.5.

As expected the high intrinsic bend of S^{γ} has a great impact on cyclization of that oligomer as it determines the most preferable direction of bending. For the lowest energy minicircles of linking number $Lk = 14$ and $Lk = 15$ the closed loop is formed exactly in the direction of the bend (see Figures P2.3.4d and Figures P2.3.4e). In the corresponding second lowest energy solutions the bent region is turned "inside out" (Figures P2.3.4b and Figures P2.3.4c). This can be inferred from the distribution of the norm of the moment $|\mathbf{m}|$, represented by the colour of the backbone tubes in the figures. This fact is related to the conjecture of Section P2.2.2 stating that the two lower energy solutions are local minima, while the higher energy ones are saddle points. This is known to be true in analogous rod computations.

In case of $S^{\lambda\prime\prime}$ the most preferred direction of bending for the oligomer is much less clear. As a result the moment distribution along the oligomer in the 3D configurations presented in Figure P2.3.5 is much more uniform than for S^{γ}. Also for $S^{\lambda\prime\prime}$ the energy differences between the two solutions of the same linking number are much lower than for S^{γ} (see Table P2.3.1).

	S^{γ}				$S^{\lambda\prime\prime}$			
	$Lk = 14$		$Lk = 15$		$Lk = 14$		$Lk = 15$	
Solution	(d)	(b)	(e)	(c)	(d)	(b)	(e)	(c)
$E\ [k_BT]$	21.08	35.31	17.10	34.67	24.54	32.27	21.26	24.44
$\Delta E\ [k_BT]$	14.24		17.57		7.73		3.18	
$\Delta E\ [\%]$	67.5		102.8		31.5		15.0	

Table P2.3.1. *The energies of the four lowest energy minicircles of oligomers S^{γ} and $S^{\lambda\prime\prime}$.*

Note also that, as already discussed in Section P2.2.2, the connectivity of the components of the solution sets is very different for the two oligomers. The blue and red branches of the bifurcation diagram for S^{γ} in Figure P2.3.4a are analogous to the blue and red branches of Figure B.3.3. However, in the case of $S^{\lambda\prime\prime}$ the lowest energy part of the solution set is a single closed component. It is shown in Figure P2.3.5a in its entirety.

Conclusions

In an attempt to develop tools that allow the study of sequence-dependent mechanics of DNA we have focused our attention on two particular coarse-grained models. In part P1 we have addressed matters regarding *cgDNA* – the discrete, rigid-base, nearest neighbour model of [Pet2012; GonPetMad2013; Pet2012]. Part P2 has been dedicated to the continuum elastic birod model of DNA of [Gra2016], originally proposed in [MoaMad2005]. Necessary background material concerning both models is outlined in part B.

Part P1 begins with more theoretical considerations concerning improvements and extensions of the *cgDNA* model itself and moves on to a more applied discussion of methods of analysing mechanical properties of DNA within the model.

Chapter P1.1 presents a simple procedure of constructing maximum entropy fits for covariance matrices with particular overlapping squares sparsity patterns illustrated in the top left panel of Figure P1.1.2. The scheme of building the inverse covariance matrix of the fit (presented schematically in Figure P1.1.3) involves summation of local inverses of the overlapping diagonal sub-blocks defining the sparsity and of local inversions of their overlaps. A recursive algorithm for maximum entropy completion of such partially specified covariance matrices is also outlined and illustrated in Figure P1.1.4. Arguments of [GonPetPas] have been quoted that the parameter set of the *cgDNA* model obtained using the presented maximum entropy fit, labelled *cgDNAparamset2*, provides better predictive capabilities than the original *cgDNAparamset1* of [Pet2012; GonPetMad2013]. This is also supported by the findings of Chapter P1.4, presented in Figure P1.4.4, where the sequence averaged persistence length with *cgDNAparamset2* is significantly closer to the consensus value than that of *cgDNAparamset1*. For those reasons *cgDNAparamset2* has been used to obtain all results of the hereby thesis.

In Chapter P1.2 we have developed a way to construct what we call a periodic *cgDNA* stiffness matrix $\boldsymbol{K}_p(S)$ and ground state configuration vector $\widehat{\boldsymbol{w}}_p(S)$ for a given DNA sequence S that characterize the energy of a long linear DNA tandem repeat $S_M = \underbrace{SS\ldots S}_{M}$ (with $M \to \infty$) of that sequence. We have laid out a rigorous argument that the standard *cgDNA* energy of any finite tandem repeat can be approximated using the periodic

coefficients up to a constant that depends on the magnitude of the end effects. A numerical study of a large ensemble of tandem repeats of fragments of various sizes has led to the conclusion that in fact $\widehat{\boldsymbol{w}}_p(S)$ for any sequence S well approximates $\widehat{\boldsymbol{w}}(S)$ five or more base pairs away from either end of the molecule. We have argued that the same coefficients can be used to characterize covalently bonded loops of DNA, as they preserve the periodicity of the loop. As a consequence the periodic ground state configuration vector has been used to parametrize the example computations of closed loops of DNA in Section P2.3.2.

Chapter P1.3 describes a method of characterizing the superhelices formed by base pair positions of ground state configurations of DNA tandem repeats in the *cgDNA* model with periodic coefficients. The few special cases of 3D configurations for which the method could fail have been pointed out, but were deemed very unlikely in case of real DNA oligomers. We have shown that the quantities such as the radius and pitch as computed by the method do not depend on the number M of repeats of the basal sequence S in the tandem repeat S_M. It has been also proven that the same values of those quantities are computed for any cyclic shift of the sequence S, provided that the periodic *cgDNA* coefficients are used.

The method has been shown to calculate pitch and radius of the primary DNA double helix instead of the superhelix for some (but not all) of the very straight fragments, depending on subtleties of the 3D configuration. This has been observed in cases where the superhelix could not be distinguished from the primary double helix. Such fragments have been named atypical, while all others were called typical.

Using our approach we have analysed superhelices formed by intrinsic shapes of all possible basal sequences of length up to 12 bp. The outcome of the study is summarized in the scatter plots of Figure P1.3.5. All oligomers of length up to 7 were found to be atypical and so extremely close to straight. We have shown that all typical superhelices of basal sequences of length under 11 bp are left-handed, while those of oligomers 12 bp long are all right-handed. 3D configurations for oligomers with extreme pitches and radii among those of length 8, 9, 10 and 12 are presented in Figure P1.3.6.

Undecanucleotides were found to be exceptional in many aspects, in particular compared to both decanucleotides and dodecanucleotides. Amongst them 15% formed left handed helices, while the rest are right-handed. Also the ranges of radius and pitch in this group have been found to be several orders of magnitude greater than those for decanucleotides and dodecanucleotides. This uniqueness of undecanucleotides could be explained by a conjecture that the helical DNA repeat in the *cgDNA* model with *cgDNAparamset2* is just under 11 bp. As depicted clearly in Figure 1 of [DubBedFur1994] this could also the reason why many of the oligomers in this group form helices with a very low pitch to radius ratio. The 3D configuration of the superhelix closest to circular amongst all studied sequences, which happens to be an undecanucleotide, is shown in Figure P1.3.7.

Chapter P1.4 presents results of computations of DNA persistence lengths using *cgDNAmc* – an efficient Monte Carlo code for the *cgDNA* model. Details of the design choices in the *cgDNA* code are presented, *e.g.* the application of Cholesky factorization in direct Monte Carlo sampling that takes advantage of the sparsity of the *cgDNA* model. The efficiency of the code has been evaluated. Presented statistics of performed simulations indicate strong sequence-dependence of the persistence lengths. The observed sequence-averaged values of persistence lengths of 53.5 nm (in the sense of Flory) and 160 bp (in the sense of apparent tangent-tangent correlation), computed using *cgDNAparamset2*, are in notably better agreement with the accepted values of 50 nm and 150 bp, respectively, than those of the original *cgDNAparamset1*.

In part P2 we turn our attention to the continuum DNA birod model. Originally introduced in [MoaMad2005], it was further studied in [Gra2016]. Most specifically it has been equipped with a Hamiltonian formulation that can be parametrized from the *cgDNA* model. The remainder of the thesis is therefore devoted to describing a software tool called *bBDNA* and to the adaptation of numerical techniques introduced in the context of elastic rods to the computation of stationary solutions of the birod DNA model.

In Chapter P2.1 we have addressed two issues of using the DNA coefficients of [Gra2016] of the birod Hamiltonian system in parameter continuation computations using the latest implementation of *AUTO* package [DoeKelKer1991a; DoeKelKer1991b] called *AUTO-07p* [DoeChaDer2009] introduced briefly in Chapter B.2.

First we have described the problem of failing bifurcation detection in *AUTO-07p* observed when any parameter continuation was run for the DNA birod model – nearly every computed solution was reported as a bifurcation point. The problem was identified as numerical stiffness of the birod system in the particular case of the DNA coefficients, which affected the method of evaluation of the so called bifurcation function. The bifurcation function is used by *AUTO-07p* to assess when the Jacobian matrix of the discretized system becomes singular for a given solution. Such solutions are recognized as bifurcation points. We have proposed an alternative definition of the bifurcation function directly related to a standard method of evaluating determinants, and have implemented it in *AUTO-07p*. The modified version of the continuation package have been used with success to compute all the stationary DNA birod solutions of this part of the thesis as well those presented in [Gra2016].

The other issue with birod DNA computations in *AUTO-07p* was the necessity of excessive discretization of the system. We found that $\sim 4N$ discretization points are required for the solver to converge for an oligomer N base pairs long. As described in Section B.3.3.3 the DNA coefficients of the birod system are only piecewise continuous with discontinuities at every base pair. In practice we observed that the discontinuities are very pronounced which is the reason that a fine discretization is required. We pointed out that despite the apparent excessive discretization required, computations in the DNA birod model are still

rather efficient. Additionally, the only technique currently known of obtaining stationary solutions in the *cgDNA* model is through discrete solves of the finite dimensional system described in [Gra2016, sec. 6.3], initialized with approximate solutions coming from discretized birod DNA solutions. Nevertheless the criticism of applying a continuum DNA model that requires more discretization points than the number of base pairs seem to be justified. Hence we have proposed a coefficient homogenization technique based on the ideas of [Gra2016, sec. 7.3] presented in the context of elastic rods. The method is a three step process. At first the intrinsic piecewise helix is factored out. Subsequently a window averaging procedure is applied. Finally a constant helix is factored in. This constant helix represents the relation between first and last frame of the factored out intrinsic configuration. The final step is crucial to ensure that boundary condition on the average rod configuration of the original system have the same meaning as in the homogenized one.

To validate our proposition we have compared a number of solutions of the homogenized birod DNA model with corresponding solutions of the original, non-homogenized system. We have observed only minor differences between corresponding configurations (the results are summarized in Table P2.1.1). The solutions for the homogenized birod DNA system have been found also to serve as good approximations of *cgDNA* stationary solutions when used to provide initial guesses for the aforementioned discrete *cgDNA*solver (the results are summarized in Table P2.1.2). Therefore all birod DNA solutions presented in the following chapters have been computed using the suggested homogenization of coefficients. We have to stress here, however, that although the presented results of coefficient homogenization suffice for the purposes of our presentation they are still preliminary and warrant further investigation, *e.g.* to identify optimal window size necessary for convergence.

Chapter P2.2 describes the *bBDNA* framework for performing parameter continuation in the birod model of DNA. The complete procedure (written in *MATLAB*®) of computing coefficients of a given DNA oligomer for the birod *AUTO* problem script has been presented, as well as the details of the ANSI C implementation of the *AUTO* problem script itself. The structure of the functional part of the *bBDNA* software has been outlined and the user interface of the application has been presented. Finally the procedure of symmetry breaking from straight, inextensible, unshearable, uniform, transversely isotropic rods directly to DNA sequence-dependent birods has been described, as implemented in the *bBDNA* framework. This particular technique of symmetry breaking was introduced in the context of elastic rods [LiMad1996; DicLiMad1996] and was used also for modelling of DNA [ManMadKah1996; FurManMad2000]. Example results of using this approach within the DNA birod model are presented in Chapter P2.3.

The last Chapter P2.3 presents results of example computations in the birod model of DNA performed using the *bBDNA* software. Section P2.3.1 is devoted to the pulling and twisting boundary value problem. Reaction to under- and over-twisting has been studied for three different sequences of length $\sim 260\,\mathrm{bp}$: one intrinsically close to straight ($S^{\lambda\prime}$)

and two with distinct superhelical structure of the intrinsic shape (left-handed for S^a and right-handed for S^c). The oligomers have been stretched with a pulling force of up to 2 pN applied in the direction of the end-to-end vector of their intrinsic configurations. Subsequently positive and negative twists of up to ±450 pNÅ have been applied (see Figures P2.3.2 and Figures P2.3.3). As expected the three sequences have been found to exhibit notably different behaviour. For the intrinsically straight $S^{\lambda\prime}$ fragment the effect of the loads on the 3D configuration was the least pronounced, although a slight extension has been observed in the case of moderate over-twisting and shortening for under-twist. Analogous relations between twist and extension have been observed for the left-handed superhelix of S^a, yet the magnitude of the effect was nearly nine times higher than for $S^{\lambda\prime}$. Under positive twist the superhelix was unravelled and stretched while for negative twist, tightening of the superhelix caused its shortening. The most interesting case was the reaction of the right-handed superhelix S^a. For small pulling forces (< 1 pN) as the superhelix was tightened through a positive twist it extended and *vice versa*, while for the pulling force of 2 pN the superhelix tightened by the twist of 150 pNÅ was shorter than the unravelled one for −150 pNÅ.

Section P2.3.2 presents examples of the closed loop boundary value problem. The considerations begin with a description of the boundary conditions appropriate for modelling minicircles, *i.e.* covalently bonded loops of DNA. A small modification of the DNA coefficients defined in [Gra2016] has been proposed, which uses periodic *cgDNA* coefficients of Chapter P1.2 and reflects periodicity of covalently bonded minicircles. Results of symmetry breaking for two sequences of 158 bp have been analysed. One of the sequences (S^γ) has been designed to have a large intrinsic bend [KahCro1992]. The other one ($S^{\lambda\prime\prime}$) is an initial fragment of the intrinsically close to straight $S^{\lambda\prime}$ used in P2.3.1. The automatic symmetry breaking script of *bBDNA* has been used to generate the low energy part of the solution set for the two sequences (see Figures P2.3.4 and P2.3.5). The sequence dependence of the results is clearly visible in the bifurcation diagrams of Figures P2.3.4a and P2.3.5a. In case of S^γ the geometry of the blue and red branches is rather simple. The correspondence to the blue and red branches of the bifurcation diagram for rods of Figure B.3.3b is clear. The energy differences between lowest and second lowest energies of loops of linking number 14 and 15 are considerable. This is due to the large intrinsic bend in the sequence, which clearly dictates the preferable direction of bending of the oligomer. For the intrinsically straight sequence S^γ the respective energy differences are much smaller and the bifurcation diagram is more complicated. This can be attributed to the fact that the oligomer is much closer to having the register symmetry of the ideal, transversely isotropic, rod case.

The two examples presented in this chapter are an illustration of how the *bBDNA* software of Chapter P2.2, together with the numerical methods proposed for treatment of the birod DNA model can be used to study sequence-dependent mechanical properties of DNA. In the case of the pulling and twisting boundary value problem we showed that the response (extension or shortening) of DNA under moderate under- and over-twisting can

strongly depend on the intrinsic shape of the given fragment. In the context of modelling minicircles we have also found strong dependence on the intrinsic shape of the closed DNA configurations as well as the connectivity of the solution set.

In an overall summary we have presented ideas, frameworks and software tools to facilitate the study of the sequence-dependent statistical mechanical properties of DNA. Of course both the tools and their underlying models could be further improved and extended. One of the most important restrictions is that currently our computations are limited to studying the first order conditions governing stationary solutions, and a formulation of the second variation governing stability properties within the birod DNA model, as well as efficient methods of its evaluation, are still a topic of active research. A complete such theory in terms of Jacobi fields is available for rods, but the stiffness of the birod equations suggest that the generalization to birods is not immediate. Nevertheless we believe that the results presented here provide a valuable contribution to the field of modelling of sequence-dependent DNA mechanics. In particular the *bBDNA* software provides a first step towards practical application of the continuum birod formalism to modelling of scientifically important aspects of DNA mechanics.

Outwith the field of DNA mechanics, and although the numerical stiffness in the birod DNA model is not yet fully understood, the proposed modification of the bifurcation function of the *AUTO-07p* [DoeChaDer2009] code resolves issues with bifurcation detection, which might also be useful in boundary value problems arising in other contexts. Similarly, we consider the maximum entropy procedure for fitting Gaussian stiffness matrices of a prescribed overlapping squares sparsity patterns, as described in Chapter P1.1, deserves to be more widely known.

Appendices

A.1 Algebra of 3D transformations

A.1.1 3D Rotations

We begin by recalling Euler's rotation theorem [Eul1776, pp. 201–203], which states that every thee dimensional displacement of a rigid body that does not change the position of the reference point of the body can be expressed as a right-handed rotation by a given angle $\theta \in [0, \pi]$ around a given axis. For a rotation axis given as a unit vector $\hat{\boldsymbol{e}}$ a right-handed rotation by θ is equivalent to a left-handed rotation by $2\pi - \theta$ in the opposite direction. To remove ambiguity from here on only right-handed (with respect to $\hat{\boldsymbol{e}}$) rotations by $\theta \in [0, \pi]$ will be considered. Note that rotations through π are a special singular case in the sense that a right-handed and left-handed rotation by π yields the same result. For that reason rotations by π will require special treatment in parts of what follows.

Rotations in three dimensions (under composition) form a group $\mathcal{SO}(3)$, which can be parametrized in multiple ways. Each of those parametrizations has its strengths and weaknesses, and different applications require a different choice. This short (and far from exhaustive) summary underlines three particular parametrizations used for computations in the thesis. These parametrizations are: rotation matrices, quaternions (e.g. [Kui1999; Han2006])) and Cayley vectors (in the specific sense defined in [LanGonHef2009; Pet2012]). In this review we focus on relations that will allow for efficient and accurate numerics, as detailed in Section P1.4.2.3.

A.1.1.1 Rotation matrices

A very natural choice for parametrizing rotations in three dimensions is the group of orthogonal 3×3 matrices with determinant $+1$ that is isomorphic to $\mathcal{SO}(3)$. Indeed in this parametrization composition of rotations is achieved using regular matrix multiplication:

$$\boldsymbol{R}_2 \circ \boldsymbol{R}_1 = \boldsymbol{R}_2\boldsymbol{R}_1 \tag{A.1.1}$$

The unique inverse rotation is represented by the inverse of the rotation matrix, which in the case of orthogonal matrices is equivalent to transposition.

It can be shown that the set of eigenvalues of each $\boldsymbol{R} \in \mathcal{SO}(3)$ can be written as $\sigma(\boldsymbol{R}) = \{1, e^{i\theta}, e^{-i\theta}\}$. The the unit eigenvector $\hat{\boldsymbol{e}}$ for the eigenvalue 1 defines the axis of rotation, which is invariant under that rotation. The argument θ of the exponentials can be interpreted as the rotation angle.

Note that if $\theta \equiv 0 \pmod{\pi}$ all the eigenvalues are real and the matrix is diagonalizable in $\mathbb{R}$. It is easy to see that an orthogonal matrix is diagonalizable if and only if it is also symmetric, and as a result is its own inverse. There are two particular cases of such symmetric rotation matrices:

a) $\theta = 2k\pi$, $(k \in \mathbb{Z})$ when $\boldsymbol{R}$ is the identity matrix and has one real eigenvalue $\lambda = 1$ of geometric multiplicity 3; In this case the rotation axis is not well defined because the eigenspace of the eigenvalue 1 is the whole of $\mathbb{R}^3$;

b) $\theta = (2k+1)\pi$, $(k \in \mathbb{Z})$ when $\boldsymbol{R}$ has two distinct eigenvalues: 1 with multiplicity $\lambda_1 = 1$ and $\lambda_2 = -1$ with multiplicity 2 (a rotation through π). Note that the same matrix represents the rotation by π around $\hat{\boldsymbol{e}}$ and around $-\hat{\boldsymbol{e}}$.

A rotation represented by a matrix $\boldsymbol{R} \in \mathcal{SO}(3)$ can be applied to a vector $\boldsymbol{v}$ through the regular matrix vector product:

$$\mathrm{rot}\,(\boldsymbol{R}, \boldsymbol{v}) := \boldsymbol{R}\boldsymbol{v} \quad . \tag{A.1.2}$$

A.1.1.2 Quaternions

Quaternion algebra is discussed at length e.g. in [Kui1999, ch. 5]). Here only basic concepts will be mentioned. As originally introduced by Hamilton [Ham1844] the quaternion algebra was defined as an extension to the complex number algebra to rank 4. A quaternion can be seen as:

$$\mathbb{R}^4 \ni \boldsymbol{q} = q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} + q_4\boldsymbol{1} \tag{A.1.3}$$

with the basis:

$$\boldsymbol{i} = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}^T, \quad \boldsymbol{j} = \begin{bmatrix}0 & 1 & 0 & 0\end{bmatrix}^T, \quad \boldsymbol{k} = \begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}^T, \tag{A.1.4a}$$

$$\boldsymbol{1} = \begin{bmatrix}0 & 0 & 0 & 1\end{bmatrix}^T \tag{A.1.4b}$$

For any quaternion $\boldsymbol{q} = (q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} + q_4\boldsymbol{1})$ define:

$$\mathrm{Im}(\boldsymbol{q}) := \begin{bmatrix}q_1 & q_2 & q_3\end{bmatrix}^T \in \mathbb{R}^3 \tag{A.1.5a}$$

$$\mathrm{Re}(\boldsymbol{q}) := q_4 \in \mathbb{R} \tag{A.1.5b}$$

as the imaginary (vector) part and real (scalar) part of the quaternion in a manner analogous to complex numbers.

For any vector $\boldsymbol{v} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}^T \in \mathbb{R}^3$ define the pure imaginary quaternion:

$$\boldsymbol{v}^q := \begin{bmatrix} v_1 & v_2 & v_3 & 0 \end{bmatrix}^T \tag{A.1.6}$$

Let $\boldsymbol{p} = (p_1\boldsymbol{i} + p_2\boldsymbol{j} + p_3\boldsymbol{k} + p_4\mathbf{1})$ and $\boldsymbol{q} = (q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} + q_4\mathbf{1})$ be two quaternions. Equality of quaternions is defined as equality of all the components and addition is defined component by component:

$$\boldsymbol{p} = \boldsymbol{q} \quad \Longleftrightarrow \quad p_1 = q_1 \ \wedge \ p_2 = q_2 \ \wedge \ p_3 = q_3 \ \wedge \ p_4 = q_4 \tag{A.1.7a}$$

$$\boldsymbol{p} + \boldsymbol{q} = (p_1 + q_1)\boldsymbol{i} + (p_2 + q_2)\boldsymbol{j} + (p_3 + q_3)\boldsymbol{k} + (p_4 + q_4)\mathbf{1} \quad . \tag{A.1.7b}$$

The Quaternion product is defined by the famous Hamilton formulae [Ham1844]:

$$\boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -\mathbf{1} \quad , \tag{A.1.8}$$

which lead to

$$\begin{aligned} \boldsymbol{pq} &= (p_1\boldsymbol{i} + p_2\boldsymbol{j} + p_3\boldsymbol{k} + p_4\mathbf{1})(q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} + q_4\mathbf{1}) \\ &= (p_1q_4 + p_2q_3 - p_3q_2 + p_4q_1)\boldsymbol{i} \\ &\quad + (-p_1q_3 + p_2q_4 + p_3q_1 + p_4q_2)\boldsymbol{j} \\ &\quad + (p_1q_2 - p_2q_1 + p_3q_4 + p_4q_3)\boldsymbol{k} \\ &\quad + (-p_1q_1 - p_2q_2 - p_3q_3 + p_4q_4)\mathbf{1} \qquad \text{(A.1.9)} \\ &= \begin{bmatrix} | \\ q_4\mathrm{Im}(\boldsymbol{p}) + p_4\mathrm{Im}(\boldsymbol{q}) + \mathrm{Im}(\boldsymbol{p}) \times \mathrm{Im}(\boldsymbol{q}) \\ | \\ p_4q_4 + \mathrm{Im}(\boldsymbol{p}) \cdot \mathrm{Im}(\boldsymbol{q}) \end{bmatrix} \qquad \text{(A.1.10)} \end{aligned}$$

Conjugation in quaternion algebra is defined as:

$$\overline{\boldsymbol{q}} = -q_1\boldsymbol{i} - q_2\boldsymbol{j} - q_3\boldsymbol{k} + q_4\mathbf{1} \quad . \tag{A.1.11}$$

The norm af a quaternion $\boldsymbol{q}$ is defined as:

$$|\boldsymbol{q}| = \sqrt{\boldsymbol{q}\overline{\boldsymbol{q}}} = \sqrt{q_1^2 + q_2^2 + q_3^2 + q_4^2} \quad . \tag{A.1.12}$$

From Equations (A.1.9) and (A.1.11) it is easy to see that the inverse $\boldsymbol{q}^{-1}$ of a quaternion $\boldsymbol{q} \neq \mathbf{0}$ (such that $\boldsymbol{q}\boldsymbol{q}^{-1} = \boldsymbol{q}^{-1}\boldsymbol{q} = \mathbf{1}$) can be written as:

$$\boldsymbol{q}^{-1} = \frac{\overline{\boldsymbol{q}}}{|\boldsymbol{q}|} \quad . \tag{A.1.13}$$

It can be shown that the set of quaternions constrained to have unit norm equipped with the quaternion product forms the special unitary group $\mathcal{SU}(2)$ (e.g. [Han2006, ch. 7])). It can also be shown that $\mathcal{SU}(2)$ is a double covering of $\mathcal{SO}(3)$. The two quaternions representing a rotation around a unit axis $\hat{\boldsymbol{e}}$ by the angle $\theta \in [0, \pi]$ can be written as:

$$\begin{cases} \boldsymbol{q}_+ &= \left[\sin\left(\frac{\theta}{2}\right)\hat{\boldsymbol{e}} \quad \cos\left(\frac{\theta}{2}\right)\right]^T \\ \boldsymbol{q}_- &= -\boldsymbol{q}_+ \end{cases} \tag{A.1.14}$$

In this representation the composition of rotation is expressed by the quaternion product (A.1.9):

$$\boldsymbol{q}_2 \circ \boldsymbol{q}_1 = \boldsymbol{q}_2\boldsymbol{q}_1. \tag{A.1.15}$$

Note that under the unit norm constraint the inverse (A.1.13) simply becomes the conjugation (A.1.11). After (A.1.14) the inverse (conjugate) quaternion $\boldsymbol{q}^{-1} = \overline{\boldsymbol{q}}$ can be seen as representing a rotation by the same angle around the opposite axis with respect to $\boldsymbol{q} \in \mathcal{SU}(2)$.

Note also that quaternions well represent the aforementioned special cases of rotations by $k\pi, k \in \mathbb{Z}$. It should be pointed out that the double covering of $\mathcal{SO}(3)$ by $\mathcal{SU}(2)$ reflects the fact that the quaternions are able to track the angle of rotation modulo 4π. This feature of $\mathcal{SU}(2)$ can be used to explain certain aspects of sequences of rotations that cannot be explained using e.g. rotation matrices including the Călugăreanu-White-Fuller formula $Lk = Tw + Wr$, pertinent in DNA modelling, see *e.g.* [HofManMad2003].

It can be also shown that for any unit quaternion $\boldsymbol{q} = (q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} + q_4\boldsymbol{1}) \in \mathcal{SU}(2)$ the matrix $\boldsymbol{R} \in \mathcal{SO}(3)$ representing the same rotation can be written as:

$$\boldsymbol{R} = \begin{bmatrix} q_1^2 - q_2^2 - q_3^2 + q_4^2 & 2(q_1q_2 - q_3q_4) & 2(q_1q_3 + q_2q_4) \\ 2(q_1q_2 + q_3q_4) & -q_1^2 + q_2^2 - q_3^2 + q_4^2 & 2(q_2q_3 - q_1q_4) \\ 2(q_1q_3 - q_2q_4) & 2(q_2q_3 + q_1q_4) & -q_1^2 - q_2^2 + q_3^2 + q_4^2 \end{bmatrix} \tag{A.1.16}$$

Note that from the above formula and the unit norm constraint we get:

$$\begin{cases} |q_1| = \frac{1}{2}\sqrt{1 + \boldsymbol{R}_{11} - \boldsymbol{R}_{22} - \boldsymbol{R}_{33}} \\ |q_2| = \frac{1}{2}\sqrt{1 - \boldsymbol{R}_{11} + \boldsymbol{R}_{22} - \boldsymbol{R}_{33}} \\ |q_3| = \frac{1}{2}\sqrt{1 - \boldsymbol{R}_{11} - \boldsymbol{R}_{22} + \boldsymbol{R}_{33}} \\ |q_4| = \frac{1}{2}\sqrt{1 + \mathrm{tr}(\boldsymbol{R})} \end{cases} \tag{A.1.17}$$

Using the double covering any one component of a quaternion to be computed can be assumed positive and so computed using the above formulae. In practice, for the sake of accuracy and efficiency this component is chosen to be far from zero. The others can be

recovered from (A.1.16) as:

$$q_2 = \frac{\boldsymbol{R}_{21} + \boldsymbol{R}_{12}}{4q_1} \qquad q_3 = \frac{\boldsymbol{R}_{13} + \boldsymbol{R}_{31}}{4q_1} \qquad q_4 = \frac{\boldsymbol{R}_{32} - \boldsymbol{R}_{23}}{4q_1} \tag{A.1.18a}$$

$$q_1 = \frac{\boldsymbol{R}_{12} + \boldsymbol{R}_{21}}{4q_2} \qquad q_3 = \frac{\boldsymbol{R}_{23} + \boldsymbol{R}_{32}}{4q_2} \qquad q_4 = \frac{\boldsymbol{R}_{13} - \boldsymbol{R}_{31}}{4q_2} \tag{A.1.18b}$$

$$q_1 = \frac{\boldsymbol{R}_{31} + \boldsymbol{R}_{13}}{4q_3} \qquad q_2 = \frac{\boldsymbol{R}_{23} + \boldsymbol{R}_{32}}{4q_3} \qquad q_4 = \frac{\boldsymbol{R}_{21} - \boldsymbol{R}_{12}}{4q_3} \tag{A.1.18c}$$

$$q_1 = \frac{\boldsymbol{R}_{32} - \boldsymbol{R}_{23}}{4q_4} \qquad q_2 = \frac{\boldsymbol{R}_{13} - \boldsymbol{R}_{31}}{4q_4} \qquad q_3 = \frac{\boldsymbol{R}_{21} - \boldsymbol{R}_{12}}{4q_4} \tag{A.1.18d}$$

The above procedure together with the geometric interpretation (A.1.14) provides a way to extract the rotation axis for a given rotation matrix whenever it is well defined (i.e. for all rotations except the identity – see considerations about identity in Section A.1.1.1).

Considered the following triple quaternion product with $\boldsymbol{q} \in \mathcal{SU}(2)$ and $\boldsymbol{v} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}^T \in \mathbb{R}^3$

$$\boldsymbol{w} = \boldsymbol{q}\boldsymbol{v}^q\overline{\boldsymbol{q}} \tag{A.1.19}$$

It can be shown that $\boldsymbol{w}$ is a pure imaginary quaternion and $\mathrm{Im}(\boldsymbol{w})$ represents the vector that is a result of applying the rotation $\boldsymbol{q}$ to the vector $\boldsymbol{v}$. Given (A.1.10) Equation (A.1.19) leads to the following definition of the quaternion rotation operator [Kui1999, ch. 5]:

$$\begin{aligned} \mathrm{rot}\,(\boldsymbol{q}, \boldsymbol{v}) := {} & (q_4^2 - \mathrm{Im}(\boldsymbol{q}) \cdot \mathrm{Im}(\boldsymbol{q}))\boldsymbol{v} \\ & + 2(\boldsymbol{v} \cdot \mathrm{Im}(\boldsymbol{q}))\mathrm{Im}(\boldsymbol{q}) \\ & + 2q_4(\mathrm{Im}(\boldsymbol{q}) \times \boldsymbol{v}) \in \mathbb{R}^3 \end{aligned} \tag{A.1.20}$$

For a more complete characterization of quaternions see e.g. [Kui1999; Han2006].

A.1.1.3 Cayley vectors

In this section only rotations by π will be excluded as a singular case of the transformations used below. Rotations around $\hat{\boldsymbol{e}}$ by angles $\alpha \in [\pi, 2\pi]$ are represented as rotations by $2\pi - \alpha$ around $-\hat{\boldsymbol{e}}$.

Before introducing the Cayley vector representation of rotations it will be convenient to define the set of skew symmetric (antisymmetric) matrices:

$$\mathcal{A} = \{A \in \mathbb{R}^{3\times 3} : A = -A^T\} \quad . \tag{A.1.21}$$

It is easy to see that $\mathcal{A}$ is isomorphic to $\mathbb{R}^3$.

In fact for each vector $\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}^T$ define:

$$[\boldsymbol{a}^\times] = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \in \mathcal{A} \tag{A.1.22}$$

which can be interpreted as a cross product operator associated with the vector $\boldsymbol{a}$ as follows:

$$[\boldsymbol{a}^\times]\,\boldsymbol{b} = \boldsymbol{a} \times \boldsymbol{b}, \quad \boldsymbol{b} \in \mathbb{R}^3 \quad . \tag{A.1.23}$$

Similarly, for each matrix $A \in \mathcal{A}$ define:

$$\text{vec}\,(A) = \begin{bmatrix} A_{32} & A_{13} & A_{21} \end{bmatrix}^T \in \mathbb{R}^3 \quad . \tag{A.1.24}$$

The Rodrigues' rotation formula that rotates a vector $\boldsymbol{v}$ around the axis $\hat{\boldsymbol{e}}$ by the angle θ [Rod1840] reads:

$$\boldsymbol{R}\boldsymbol{v} = \boldsymbol{v}\cos\theta + (\hat{\boldsymbol{e}} \times \boldsymbol{v})\sin\theta + \hat{\boldsymbol{e}}(\hat{\boldsymbol{e}} \cdot \boldsymbol{v})(1 - \cos\theta) \tag{A.1.25}$$

which is equivalent to:

$$\boldsymbol{R} = \boldsymbol{I}\cos\theta + [\hat{\boldsymbol{e}}^\times]\sin\theta + \hat{\boldsymbol{e}}\hat{\boldsymbol{e}}^T(1 - \cos\theta) \tag{A.1.26a}$$

$$= \boldsymbol{I}\cos\theta + [\hat{\boldsymbol{e}}^\times]\sin\theta + \left([\hat{\boldsymbol{e}}^\times]^2 + \boldsymbol{I}\right)(1 - \cos\theta) \tag{A.1.26b}$$

$$= \boldsymbol{I} + [\hat{\boldsymbol{e}}^\times]\sin\theta + [\hat{\boldsymbol{e}}^\times]^2(1 - \cos\theta) \tag{A.1.26c}$$

where $\boldsymbol{R} \in \mathcal{SO}(3)$ is the respective rotation matrix applied to $\boldsymbol{v}$. Using the identities:

$$\cos\theta = \frac{1 - \tan^2\left(\frac{\theta}{2}\right)}{1 + \tan^2\left(\frac{\theta}{2}\right)} \qquad \Rightarrow \quad 1 - \cos\theta = \frac{2\tan^2\left(\frac{\theta}{2}\right)}{1 + \tan^2\left(\frac{\theta}{2}\right)} \tag{A.1.27a}$$

$$\Rightarrow \quad 1 + \cos\theta = \frac{2}{1 + \tan^2\left(\frac{\theta}{2}\right)} \tag{A.1.27b}$$

$$\sin\theta = \frac{2\tan\left(\frac{\theta}{2}\right)}{1 + \tan^2\left(\frac{\theta}{2}\right)} \tag{A.1.27c}$$

Equation (A.1.26c) can be rewritten as:

$$\boldsymbol{R} = \boldsymbol{I} + [\hat{\boldsymbol{e}}^\times]\frac{2\tan\left(\frac{\theta}{2}\right)}{1 + \tan^2\left(\frac{\theta}{2}\right)} + [\hat{\boldsymbol{e}}^\times]^2\frac{2\tan^2\left(\frac{\theta}{2}\right)}{1 + \tan^2\left(\frac{\theta}{2}\right)} \tag{A.1.28}$$

$$= \boldsymbol{I} + \frac{4}{4 + |\boldsymbol{\eta}|^2}\left([\boldsymbol{\eta}^\times] + \frac{1}{2}[\boldsymbol{\eta}^\times]^2\right) \quad . \tag{A.1.29}$$

where

$$\boldsymbol{\eta} = 2\tan\left(\tfrac{\theta}{2}\right)\hat{\boldsymbol{e}} \tag{A.1.30}$$

is the Cayley vector representation of the rotation $\boldsymbol{R}$, as defined *e.g.* in [LanGonHef2009; Pet2012].

From Equation (A.1.26a) and the fact that I and $[\boldsymbol{\eta}^\times]^2$ are symmetric, while $[\boldsymbol{\eta}^\times]$ is skew symmetric we get:

$$\operatorname{tr}(\boldsymbol{R}) = 3\cos\theta + 0 + |\hat{\boldsymbol{e}}|(1-\cos\theta) \tag{A.1.31}$$

$$\operatorname{tr}(\boldsymbol{R}) = 1 + 2\cos\theta \quad . \tag{A.1.32}$$

From Equations (A.1.27c) and (A.1.32) we have for $\theta \in [0, \pi)$:

$$2\tan\left(\tfrac{\theta}{2}\right) = \frac{2\sin\theta}{1+\cos\theta} = \frac{4\sin\theta}{\operatorname{tr}(\boldsymbol{R})+1} \quad . \tag{A.1.33}$$

Finally we can combine (A.1.33) with (A.1.29) to invert it:

$$\boldsymbol{R} - \boldsymbol{R}^T = [\hat{\boldsymbol{e}}^\times] 2\sin\theta \tag{A.1.34a}$$

$$\frac{2}{\operatorname{tr}(\boldsymbol{R})+1}(\boldsymbol{R} - \boldsymbol{R}^T) = [\hat{\boldsymbol{e}}^\times]\frac{4\sin\theta}{\operatorname{tr}(\boldsymbol{R})+1} \tag{A.1.34b}$$

$$[\boldsymbol{\eta}^\times] = \frac{2}{\operatorname{tr}(\boldsymbol{R})+1}(\boldsymbol{R} - \boldsymbol{R}^T) \tag{A.1.34c}$$

Note that in case of a rotation by $\theta = \pi$ we have $\sin\theta = 0$ and $\boldsymbol{R} - \boldsymbol{R}^T = \boldsymbol{0}$ (because $\boldsymbol{R}$ is symmetric, as discussed in Section A.1.1.1). As a result the formula (A.1.34a) holds ($\boldsymbol{0} = \boldsymbol{0}$), but cannot be used to extract the rotation axis even though the axis is well defined. The formula (A.1.33) is undefined because $\operatorname{tr}(\boldsymbol{R}) = 1 + 2\cos(\pi) = -1$.

After Equations (A.1.29) and (A.1.34c) we define the notation (used *e.g.* in [LanGonHef2009]):

$$\operatorname{cay}(\boldsymbol{\eta}) := \boldsymbol{I} + \frac{4}{4+|\boldsymbol{\eta}|^2}\left([\boldsymbol{\eta}^\times] + \frac{1}{2}[\boldsymbol{\eta}^\times]^2\right) \in \mathcal{SO}(3), \qquad \text{for } \boldsymbol{\eta} \in \mathbb{R}^3 \tag{A.1.35}$$

and

$$\operatorname{cay}^{-1}(\boldsymbol{R}) := \frac{2}{\operatorname{tr}(\boldsymbol{R})+1}\operatorname{vec}\left(\boldsymbol{R} - \boldsymbol{R}^T\right) \in \mathbb{R}^3 \qquad \text{for } \boldsymbol{R} \in \mathcal{SO}(3). \tag{A.1.36}$$

Note that (A.1.35) is well defined for any vector $\boldsymbol{\eta} \in \mathbb{R}^3$. The expression (A.1.36) is well defined for the identity matrix, whose Cayley vector representation is $\operatorname{cay}^{-1}(I) = \boldsymbol{0}$. It is undefined, however, for any rotation through π, with:

$$|\operatorname{cay}^{-1}(R_\theta)| = \left|2\tan\left(\tfrac{\theta}{2}\right)\right| \to +\infty \quad \text{for} \quad \theta \to \pi. \tag{A.1.37}$$

for R_θ a rotation by angle θ around any given axis. As a result in what follows Cayley vectors will only be used to represent relative (i.e. smaller than π) rotations.

The name Cayley vector comes from the Cayley transform that, in its original description [Cay1846], provides a mapping between $\mathcal{SO}(3)$ and the set of skew symmetric matrices $\mathcal{A}$. With the definition (A.1.35) of Cayley vector $\boldsymbol{\eta}$ (which is adapted from [LanGonHef2009]) the Cayley transform gives an alternative form of the formula (A.1.35):

$$\mathrm{cay}(\boldsymbol{\eta}) = \left(I + \frac{1}{2}[\boldsymbol{\eta}^\times]\right)\left(I - \frac{1}{2}[\boldsymbol{\eta}^\times]\right)^{-1} \tag{A.1.38}$$

where I is the identity matrix. Note that the matrix $(I+M)$ with $M \in \mathbb{R}^{3\times3}$ is non-singular if and only if -1 is not an eigenvalue of M. Any matrix $A \in \mathcal{A}$ has a single eigenvalue 0, hence the formula (A.1.38) is well defined.

From Equation (A.1.38) we have:

$$\mathrm{cay}(\boldsymbol{\eta})\left(I - \frac{1}{2}[\boldsymbol{\eta}^\times]\right) = I + \frac{1}{2}[\boldsymbol{\eta}^\times] \tag{A.1.39a}$$

$$2(\mathrm{cay}(\boldsymbol{\eta}) - I) = (I + \mathrm{cay}(\boldsymbol{\eta}))\,[\boldsymbol{\eta}^\times] \tag{A.1.39b}$$

so that the inverse transform can be computed provided that $(I + \mathrm{cay}(\boldsymbol{\eta}))$ is non-singular (*i.e.* for rotations by $\theta \in [, \pi)$ – see considerations about spectral decomposition of rotation matrices). Therefore for any $\boldsymbol{R} \in \mathcal{SO}(3)$ with the exclusion of rotations through π we can write formula (A.1.36) in the alternative form:

$$\mathrm{cay}^{-1}(\boldsymbol{R}) = 2\mathrm{vec}\left((\boldsymbol{R} + I)^{-1}(\boldsymbol{R} - I)\right) \tag{A.1.40}$$

Formulae (A.1.14) and (A.1.35) allow the definition of the relation between a Cayley vector $\boldsymbol{\eta} = \begin{bmatrix}\eta_1 & \eta_2 & \eta_3\end{bmatrix}$ and a quaternion representing the same rotation:

$$\mathrm{quat}(\boldsymbol{\eta}) := \sqrt{\frac{4}{|\boldsymbol{\eta}|^2 + 4}}\left(\frac{\eta_1}{2}\boldsymbol{i} + \frac{\eta_2}{2}\boldsymbol{j} + \frac{\eta_3}{2}\boldsymbol{k} + \mathbf{1}\right) \in \mathcal{SU}(2) \quad . \tag{A.1.41}$$

Note that this definition is well defined for any Cayley vector and always computes the q_+ variant of Equation (A.1.14). For any quaternion $\boldsymbol{q} = (q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} + q_4\mathbf{1}) \in \mathcal{SU}(2)$ such that $q_4 \neq 0$ (not a rotation by π) define also the Cayley vector representing the same rotation as:

$$\mathrm{quat}^{-1}(\boldsymbol{q}) := \begin{bmatrix}2\frac{q_1}{q_4} & 2\frac{q_2}{q_4} & 2\frac{q_3}{q_4}\end{bmatrix}^T \quad . \tag{A.1.42}$$

The Cayley vector $\boldsymbol{\eta}_2 \circ \boldsymbol{\eta}_1$ representing the composition of rotations represented as Cayley vectors $\boldsymbol{\eta}_2$ with $\boldsymbol{\eta}_1$ can be e.g. expressed using formula (A.1.1) as:

$$\boldsymbol{\eta}_2 \circ \boldsymbol{\eta}_1 = \mathrm{cay}^{-1}(\mathrm{cay}(\boldsymbol{\eta}_2)\,\mathrm{cay}(\boldsymbol{\eta}_1)) \tag{A.1.43}$$

or alternatively, by formula (A.1.15) as

$$\boldsymbol{\eta}_2 \circ \boldsymbol{\eta}_1 = \mathrm{quat}^{-1}(\mathrm{quat}(\boldsymbol{\eta}_2) \circ \mathrm{quat}(\boldsymbol{\eta}_1)) \tag{A.1.44}$$

A.1.1.4 Half rotations

Given a rotation it is sometimes necessary to compute the half rotation (rotation by half the angle – see e.g. Section B.1.3). For a rotation matrix $\boldsymbol{R} \in \mathcal{SO}(3)$ a half rotation can be computed as the principal square root of the matrix $\sqrt{\boldsymbol{R}}$.

Note that for $\theta \in [0, \pi]$ we can invert the trigonometric identities:

$$\cos\left(\tfrac{\theta}{2}\right) = 2\cos^2\left(\tfrac{\theta}{4}\right) - 1 \qquad \Rightarrow \quad \cos\left(\tfrac{\theta}{4}\right) = \sqrt{\frac{\cos\left(\tfrac{\theta}{2}\right)+1}{2}} = \frac{\cos\left(\tfrac{\theta}{2}\right)+1}{\sqrt{2(\cos\left(\tfrac{\theta}{2}\right)+1)}} \tag{A.1.45}$$

$$\Downarrow$$

$$\sin\left(\tfrac{\theta}{2}\right) = 2\sin\left(\tfrac{\theta}{4}\right)\cos\left(\tfrac{\theta}{4}\right) \quad \Rightarrow \qquad \sin\left(\tfrac{\theta}{4}\right) = \frac{\sin\left(\tfrac{\theta}{2}\right)}{2\cos\left(\tfrac{\theta}{4}\right)} = \frac{\sin\left(\tfrac{\theta}{2}\right)}{\sqrt{2(\cos\left(\tfrac{\theta}{2}\right)+1)}} \quad . \tag{A.1.46}$$

As a result for a quaternion $\boldsymbol{q} = (q_1\boldsymbol{i}+q_2\boldsymbol{j}+q_3\boldsymbol{k}+q_4\mathbf{1}) \in \mathcal{SU}(2)$ the geometric interpretation (A.1.14) allows for a simple explicit formula for the quaternion of the half rotation:

$$\sqrt{\boldsymbol{q}} = \frac{\boldsymbol{q} + \mathrm{sgn}^*(q_4)\mathbf{1}}{\sqrt{2(|q_4|+1)}} = \frac{1}{\sqrt{2(|q_4|+1)}}(q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k} + (q_4 + \mathrm{sgn}^*(q_4))\mathbf{1}) \tag{A.1.47}$$

where

$$\mathrm{sgn}^*(a) = \begin{cases} 1 & \text{for } a \geq 0 \\ -1 & \text{for } a < 0 \end{cases} \tag{A.1.48}$$

The norm of q_4 has to be taken and the multiplication by $\mathrm{sgn}^*(q_4)$ has to be added in (A.1.47) compared to [Han2006, ap. F.2] to make the formula true not only for $\boldsymbol{q}_+$ but also $\boldsymbol{q}_-$ of (A.1.14).

To define an analogous formula for Cayley vectors for angles $\theta \in [0, \pi)$ we use (A.1.45) and (A.1.46) and the identities:

$$\cos\left(\tfrac{\theta}{2}\right) = \frac{1}{\sqrt{\tan^2\left(\tfrac{\theta}{2}\right)+1}} \tag{A.1.49a}$$

$$\sin\left(\tfrac{\theta}{2}\right) = \frac{\tan\left(\tfrac{\theta}{2}\right)}{\sqrt{\tan^2\left(\tfrac{\theta}{2}\right)+1}} \tag{A.1.49b}$$

to get:

$$\tan\left(\tfrac{\theta}{4}\right) = \frac{\sin\left(\tfrac{\theta}{4}\right)}{\cos\left(\tfrac{\theta}{4}\right)} = \frac{\sin\left(\tfrac{\theta}{2}\right)}{\cos\left(\tfrac{\theta}{2}\right)+1} = \frac{\tan\left(\tfrac{\theta}{2}\right)}{1+\sqrt{\tan^2\left(\tfrac{\theta}{2}\right)+1}} \quad . \tag{A.1.50}$$

This, together with the definition (A.1.30), finally implies that the half rotation Cayley vector can be expressed as a function of the full rotation $\boldsymbol{\eta}$ as:

$$\sqrt{\boldsymbol{\eta}} = \frac{2}{2 + \sqrt{4 + |\boldsymbol{\eta}|^2}} \boldsymbol{\eta} \quad . \tag{A.1.51}$$

A.1.2 Homogeneous coordinates and rigid body motions

A position and orientation of a rigid body in three dimensional space can be described using an element of the special Euclidean group $\mathcal{SE}(3)$ (also called the group of rigid body motions). Each element of $\mathcal{SE}(3)$ can be seen as a pair $\left(\boldsymbol{R}, \boldsymbol{r}\right)$ where $\boldsymbol{R} \in \mathcal{SO}(3)$ describes the orientation of a rigid body and $\boldsymbol{r} \in \mathbb{R}^3$ the position of its reference point.

In this section we describe the *homogeneous coordinates* for $\mathbb{R}^3$, introduced by August Ferdinand Möbius in the context of projective geometry [Möb1827], that allow for a very convenient parametrization of $\mathcal{SE}(3)$. In homogeneous coordinates a point $\begin{bmatrix} x & y & z \end{bmatrix} \in \mathbb{R}^3$ is represented as the tuple $\begin{bmatrix} x & y & z & 1 \end{bmatrix} \in \mathbb{R}^4$. It can be verified by direct calculation that in these coordinates the rotation operator $\mathcal{R}$ for the rotation $\boldsymbol{R} \in \mathcal{SO}(3)$ and translation operator $\mathcal{T}$ for a translation by $\boldsymbol{v} \in \mathbb{R}^3$ can be written as the 4×4 matrices:

$$\mathcal{R} := \begin{bmatrix} \boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0}^T & 1 \end{bmatrix}, \qquad \mathcal{T} := \begin{bmatrix} \boldsymbol{I} & \boldsymbol{v} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \tag{A.1.52}$$

where $\boldsymbol{0} \in \mathbb{R}^3$ and I is the 3×3 identity matrix. The inverse operators write:

$$\mathcal{R}^{-1} = \begin{bmatrix} \boldsymbol{R}^T & \boldsymbol{0} \\ \boldsymbol{0}^T & 1 \end{bmatrix}, \qquad \mathcal{T}^{-1} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{v} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \quad . \tag{A.1.53}$$

Any rigid body motion $\mathcal{Q} \in \mathcal{SE}(3)$ of a rotation $\boldsymbol{R} \in \mathcal{SO}(3)$ and a translation by $\boldsymbol{v} \in \mathbb{R}^3$ can be parametrized using a 4×4 matrix as:

$$\mathcal{Q} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{v} \\ \boldsymbol{0}^T & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{v} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \quad . \tag{A.1.54}$$

The group inverse $\mathcal{Q}^{-1}$ can be written as:

$$\mathcal{Q}^{-1} = \begin{bmatrix} \boldsymbol{R}^T & \boldsymbol{v} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{v} \\ \boldsymbol{0}^T & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}^T & -\boldsymbol{R}^T\boldsymbol{v} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \quad . \tag{A.1.55}$$

For more details see *e.g.*. [Kui1999, ch 14].

A.2 Supplementary material for Chapter P1.4

A.2.1 Downloading the software

The C++03 code *cgDNAmc*, along with two libraries it depends upon, *algebra3d* and *cgDNArecon*, is freely available with online instructions on how to download, compile, and run it.[1] The user can supply any desired problem-specific, post-processing code fragments.

A.2.2 DNA sequences

A.2.2.1 λ-phage genome

The sequence S^{λ} consists of base pairs 36901–37200 of the λ-phage genome of Sanger et al. [SanCouHon1982]. It has been chosen as the one with median value of $\ell_p^{[0]}$ among all consecutive fragments of length 300 bp of the genome. The full sequence is available online[2]. A single repeat was used for $\ell_p^{[0]}$ computations, 5 repeats for ℓ_F.

```
TAGAGCGATT TATCTTCTGA ACCAGACTCT TGTCATTTGT TTTGGTAAAG
AGAAAAGTTT TTCCATCGAT TTTATGAATA TACAAATAAT TGGAGCCAAC
CTGCAGGTGA TGATTATCAG CCAGCAGAGA ATTAAGGAAA ACAGACAGGT
TTATTGAGCG CTTATCTTTC CCTTTATTTT TGCTGCGGTA AGTCGCATAA
AAACCATTCT TCATAATTCA ATCCATTTAC TATGTTATGT TCTGAGGGGA
GTGAAAATTC CCCTAATTCG ATGAAGATTC TTGCTCAATT GTTATCAGCT
```

[1] see http://lcvmwww.epfl.ch/cgDNA
[2] http://www.ncbi.nlm.nih.gov/nuccore/215104

A.3 Supplementary material for Chapter P2.3

A.3.1 DNA sequences

A.3.1.1 Kahn and Crothers [KahCro1992] c11t15 (S^{γ})

A 158 bp long sequence designed to be intrinsically bent through the introduction of six phased *A*-tracts, originally used for *in vitro* cyclization experiments.

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

Bibliography

[AndBaiBis1999] E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, Third Edit. Philadelphia: Society for Industrial and Applied Mathematics, 1999.

[BecEve2007] N. B. Becker and R. Everaers, "From rigid base pairs to semi-flexible polymers: Coarse-graining DNA", *Physical Review E*, vol. 76, no. 2, p. 021 923, 2007.

[BedFurKat1995] J. Bednar, P. Furrer, V. Katritch, A. Z. Stasiak, J. Dubochet, and A. Stasiak, "Determination of DNA Persistence Length by Cryo-electron Microscopy. Separation of the Static and Dynamic Contributions to the Apparent Persistence Length of DNA", *Journal of Molecular Biology*, vol. 254, no. 4, pp. 579–594, 1995.

[Ben1977] C. J. Benham, "Elastic model of supercoiling.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 6, pp. 2397–2401, 1977.

[Ben1979] ——, "An elastic model of the large-scale structure of duplex DNA", *Biopolymers*, vol. 18, pp. 609–623, 1979.

[BevBarByu2004] D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham III, S. B. Dixit, E. Giudice, F. Lankaš, R. Lavery, J. H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. M. Thayer, P. Varnai, and M. A. Young, "Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(C(p)G) steps.", *Biophysical Journal*, vol. 87, no. 6, pp. 3799–3813, 2004.

[BugFuj1969] P. Bugl and S. Fujita, "Dynamics of a Long Polymer Backbone", *The Journal of Chemical Physics*, vol. 50, no. 8, p. 3137, 1969.

[CalDreLuiTra2004] C. R. Calladine, H. R. Drew, B. F. Luisi, and A. A. Travers, *Understanding DNA: The Molecule and How it Works*, Third Edit. Elsevier Science, 2004.

[Cay1846] A. Cayley, "Sur quelques propriétés des déterminants gauches.", *Journal für die reine und angewandte Mathematik (Crelles Journal)*, no. 32, pp. 119–123, 1846.

[ChoGorMad2006] N. Chouaieb, A. Goriely, and J. H. Maddocks, "Helices.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 25, pp. 9398–9403, 2006.

[ChoMad2004] N. Chouaieb and J. H. Maddocks, "Kirchhoff's problem of helical equilibria of uniform rods", *Journal of Elasticity*, vol. 77, no. 3, pp. 221–247, 2004.

[CosCos1909] E. Cosserat and F. Cosserat, *Théorie des corps déformables*. Paris: Librairie Scientifique A. Hermann et Fils, 1909.

[CzaSwiOls2006] L. Czapla, D. Swigon, and W. K. Olson, "Sequence-dependent effects in the cyclization of short DNA", *Journal of Chemical Theory and Computation*, vol. 2, no. 3, pp. 685–695, 2006.

[Dem1972] A. Dempster, "Covariance Selection", *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[DicLiMad1996] D. J. Dichmann, Y. Li, and J. H. Maddocks, "Hamiltonian Formulations and Symmetries in Rod Mechanics", in *Mathematical Approaches to Biomolecular Structure and Dynamics*, ser. The IMA Volumes in Mathematics and its Applications, J. P. Mesirov, K. Schulten, and D. W. Sumners, Eds., vol. 82, New York, NY: Springer New York, 1996.

[DixBevCas2005] S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. Cheatham III, E. Giudice, F. Lankaš, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar, K. M. Thayer, and P. Varnai, "Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps", *Biophysical Journal*, vol. 89, no. 6, pp. 3721–3740, 2005.

[DoeChaDer2009] E. J. Doedel, A. R. Champneys, F. Dercole, T. F. Fairgrieve, Y. Kuznetsov, B. Oldeman, R. C. Paffenroth, B. Sandstede, X. Wang, and C. Zhang, "AUTO-07p: Continuation And Bifurcation Software For Ordinary Differential Equations", Tech. Rep., 2009.

[DoeKelKer1991a] E. Doedel, H. B. Keller, and J. P. Kernevez, "Numerical Analysis and Control of Bifurcation Problems (I): Bifurcation in Finite Dimensions", *International Journal of Bifurcation and Chaos*, vol. 01, no. 03, pp. 493–520, 1991.

[DoeKelKer1991b] ——, "Numerical Analysis and Control of Bifurcation Problems (II): Bifurcation in Infinite Dimensions", *International Journal of Bifurcation and Chaos*, vol. 01, no. 04, pp. 745–772, 1991.

[DoiEdw1986] M. Doi and S. F. Edwards, *The theory of polymer dynamics*. Clarendon Press, 1986.

[DubBedFur1994] J. Dubochet, J. Bednar, P. Furrer, A. Z. Stasiak, A. Stasiak, and A. A. Bolshoy, "Determination of the DNA helical repeat by cryo-electron microscopy", *Nature Structural Biology*, vol. 1, no. 6, pp. 361–363, 1994.

[ĐurGorMad2013] B. Đuričković, A. Goriely, and J. H. Maddocks, "Twist and Stretch of Helices Explained via the Kirchhoff-Love Rod Model of Elastic Filaments", *Physical Review Letters*, vol. 111, no. 10, p. 108 103, 2013.

[Eul1776] L. Euler, "Formulae generales pro translatione quacunque corporum rigidorum", *Novi Commentarii academiae scientiarum Petropolitanae*, vol. 20, pp. 189–207, 1776.

[FatEslEjt2012] A. Fathizadeh, B. Eslami-Mossallam, and M. R. Ejtehadi, "Definition of the persistence length in the coarse-grained models of DNA elasticity", *Physical Review E – Statistical, Nonlinear, and Soft Matter Physics*, vol. 86, no. 5, p. 051 907, 2012.

[Flo1973] P. J. Flory, "Moments of the End-to-End Vector of a Chain Molecule, Its Persistence and Distribution", *Proceedings of the National Academy of Sciences*, vol. 70, no. 6, pp. 1819–1823, 1973.

[Flo1969] P. Flory, *Statistical Mechanics of Chain Molecules*. New York: Interscience, 1969.

[FurManMad2000] P. B. Furrer, R. S. Manning, and J. H. Maddocks, "DNA rings with multiple energy minima.", *Biophysical Journal*, vol. 79, no. 1, pp. 116–36, 2000.

[GegVol2010] S. Geggier and A. Vologodskii, "Sequence dependence of DNA bending rigidity.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 35, pp. 15 421–15 426, 2010.

[GolVan1996] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, 1996.

[GonPetMad2013] O. Gonzalez, D. Petkevičiūtė, and J. H. Maddocks, "A sequence-dependent rigid-base model of DNA", *The Journal of Chemical Physics*, vol. 138, no. 5, p. 055 102, 2013.

[GonPetPas] O. Gonzalez, D. Petkevičiūtė, M. Pasi, J. Glowacki, and J. H. Maddocks, "Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA", *in preparation*,

[Gra2016] A. Grandchamp, "On the statistical physics of chains and rods, with application to multi-scale sequence dependent DNA modelling", PhD thesis, EPFL, 2016.

[Hag1988] P. J. Hagerman, "Flexibility of DNA.", *Annual review of biophysics and biophysical chemistry*, vol. 17, pp. 265–286, 1988.

[Ham1844] W. R. Hamilton, "On a new Species of Imaginary Quantities Connected with a Theory of Quaternions", *Proceedings of the Royal Irish Academy*, vol. 2, pp. 424–434, 1844.

[Han2006] A. J. Hanson, *Visualizing Quaternions*. Elsevier Inc., 2006.

[HinFreWhidPab2013] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo, "An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization", *The Journal of Chemical Physics*, vol. 139, no. 14, p. 144 903, 2013.

[HofManMad2003] K. A. Hoffman, R. S. Manning, and J. H. Maddocks, "Link, twist, energy, and the stability of DNA minicircles.", *Biopolymers*, vol. 70, no. 2, pp. 145–57, 2003.

[JohLun1998] C. R. Johnson and M. Lundquist, "Local inversion of matrices with sparse inverses", *Linear Algebra and its Applications*, vol. 277, no. 1-3, pp. 33–39, 1998.

[KahCro1992] J. D. Kahn and D. M. Crothers, "Protein-induced bending and DNA cyclization.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 14, pp. 6343–6347, 1992.

[KehMad2000] S. Kehrbaum and J. H. Maddocks, "Effective properties of elastic rods with high intrinsic twist", in *IMACS World Congress*, vol. 16, 2000.

[KnoRatSchdPab2007] T. A. Knotts, N. Rathore, D. C. Schwartz, and J. J. de Pablo, "A coarse grain model for DNA", *The Journal of Chemical Physics*, vol. 126, no. 8, p. 084 901, 2007.

[KraPor1949] O. Kratky and G. Porod, "Röntgenuntersuchung gelöster Fadenmoleküle", *Recueil des Travaux Chimiques des Pays-Bas*, vol. 68, no. 12, pp. 1106–1122, 1949.

[Kui1999] J. B. Kuipers, *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace and Virtual Reality*. Princeton University Press, 1999.

[KulLei1951] S. Kullback and R. A. Leibler, "On Information and Sufficiency", *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[La Tay2010] A. R. La Spada and J. P. Taylor, "Repeat expansion disease: progress and puzzles in disease pathogenesis", *Nature Reviews Genetics*, vol. 11, no. 4, pp. 247–258, 2010.

[LanGonHef2009] F. Lankaš, O. Gonzalez, L. M. Heffler, G. Stoll, M. Moakher, and J. H. Maddocks, "On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations", *Physical Chemistry Chemical Physics*, vol. 11, no. 45, p. 10 565, 2009.

[Lau1996] S. L. Lauritzen, *Graphical Models*. Clarendon Press, 1996.

[LavMoaMad2009] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkevičiūtė, and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: Curves+", *Nucleic Acids Research*, vol. 37, no. 17, pp. 5917–5929, 2009.

[LavZakBev2010] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. a. Case, T. E. Cheatham III, S. Dixit, B. Jayaram, F. Lankaš, C. Laughton, J. H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer, "A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA", *Nucleic Acids Research*, vol. 38, no. 1, pp. 299–313, 2010.

[LiMad1996] Y. Li and J. H. Maddocks, "On the Computation of Equilibria of Elastic Rods Part I : Integrals , Symmetry and a Hamiltonian Formulation", *preprint*, 1996.

[MacSpaLiwSch2014] M. Maciejczyk, A. Spasic, A. Liwo, and H. A. Scheraga, "DNA Duplex Formation with a Coarse-Grained Model", *Journal of Chemical Theory and Computation*, vol. 10, no. 11, pp. 5020–5035, 2014.

[Mad1987] J. H. Maddocks, "Stability and folds", *Archive for Rational Mechanics and Analysis*, vol. 99, no. 4, pp. 301–328, 1987.

[MadManPaf1997] J. H. Maddocks, R. S. Manning, R. C. Paffenroth, K. A. Rogers, and J. A. Warner, "Interactive Computation, Parameter Continuation, and Visualization", *International Journal of Bifurcation and Chaos*, vol. 7, no. 8, pp. 1699–1715, 1997.

[ManMad1999] R. S. Manning and J. H. Maddocks, "Symmetry breaking and the twisted elastic ring", *Computer Methods in Applied Mechanics and Engineering*, vol. 170, no. 3-4, pp. 313–330, 1999.

[ManMadKah1996] R. S. Manning, J. H. Maddocks, and J. D. Kahn, "A continuum rod model of sequence-dependent DNA structure", *The Journal of Chemical Physics*, vol. 105, no. 13, p. 5626, 1996.

[MarSig1994] J. F. Marko and E. D. Siggia, "Bending and twisting elasticity of DNA", *Macromolecules*, vol. 27, no. 4, pp. 981–988, 1994.

[MarOls1988] R. C. Maroun and W. K. Olson, "Base sequence effects in double-helical DNA. III. Average properties of curved DNA", *Biopolymers*, vol. 27, no. 4, pp. 585–603, 1988.

[Mar2003] G. Marsaglia, "Xorshift RNGs", *Journal of Statistical Software*, vol. 8, no. 14, pp. 1–6, 2003.

[MarTsa2000] G. Marsaglia and W. W. Tsang, "The Ziggurat Method for Generating Random Variables", *Journal of Statistical Software*, vol. 5, no. 8, pp. 1–7, 2000.

[Maz2006] A. K. Mazur, "Evaluation of elastic properties of atomistic DNA models.", *Biophysical Journal*, vol. 91, no. 12, pp. 4507–4518, 2006.

[MetRosRos1953] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, vol. 21, no. 6, p. 1087, 1953. arXiv: 5744249209.

[MoaMad2005] M. Moakher and J. H. Maddocks, "A Double-Strand Elastic Rod Theory", *Archive for Rational Mechanics and Analysis*, vol. 177, no. 1, pp. 53–91, 2005.

[Möb1827] A. F. Möbius, *Der barycentrische Calcül*. Leipzig: Johann Ambrosius Barth Verlag, 1827.

[NoyGol2012] A. Noy and R. Golestanian, "Length Scale Dependence of DNA Mechanical Properties", *Physical Review Letters*, vol. 109, no. 22, p. 228 101, 2012. arXiv: arXiv:1210.7205v1.

[OlsGorLu1998] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes", *Proceedings of the National Academy of Sciences*, vol. 95, no. 19, pp. 11 163–11 168, 1998.

[OlsBanBur2001] W. K. Olson, M. Bansal, S. K. Burley, R. E. Dickerson, M. Gerstein, S. C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C. S. Tung, E. Westhof, C. Wolberger, and H. M. Berman, "A standard reference frame for the description of nucleic acid base-pair geometry.", *Journal of molecular biology*, vol. 313, no. 1, pp. 229–237, 2001.

[PadZelGas2015] J. Padeken, P. Zeller, and S. M. Gasser, "Repeat DNA in genome organization and stability", *Current Opinion in Genetics & Development*, vol. 31, pp. 12–19, 2015.

[Paf1999a] R. C. Paffenroth, "Mathematical Visualisation, Parameter Continuation and Steered Computations", PhD thesis, University of Maryland, 1999.

[Paf1999b] ——, "VBM and MCCC - Packages for Objected Oriented Visualization and Computation of Bifurcation Manifolds", in *Object oriented methods for interoperable scientific and engineering computing*, 1999, pp. 256–265.

[PasMadBev2014] M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. a. Case, T. E. Cheatham III, P. D. Dans, B. Jayaram, F. Lankaš, C. Laughton, J. Mitchell, R. Osman, M. Orozco, A. Perez, D. Petkevičiūtė, N. Spackova, J. Sponer, K. Zakrzewska, and R. Lavery, "μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA", *Nucleic Acids Research*, vol. 42, no. 19, pp. 12 272–12 283, 2014.

[PérLanLuqOro2008] A. Pérez, F. Lankaš, F. J. Luque, and M. Orozco, "Towards a molecular dynamics consensus view of B-DNA flexibility", *Nucleic Acids Research*, vol. 36, no. 7, pp. 2379–2394, 2008.

[PetMah2010] J. P. Peters and L. J. Maher, "DNA curvature and flexibility in vitro and in vivo", *Quarterly Reviews of Biophysics*, vol. 43, no. 01, pp. 23–63, 2010.

[PetPasGonMad2014] D. Petkevičiūtė, M. Pasi, O. Gonzalez, and J. H. Maddocks, "cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA", *Nucleic Acids Research*, vol. 42, no. 20, e153–e153, 2014.

[Pet2012] D. Petkevičiūtė, "A DNA Coarse-Grain Rigid Base Model and Parameter Estimation from Molecular Dynamics Simulations", PhD thesis, EPFL, 2012.

[PreTeuVetFla2007] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd ed. Cambridge University Press, 2007.

[ReyMad2000] S. Rey and J. H. Maddocks, "Buckling of an Elastic Rod with High Intrinsic Twist", in *Proceedings of the 16th IMACS World Congress 2000*, M. Deville and R. Owens, Eds., 2000.

[RitGilKoo2009] M. Rittman, E. Gilroy, H. Koohy, A. Rodger, and A. Richards, "Is DNA a worm-like chain in Couette flow?: In search of persistence length, a critical review", *Science Progress*, vol. 92, no. 2, pp. 163–204, 2009.

[Rod1840] O. Rodrigues, "Des lois géométriques qui régissent les déplacements d'un corps solide dans l'espace, et de la variation des coordonnées provenant de ces déplacements considérés indépendammet des causes qui peuvent les produire", *Journal de mathématiques pures et appliquées*, vol. 5, pp. 380–340, 1840.

[SanCouHon1982] F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen, "Nucleotide sequence of bacteriophage λ DNA", *Journal of Molecular Biology*, vol. 162, no. 4, pp. 729–773, 1982.

[SavPap2010] A. Savelyev and G. A. Papoian, "Chemically accurate coarse graining of double-stranded DNA", *Proceedings of the National Academy of Sciences*, vol. 107, no. 47, pp. 20 340–20 345, 2010.

[SayAvsKab2010] M. Sayar, B. Avsaroglu, and A. Kabakcıoglu, "Twist-writhe partitioning in a coarse-grained DNA minicircle model", *Physical Review E*, vol. 81, no. 4, p. 041 916, 2010. arXiv: 0912.0870.

[Sch1974] J. A. Schellman, "Flexibility of DNA", *Biopolymers*, vol. 13, no. 1, pp. 217–226, 1974.

[SchHar1995] J. A. Schellman and S. C. Harvey, "Static contributions to the persistence length of DNA and dynamic contributions to DNA curvature", *Biophysical Chemistry*, vol. 55, no. 1-2, pp. 95–114, 1995.

[SpeKii1986] T. P. Speed and H. T. Kiiveri, "Gaussian Markov Distributions over Finite Graphs", *The Annals of Statistics*, vol. 14, no. 1, pp. 138–150, 1986.

[StrNgu2004] G. Strang and T. Nguyen, "The Interplay of Ranks of Submatrices", *SIAM Review*, vol. 46, no. 4, pp. 637–646, 2004.

[SulRomOul2012] P. Sulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis, "Sequence-dependent thermodynamics of a coarse-grained DNA model", *The Journal of Chemical Physics*, vol. 137, no. 13, p. 135 101, 2012. arXiv: arXiv:1207.3391v1.

[ThéCouLe Rév1988] B. Théveny, D. Coulaud, M. Le Bret, and B. Révet, "Local Variations of Curvature and Flexibility Along DNA Molecules Analyzed from Electron Micrographs", in *Structure and expression: proceedings of the Fifth Conversation in the Discipline Biomolecular Stereodynamics held at the State University of New York at Albany, June 2-6, 1987, Volume 3*, W. K. Olson, M. H. Sarma, R. H. Sarma, and M. Sundaralingham, Eds., Schenectady: Adenine Press, 1988.

[VirBerHen2004] J. Virstedt, T. Berge, R. M. Henderson, M. J. Waring, and A. a. Travers, "The influence of DNA stiffness upon nucleosome formation", *Journal of Structural Biology*, vol. 148, no. 1, pp. 66–85, 2004.

[VolVol2002] M. Vologodskaia and A. Vologodskii, "Contribution of the intrinsic curvature to measured DNA persistence length", *Journal of Molecular Biology*, vol. 317, no. 2, pp. 205–213, 2002.

[WalGonMad2010] J. Walter, O. Gonzalez, and J. H. Maddocks, "On the Stochastic Modeling of Rigid Body Systems with Application to Polymer Dynamics", *Multiscale Modeling & Simulation*, vol. 8, no. 3, pp. 1018–1053, 2010.

[Yam1976] H. Yamakawa, "Statistical mechanics of helical wormlike chains. I. Differential equations and moments", *The Journal of Chemical Physics*, vol. 64, no. 12, p. 5222, 1976.

[Yam1997] ——, *Helical Wormlike Chains in Polymer Solutions*. Berlin, Heidelberg: Springer, 1997.

Curriculum Vitæ

I was born on the 28th of April 1986 in Zabrze, Poland, but received all my undergraduate education in Krakow. I obtained a Master Degree in Computer Science at the AGH University of Science and Technology in Krakow in 2010. In the course of my undergraduate studies I had the opportunity to spend 4 months in 2008 at the Rutherford Appleton Laboratory where I worked on data management software for high power lasers of the Central Laser Facility. In the following year I did a 3 month internship at the Shrivenham Campus of the Defence Academy of the United Kingdom were I worked on a component of a military training software. In September of 2010 I started my PhD research project with Prof. John H. Maddocks at EPFL.