# A sequence-dependent coarse-grain model of B-DNA with explicit description of bases and phosphate groups parametrised from large scale Molecular Dynamics simulations

**Thèse N° 9552**

**Présentée le 23 août 2019**

**à la Faculté des sciences de base**
**Chaire d'analyse appliquée**
**Programme doctoral en mathématiques**

**pour l'obtention du grade de Docteur ès Sciences**

**par**

**Alessandro Samuele PATELLI**

**Acceptée sur proposition du jury**

**Prof. T. Mountford, président du jury**
**Prof. J. Maddocks, directeur de thèse**
**Prof. C. Benham, rapporteur**
**Dr B. Eslami , rapporteur**
**Dr S. Bonella, rapporteuse**

**2019**

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

[…]
chi sei tu che turbi le mie notti
e invadi la mia mente
che impedisce alla notte
di avvolgermi nel suo manto
di stelle
che impedisci alla mia mente
di raccogliersi in silenzio
ad ascoltare il creato
chi sei tu
[…]
— Miranda Patelli

# Acknowledgements

I would like to thank Prof. J.H. Maddocks for giving me the opportunity to work in his group, for teaching me the fine art of DNA modelling, and for all the jokes and funny anecdotes. I will really treasure them. I also want to thank the jury members: Prof T. Mountford, Dr S. Bonella, Prof. C. Benham, and Dr. B. Eslami for the useful comments and discussions that have arisen during the exam. Many thanks go to Prof. O. Gonzalez for the big help in getting the first cgDNA+ parameter set and to Daiva for teaching me all the detail about MD simulations. I am very grateful to Carine Tschanz for her kindness, and for having managed any (scary) administrative procedure in a quick and professional way. Thank to all the LCVMM members: Pauline, Alex, Jarek, Thomas L., Thomas Z., Lennart, Alastair, and Philippe, for all the scientific discussions and all the pleasant moments. There would have been no data to modelling without the help of SCITAS and in particular without the help of Gilles.
In the past 11 years I had the luck to meet a lot of friends that helped me through this journey. In particular I would like to thank my faithful comrades: Adrien, Christoph, Jacques, David, Arnaud, Loic and Jim for all the unforgettable moments we spend together, for the uncountable laughs we had, and for all the "assiette de la soir" we shared. My liver and myself would like to bless the FermentaTori: Giacomo, Dave, and Riccardo, for helping me wash away the stress of the life with ONE beer (at a time). I am grateful for your friendship.
Ringrazio i miei genitori per il sostengo e l'amore incondizionato che mi ha permesso di portare a termine questo lungo percorso iniziato molti anni fa. Se sono l'uomo (affascinante) che sono lo devo solo a voi. Vi dedico questo lavoro.
Grazie a mia sorella Barbara per essere stata il mio punto di riferimento qui a Losanna sin dal giorno uno. Grazie alla piccola Elisa per aver portato ancora più gioia nella famiglia. Grazie anche a mio fratello Gabriele per la sua amicizia, il suo genio e la sua sregolatezza. I momenti passati insieme a voi sono unici e preziosi.
Infine vorrei ringraziare la mia Ziollina per essere sempre stata presente con il suo sorriso ed il suo affetto e per essersi presa cura di me nei giorni di tristezza e di frustrazione. Potermi svegliare ogni giorno accanto a te è il mio più grande regalo.
This work was supported by the Swiss National Science Foundation, project numbers 200020-143613 and 200020-163324.

*Lausanne, 16 May 2019* Alessandro S. Patelli.

# Abstract

We introduce a sequence–dependent coarse–grain model of double–stranded DNA with an explicit description of both the bases and the phosphate groups as interacting rigid–bodies. The model parameters are trained on extensive, state–of–the–art large scale molecular dynamics (MD) simulations. The model paradigm relies on three main approximations: 1) nucleic acid bases and phosphate groups are rigid, 2) interactions are nearest–neighbour and can be modelled with a quadratic energy, 3) model parameters have dimer sequence dependence. For an arbitrary sequence, the model predicts a sequence–dependent Gaussian equilibrium probability distribution. The parameter set comprises dimer–based elements, which are used to reconstruct mean configurations, called ground–states, which can have strong non–local sequence dependence, and precision matrices, or stiffness matrices, for any sequence of any length. This prediction step is sufficiently efficient that it is straightforward to construct probability density functions for millions of fragments each of length a few hundred base–pairs. The estimation of a parameter set consists in minimising the sum of Kullback–Leibler divergences between Gaussians predicted by the model and analogous Gaussians estimated directly from MD simulations of a training library of sequences. The training library comprises a short list of short palindromic DNA sequences. We designed the palindromic library using an ad hoc algorithm to include multiple instances of all independent tetramer sub–sequences. We exploit palindromic symmetry properties to study the convergence of the statistics extracted from MD simulations of palindromes and to define palindromically symmetrised estimators of first and second centred moments. The computation of the parameter set is delicate and needs the use of sophisticated numerics. We present an efficient and reliable procedure for estimating a complete parameter set which involves a generalisation of the classic Fisher information matrix and its relationship to the relative entropy, or Kullback–Leibler divergence. The model is a computationally efficient tool that allows the study of the mechanical properties of double–stranded DNA of arbitrary length and sequence. We use the model to study the sequence–dependent rigidity of DNA and we compute sequence–dependent apparent and dynamic persistence lengths. The explicit treatment of the phosphate group also allows computation of sequence–dependent grooves widths. Moreover, with fine–grained representation of predicted ground–states, we can also study sequence–dependence of sugar puckering modes and BI–BII backbone conformations.

**Keywords:** coarse–grain DNA, phosphate group, palindromic sequence, molecular dynamics simulation, Kullback–Leibler divergence, Fisher information, parameter set estimation, sugar puckering, BI–BII conformations, groove width.

# Résumé

Dans cette thèse, nous introduisons un modèle gros–grains d'ADN à deux brins qui dépend de la séquence, avec une description explicite des bases et des groupes phosphate en tant que corps rigides en interaction. Les paramètres du modèle sont déterminés par des simulations approfondies de dynamique moléculaire (DM) à large échelle. Le paradigme du modèle repose sur trois approximations principales : 1) les bases d'acide nucléique et les groupes phosphate sont rigides, 2) les interactions sont locales et de type "plus proche voisin", et peuvent être modélisées par une énergie quadratique, 3) les paramètres du modèle ont une dépendance au niveau des dimères de la séquence. Pour une séquence donnée, le modèle prédit une densité de probabilité à l'équilibre gaussienne qui dépend de cette séquence. L'ensemble de paramètres du modèle inclut des éléments liés aux dimères, qui sont utilisés pour reconstituer, pour des séquences de longueur arbitraire, d'une part des configurations moyennes, appelées états de base, et qui peuvent avoir une forte dépendance non locale au sein de la séquence, et d'autre part des matrices de précision, ou matrices de rigidité. Cette étape de prédiction est suffisamment efficace pour qu'il soit aisé de construire des fonctions de densité de probabilité pour des millions de fragments qui soient chacun d'une longueur de quelques centaines de paires de bases. L'estimation de l'ensemble des paramètres consiste en une minimisation de la somme des divergences de Kullback-Leibler entre des gaussiennes prédites par le modèle et des gaussiennes obtenues directement par des simulations de DM sur une librairie de séquences. La librairie comprend une brève liste de courtes séquences palindromiques d'ADN. La librairie palindromique est générée en utilisant un algorithme ad-hoc afin d'inclure de multiples exemples de toutes les sous-séquences indépendantes de tetramères. On exploite les propriétés de symétrie palindromique pour étudier la convergence des estimateurs statistiques extraits des simulations de DM des palindromes et pour définir des estimateurs palindromiquement symétrisés des premier et des second moments centrés. Le calcul de l'ensemble des paramètres est délicat et nécessite l'utilisation de techniques sophistiquées d'analyse numérique. Nous présentons une façon efficace et fiable pour estimer un ensemble complet de paramètres, qui fait intervenir une généralisation de la matrice classique d'information de Fisher ainsi que sa relation avec l'entropie relative, ou divergence de Kullback-Leibler. Le modèle est un outil, computationnellement robuste, qui permet l'étude des propriétés mécaniques d'un double brin d'ADN de longueur et de séquence arbitraire. Ce modèle est utilisé pour

étudier la rigidité en fonction de la séquence d'ADN ; en particulier, nous calculons les longueurs de persistance apparente et dynamique, dépendant de la séquence. Le traitement explicite des groupes phosphate permet aussi un calcul de la largeur des sillons, une quantité qui dépend de la séquence. De plus, la représentation à grains fins des configurations moyennes prédites par le modèle permet une étude des différents "puckerings" du sucre et des conformations du squelette BI-BII.

# Riassunto

In questa tesi introduciamo un modello “coarse-grain” della doppia elica del DNA con dipendenza a livello della sequenza e con descrizione esplicita delle basi azotate e dei gruppi fosfati. I parametri del modello sono determinati a partire da delle simulazioni di dinamica molecolare. Il paradigma del modello si base su tre ipotesi principali: 1) le basi azotate e i gruppi fosfati sono unità rigide 2) le interazioni fisiche sono locali e di tipo “più vicine” 3) i parametri del modello hanno dipendenza dalla sequenza a livello dei dimeri. Per una sequenza data, il modello predice una funzione di densità di probabilità di Gauss. L’insieme dei parametri che sono utilizzati per ricostruire il vettore della media, detto anche stato fondamentale, e la inversa della matrice di covarianza, detta anche matrice di rigidezza. Lo stato fondamentale dipende non localmente rispetto alla sequenza, mentre la matrice di rigidezza ha una dipendenza locale. La ricostruzione dei parametri della distribuzione di Gauss è sufficientemente efficace, permettendo di ricostruire milioni di funzioni di distribuzione di probabilità per sequenze di lunghezza di alcune centinaia di basi. La stima dei parametri del modello consiste nel minimizzare la somma di divergenze Kullback–Leibler tra gaussiane predette dal modello e gaussiane empiriche. Quest’ultime sono ottenute da simulazione di dinamica molecolare di una libreria di sequenze. Una libraria consiste in una lista di corte sequenze palindromiche. La libreria palindromica è stata generata da un algoritmo appositamente sviluppato per includere tutte le sotto sequenze di lunghezza pari a quattro paia di basi. La simmetria palindromica è in seguito usata per studiare la convergenza degli stimatori associati alle simulazione di dinamica molecolare di queste sequenze. Ottenere l’insieme dei parametri del modello è un processo delicato che necessita l’utilizzo di metodi numerici sofisticati. In questo lavoro presentiamo un metodo efficace e robusto che utilizza una generalizzazione della matrice d’informazione di Fisher e la sua relazione con la divergenza di Kullback–Leibler. Il modello è uno strumento computazionalmente efficiente che permette lo studio delle proprietà meccaniche di filamenti di DNA di lunghezza e sequenza arbitrarie. Quest’ultimo può essere impiegato per lo studio della rigidità del DNA (sequenza–dipendente) ed in particolare per il calcolo della lunghezza persistente apparente e dinamica (anch’esse sequenza–dipendenti). Il trattamento esplicito dei gruppi fosfati permette, per sequenze arbitrarie, il calcolo del loro solco maggiore e minore. Inoltre, per ogni configurazione media predetta dal modello, è possibile derivare una rappresentazione atomistica della molecola per studiare il “puckering”

degli zuccheri pentosi e le conformazioni dello scheletro fosfato–deossiribosio, dette BI–BII.

# Introduction

Deoxyribonucleic acid, or DNA, is the molecule that is responsible for much of the functioning of the cell of any living organism. DNA is a molecule consisting of two chains, called strands, which are attached one to another via hydrogen bonding between the nucleic acid bases. The interaction between the two strands leads to the typical double helix. Each backbone is composed of alternating phosphate groups and sugar rings with attached nucleic bases. There are four standard nucleic acid bases Adenine, Thymine, Guanine, and Cytosine, respectively abbreviated by A,T,G, and C. In the double helix the two strands interact at the level of the base–pair, and the standard Crick–Watson pairing rule states that A and T always pair, and G and C always pair. A significant feature of DNA is that its shape has been observed to have a strong sequence–dependence. For example, it has been noticed that some DNA sequences containing a specific succession of bases have an intrinsic bend [76, 43, 22]. A particular example showing the relation between sequence and shape are phased runs of three to six Adenine nucleic acid base, or A–tracts, repeated with a helical periodicity which leads to a significant global curvature of the double helix molecule [22]. A second significant physical property of the DNA molecule is its local rigidity, or stiffness, that also has been shown to have a complex sequence–dependent behaviour [51, 71]. A combination of both intrinsic shape and local stiffness properties, characterises the overall deformations or fluctuations of the DNA which consequently show a considerable variation between different sequences. In particular, it remains an interesting, yet non trivial, problem to fully quantify these properties.
From an experimental point the rigidity of naked DNA is estimated in different ways. Here we mention two classic approaches: cyclisation experiments, for fairly short fragments, and single molecule tweezers, for longer fragments. The first experimental method consists in the quantification of the probability of closed loop formation starting from multiple copies of a linear piece for the same DNA with cohesive, or sticky, ends. In the second technique a bead is attached to an end of a single linear molecule of DNA by magnetic or optical tweezers or a micropipette then the molecule is pulled and twisted. However, for both experiments, a mechanical model is needed to rationalise the outcomes and to use of the results as a prediction tools.
Modelling the mechanical properties of DNA is strongly related to the length scales of interest. Different length scales lead to different models in terms of the number of model parameters, in terms of data to be fitted, and in terms of target applications of

the model. For long length scales, the most widely used models have a comparatively low number of parameters and just model the DNA as a single uniform sequence–independent rod or wormlike chain, see for example [31]. At shorter length scales a full atomistic approach leads to a very detailed mechanical model of DNA, which can be used in the study of local properties of the molecule. The main drawback of the full atomistic model is its high computational cost due to the large numbers of degrees of freedom and associated parameters involved in the model particularly if the solvent is treated explicitly. which is often thought to be necessary to capture the detailed electrostatics of DNA. A third method for modelling the mechanical property of the DNA is based on coarse–graining fully atomistic models by introducing additional assumptions that control the level of detail present in the model. The rigid–base–pair model [51, 78] identifies each base–pair as a rigid unit resulting in a single, rigid–body chain representation of the DNA molecule. The sequence dependence is typically taken into account at the dimer sequence level and the resulting model is local in both shape and stiffness. On the other hand, a rigid–base [55, 19] model identifies each base as a rigid unit, which consequently leads to a more detailed model, and the DNA structure is represented as the interactions between two single chains, one for each strand.
The cgDNA rigid–base model presented in [55, 19] is the starting point for this work. The cgDNA model is a coarse–grain sequence–dependent rigid–base model of DNA in solution. For an arbitrary DNA sequence composed in the standard alphabet $\{A, T, G, C\}$, the cgDNA model predicts a Gaussian stationary or equilibrium distribution for the underlying dynamics by reconstructing a mean configuration, or configuration of the minimal energy state, and a stiffness matrix. One of the main properties of the model is that an arbitrary DNA fragment has an intrinsic shape that has non–local sequence–dependence as a consequence of the fact that each base cannot minimize the interaction energy between all of its neighbours. The latter phenomenon is called frustration, and is an implication of the specific banded, but not block diagonal, pattern of the stiffness matrix. The banded sparsity pattern of the stiffness matrix corresponds to each base interacting with five nearest neighbours. Consequently the configuration of a minimal state of a given sequence exhibits a non–local behaviour under single letter mutation in the sequence, although the stiffness matrix changes locally. cgDNA can be used for model–based analysis of different features of the DNA. For example in [45], the authors exploited the cgDNA predicted Gaussian to compute the apparent and dynamic persistence lengths for a large number of sequences, and in particular, they introduced and computed sequence–averaged apparent and dynamic persistence lengths.

This work is divided into four parts. Part I is dedicated to background material. In Part II we estimate and compare different cgDNA parameter sets extracted from different MD protocols and introduce some enhancements to the cgDNA model. Part III is the central contribution of the thesis. It presents the cgDNA+ model, which is a refinement

of cgDNA that introduces additional degrees of freedom associated with an explicit description of the configurations of the phosphate groups. Part IV contains some illustrative applications of the cgDNA+ model.

Chapter 1 has as its primary goal the presentation of all the basic notions that are useful for the complete understanding of the cgDNA and cgDNA+ models. We start with the standard chemical detail about DNA. The text presents the primary, secondary, and tertiary structures of DNA by giving the main definitions and notation that are used throughout this work. In addition to the classic DNA notions associated to bases, the torsional angles and sugar ring puckering will be important in the applications of the cgDNA+ model, while ideal bases will be the starting point for the coarse–grain process of modelling the full atomistic representation of the DNA.
Chapter 2 presents the core mathematical ideas that will be useful in the development of our coarse–grain models. Section 2.1 presents the properties of the Lie groups SO(3) and SE(3) which are the main geometrical objects involved in the modelling procedure. In particular, these Lie matrix groups are important for the formalisation of the bichain interpretation of DNA, which leads to a set of internal coordinates leading to a tractable and apparently rather accurate quadratic energy in our model.
Chapter 3 briefly introduces the state–of–the–art molecular dynamics simulations that we have used to train our coarse–grain models. In particular we present the computational workflow that we implement for post-processing the large scale trajectory data set produced by the full atomistic computations. From time series of atom positions, we extract a time series of bichain internal coordinates. First and second (centred) moments are then computed using standard estimators. We also present an algorithm, as developed in [18, 20], for estimating a stiffness, or precision, matrix with a prescribed sparsity pattern from a full covariance matrix with dense inverse.
Chapter 4 presents the cgDNA model [55, 19]: from the main assumptions underpinning the coarse–grain model to its applications to the study of sequence–dependent persistence lengths. The importance of this chapter to this thesis relies on the multiple concepts and mathematical notions that form the core of the dogma of the cgDNA coarse–graining methodology, and which will be the starting point for the development of the enhanced cgDNA+ model.
The last chapter of the background part, Chapter 5, is entirely dedicated to the estimation of the cgDNA parameter sets from MD data. In this chapter, we recall the Kullback–Leibler divergence and present its properties. Then we introduce the detail of cgDNA parameter set extraction procedures along with the definition of a positive–definite best–fit parameter set.

Part II of the thesis is dedicated to the comparison of different best–fit cgDNA parameter sets computed for various different MD simulations protocols, which in this work are all modifications of the ABC protocol [8, 15].
In Chapter 6 we compare cgDNA parameter sets extracted from MD data sets based

on protocols with different simulation durations, sequence libraries, force fields, and ion types and concentrations. An elementary one–at–a–time sensitivity analysis is then performed to compare the coarse–grain best–fit parameter sets. We present a parameter continuation algorithm that we use to compute all the different cgDNA parameter sets based on different MD training data. An average per degree of freedom Kullback–Leibler divergence is used to quantify the differences between the various cgDNA parameter sets and the MD data set used in the fitting procedure. Moreover, persistence lengths are computed for all the parameter sets to study the impact of the MD protocol on the overall rigidity of the DNA.
In Chapter 7 we focus our attention on the design of MD training libraries. In particular we present an algorithm that we have developed which searches for a library composed of only palindromic sequences. In particular, the sequence library, which we call the palindromic library, is composed of 16 sequences each of length 24 base–pairs. The algorithm searches for a palindromic library which contains at least two copies of all non–palindromic independent tetramers. We recall that a palindrome has the property that the sequence on the reading strand matches the complementary sequence when both strands are read in the $5' \to 3'$ direction. In the context of the prediction of cgDNA ground–states, we have a simple linear relation between the internal coordinates of a sequence read on one strand and the internal coordinates read from its complementary strand. In particular, for a palindromic sequence, the latter relation states that the internal coordinates of a ground–state of a palindrome is invariant under a change of reading strand. This property leads to the idea of quantifying the lack of convergence in the MD simulations of a palindromic sequence by computing the error between the mean estimator and its palindromically symmetrized version. We then apply the same idea to compute the convergence error for the estimated covariance matrix. With this approach, we have a quantifiable way of testing the quality of the training library data, and also a way of estimating the palindromic symmetrised first and (centred) second moment that will be used in the parameter set extraction. We complete Chapter 7 by introducing a new cgDNA parameter set format which has dimer dependent blocks for the ten independent interior blocks, and sixteen end dimer blocks. The new format of the parameter set leads to a non uniqueness of the best–fit parameter set, although the reconstructions of the mean and the stiffness matrix of any sequence remain unique. Having additional dedicated dimer–dependent end blocks seems to lead to a significant improvement in the accuracy of the cgDNA model.

The central part of this work is Part III where we introduce the coarse–grain model we call cgDNA+ which add an explicit treatment of the phosphate groups. The modelling dogma behind cgDNA+ is similar to the one of cgDNA. In fact, in the cgDNA+ model we add extra degrees of freedom based on the assumption of rigidity of the phosphate groups atoms.
In Chapter 8 we introduce the mathematical background underlying the cgDNA+ model. More precisely, we generalize the concept of double chains, and the bichain

representation introduced in [21] that are pertinent to the cgDNA model to double interacting strands and tetrachains. To treat the phosphate groups explicitly, we consider an extra set of six degrees of freedom between each base and its associated phosphate group. We give a general definition of internal coordinates for the tetrachains, and derive the Crick–Watson symmetry transformation describing the relation between the internal coordinates of a sequence and the internal coordinates read with respect to its complementary strand. Then we introduce a nearest–neighbour energy for the tetrachain model and compute its first variation. Consequently, we can compute the expression for the total external load acting on a single phosphate group necessary to hold it in equilibrium in any configuration. We give expressions for the variations in both coordinates free and coordinate specific cases.
Chapter 9 contains all the detail about the cgDNA+ model. The internal coordinates are defined by base–to–phosphate degrees of freedom in the relatively rigid–body motion from the base to its $5'$ phosphate group. Once the definition of internal coordinates is chosen we investigate, using the palindromic MD data set, the sparsity pattern of the observed stiffness matrices. Remarkably a block structure appears which leads to the definition of the assumed cgDNA+ model sparsity pattern. We then compute the palindromic error in the mean and covariance estimators for the cgDNA+ internal coordinates and the palindromic data set in order to understand the convergence rate of the base–to–phosphate degrees of freedom. After introducing all the model assumptions, we define the cgDNA+ parameter set format which contains only dimer–dependent stiffness matrices and sigma vectors. As already done for the modified cgDNA model described in Chapter 7, we allow dimer specific blocks for the ten independent interior dimers and sixteen additional for the end ones. In the case of cgDNA+ this is necessary because the end blocks have a different dimension to the interior blocks because there is no $5'$ phosphates associated to the bases forming the first and last base pairs. We then show how to compute a first cgDNA+ parameter set trained on the palindromic data set. This step required the introduction of a new computational approach due to the large number of parameters to be estimated. To that end, we introduce the Fisher information matrix and its relationship with the second derivatives of Kullback–Leibler divergence. We then show how to take advantage of the relation between Fisher information and Kullback–Leibler divergence to compute a good initial guess for the fitting optimization problem. In collaboration with O. Gonzalez, we introduce a Fisher–informed gradient flow which shows very good performance in numerically solving the fitting problem. Once the first cgDNA+ parameter set has been computed, we show how to prove that the best–fit parameter set is, in fact, positive definite, meaning that for any arbitrary sequence the predicted stiffness matrix is positive–definite. This exercise is not trivial as the format of the parameter set leads to a non–injective reconstruction scheme due to a freedom in the overlaps. Thus, the best–fit parameter set is not unique and thanks to this feature we can take advantage of the null–space in order to prove the positiveness of a cgDNA+ parameter set. We continue by illustrating the performance of the best–fit cgDNA+

parameter set in approximating the observed ground–states and the tangent–tangent correlations for the Palindromic data set. Finally, we discuss the number of degrees of freedom introduced in the cgDNA+ model compared to the improvement in the level of approximation of the observed data. In particular, we use the Akaike information criterion to quantify the actual gains consequent upon introducing the extra degree of freedoms in cgDNA+. Moreover, we discuss a possible further extension of the model by allowing additional blocks in the imposed sparsity pattern of the stiffness matrix corresponding to local interactions beyond nearest neighbour.

Part IV is dedicated to four applications of the cgDNA+ model, and in particular to applications which are only possible thanks to the explicit treatment of the phosphate groups.
In Chapter 10, we start by computing the spectra of persistence lengths (apparent and dynamic). In Chapter 11, we study packing forces computed for two different crystal structures of the Drew–Dickerson dodecamer sequence by considering only the external load acting on the phosphate group. Chapter 12 is all about the backbone and in particular sugar ring puckering. In fact, from a cgDNA+ predicted ground–state we can embed the ideal atoms for each base and each phosphate group. Then, we can recompute the position of the sugar ring atoms which are not explicitly considered in the cgDNA+ degrees of freedom. Consequently, we can compute the torsional endocyclic angles of the sugar ring and categorise its conformation using the value of a pseudorotation phase angle. Then we can compute the backbone torsional angles (conventionality named $\varepsilon$ and $\zeta$) which can be used to compute the conformation of the backbone as being call BI or BII. We can then study the repartition of BI and BII backbone configurations for each dimer according to all possible sequence contexts. The final application in Chapter 13 is about a groove width computation. In particular, we show how to compute and identify the major and minor grooves width for an arbitrary sequence by mimicking the methodology proposed in [38] but now within the cgDNA+ coarse–grain model, which allows many possible sequence to be considered. In particular we can study the sequence context dependence of both grooves using a simplified yet faster method for detecting major and minor grooves widths.

The thesis is closed with a discussion of our conclusions, and outlook for further model development and applications

# Contents

# List of Figures

# List of Tables

# Part I

# Background

# 1 Introduction to DNA

In this chapter we introduce basic facts about right–handed B–form DNA. In particular we cover the basic aspects of the chemical structure of the DNA that will be useful for our modelling, as well as the main features of the double helix such as the grooves.
A DNA molecule has four *nucleic bases*, two purines A and G and two pyrimidines T and C. The purines, or simply R, are bigger (two very rigid rings), while the pyrimidines, or Y, are smaller (one rigid ring). In standard DNA, A pairs with T with two hydrogen bonds, while C pairs with G with three hydrogen bonds, in both cases forming a *base–pair*. In this work we will only consider these four base types along with the aforementioned pairing, also called *Crick–Watson pairing*. The double helix of the DNA molecule is composed of two interacting anti–parallel strands. Each strand is itself composed of repetitions of a unit called a *nucleotide* formed by a base, a sugar ring, and a phosphate group. The chain formed by the repeated pattern of a sugar ring and a phosphate group is called a *backbone*. It has a specific direction, given by the sugar group, called the $5' \to 3'$ direction. A DNA molecule is associated to a list of bases $X \in \{A, T, G, C\}$ called a *sequence*, denoted here by $\mathcal{S} = X_1 X_2 \cdots X_N$, where $N \in \mathbb{N}$ is the length of sequence counted in number of base–pairs. The sequence is actually the list of bases composing one of the two strands, written in the $5' \to 3'$ direction. That strand is called the *reading strand* while its anti–parallel is called the *complementary strand*. In this work we also refer to the reading strand as the *Watson strand* and to the complementary strand as the *Crick strand*. The sequence of the complementary strand is denoted by $\overline{\mathcal{S}} = \overline{X}_1 \overline{X}_2 \cdots \overline{X}_N$ where $\overline{X}_n$ is the Crick–Watson complement of $X_{N-n+1}$. Sequences that satisfy $\mathcal{S} = \overline{\mathcal{S}}$ are called *palindromes* and will play an important role in this work.
In an idealized, first, approximation B–form DNA is a uniform double helical structure with a straight centreline. The double helix has one full turn every 10.5 base pairs or so, i.e. the DNA double helix as an high intrinsic twist. Moreover, the average distance between two consecutive base–pair is about 3.4 Å. The distance between the two backbone is not constant and forms two distinct regions called the *minor groove* and the *major groove* . The DNA grooves could play an important role in the readout process of the sequence from the proteins. Now, the B–form DNA is far from being

a simple uniform double helical structure, in fact the different stacking interactions between purines and pyrimidines, and because of the different numbers of hydrogen bonds in the Crick–Watson pairing, the sequence modulates both the intrinsic shape, and local rigidity of the molecule in a biologically significant way.
Recapping we have presented three nucleic acid structures of the DNA:

- **Primary structure**: The single stranded linear chain of a string or word in the alphabet $\{A, T, G, C\}$ nucleotides represented by a DNA sequence
- **Secondary structure**: Crick–Watson pairing defining the interactions between bases on different strands. In case of B–form DNA two strands with an overall structure is a double helix.
- **Tertiary structure**: The physical properties of the double helix such as its intrinsic shape and its rigidity as a function of the sequence.

There is also a quaternary structure that refers to DNA–protein complexes and their interactions. In this work the nucleic acid tertiary structure is of main interest and its study and understanding is our major goal.
In the next sections we focus on two further different notions related to the backbone and its chemical structure that will be of importance.

## 1.1 Torsional angles

The backbone is a chain composed by two main units repeated in an alternating way. The first unit is the sugar ring and the second is the phosphate group. In figure 1.1 we show a schematic representation of part of a backbone composed by a sugar ring and two phosphate groups. As a first remark one can notice that the sugar ring is separated from the phosphate group. On the right–hand side, by a single covalent bond while it is connected by two bonds to the left hand side phosphate group. This asymmetry identifies the $5' \to 3'$ orientation using the schematic representation. In the figure we label only the oxygen atoms *O* and the phosphorous atoms *P* while the non–labelled atoms are carbons *C* and hydrogen are not shown at all. Now starting from the oxygen atom we count the carbon atoms clockwise. In this manner the third and the fifth carbon atoms define the $5' \to 3'$ direction. We have that, with respect to the sugar ring, the left-hand side phosphate group is the $5'$ phosphate group while the right–hand side one is the $3'$ phosphate group. In figure 1.1 we show also six angles called the *backbone torsion angles* which are related to four consecutive covalently bonded atoms. In general, let $\{A, B, C, D\}$ be a group of four atoms, the torsion angle $\theta$ associated to the group of atoms is define as the angle between the plane passing through $\{A, B, C\}$ and the plane passing through $\{B, C, D\}$. In table 1.1 for each torsion angle we report the associated group of atoms. For sake of completeness in table 1.1 we report also

| Torsion angle | Group of atoms |
|---|---|
| $\alpha$ | $O'_3 - P - O'_5 - C'_5$ |
| $\beta$ | $P - O'_5 - C'_5 - C'_4$ |
| $\gamma$ | $O'_5 - C'_5 - C'_4 - C'_3$ |
| $\delta$ | $C'_5 - C'_4 - C'_3 - O'_3$ |
| $\varepsilon$ | $C'_4 - C'_3 - O'_3 - P$ |
| $\zeta$ | $C'_3 - O'_3 - P - O'_5$ |
| $\chi$ | $O'_4 - C'_1 - N_{1/\mathbf{9}} - C_{2/\mathbf{4}}$ |

Table 1.1 – Torsional angle and related group of atoms. For the torsion angle $\chi$ two choices are possible for the third and fourth atoms depending on whether the base is R or Y. In italics we report the atom number in case of a base Y and in bold in case of a base R.

the torsion angle $\chi$ which is related to the relative orientation between a sugar group and its base. As the chemical structure of the base changes between pyrimidine and purine, the $\chi$ torsion angle is defined for two different atoms groups, both reported in table 1.1.
The $\varepsilon$ and $\zeta$ torsional angles have been associated to the so called BI–BII junction conformations which are characterised by two distinct positions of the phosphate group [23]. The value of the difference $\varepsilon - \zeta$ identifies the conformation of the junction. More precisely if the difference is negative the junction is in the BI conformation, which on the contrary, if the difference is positive the junction is in the BII conformation.

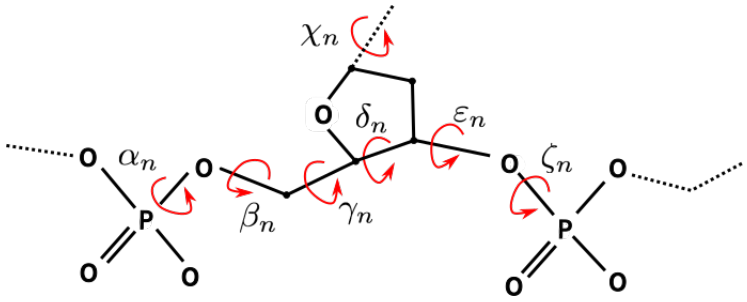



Figure 1.1 – Torsional angle of the backbone

## 1.2 Sugar ring puckering

We will also need more specific features of the sugar group. One of the main properties of the sugar ring is that, due to its chemical structure, its configuration cannot be planar. In fact, the spatial conformations of the sugar ring are of two kinds called *envelope* and *twist*. The envelope configuration is characterised by four atoms being planar and one being out of plane, while the twist conformation is associated to three atoms being planar and two being out of plane one opposite to the other. In figure 1.2 we show a schematic representation of an envelope and a twist configuration. In

general we refer to sugar ring *puckering* any sugar ring conformation.

The sugar ring puckering can be completely characterise by five *endocycling tor-*


Figure 1.2 – Three possible configuration of the sugar ring: the planar which never occurs, the envelope, and the twist. The red dots do not lay in the same plane formed by the black dots.

*sion angles* normally called $v_n,\ n = 0,\ldots,4$. The atoms forming the sugar ring are $\{O'_4, C'_1, C'_2, C'_3, C'_4\}$ and the relation to the torsion angles are reported in table 1.2. Using the torsion angle we can compute a *pseudorotation* parameter that can be used to infer the sugar puckering mode: the pseudorotation phase angle $P$. There are at least two slightly different ways for computing these pseudorotation parameters:

$$\tan P = \frac{\sum_{i=1}^{5} -\theta_i \sin(\frac{4}{5}\pi(i-1))}{\sum_{i=1}^{5} \theta_i \cos(\frac{4}{5}\pi(i-1))} \text{ [3]}, \tag{1.1}$$

$$\tan P = \frac{(v_4 + v_1) - (v_3 + v_0)}{2v_2\left(\sin\left(\frac{1}{5}\pi\right) + \sin\left(\frac{2}{5}\pi\right)\right)} \text{ [77]}. \tag{1.2}$$

Once $P$ is computed using the arctangent in both cases it is converted to degrees. For

| Name | Definition |
|---|---|
| $v_0$ or $\theta_4$ | $C'_4 - O'_4 - C'_1 - C'_2$ |
| $v_1$ or $\theta_5$ | $O'_4 - C'_1 - C'_2 - C'_3$ |
| $v_2$ or $\theta_1$ | $C'_1 - C'_2 - C'_3 - C'_4$ |
| $v_3$ or $\theta_2$ | $C'_2 - C'_3 - C'_4 - O'_4$ |
| $v_4$ or $\theta_3$ | $C'_3 - C'_4 - O'_4 - C'_1$ |

Table 1.2 – Naming and definition of the endocyclic sugar torsion angles.

the second definition if $P < 0$ then $P = P + 360°$. Finally by using the pseudo–rotation cycle in figure 1.3 the value of the pseudo–rotation phase angle is used to label the sugar puckering modes. In either envelope or twist conformations the atoms with the biggest displacement names the configuration and the labels *endo* and *exo* indicate in which direction the atoms is displaced.

## 1.3 Rigid–body configuration of ideal nucleic acid bases

It as been observed that the bases $\{A, T, G, C\}$ are incredibly close to being rigid. Thus it is common to assume that the atoms forming a nucleic base lie in the same plane. During the so called Tsukuba meeting, the participants established a reference system of the bases. For the aim of this work the *Tsukuba convention* [50] is the definition

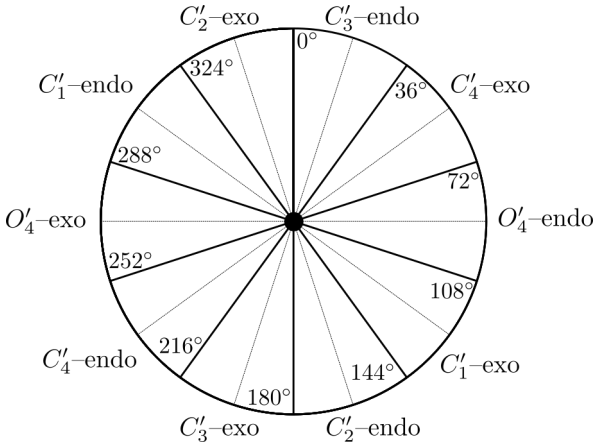



Figure 1.3 – Pseudorotation phase angle wheel. Each $36°$ the sugar puckering mode changes. The label of the puckering mode is given by the name of the atoms with the biggest displacement and the direction of the displacement is named *endo* or *exo*.

of idealised nucleic acid bases described by a pair composed by an atom type and associated three dimensional Cartesian coordinates. For the sake of completeness in the appendix we report in table A.1 the detailed Tsukuba convention.
The rigidity assumption implies that a single base can be interpreted as a box containing the atoms. A box can be described by a single point called the *reference point* and its *orientation*. In the context of the ideal bases of DNA [38] described a procedure to compute the ideal base reference point and the ideal base orientation. For completeness we report the entire procedure in [38], but we first make two remarks. The first is that the procedure is mathematically the same for purine and pyrimidine bases, but the atoms to be used vary between the two base types. In table 1.3 we report the parameters used in the computations. The second remark is about mathematical notation of the configuration of a rigid object. In particular its orientation is expressed by a proper rotation matrix $R \in \mathbb{R}^{3\times3}$ and its position is given by a three dimensional vector $r \in \mathbb{R}$. In this work we will denote by $\mathbf{g} = (R, r)$ a *rigid–body configuration* of a rigid object. The mathematical object $\mathbf{g}$ will be better introduced and discussed in the next chapter, as well as the definition of proper rotation matrix. In the following paragraph we present how to compute the rigid–body configuration of an ideal basis, whose atoms coordinates are given by the Tsukuba convention.
Let $a, b \in \mathbb{R}^3$ and $d \in \mathbb{R}$, compute $\tilde{R}_3 = a \times b$, where $\times$ is the vector product, then define $R_3 = \frac{\tilde{R}_3}{\|\tilde{R}_3\|}$ and $c = d\frac{a}{\|a\|}$, where $d \in \mathbb{R}$ and $\|\cdot\|$ is the euclidean norm. Compute $r = Q(R_3, \tau_1)c$ where $Q(R_3, \tau_1)$ is a matrix which rotates the vector $c$ around the unitary axis $R_3$ thought the angle $\tau_1$. In section 2.1 we give explicit formula for computing $Q(R_3, \tau_1)$. Next we compute $R_2 = Q(R_3, \tau_2)c$, $\tilde{R}_1 = R_2 \times R_3$, and $R_2 = \frac{\tilde{R}_2}{\|\tilde{R}_2\|}$. The base reference position is given by $r \in \mathbb{R}$ while the base reference orientation is given by the matrix $R$ which column are the unitary vectors $R_n,\ n = 1, 2, 3$, denoted by $R = (R_1|R_2|R_3)$.

| parameter | definition |
|---|---|
| a | $x(N_{1/\mathbf{9}}) - x(C_1')$ |
| b | $x(N_{1/\mathbf{9}}) - x(C_{2/\mathbf{4}})$ |
| d | 4.702 Å |
| $\tau_1$ | 2,4691 rad ( 141.47°) |
| $\tau_1$ | -0,9496 rad ( -54.41°) |

Table 1.3 – Definition of the parameters used for the computation of the base reference point and base reference orientation. The notation $x(A)$ stands for the Cartesian coordinates of the atoms $A$ given by the Tsukuba convention [50] table A.1.

Finally in table 1.3 we report the definition of all the parameters that have been used in the above method. Finally let $\mathbf{a}^X \in \mathbb{R}^{3n_X}$ be the set of $n_X$ ideal atom coordinates for the base $X \in \{A, T, G, C\}$. With the above procedure we can compute the rigid–body configuration $\mathbf{g}^X = (R^X, r^X)$, thus for each base we have the following couple $(\mathbf{a}^X, \mathbf{g}^X)$ of ideal coordinates and ideal rigid–body configuration.

# 2 Mathematics behind coarse-grain DNA models

This chapter is dedicated to the main mathematical notions that will be used throughout this work. More precisely we first introduce basic mathematical notation for matrix and probability calculus. Then we focus on the special euclidean group $SE(3)$ for which we describe the structure and most important properties that will be useful here. We then apply the $SE(3)$ group to the coarse–graining process of double stranded DNA by presenting briefly a classic polymer physics model, and an introduction to persistence length. We continue to the mathematical modelling of more realistic DNA by presenting the bichain representation of double stranded DNA introduced in [21]. We denote $A, B \in \mathbb{R}^{n\times n}$ two real $n$ times $n$ matrix, and use the following notation

$$A : B = \text{trace}(B^T A) = \sum_{i,j=1}^{n} A_{ij} B_{ji}, \tag{2.1}$$

for the *Frobenius inner product*, where the subscripts indicate the $ij$ entries of the matrices, the superscript $^T$ indicate the transpose matrix, and trace is the usual trace of the matrix. The norm induced by the Frobenius inner product will be denoted

$$\|A\| = \sqrt{A : A}, \tag{2.2}$$

and will be called the *Frobenius norm*. The Frobenius inner product (2.1) is related to the Euclidean inner product by *vectorising* both matrices $A, B \in \mathbb{R}^{n\times n}$ using the

following rule

$$\mathbf{vec}(A) = \begin{bmatrix} A_{11} \\ A_{12} \\ \vdots \\ A_{1n} \\ A_{21} \\ \vdots \\ A_{nn} \end{bmatrix}, \tag{2.3}$$

so that we have the equivalence of inner products

$$A : B = \mathbf{vec}(A) \cdot \mathbf{vec}(B). \tag{2.4}$$

The determinant of the square matrix A is denoted simply $|A|$. The principal square root of a matrix $A$ is defined by $A^{\frac{1}{2}}$. The expectation of an observable $F : \mathbb{R}^n \to \mathbb{R}^m$ with respect to a probability density function is

$$\langle F(w) \rangle_p = \int_{\mathbb{R}^n} F(w) p(w) dw. \tag{2.5}$$

A multivariate normally distributed random variable in $\mathbb{R}^n$ will be noted by $X \approx \mathcal{N}(\mu, C)$, where $\mu \in \mathbb{R}^n$ is the *mean* and $0 < C = C^T \in \mathbb{R}^{n \times n}$ is the *covariance matrix*. The probability density function of $X$ will be denoted by

$$\rho(w; \mu, K) = \frac{1}{Z} \exp\{-\frac{1}{2} (w - \mu) \cdot K (w - \mu)\}, \tag{2.6}$$

$$Z = |2\pi C|^{-\frac{1}{2}}, \tag{2.7}$$

where $K = C^{-1}$ is called the *precision matrix*, or, in the context of this work, the*stiffness matrix*. Depending on need, sometimes a more pertinent parametric notation for $\rho$ will be used, more precisely,

$$\rho(w; \mu, K) = \rho(w; \theta), \tag{2.8}$$

where $\theta \in \mathbb{R}^{n+n^2}$ is a vector containing all the entries of $\mu$ and $K$. We will use the following notation

$$\theta := \text{param}(\mu, \mathbf{vec}(K)) = \begin{bmatrix} \mu \\ \mathbf{vec}(K) \end{bmatrix} \in \mathbb{R}^{n+n^2}. \tag{2.9}$$

For the sake of simplicity in this work the vector $\theta$ contains all the entries in $K$ even if in the case of Gaussian probability distributions the precision matrix is symmetric.

## 2.1 The groups SO(3) and SE(3)

We start by introducing the *special orthogonal matrix group* denoted $SO(3)$ that is defined by

$$SO(3) = \{R \in \mathbb{R}^{3\times 3} | R^T R = R R^T = I,\ |R| = 1\}, \tag{2.10}$$

where $I \in R^{3\times 3}$ is the identity matrix. The group $SO(3)$ represents the group of all proper rotations in euclidean space. *Rodrigues' rotation formula* [61] characterises a *right–handed* rotation around the unit vector $x$ through an angle $0 \leq \theta leq \pi$:

$$R(x,\theta) := I + \sin(\theta)[x\times] + (1-\cos(\theta))[x\times]^2 \in SO(3), \tag{2.11}$$

where $[\cdot\times] : \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ denotes the linear mapping defined by

$$[x\times] = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}, \tag{2.12}$$

for any arbitrary vector $x \in \mathbb{R}^3$, called the *skew–operator* . Clearly, if $S$ is a skew matrix, we have that $S = [u\times]$ with $u = (S_{32}\ S_{13}\ S_{21}) \in \mathbb{R}^3$. As $SO(3)$ is a Lie group it admits a linear algebra, denoted by $so(3)$, defined by

$$so(3) = \{S \in \mathbb{R}^{3\times 3} | S = [x\times],\ \text{with } x \in \mathbb{R}^3\}. \tag{2.13}$$

Hence, $so(3)$ is the set of all three dimensional skew matrices. In Lie group theory the Lie group and the Lie algebra are related by two transformations called the *exponential* and the *logarithm*. The exponential map, $\exp : so(3) \to SO(3)$ is defined by (2.11) where for any $u \in \mathbb{R}^3$, $\exp([u\times]) = R(x,\theta)$, where $\theta$ is the norm of $u$ and $x = u/\theta$. We stress the fact that the mapping $\exp$ is defined for any arbitrary vector $u$ but $\exp([u\times]) = \exp([\tilde{u}\times])$ where $|\tilde{u}| = |u| \text{mod}(2\pi)$. The logarithm map is, instead, defined by $\log : SO(3) \to so(3)$:

$$\log(R) = \frac{\theta}{2\sin(\theta)}(R - R^T), \tag{2.14}$$

where $\theta$ satisfies $1+2\cos(\theta) = \text{trace}(R)$ and $\log(R) = [u\times]$ with $\|u\| = \theta$. For $SO(3)$ both mappings are in fact the actual matrix exponential and matrix logarithm functions. Another parametrization for rotations of interest in this work, is the *Cayley vector representation* in the specific sense defined in [36] [55] . Let $cay : \mathbb{R}^3 \to SO(3)$,

$$cay_\alpha(\eta) = I + \frac{4\alpha^2}{4\alpha^2 + |\eta|^2}\left(\frac{1}{\alpha}[\eta\times] + \frac{1}{2\alpha^2}[\eta\times]^2\right),\ \alpha \in \mathbb{R}. \tag{2.15}$$

For $\eta \in \mathbb{R}^3$ with $\|\eta\| = 2\alpha \tan(\frac{\theta}{2})$ we have that $cay_\alpha(\eta) = R(x,\theta)$ where $x = \frac{\eta}{|\eta|}$. The scalar factor $\alpha$ in (2.15) is in general equal to one but in the context of DNA $\alpha = 5$. The inverse transformation $cay^{-1} : SO(3) \to \mathbb{R}^3$ is

$$cay_\alpha^{-1}(R) = \frac{2\alpha}{1 + \text{trace}(R)} \text{vec}(R - R^T). \tag{2.16}$$

Depending on the context we will use both parametrisations of the rotation group. The matrix group $SO(3)$, being a Lie group, is also equipped with a differential structure and thus, for a differentiable curve $R(t) \subset SO(3)$, for $t \in \mathbb{R}$ we have that

$$\frac{d}{dt}R(t) = [\phi_g(t)\times]R(t) \tag{2.17}$$

$$= R(t)[\phi_d(t)\times], \tag{2.18}$$

where $\phi_g(t), \phi_d(t) \in \mathbb{R}^3$ are the *Darboux vectors*. The Darboux vector $\phi_g(t)$ satisfying (2.17) for all $t$ is called the *left infinitesimal generator* while $\phi_d(t)$ satisfying (2.18) is called the *right infinitesimal generator*. For the SO(3) Lie group, left and right infinitesimal generators satisfy the following relationship

$$R(t)\phi_d(t) = \phi_g(t), \ \forall t. \tag{2.19}$$

The homogeneous matrix representation of the *special euclidean group*, denoted by $SE(3)$, is the Lie group of *rigid body transformations* defined by

$$SE(3) = \left\{ \mathbf{g} \in \mathbb{R}^{4\times4} | \mathbf{g} = \begin{bmatrix} R & r \\ 0 & 1 \end{bmatrix} = (R,r) \text{ with } R \in SO(3) \text{ and } r \in \mathbb{R}^3 \right\}, \tag{2.20}$$

where, for an arbitrary element $\mathbf{g}$, the rotational part is represented by the matrix $R$ and the translational part is described by the vector $r$. The product of two element in $SE(3)$ is given by standard matrix multiplication, i.e, for $\mathbf{g}_1, \mathbf{g}_2 \in SE(3)$,

$$\mathbf{g}_1\mathbf{g}_2 = \begin{bmatrix} R_1R_2 & R_1r_2 + r_1 \\ 0 & 1 \end{bmatrix}, \tag{2.21}$$

while the inverse of a rigid body transformation $\mathbf{g}$ is given by

$$\mathbf{g}^{-1} = \begin{bmatrix} R^T & -R^Tr \\ 0 & 1 \end{bmatrix}. \tag{2.22}$$

The Lie algebra, denoted $se(3)$, is defined by

$$se(3) = \left\{ T \in \mathbb{R}^4 | T = T(\phi) = \begin{bmatrix} [u\times] & v \\ 0 & 0 \end{bmatrix}, \text{ with } \phi = (u,v) \text{ and } u,v \in \mathbb{R}^3 \right\}, \tag{2.23}$$

where we can introduce the operator $\mathcal{T} : \mathbb{R}^6 \to \mathbb{R}^{4\times4}$ defined by

$$\mathcal{T}\phi = \begin{bmatrix} [u\times] & v \\ 0 & 0 \end{bmatrix}, \text{ with } \phi = (u, v), \tag{2.24}$$

so that an arbitrary element $T = T(\phi) \in se(3)$ can be simply written as $T(\phi) = \mathcal{T}\phi$. The operator (2.24) is called the *tangent operator* and admits a unique *adjoint operator* $\mathcal{T}^* : \mathbb{R}^{4\times4} \to \mathbb{R}^6$ defined by

$$\mathcal{T}^* \begin{pmatrix} X & x \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \text{Vect}(X) \\ x \end{pmatrix}, \tag{2.25}$$

where

$$\text{Vect}(X) = \begin{bmatrix} X_{32} - X_{23} \\ X_{13} - X_{31} \\ X_{21} - X_{12} \end{bmatrix}, \ \forall X \in \mathbb{R}^{3\times3}. \tag{2.26}$$

Let $\mathbf{g} \in$ SE(3); then we define the right infinitesimal generator as

$$\frac{d}{dt}\mathbf{g} = \mathbf{g}\mathcal{T}\phi_d = \begin{bmatrix} R(t)[\phi_d^R(t)\times] & R(t)\phi_d^r(t) \\ 0 & 0 \end{bmatrix}, \tag{2.27}$$

and the left infinitesimal generator of $\mathbf{g}$ as

$$\frac{d}{dt}\mathbf{g} = \mathcal{T}\phi_g(t)\mathbf{g} = \begin{bmatrix} [\phi_g^R(t)\times]R(t) & [\phi_g^R(t)\times]r(t) + \phi^r(t) \\ 0 & 0 \end{bmatrix}. \tag{2.28}$$

Just as for the $SO(3)$ group we have a linear relationship between left and right infinitesimal generators

$$\phi_g(t) = \text{Ad}_{\mathbf{g}(t)}\phi_d(t), \text{ with } \text{Ad}_{\mathbf{g}(t)} = \begin{bmatrix} R(t) & 0 \\ [r(t)\times]R(t) & R(t) \end{bmatrix}, \ \forall t. \tag{2.29}$$

The Ad operator has two useful properties

$$\text{Ad}_{\mathbf{g}_1\mathbf{g}_2} = \text{Ad}_{\mathbf{g}_1}\text{Ad}_{\mathbf{g}_2}, \tag{2.30}$$

$$\text{Ad}_{\mathbf{g}}^{-1} = \text{Ad}_{\mathbf{g}^{-1}} = \begin{bmatrix} R^T & 0 \\ -R^T[r\times] & R^T \end{bmatrix}. \tag{2.31}$$

Let now $\mathbf{g}_1(t), \mathbf{g}_2(t) \in SE(3)$ with left infinitesimal generator defined by $\phi_{(g,1)}(t), \phi_{(g,2)}(t)$, i.e

$$\frac{d}{dt}\mathbf{g}_i = \mathcal{T}\phi_{(g,i)}(t)\mathbf{g}_i(t), \ i = 1, 2. \tag{2.32}$$

Assume now that $\mathbf{g}(t) \in SE(3)$ satisfies $\frac{d}{dt}\mathbf{g}(t) = \mathcal{T}\phi_g(t)\mathbf{g}(t)$ and that it can be written as $\mathbf{g}(t) = \mathbf{g}_1(t)\mathbf{g}_2(t)$. We can relate $\phi_g(t)$ to $\phi_{(g,i)}$ via the equation

$$\phi_g(t) = \phi_{(g,1)}(t) + \mathrm{Ad}_{\mathbf{g}_1(t)}\phi_{(g,2)}(t), \tag{2.33}$$

and by using equation (2.29) we can rewrite the above expression for the right infinitesimal generator, namely

$$\phi_d(t) = \mathrm{Ad}_{\mathbf{g}_2^{-1}(t)}\phi_{(d,1)}(t) + \phi_{(d,2)}(t). \tag{2.34}$$

We now briefly introduce the natural exponential map, $\mathrm{Exp} : se(3) \to SE(3)$ as defined in [12]:

$$\mathrm{Exp}(\mathcal{T}\phi) = \begin{bmatrix} \exp(u) & \left(I + \frac{(1-\cos(\theta))}{\theta}[x\times] + \frac{(\theta-\sin(\theta))}{\theta}[x\times]^2\right)v \\ 0 & 1 \end{bmatrix}, \tag{2.35}$$

for $\phi = (u, v)$ and $\theta = |u|$, $x = \frac{u}{\theta}$, and $\exp(u)$ as defined in (2.11) and (2.35) is in fact equivalent to the standard exponential for matrices of the form $\mathcal{T}\phi$. The standard definition is used to define the neighbour rigid body transformation of a given one, noted $\overline{\mathbf{g}}$, by truncating the power series at a chosen order, for example the *second order* (left) approximation of a neighbour of $\overline{\mathbf{g}}$ is defined by

$$\mathbf{g} = \left(I + \mathcal{T}\phi_g + \frac{1}{2}(\mathcal{T}\phi_g)^2\right)\overline{\mathbf{g}} + \mathcal{O}(|\phi_g|^2). \tag{2.36}$$

We refer to [21] for more detail about matrix calculus.

## 2.2 Polymer physics and persistence length

A *polymer* can be modelled by a linear chain of rigid bodies represented by a set of rigid body configurations $\mathbf{g} = \{\mathbf{g}_n\}_{n=1}^N \in SE(3)^N$, with $\mathbf{g}_n = (R_n, r_n) \in SE(3)$. One classic observable in polymer physics is the relative rigid body displacement

$$F(\mathbf{g}, n) = \mathbf{g}_0^{-1}\mathbf{g}_n = \begin{bmatrix} R_0^T R_n & R_0^T(r_n - r_0) \\ 0 & 1 \end{bmatrix} \tag{2.37}$$

where $\mathbf{g}_0 = (R_0, r_0)$ is a reference frame that usually is taken to be away from the end of the chain. It is common to define from (2.37) two different expectations:

$$\text{the Flory vector [17]} : \langle R_0^T(r_n - r_0)\rangle, \tag{2.38}$$

$$\text{tangent–tangent correlation} : \langle (R_0^T R_n)_{(3,3)}\rangle = \langle \mathbf{t}_0 \cdot \mathbf{t}_n\rangle, \tag{2.39}$$

where $\langle\cdot\rangle$ denotes the expectation with respect to an underlying equilibrium distribution. We observe that both expectations are functions of the chain index $n \geq 1$ and are

respectively a vector and a scalar.
A classic simple polymer model is the discrete version of the *Kratky-Porod wormlike–chain* (WLC) [64] [32] where the chain is assumed to be composed by rigid links and with the same length $b$, in such a way that the configuration of an $N$ long polymer chain can be described by means of a set of unitary tangent vector $\{\mathbf{t}_n\}_{n=1}^{N-1}$, where $\mathbf{t}_n = \frac{1}{b}(r_{n+1} - r_n)$. The free energy associated to the WLC is assumed to be

$$E(\mathbf{g}) = \frac{B}{b}\sum_{n=1}^{N-1}(1 - \mathbf{t}_n \cdot \mathbf{t}_{n+1}), \tag{2.40}$$

where $B$ is a constant *bending rigidity* parameter. We remark that the ground–state, or state of minimal energy, of such a model is intrinsically straight, meaning that all the unit tangent vectors are aligned. We further assume that the equilibrium distribution of the WLC is Boltzmann, i.e, it can be written as $\rho(\mathbf{g}) = \exp\{\beta E(\mathbf{g})\}$ with $\beta^{-1} = k_b T$. With this simple model, both the Flory vector and the tangent–tangent correlation functions, can be computed analytically:

$$\langle R_0^T(r_n - r_0)\rangle_\rho = b\ell_p\left(1 - \exp\left(-\frac{n}{\ell_p}\right)\right)\mathbf{e}_3, \tag{2.41}$$

$$\langle \mathbf{t}_0 \cdot \mathbf{t}_n\rangle_\rho = \exp\left(-\frac{n}{\ell_p}\right), \tag{2.42}$$

where $\mathbf{e}_3 = (0,0,1)^T$, and the exponential decay parameter $\ell_p$ is called the *persistence length*, here expressed in base–pairs. Moreover, in the WLC model we naturally find that $b\ell_p = \beta B$ which represents the persistence length in arc length units of $b$. As a last comment we stress that both expressions in (2.41) are exact in the continuous limit of the WLC model, for which the quantity $b\ell_p$ stays constant while $b \to 0$, $N \to \infty$, $Nb \to L$, and $nb \to s \in [0\ L]$, where $L$ is the length of the polymer in arc–length units.
In the context of DNA the chemical composition of the polymer chain is a function of a specific sequence $\mathcal{S}$ in the $\{A,T,G,C\}$ alphabet which implies that both the Flory vector and the tangent–tangent correlation function will be function also of $\mathcal{S}$. For more detail we refer to [18] [45]. Hereafter we briefly introduce the sequence dependent and sequence average generalization of (2.41):

$$\ell_F(\mathcal{S}) = \lim_{n\to\infty}\|\langle R_0^T(r_n - r_0)\rangle\|, \qquad \exp\left(-\frac{n}{\ell_p(\mathcal{S})}\right) \approx \langle \mathbf{t}_n \cdot \mathbf{t}_0\rangle, \tag{2.43}$$

$$\overline{\ell_F} = \lim_{n\to\infty}\|\{R_0^T(r_n - r_0)\}\|, \qquad \exp\left(-\frac{n}{\overline{\ell_p}}\right) \approx \{\langle \mathbf{t}_n \cdot \mathbf{t}_0\rangle\}, \tag{2.44}$$

where $\{\cdot\}$, is an average over an ensemble of sequences, $\ell_F(\mathcal{S})$ is the *sequence–dependent Flory persistence length*, $\overline{\ell_F}$ is the *sequence average Flory persistence length*, $\ell_p(\mathcal{S})$ is the *sequence–dependent persistence length*, $\overline{\ell_p}$ is the *sequence average persistence length*, and $\approx$ signifies that $\ell_p(\mathcal{S})$ is the negative reciprocal of the slope of the linear fit through

the origin of $\log(\langle \mathbf{t}_n \cdot \mathbf{t}_0 \rangle)$ vs $n$. For sake of simplicity $\ell_p(\mathcal{S})$ will denote the tangent–tangent correlation persistence length and we do not introduce a specific notation for that. But, it should be remarked that, $\ell_F(\mathcal{S})$ and $\ell_p(\mathcal{S})$ are now not in principle the same due to the non-trivial intrinsic shape of DNA. For example it is known that sequences containing phased A–tracts [44] [65][60] tend to have a high intrinsic curvature which implies a non–linear decay in the log–tangent–tangent correlation leading to a poor approximation of $\ell_p(\mathcal{S})$. In section (4.3), we will present a few examples of computations of $\ell_p(\mathcal{S})$. Furthermore, in [18] [45], the concept of dynamic persistence length is presented and studied. In [74] the authors proposed the following sequence–averaged decomposition of the persistence length:

$$\frac{1}{\overline{\ell}_p} = \frac{1}{\overline{\ell}_s} + \frac{1}{\overline{\ell}_d}, \tag{2.45}$$

where, following [45], $\overline{\ell}_p$ is renamed to the *apparent persistence length*, $\overline{\ell}_s$ is the *static persistence length*, and $\overline{\ell}_d$ is the *dynamic persistence length*. Equation (2.45) states that the apparent persistence length can be decomposed into two contribution, a static one related to intrinsic shape, and a dynamic one related to thermal fluctuations. The static contribution can be computed as the sequence ensemble average of the deterministic form of $(2.44)_2$, defined by

$$\exp\left(-\frac{n}{\overline{\ell}_s}\right) \approx \{\widehat{\mathbf{t}}_n \cdot \widehat{\mathbf{t}}_0\}, \tag{2.46}$$

where $\widehat{\mathbf{t}}_n$ is evaluated on the ground–state configuration for each sequence in the ensemble. Finally in [45] the authors generalized the sequence–averaged dynamic persistence length of [74] in the following *sequence–dependent dynamic persistence length*

$$(\widehat{\mathbf{t}}_n \cdot \widehat{\mathbf{t}}_0) \exp\left(-\frac{n}{\ell_d(\mathcal{S})}\right) \approx \langle \mathbf{t}_n \cdot \mathbf{t}_0 \rangle. \tag{2.47}$$

Hence, $\ell_d(\mathcal{S})$ can be computed from the linear fit of a plot of $\ln\langle \mathbf{t}_n \cdot \mathbf{t}_0 \rangle - \ln(\widehat{\mathbf{t}}_n \cdot \widehat{\mathbf{t}}_0)$ as function of $n$. The quality of the latter fit has been shown [18, 45] to be always better then its analogous $(2.43)_2$ which make $\ell_d(\mathcal{S})$ more robust as a proxy for the "rigidity" of different DNA sequences.

## 2.3 From atoms to rigid-bodies

In this section we present the classic approach underlying the coarse–graining of double stranded DNA molecules. The first step is to set the level of coarse graining to consider in the model, meaning that some group of atoms will be considered as part of the same unit and others will be not explicitly considered in the model. For example

we will consider that each base of the DNA molecule is an individual rigid unit. In the following we will present the standard methodology used to associate a rigid–body to each unit. Consider two set of atoms coordinates $\mathbf{a}^X = \{a_i^X\}_{i=1}^m$ and $\mathbf{p} = \{p_i\}_{i=1}^m$ with a one–to–one correspondence between $a_j$ and $p_j$, in the sense that they represent the same atom type. We will refer to **a** as ideal atoms and to **p** as observed atoms and we are interested in associating a rigid body to the observed atoms. In the DNA context the ideal atoms are fixed coordinates for each type of bases and are planar. In [50] the authors give the list of atoms type and Cartesian coordinates forming the ideal bases for the nucleic acid bases $A, T, G$, and $C$. Moreover, in [38] the authors give a way of associating a rigid body $\mathbf{g_a}$ to each ideal atoms groups **a**. In general, for the ideal atom group **a** we have a couple $(\mathbf{g_a}, \mathbf{a})$ where $\mathbf{g_a} \in SE(3)$ has the matrix form described in (2.20). Now, mathematically the assumption of rigidity of the observed atoms **p** imply that there exists a rigid–body transformation $\mathbf{g} = (R, r) \in SE(3)$ such that

$$q_i = R_i a_i + r_i, \ \forall i = 1, \dots, m, \tag{2.48}$$

with $(R, r)$ satisfying

$$(R, r) = \underset{Q \in SO(3), t \in \mathbb{R}^3}{\operatorname{argmin}} \sum_{i=1}^{m} \|q_i - p_i\|^2, \tag{2.49}$$

and finally we compute the rigid body associated to **p** by right rigid body transformation

$$\mathbf{g}^p = \mathbf{g}\mathbf{g}_a, \tag{2.50}$$

where $\mathbf{g} = (R, r)$. The least square system (2.49) can be solved by defining $S = XY^T$ where the column of $X$ and $Y$ are given by

$$X_i = a_i - \overline{a}, \ \ i = 1, \dots, m, \ \overline{a} = \frac{1}{m} \sum_{k=1}^{m} a_i,$$
$$Y_i = p_i - \overline{p}, \ \ i = 1, \dots, m, \ \overline{p} = \frac{1}{m} \sum_{k=1}^{m} p_i,$$

and by computing the singular value decomposition of $S = U \Sigma V^T$. Then the rotation matrix $R$ is computed as

$$R = VDU^T, \text{ with } D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \|VU^T\| \end{bmatrix},$$

where $\|VU^T\| = \pm 1$ guarantees that $R \in SO(3)$. Finally the translational part $r$ is

$$r = \overline{p} - R\overline{a}.$$

For all the details about the derivation of the above computation we refer to [67].
Let us consider a $N$ base–pair long DNA fragment with sequence $\mathcal{S}$ and assume that

$$\mathbf{p}(\mathcal{S}) = (p_1, \ldots, p_L)(\mathcal{S}) \in \mathbb{R}^{3L},$$

with $p_i \in \mathbb{R}^3$ is the Cartesian coordinates of the $i$–th atom and $L$ the total number of atoms in the DNA. We first coarse–grain $\mathbf{p}(\mathcal{S})$ by forming units with some of the atoms and by neglecting the rest. For example, by coarse–graining at the level of single bases, we obtain the following map

$$\mathbf{r}(\mathcal{S}) \to (\mathbf{p}_X, \mathbf{p}_{\overline{X}}) = (\mathbf{p}^{X_1}, \ldots, \mathbf{p}^{X_N}, \mathbf{p}^{\overline{X}_1}, \ldots, \mathbf{p}^{\overline{X}_N}) \in \mathbb{R}^{3\ell}, \tag{2.51}$$

where $\mathcal{S} = X_1 X_2 \cdots X_N$, $X_i \in \{A, T, G, C\}$, $\mathbf{p}^{X_i} = (p_1^{X_i}, \ldots, p_{n_i}^{X_i}) \in \mathbb{R}^{3n_i}$ are the atoms considered for the base $X_i$, $\overline{X}$ denote the Crick–Watson complementary base of $X$, and $\ell = \sum_{i=1}^{N}(n_i + \overline{n}_i)$. We can now apply (2.49) to $(\mathbf{p}_X, \mathbf{p}_{\overline{X}})$ in the following way

$$\mathfrak{R}^{\cdot\cdot}(\mathbf{p}_X, \mathbf{p}_{\overline{X}}) = (\mathbf{g}_1^+, \ldots, \mathbf{g}_N^+, \overline{\mathbf{g}}_1^+, \ldots, \overline{\mathbf{g}}_N^+) = (\mathbf{g}^+, \overline{\mathbf{g}}^+)(\mathcal{S}) \in SE(3)^{2N}, \tag{2.52}$$

where $\mathbf{g}_n^+$ and $\overline{\mathbf{g}}_n^+$ are the rigid body associated, respectively to $\mathbf{p}^{X_n}$ and $\mathbf{p}^{\overline{X}_n}$ computed by (2.50). In the context of rigid–base modelling of DNA it is in fact more convenient to work with both base frames having the $\mathbf{d}_3$ axis along approximately the same direction, see figure 2.1, thus we introduce the matrix $P \in O(3)$ defined by

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \tag{2.53}$$

and we define

$$\mathfrak{R}^{\cdot\cdot}(\mathbf{p}_X, \mathbf{p}_{\overline{X}}) = (\mathbf{g}^+, \mathbf{g}^-)(\mathcal{S}) \in SE(3)^{2N}, \tag{2.54}$$

where

$$\mathbf{g}_n^- = (R_n^-, r_n^-) = (\overline{R}_n^+ P, \overline{r}_n^+), \forall n. \tag{2.55}$$

As a final remark, we stress the fact that the mapping $\mathfrak{R}^{\cdot\cdot}$ is not invertible because a least square fitting is involved. But the atomistic resolution of only the coarse–grain units, can be retrieved approximately by using the transformation described in (2.48).

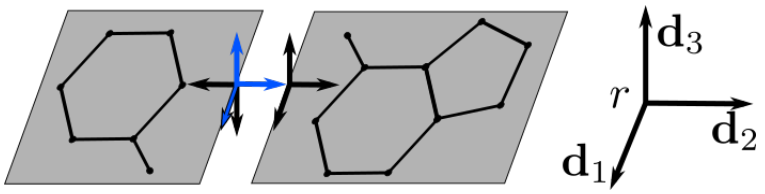



Figure 2.1 – Example of rigid body, in black, fitted to two complementary bases. In the context of rigid–base coarse–graining, it is more convenient to rotate the frame $\bar{\mathbf{g}}_n^+$ along the $\mathbf{d}_1$ axis by $P \in O(3)$, see (2.53) and (2.55), to obtain the blue frame .

## 2.4 Bichain interpretation of DNA

When coarse–grained to the level of bases, a molecule of double stranded DNA can be interpreted as a double chain of rigid bodies. In this section we will recall the basic concept and notation for rigid–body double chain configurations and we will present its relationship to the bichain representation of coarse–grain double stranded DNA. More detail can be found in [21].
Formally, a rigid–body double–chain configuration is denoted by the couple

$$(\mathbf{g}^+, \mathbf{g}^-) = (\mathbf{g}_1^+, \mathbf{g}_2^+, \ldots, \mathbf{g}_N^+, \mathbf{g}_1^-, \mathbf{g}_2^-, \ldots, \mathbf{g}_N^-) \in SE(3)^{2N}.$$

In the context of DNA $\mathbf{g}^+, \mathbf{g}^-$ represent respectively the reading strand and complementary strand where each strand is described by a single chain of rigid bodies. A more convenient way for interpreting a molecule of DNA is in fact the rigid–body bichain interpretation. A bichain configuration of a $N$ base–pair long fragment of DNA can be described by a *macrostructure* and a *microstructure* respectively noted $\mathbf{g}$ and $\mathcal{P}$. Hence $(\mathbf{g}, \mathcal{P}) \in SE(3)^{2N}$ will be used to denote a bichain configuration. In the DNA context, the macrostructure configuration $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_N) \in SE(3)^N$ describe the position and orientation of each base-pair along the molecule, while the microstructure configuration $\mathcal{P} = (\mathcal{P}_1, \ldots, \mathcal{P}_N) \in SE(3)^N$ describes the relative configuration of the complementary bases in the same base–pair level. Clearly, the microstructure configurations do not form a chain and this concept can be useful in some applications.
From coarse–grained MD trajectories we actually observe times series of double chain configurations of DNA, but there is an invertible mapping between both interpretations. Indeed, for any double chain configuration we can define the invertible mapping $\mathfrak{R} : SE(3)^{2N} \to SE(3)^{2N}$ defined by

$$\mathfrak{R}^{\parallel}(\mathbf{g}^+, \mathbf{g}^-) = (\mathbf{g}, \mathcal{P}), \tag{2.56}$$

where

$$\mathbf{g}_n = (R_n, r_n) = \begin{bmatrix} R_n^-([R_n^-]^T R_n^+)^{\frac{1}{2}} & \frac{1}{2}(r_n^+ + r_n^-) \\ 0 & 1 \end{bmatrix}, \tag{2.57}$$

where $\mathbf{g}_n$ represent the $n$th base–pair rigid body along the chain. Its definition is in fact the half point along the geodesic going from $\mathbf{g}_n^-$ to $\mathbf{g}_n^+$, when the left invariant Riemannian metrix is considered. The microstructure, instead, is defined as the rigid body displacement between the two strands at the same base–pair level, i.e,

$$\mathcal{P}_n = (P_n, p_n) = \mathbf{g}_n^- \mathbf{g}_n^+ = \begin{bmatrix} [R_n^-]^T R_n^+ & R_n^T (r_n^+ - r_n^-) \\ 0 & 1 \end{bmatrix}. \tag{2.58}$$

### 2.4.1 Internal coordinates for $(\mathbf{g}, \mathcal{P})$

We can now introduce the set of internal coordinates for a bichain configuration. In this representation the internal coordinates naturally split into coordinates for the macrostructure and coordinates for the microstructure. For defining the macrostructure coordinates, also called *inter coordinates*, we have to introduce the *inter base–pair junction displacements*

$$a_n = \begin{bmatrix} Q_n & q_n \\ 0 & 1 \end{bmatrix} \tag{2.59}$$

that satisfy $\mathbf{g}_{n+1} = \mathbf{g}_n a_n$. The inter coordinates are, in fact, a parametrization of the junction displacement. Hence, let $x_n = (u_n, v_n) \in \mathbb{R}^6$ be a set of coordinates parametrizing $a_n = a_n(x_n) = (Q_n, q_n)(x_n)$ where

$$Q_n = Q(u_n) = cay_\alpha(u_n),\ \alpha = 5, \tag{2.60}$$

$$q_n = q_n(u_n, v_n) = Q^{\frac{1}{2}}(u_n) v_n, \tag{2.61}$$

where $cay_\alpha$ is defined in (2.15). Finally we have a reconstruction rule for the chain part of the bichain representation,

$$x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^{6(N-1)} \mapsto \mathbf{g}(x) = (\mathbf{g}_1, \ldots, \mathbf{g}_N) \in SE(3)^N, \tag{2.62}$$

by using the formulas

$$\mathbf{g}_{n+1} = \mathbf{g}_n a_n = \mathbf{g}_1 \prod_{i=1}^{n} a_i. \tag{2.63}$$

The set of internal coordinates for the microstructure will be called, *intra base–pair coordinates*. They parametrise the intra displacement defined by $\mathcal{P}_n$. Let $y_n = (\eta_n, \mathrm{w}_n) \in R^6$ be the *intra coordinates* parametrising $\mathcal{P}_n$, then they satisfy

$$\mathcal{P}(y_n) = \begin{bmatrix} P(\eta_n) & P(\eta_n)^{\frac{1}{2}} \mathrm{w}_n \\ 0 & 1 \end{bmatrix}, \tag{2.64}$$

where

$$P_n = P(\eta_n) = cay_\alpha(\eta_n),\ \alpha = 5 \tag{2.65}$$

$$p_n = p_n(\eta_n, \mathbf{w}_n) = P(\eta_n)^{\frac{1}{2}}\mathbf{w}_n. \tag{2.66}$$

Finally, by defining

$$(x, y) = (y_1, x_1, y_2, \ldots, x_{N-1}, y_N) \in \mathbb{R}^{12N-6} \tag{2.67}$$

as the bichain internal coordinates we can formally write an invertible reconstruction rule for a bichain $\mathfrak{R}^\ddagger : \mathbb{R}^{12N-6} \to SE(3)^{2N}$

$$(x, y) \mapsto (\mathbf{g}, \mathcal{P})(x, y). \tag{2.68}$$

The inverse map, noted $[\mathfrak{R}^\ddagger]^{-1}$, can then be used to define a mapping between any double chain configuration and its bichain internal coordinates, which we will denote by $\mathfrak{R} : SE(3)^N \to \mathbb{R}^{12N-6}$, defined by

$$\mathfrak{R}(\mathbf{g}^+, \mathbf{g}^-) = [\mathfrak{R}^\ddagger]^{-1}(\mathfrak{R}^\parallel(\mathbf{g}^+, \mathbf{g}^-)) = (x, y). \tag{2.69}$$

### 2.4.2 Internal Energy for $(\mathbf{g}, \mathcal{P})$

In this paragraph we introduce and discuss the internal energy for bichain configuration along with its first variation, will be important later in this work. But before we present the results for the bichain we will briefly introduce the *equilibrium conditions for chains*. The equilibrium conditions for a set of rigid bodies $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_N) \in SE(3)^N$, $\mathbf{g}_n = (R_n, r_n) \in SE(3)$, can be written, in a short format, as

$$-\mu_{n+1}(\mathbf{g}) + \mu_n(\mathbf{g}) + \lambda_n = 0,\ \forall n = 1, \ldots, N \tag{2.70}$$

with

$$\mu_n = \begin{bmatrix} \mathbf{m}_n + r_n \times \mathbf{n}_n \\ \mathbf{n}_n \end{bmatrix} \in \mathbb{R}^6, \qquad \lambda_n = \begin{bmatrix} \mathbf{c}_n + r_n \times \mathbf{f}_n \\ \mathbf{f}_n \end{bmatrix} \in \mathbb{R}^6,\ n = 2, \ldots, N \tag{2.71}$$

$$\text{and } \mu_1 = \mu_{N+1} = 0.$$

where $\mathbf{m}_n, \mathbf{n}_n \in \mathbb{R}^3$ are respectively the total *internal couple* around $r_n$ and *internal force* on $\mathbf{g}_n$ from the downstream (along the chain) rigid–body $\mathbf{g}_{n-1}$. The corresponding external loads are denoted $\mathbf{c}_n \in \mathbb{R}^3$ for the total external couple, around $r_n$, and $\mathbf{f}_n \in \mathbb{R}^3$ for the total external force acting on $\mathbf{g}_n$. Moreover, we set $\mu_1 = \mu_{N+1} = 0$. The mapping $\mathbf{g} \mapsto \mu_n(\mathbf{g}),\ \forall n = 1, \ldots, N$, is called the *local chain constitutive relations* which together with equations (2.70) form the equilibrium conditions for the chain

$\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_N) \in SE(3)^N$. Hereafter we will introduce the local energy for bichains configurations, compute its first variation and relate it with the balance laws (2.70). Let us consider a configuration $(\mathbf{g}, \mathcal{P}) \in SE(3)^{2N}$. The bichain internal energy $E : SE(3)^N \to \mathbb{R}$ for $(\mathbf{g}, \mathcal{P})$ of interest is of the form

$$E(\mathbf{g}, \mathcal{P}) = \sum_{n=1}^{N-1} w_n(\mathcal{P}_n, a_n, \mathcal{P}_{n+1}), \tag{2.72}$$

in particular, the energy (2.72) has local energy contributions defined at the junction level $\mathcal{J}_n = (\mathcal{P}_n, a_n, \mathcal{P}_{n+1})$. We will refer to this type of energy, with this particular local interactions, as a nearest neighbour interaction energy. It will be shown to play a major role in the modelling of DNA, see for instance chapter 4. The first (left) variation of the internal energy (2.72) reads:

$$D_l E(\mathbf{g}, \mathcal{P})(\Theta, \theta^{\mathcal{P}}) = D_l E(\mathbf{g}, \mathcal{P}) \cdot \Theta + D_l E(\mathbf{g}, \mathcal{P}) \cdot \theta^{\mathcal{P}}, \tag{2.73}$$

where

$$D_l E(\mathbf{g}, \mathcal{P}) \cdot \Theta = \sum_{n=1}^{N-1} (-\mu_{n+1}(\mathbf{g}) + \mu_n(\mathbf{g})) \cdot \Theta_n, \tag{2.74}$$

$$D_l E(\mathbf{g}, \mathcal{P}) \cdot \theta^{\mathcal{P}} = \sum_{n=1}^{N} \mu_n^{\mathcal{P}} \cdot \theta_n^{\mathcal{P}}, \tag{2.75}$$

and

$$\begin{aligned} \delta\mathbf{g}_n &= \mathcal{T}\theta_n \mathbf{g}_n, & \theta &= (\theta_1, \ldots, \theta_N), \\ \delta\mathcal{P}_n &= \mathcal{T}\Theta_n^{\mathcal{P}} \mathcal{P}_n, & \Theta^{\mathcal{P}} &= (\Theta_1^{\mathcal{P}}, \ldots, \Theta_N^{\mathcal{P}}), \\ \mu_{n+1}(\mathbf{g}) &= \mathrm{Ad}_{\mathbf{g}_n}^{-T} \mathcal{T}^* \left(\partial_{a_n} w_n a_n^T\right) \in \mathbb{R}^6, & \mu_1 &= \mu_{N+1} = 0, \end{aligned} \tag{2.76}$$

$$\mu_n^{\mathcal{P}} = \mathcal{T}^* \left(\partial_{\mathcal{P}_n}(w_n + w_{n-1}) \mathcal{P}_n^T\right) \in \mathbb{R}^6. \tag{2.77}$$

It has been shown in [21] that the configuration $(\mathbf{g}, \mathcal{P}) \in SE(3)^{2N}$ that makes $E$ stationary, under certain geometric constraints on $(\mathbf{g}, \mathcal{P})$, satisfy for $n = 1, \ldots, N$, the following bichain balance laws

$$-\mu_{n+1}(\mathbf{g}) + \mu_n(\mathbf{g}) = 0, \tag{2.78}$$

$$\mu_n^{\mathcal{P}} = 0, \tag{2.79}$$

where the first equation is directly related to the equilibrium conditions for the chain $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_N) \in SE(3)^N$, and the second equation is the equilibrium of the microstructure represented by the intra rigid body displacement $\mathcal{P} = (\mathcal{P}_1, \ldots, \mathcal{P}_N) \in SE(3)^N$. These two conditions, and in particular the chain part, can be generalized also to allow external forces and, thus, the equilibrium conditions (2.78) can be rewrit-

ten has

$$(D_l E(\mathbf{g}, \mathcal{P}) + \lambda(g)) \cdot \theta + D_l E(\mathbf{g}, \mathcal{P}) \cdot \Theta^{\mathcal{P}} = 0, \tag{2.80}$$

$$(\mathbf{g}, \mathcal{P}) \in \mathcal{C}^{\mathbf{g}} \times \mathcal{C}^{\mathcal{P}}, \tag{2.81}$$

$$\forall \theta \in \mathrm{D}\mathcal{C}^{\mathbf{g}} \text{ and } \forall \Theta^{\mathcal{P}} \in \mathrm{D}\mathcal{C}^{\mathcal{P}} \tag{2.82}$$

where $\mathcal{C}^{\mathbf{g}}$ and $\mathcal{C}^{\mathcal{P}}$ are the set of geometric constraint that prescribe a fixed value to, respectively, a subset of rigid bodies in $\mathbf{g}$ and a subset of intra displacements of $\mathcal{P}$. The geometric constraints define then the spaces of admissible variations $\mathrm{D}\mathcal{C}^{\mathbf{g}}$ and $\mathrm{D}\mathcal{C}^{\mathcal{P}}$ where, for example,

$$\mathrm{D}\mathcal{C}^{\mathbf{g}} = \{\theta \in \mathbb{R}^{6N} | \theta_n = 0 \text{ if } \mathbf{g}_n \text{ is prescribed } \}. \tag{2.83}$$

In practice the bichain energy is in general given in term of internal coordinates $w = (x, y)$ in the following form

$$E(x, y) = \sum_{n=1}^{N-1} w_n(y_n, x_n, y_{n+1}), \tag{2.84}$$

which leads to the computationally tractable formulas

$$\mu_{n+1}(\mathbf{g}(x)) = \mathrm{Ad}_{\mathbf{g}_n}^{-T} \mathbb{L}_{x_n}^{-T} \partial_{x_n} w_n \in \mathbb{R}^6, \tag{2.85}$$

$$\mu_n^{\mathcal{P}} = \mathbb{L}_{y_n}^{-T} \partial_{y_n} (w_n + w_{n-1}) \in \mathbb{R}^6. \tag{2.86}$$

The matrix $\mathbb{L}_x$ has been derived in [21] with respect to the internal coordinates used in [55]. For $x = (u, v) \in \mathbb{R}^6$,

$$\mathbb{L}_x = \begin{bmatrix} \mathbb{P}_1(u) & 0 \\ Q^{\frac{1}{2}}[v\times]\mathbb{P}_2(u) & Q^{\frac{1}{2}}(u) \end{bmatrix}, \tag{2.87}$$

with

$$\mathbb{P}_1(u) = \frac{4\alpha^2}{4\alpha^2 + \|u\|^2} \left( \frac{1}{\alpha} I + \frac{1}{2\alpha^2} [u\times] \right), \tag{2.88}$$

$$\mathbb{P}_2(u) = \left( I + Q^{\frac{1}{2}}(u) \right)^{-1} \mathbb{P}_1(u). \tag{2.89}$$

The matrix $\mathrm{Ad}_{\mathbf{g}}^{-1}$ is defined in (2.29).

# 3 On molecular dynamics simulation protocols and analysis

## 3.1 Potential and force field

Molecular dynamics (MD) simulations are nowadays widely used for the study of both naked DNA in solution and of DNA-ligand complexes as they allow, in a comparatively short amount of time, the computation of a large number of atomic trajectories. Times series of microsecond, in simulation times, or billions of time steps are now standard thanks to the improvement from both a software and hardware engineering point of view. Given a potential energy $E : \mathbb{R}^{3L} \to \mathbb{R}$, with $L \gg 1$ representing the total number of atoms in the system, MD simulation use numerical methods to integrate Newton's second law of motion

$$-\frac{\partial}{\partial r_i} U(r_1, \dots, r_L) = m_i \frac{d^2}{dt^2} r_i, \ \forall i = 1, \dots, L, \tag{3.1}$$

where $r_i \in \mathbb{R}^3$ is the Cartesian position of the $i$th atom, $m_i \in \mathbb{R}$ its mass, and $-\frac{\partial}{\partial r_i} U(r_1, \dots, r_L) = F_i \in \mathbb{R}^3$ the force acting on it. In past years a lot of effort has been put in the derivation of accurate potentials $U$ and consequently the derivation of the *force field* $F$ [59, 24].
Let us consider a number $L$ of atoms, and let $\mathbf{r} = (r_1, \dots, r_L)$ be their Cartesian coordinates. The potential energy $U(\mathbf{r})$ is in general assumed to have two distinct contributions: an energy coming from *covalently bonded interactions*, denoted by $U_b(\mathbf{r})$, and anther potential coming from *non covalently bonded interactions*, denoted $U_{nb}(\mathbf{r})$. The potential energy $U_b(\mathbf{r})$ is in general defined by

$$U_b(\mathbf{r}) = \sum_{\text{bonds}} k_b(d_b - \widehat{d}_b)^2 + \sum_{\text{angles}} k_a(\theta_a - \widehat{\theta}_a)^2 + \sum_{\text{dihedrals}} \frac{V_d}{2}(1 + \cos(n_d \phi_d - \delta_d)), \tag{3.2}$$

where the first term is a harmonic potential modelling the elastic energy of a covalent bond, the second term is again a harmonic potential modelling the elastic energy of angles between two bonds, and the last term accounts for the contribution of dihedral, torsion, angle potentials. In figure 3.1 we show a schematic representation of the three

different terms. In [39] one can find the detailed definition of all the parameters.

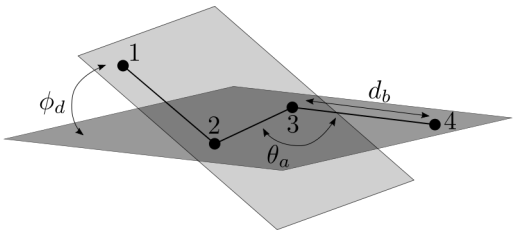



Figure 3.1 – Schematic representation of four atoms, numbered from 1 to 4. The dihedral angle between the atoms [1 2 3 4] is denoted $\phi_d$, the bond angle between atoms [2 3 4] is denoted by $\theta_a$ and the bond distance between atoms [3 4] is denoted by $d_b$.

The non-bonded energy potential in general reads

$$U_{nb}(\mathbf{r}) = \sum_{i<j} \frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}} + \sum_{i<j} \frac{q_i q_j}{\epsilon d_{ij}}, \tag{3.3}$$

where the first term is the van der Waals force accounting for the attractive and repulsive forces between a pair of atoms approximated by the Lennard–Jones potential, and the second term is related to electrostatic interactions between atoms and is represented by the Coulomb potential. Again in [39] the reader can find all the details about the parameters in (3.3). For the full atomistic simulation of DNA the most commonly used simulation programs are AMBER [53, 11] and CHARMM [10].

## 3.2 The ABC collaboration and simulation protocol

The Ascona B–DNA consortium, or ABC, was an international collaboration between groups with the aim to build a shared pool of MD trajectories of linear fragments of DNA, which we will refer to as the ABC data set. The consortium designed a set of 39 sequences of length 18 base–pairs in such a way to have multiple instances of all independent 136 tetramer sub–sequences without counting the end dimers. In table B.1 one can find the complete list of sequences. An important goal of the ABC collaboration was also to establish a single, consistent, MD simulation protocol to be used in all the 39 simulations. In table C.1 we report the most pertinent, for this work, parameters of the ABC protocol. The ABC collaboration lead to a series of four articles [52, 37, 8, 15]. The most recent analysed a set of one microsecond MD simulations of the ABC library. The main conclusions of [52] are that there are indeed strong sequence effects at the tetranucleotide level on the distributions of the standard helical parameters, as was already previously observed in shorter duration simulations, and that microsecond simulations are apparently long enough for the statistics of many observed quantities to have converged. The longer simulations also confirm a

phenomenon observed in the previous shorter duration ABC simulations [37, 55, 19], namely that within some sequence contexts, histograms along the time series of some of the standard 12 DNA helical parameters deviate in a noticeable way from Gaussian distributions. The long simulations of [52] allowed the authors to characterise the deviations from Gaussianity in terms of type of helical parameter and of dinucleotide step in the purine–pyrimidine alphabet $(R/Y)$. More precisely, only three of the helical parameters, namely shift, slide and twist, and only at junctions that are either $RR$ or $YR$ dinucleotide steps deviate from Gaussianity, but never at $RY$ steps. The non-gaussianity is related to a double-well phenomenon at both $RR$ and at $YR$ steps, with the relative occupancies of each well strongly linked to the tetranucleotide flanking sequence of the given dinucleotide step. Moreover the well occupancies are very highly correlated with a bimodal behaviour of a specific phosphate group (corresponding to the so-called BI-BII transition of the local backbone angles), with the phosphate being in the junction between $RR$ steps, and in a neighbouring junction for $YR$ steps. The BI–BII transition has been investigated in more detail in [4, 5].

## 3.3 Analysis of trajectories

The ABC data set comprises a large number of MD trajectories that can be used to compute, firstly time series of rigid body double chains, and secondly time series of internal coordinates. Hence, for a sequence $\mathcal{S}$ in the ABC data set and for each snapshot $k$ one can compute

$$\mathfrak{R}^{\cdot\cdot}\left(\mathbf{r}^{(k)}(\mathcal{S})\right) = (\mathbf{g}^{+},\mathbf{g}^{-})^{(k)}[\mathcal{S}] \in SE(3)^{2N}, \tag{3.4}$$

$$\mathfrak{R}\left((\mathbf{g}^{+},\mathbf{g}^{-})^{(k)}[\mathcal{S}]\right) = (x,y)^{(k)}[\mathcal{S}] = w^{(k)}(\mathcal{S}) \in \mathbb{R}^{12N-6}, \tag{3.5}$$

where $\mathfrak{R}^{\cdot\cdot}$ was defined in (2.54) and $\mathfrak{R}$ in (2.69). Finally for each sequence we obtain a time series of internal coordinates denoted by $\{w^{(k)}(\mathcal{S})\}_{k=1}^{M}$, $M \gg 1$. We can now compute two standard statistics, namely the first and centred second moment, or covariance matrix, by computing

$$\mu(\mathcal{S}) := \frac{1}{M}\sum_{n=1}^{M} w_n^{(k)}(\mathcal{S}), \tag{3.6}$$

$$C(\mathcal{S}) := \frac{1}{M}\sum_{n=1}^{M} (w_n^{(k)}(\mathcal{S}) - \widehat{w}(\mathcal{S}))^{T}(w_n^{(k)}(\mathcal{S}) - \widehat{w}(\mathcal{S})), \tag{3.7}$$

(the fact that the estimate (3.7) with weight $\frac{1}{M}$ is biased is inconsequential for us as $M \sim 10^6$). As both estimations are done for each sequence separately we will refer to the couple $(\widehat{w}, C)[\mathcal{S}]$ as *oligomer–based statistics*.
Two natural analyse can be done on the oligomer–based statistics: the first has been

already mentioned in section 3.2, namely the study of the one–dimensional marginal histogram for each component of the helical parameters, the second is the study of the inverse of the covariance matrix, or precision matrix or, for us stiffness matrix. For both cases we show an example in figures 3.2 and 3.3. The sparsity pattern of the observed inverse covariance is close to 18 times 18 block diagonal with 6 times 6 overlaps. This behaviour can be observed for any arbitrary sequence, and played a central role in the derivation of the cgDNA model, see [55, 19] and chapter 4 of this work. In the next section we will show how to impose the pattern to the observed stiffness matrix. The one-dimensional histograms reveals that some of the helical parameters deviate noticeably from Gaussian. As already presented in section (3.2) the non-gaussianity appears only in the inter–base–pair parameters and for just for some of the helical parameters. For those parameters the deviation from Gaussianity can be characterised in terms of its junction step expressed in the purine-pyrimidine alphabet, and in term of the flanking sequences. The histograms are then quite consistent independent of the location of the given tetranucleotide sequence contest provided it is sufficiently far from an ends.

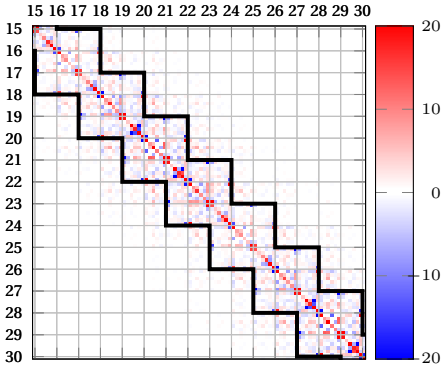



Figure 3.2 – Portion of the inverse covariance matrix observed from $1\mu s$ simulations for sequence $\mathcal{S}_4$, see table B.1. The black lines highlights the sparsity pattern corresponding to nearest neighbour interactions.

The computational estimation of higher moments in a high–dimensional space, such as $\mathbb{R}^{12N-6}$ with $N = 18$ is far from being a trivial exercise. We therefore, consider only the two first moments; the mean and the covariance. For example for a sequence $\mathcal{S}$ in the ABC training set we have, thanks to the maximum entropy principle [42, 25, 26, 27, 20], that the Gaussian distribution $\rho(w, \theta(\mathcal{S}))$ with $\theta(\mathcal{S}) = \text{param}(\mu, C^{-1})[\mathcal{S}]$ is the distribution that maximizes the entropy function $S$, defined by

$$S(p) = -\int_{\mathbb{R}^{12N-6}} p(w) \log p(w) dw, \tag{3.8}$$

under the constraint that the distribution matches first and second observed moment,

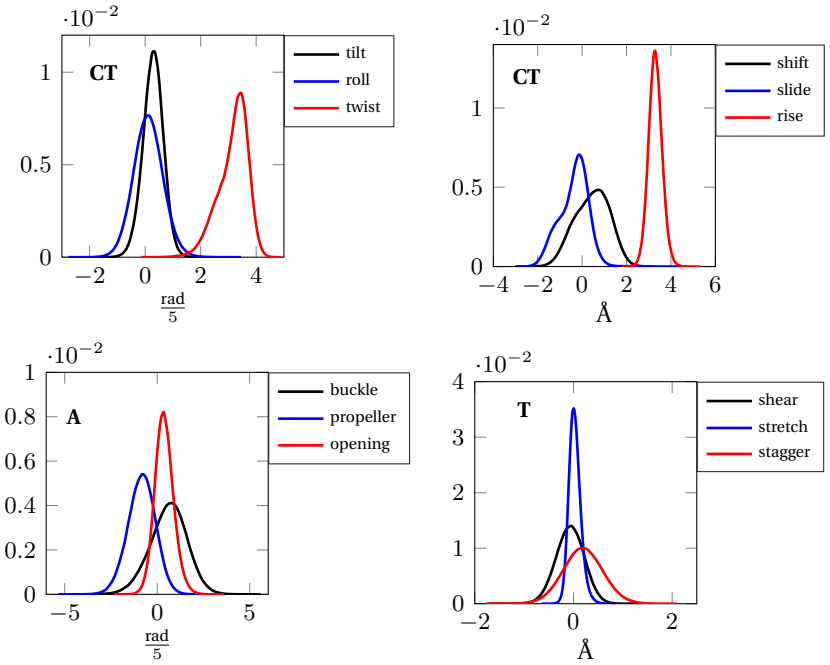



Figure 3.3 – We present four examples of histogram computed for a one dimensional marginal for sequence $\mathcal{S}_4$, see table B.1. The top line corresponds to inter–base–pair coordinates, while the second are intra–base–pair coordinates. In the first column we show the rotational coordinates while in the second the translations.

i.e, under the constraint that the distribution belongs to

$$\mathcal{C}(\mathcal{S}) = \left\{ p(w) |\, \langle 1 \rangle_p = 1 \,,\; \langle w \rangle_p = \mu \,,\; \langle (w-\mu)^T (w-\mu) \rangle_p = C \right\}. \tag{3.9}$$

More precisely

$$\rho(w, \theta(\mathcal{S})) = \underset{p \in \mathcal{C}(\mathcal{S})}{\operatorname{argmax}}\, S(p). \tag{3.10}$$

In an equivalent way one can also estimate the distribution $\rho(w, \theta(\mathcal{S}))$ by means of maximum likelihood estimation. The main difference between the two estimation procedures is that in the maximum entropy principle we do not assume a priori that the solution is a Gaussian distribution, while for maximum likelihood we assume a normal distribution in order to have a tractable problem.

### 3.3.1 Hydrogen bond filtering

First and second moments are estimated from a time series of internal coordinates (2.67) extracted from MD simulations. We recall that the rotational components are

computed using the Cayley vector transformation (2.15)). If the relative rigid–body rotation is closely to being a rotation through $\pi$, the norm of the corresponding Cayley vector tends to the infinity. Such problems arise in the MD simulations of DNA especially close to the ends of the molecule, where, depending on the end dimer, the two strands can open and the base–pair hydrogen bonds are consequently broken quite frequently. In order to avoid significant bias in the estimations of the moments,

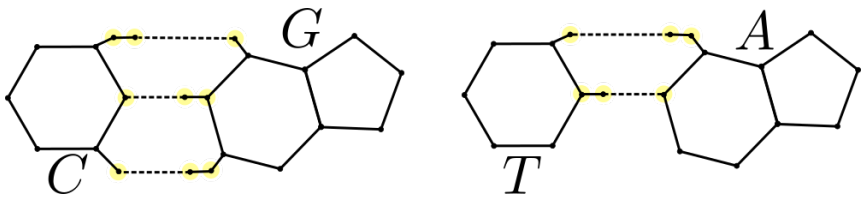



Figure 3.4 – Schematic representation of the hydrogen bonds between $G$ and $C$, left, and between $A$ and $T$, right. The yellow circles highlight the atoms that are considered in the filtering process.

we introduce a filtering step in post–processing the time series data which detects and discard snapshots with fraying ends or any other anomaly which could lead to very large Cayley vectors for that configuration. The procedure simply computes the distances $d_{HB}$ and the bond angles $\theta_{HB}$ between the atoms forming the base–pair hydrogen bonds of all the base–pairs in the molecule. Then a hydrogen bond of a base–pair is declared broken if 1) $d_{HB} > 4$ Å or 2) $\theta_{HB} > 120°$. We discard all the snapshots that have one or more broken hydrogen bonds in any base–pair. In figure 3.4 we show a schematic representation of the atoms considered for the hydrogen bond filtering for both Crick–Watson base–pair pairing.

## 3.4 Estimation of banded stiffness matrix

In the previous section we have presented a way of estimating first and (centred) second moments from a time series of internal coordinates. Moreover, in figure 3.2 we showed an example of a stiffness matrix estimated from a MD time series of internal coordinates for sequence $\mathcal{S}_4$ of the ABC training library, see table B.1. In this figure it is clear that the matrix is not banded, but a consistent block structured pattern, highlighted in figure 3.2, suggests that the assumption of nearest–neighbour interactions in the energy corresponds to overlapping $18x18$ blocks is a quite accurate approximation. Thus, the need to estimate banded stiffness matrices from an observed dense one. We describe now a simple and yet elegant way of doing so in the specific case of the cgDNA sparsity pattern, i.e, $18 \times 18$ diagonal blocks with $6 \times 6$ overlaps. The details and the general proof for what follows can be found in [18].
Let $C(\mathcal{S}) \in \mathbb{R}^{(12N-6)\times(12N-6)}$ be an observed covariance matrix and let us rename the blocks forming the cgDNA sparsity pattern by $C_i$, $i$ odd, for the $i$th $18 \times 18$ block and $C_j$, $j$ even, the $j$th $6 \times 6$ overlapping block. In figure 3.5 we show a schematic represen-

tation of $C(\mathcal{S})$ for helping the reader. Moreover, to each block $C_i$ we associate a set of indices $\text{ind}_i \in \mathbb{N}^{n_i}$ indicating its location in the matrix $C(\mathcal{S})$, where $\#n_i = 6$, for $i$ even and $\#n_i = 18$, for $i$ odd. For example $\text{ind}_1 = 1, \ldots, 18$ and thus $(C(\mathcal{S}))_{(\text{ind}_1,\text{ind}_1)} = C_1$. The procedure to construct a banded stiffness matrix, denoted $K_{\text{band}}(\mathcal{S})$ is given in

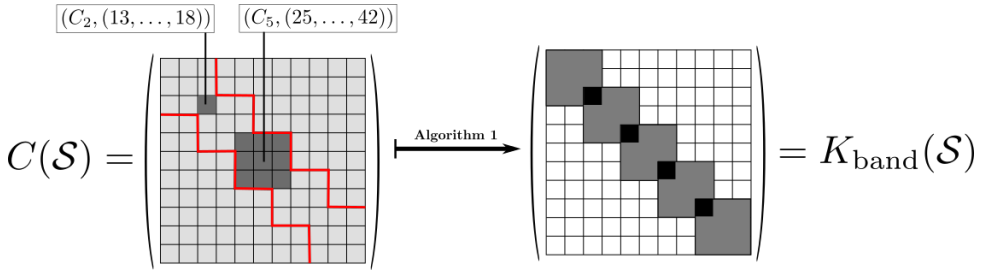



Figure 3.5 – Schematic representation of algorithm 1

algorithm 1 and will only use the blocks $C_i$ and $C_j$ of $C(\mathcal{S})$. The theory behind the

**Algorithm 1** Banded stiffness estimation

**Given:** $C(\mathcal{S}) \in \mathbb{R}^{(12n-6)\times(12n-6)} \leftrightarrow \{(C_1, \text{ind}_1), \ldots, (C_n, \text{ind}_n)\}$, $n = 2N - 3$,
**for** $k = 1 : 2N - 3$ **do**
  Compute $K_{tmp} = [C_k]^{-1}$
  **if** k is even **then**
    $K_{tmp} = -K_{tmp}$
  **end if**
  $(K_{\text{band}})_{(\text{ind}_k,\text{ind}_k)} = (K_{\text{band}})_{(\text{ind}_k,\text{ind}_k)} + K_{tmp}$
**end for**
**Output:** $K_{\text{band}} \in \mathbb{R}^{(12N-6)\times(12N-6)}$, $K_{\text{band}} = K_{\text{band}}^T > 0$.

algorithm in fact provide a characterisation of a banded matrix in terms of the entries of its inverse. More precisely, let us consider $\widetilde{C} = (K_{\text{band}})^{-1}$: the theorem that lead to the derivation of the algorithm 1 reveals that the block outside the stencil of $\widetilde{C}$ are all functions of the blocks inside its stencil. Again more detail can be founded in [18, 20].

# 4 cgDNA: a sequence–dependent rigid–base model for DNA

In [55, 19] the authors introduced a sequence–dependent, rigid–base, coarse–grain model of B–form DNA called the cgDNA model. The cgDNA model is parametrized from full atomistic molecular dynamics simulations of a set of sequences of short length. Given a parameter set $\mathcal{P}$ and an arbitrary DNA sequence $\mathcal{S}$, cgDNA predicts a Gaussian equilibrium probability density function in configuration space, by reconstructing the mean $\mu \equiv \mu(\mathcal{P},\mathcal{S}) \in \mathbb{R}^N$, or ground–state, and the precision matrix $K \equiv K(\mathcal{P},\mathcal{S}) \in \mathbb{R}^{N\times N}$, or stiffness matrix:

$$\rho(w;\mathcal{P},\mathcal{S}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}(w-\mu)\cdot K(w-\mu)\right\}. \tag{4.1}$$

In this section we briefly review the main assumptions underpinning the cgDNA model and the reconstruction of the density (4.1). We will then discuss the structure and properties of the parameter set $\mathcal{P}$ and compare predictions of the model and MD observables. In particular we will present a study on sequence–dependent persistence length of DNA using the cgDNA model.

## 4.1 Main assumptions underlying the cgDNA model

The cgDNA model is based on many assumption that will be listed and briefly discussed hereafter. We start with the assumptions on the chemical structure of the molecule of DNA and in particular on the form of the DNA. The cgDNA model [55, 19] is based on molecular dynamics simulations of double stranded B–form DNA fragments and consider only Crick-Watson pairing for the bases, and the standard alphabet of bases $\{A,T,G,C\}$. The final assumption on the chemical structure is about the rigidity of the bases, which was already discussed in section 1.3. The Curves+ software [38] is used in the fitting procedures. The convention about ideal atom coordinates are reported in appendix A.
As already mentioned the cgDNA model is a rigid–base model of DNA, fixing the level

of coarse graining to single bases. More than an assumption, the latter is a modelling decision based on observation. In section 2.4 we presented the mathematics behind bichains and in particular the relation between the rigid body representation and internal coordinates. When considering a bichain representation of a $n$ base–pair long DNA sequence $\mathcal{S}$, the internal coordinates $w(\mathcal{S}) \in \mathbb{R}^{12n-6}$ satisfy the following physical property related to the Crick-Watson symmetry,

$$w(\mathcal{S}) = E_{2n-1} w(\overline{\mathcal{S}}), \tag{4.2}$$

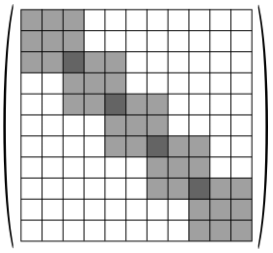
where $\overline{\mathcal{S}}$ is the complementary sequence of $\mathcal{S}$ and $E_{2n-1} \in R^{(12n-6)\times(12n-6)}$ is a block, trailing diagonal matrix composed by $2n-1$ copies of $E = \text{diag}(-1,1,1,-1,1,1) \in R^{6\times 6}$

$$E_{2n-1} = \begin{bmatrix} & & & E \\ & & E & \\ & \iddots & & \\ E & & & \end{bmatrix}, \tag{4.3}$$

where $E_{2n-1} = E_{2n-1}^T = E_{2n-1}^{-1}$.
The next assumption is motivated by the modelling choice of coarse–graining to the level of individual rigid bases. In what follow $\mathcal{S} = X_1 X_2 \dots X_n$, $X_i \in \{A,T,G,C\}$ is a sequence. The internal energy for $\mathcal{S}$ will assume to be a shifted quadratic function in the internal coordinates $w = (y_1, x_1, y_2, \dots, y_n) = (x, y)$

$$U_{\text{tot}}(w;\mathcal{S}) = \frac{1}{2}(w-\mu)\cdot K(w-\mu), \tag{4.4}$$

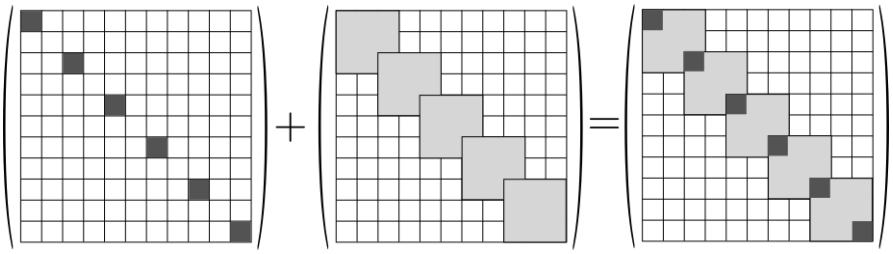
where $\mu \equiv \mu(\mathcal{S}) \in \mathbb{R}^{12n-6}$ is the ground-state of $\mathcal{S}$, or its minimal configuration energy, and $K = K^T \equiv K(\mathcal{S}) \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a positive definite matrix called the stiffness. Based on observation of statistics estimated from MD trajectories, see for instance figure 3.2, we assume that the stiffness matrix $K$ is banded, i.e is a sparse matrix in which non–zeros entries are all close to a diagonal band. More precisely the sparsity pattern of $K$ is $18 \times 18$ block diagonal with $6 \times 6$ overlaps:

K= 

The cgDNA model further assumes that the free energy has local contributions of the form:

$$U_{\text{local}}(w;\mathcal{S}) = \sum_{i=1}^{n-1} \frac{1}{2}(w_i - \mu_i^{X_iX_{i+1}}) \cdot K^{X_iX_{i+1}} (w_i - \mu_i^{X_iX_{i+1}}) + \sum \frac{1}{2}(y_i - \mu^{X_i}) \cdot K^{X_i}(y_i - \mu^{X_i}) \tag{4.5}$$

where, $w_i = (y_i, x_i, y_{i+1})$ and the two kind of contributions are: *dimer–based* and *monomer-based*. Now, $U_{\text{local}}$ can be written as a single shifted quadratic form and can be compared to $U_{\text{tot}}$. What one finds is that the stiffness matrix $K$ is simply computed as the sum of the dimer and monomer blocks, as shown in the next scheme,



and the ground–state is equal to

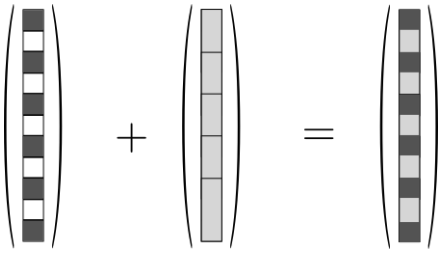
$$\mu = K^{-1}\sigma, \tag{4.6}$$

where $\sigma$ is a vector which has dimer and monomer contributions defined by

$$\sigma^{X_iX_{i+1}} = K^{X_iX_{i+1}} \mu_i^{X_iX_{i+1}}, \tag{4.7}$$

$$\sigma^{X_i} = K^{X_i} \mu_i^{X_i}, \tag{4.8}$$

with the definition of $\sigma$ summarized using the following scheme



We want now to focus on one important point: $\mu$ has a non local dependence on the entries of $\mu^{X_iX_{i+1}}$ and $\mu^{X_i}$ due to $K^{-1}$, as the inverse of a band matrix with overlaps being in general non banded. Moreover, by completing the square in (4.5) a non zero constant term $\widehat{U}$ will naturally appear reflecting the fact that in the ground–state all the interactions of each base cannot simultaneously vanish. Hence, the oligomer–based

local energy (4.5) is also a model for frustration. More detail can be found for example in chapter 6 of [55].

## 4.2 The cgDNA parameter set

The cgDNA parameter set is the set of weighted shape vectors and symmetric matrices of the form

$$\mathcal{P} = \{\sigma^{\alpha}, \sigma^{\alpha\beta}, K^{\alpha}, K^{\alpha\beta}\}_{\alpha,\beta\in\{A,T,G,C\}}. \tag{4.9}$$

By assuming Crick-Watson symmetries we can reduce the size of $\mathcal{P}$ to contain only independent elements. For the monomer-dependent elements $\alpha \in \mathrm{M} = \{A, G\}$ and $\alpha\beta \in \mathrm{D}$ where D contains the four palindromic dimers and six independent non palindromic dimers. In particular D must satisfy:

$$\text{if } \alpha\beta, \overline{\alpha\beta} \in \mathrm{D} \Rightarrow \alpha\beta = \overline{\alpha\beta}, \text{and} \tag{4.10}$$

$$\text{if } \alpha\beta \notin \mathrm{D} \Rightarrow \overline{\alpha\beta}. \in D \tag{4.11}$$

Now, given an arbitrary DNA sequence $\mathcal{S} = X_1X_2, \ldots, X_n$ we can use $\mathcal{P}$ to reconstruct the ground–state $\mu \equiv \mu(\mathcal{P}, \mathcal{S})$ and the stiffness matrix $K \equiv K(\mathcal{P}, \mathcal{S})$ using the following reconstructions rules:

$$K(\mathcal{P}, \mathcal{S}) = P_d^T K_d P_d + P_m^T K_m P_m, \tag{4.12}$$

$$\sigma(\mathcal{P}, \mathcal{S}) = P_d^T \sigma_d + P_m^T \sigma_m \tag{4.13}$$

$$\mu(\mathcal{P}, \mathcal{S}) = K(\mathcal{P}, \mathcal{S})^{-1}\sigma(\mathcal{P}, \mathcal{S}), \tag{4.14}$$

where

$$\begin{aligned} K_d &= \mathrm{diag}(K^{X_1X_2}, \ldots, K^{X_{n-1}X_n}), \\ K_m &= \mathrm{diag}(K^{X_1}, \ldots, K^{X_n}), \\ \sigma_d &= (\sigma^{X_1X_2}, \ldots, \sigma^{X_{n-1}X_n}) \\ \sigma_m &= (\sigma^{X_1}, \ldots, \sigma^{X_n}). \end{aligned}$$

The matrices $P_d \in \mathbb{R}^{18(n-1)\times(12n-6)}$ and $P_m \in \mathbb{R}^{6n\times(12n-6)}$ take the following form:

$$P_d = \begin{bmatrix} I & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & 0 & 0 & & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & 0 & I & 0 & & 0 \\ 0 & 0 & 0 & 0 & I & & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & & I \end{bmatrix}, \; P_m = \begin{bmatrix} I & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & I & 0 & 0 & & 0 \\ 0 & 0 & 0 & 0 & I & & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & & I \end{bmatrix}. \tag{4.15}$$

In chapter 5 we will present the methodology used in [55, 19] to derive a best–fit cgDNA parameter set $\mathcal{P}$ trained on the ABC data set. In this chapter we will just show an example of comparison between prediction of the cgDNA model and observations from simulations in the ABC data set. In particular we have (randomly) selected sequence $\mathcal{S}_4$, and its MD observed ground–state and stiffness matrix. In figures 4.1 we compare the predicted cgDNA ground–state with the observed ground–state of $\mathcal{S}_4$. For visualisation purpose we have divided the shape vectors into rotational and translational components of intra– and inter–base–pair variables and plot each component individually. We remark an excellent agreement between prediction and observation. In order to compare the stiffness matrix we choose to use the tangent–tangent correlation as way of comparison.

## 4.3 Study of persistence length using the cgDNA model

In section 2.2 we introduced the concept of sequence–dependent apparent and dynamic persistence length. Evaluation of the persistence length is one way to compare stiffness, but first some remarks should be made. We first stress that the sequence–dependent Flory persistence length is computed as the limit of norm of the Flory persistence vector. In practice the sequence–dependent Flory persistence length can be computed for DNA fragments with length of order $10^3$. The length of the sequence simulated in the MD and in particular in the ABC project vary between 12 and 18 base–pairs. Thus, in what follow we will consider only persistence lengths computed by means of the tangent–tangent correlation function, and we will drop the "sequence–dependent" adjective in front of persistence length as we are in the framework of the cgDNA sequence–dependent model.
We reconsider again $\mathcal{S}_4$ and in particular its cgDNA reconstructed Gaussian $\rho(w;\mathcal{P},\mathcal{S}_4) \equiv \rho_m(w)$, and the observed banded Gaussian, $\rho(w;\mathcal{S}_4) = \rho_o(w)$. We can now numerically compute the tangent–tangent correlation with respect to $\rho_m$ and $\rho_o$ along with the static persistence lengths of both ground–states and plot the results. In figure 4.2, left, we show the comparison for $\mathcal{S}_4$ and we can observe that the prediction of the model

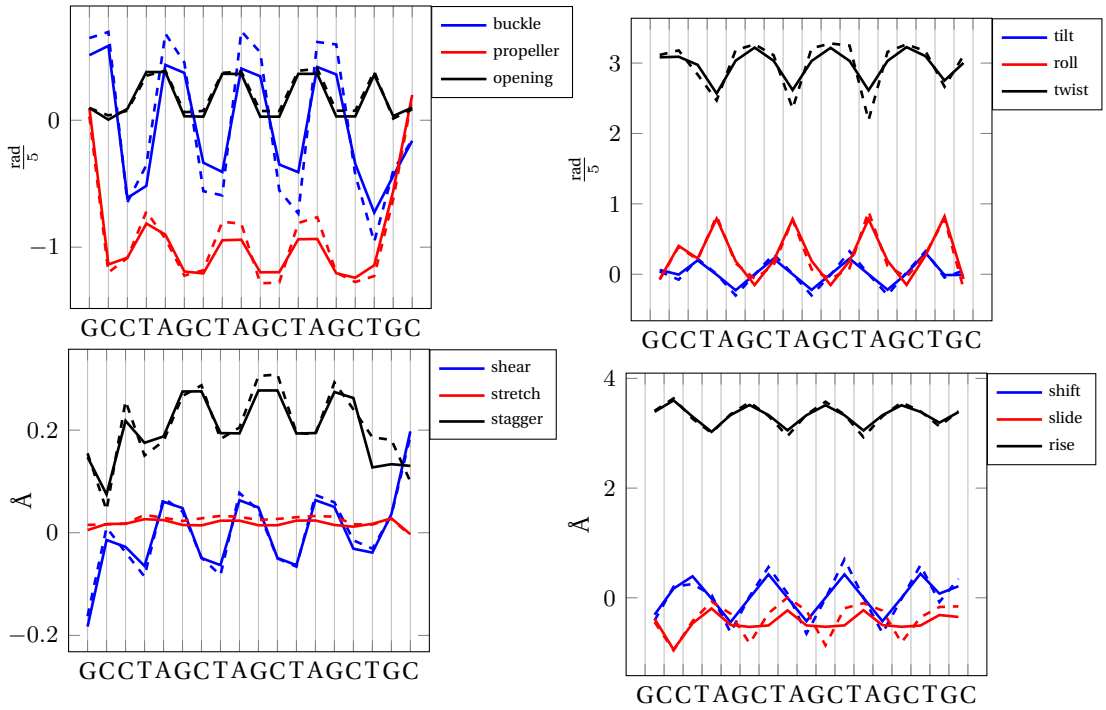



Figure 4.1 – Comparison between cgDNA (*solid*) and MD (*dashed*) ground–states for sequence $\mathcal{S}_4$.

is again in good agreement with the data. The value of the apparent and dynamic persistence length are respectively **130.4**/129.9 base-pairs (bp), **157.3**/156.4 bp, where the bold values are the prediction of cgDNA. Compared to the comparison of the ground–states done in the previous section, both persistence lengths are non–trivial functions of the internal coordinate and the stiffness matrix which enhances the difficulty for this kind of comparison. We also recall that the sequence $\mathcal{S}_4$ was in the training library used in parameter extraction procedure for obtaining $\mathcal{P}$, but no fitting of the persistence length was done. Finally, we refer again to [18, 45] for the details about the convergence of the tangent–tangent correlation as function of number of drawn configurations. In figure 4.2, right, we present another example of tangent–tangent correlation computations done using cgDNA reconstructed Gaussian for the sequence $\mathcal{S}_{\text{A-tr}} = (A_6CGCGA_6CGGGC)_3$, where $X_n$ means $n$ repetitions of the sequence $X$, see figure 4.3. As already mentioned in section 2.2, the approximation of the apparent persistence length from the log-ttc plot is in general poor for sequences with a high intrinsic bend, as for example, phased $A$–tracts sequences. In figure 4.2, right, we also illustrate the efficiency of the shape factorization which leads to a more linear decay and thus, to a good and robust approximation of the dynamic persistence length. For $\mathcal{S}_{\text{A-t}}$ we obtained the following values for the persistence length: $\ell_p = 66$ bp, $\ell_d = 196$ bp. We stress again that for bent sequences the apparent persistence length will be in general an under estimate due to the bad quality of the linear fit in a semi–log plot.

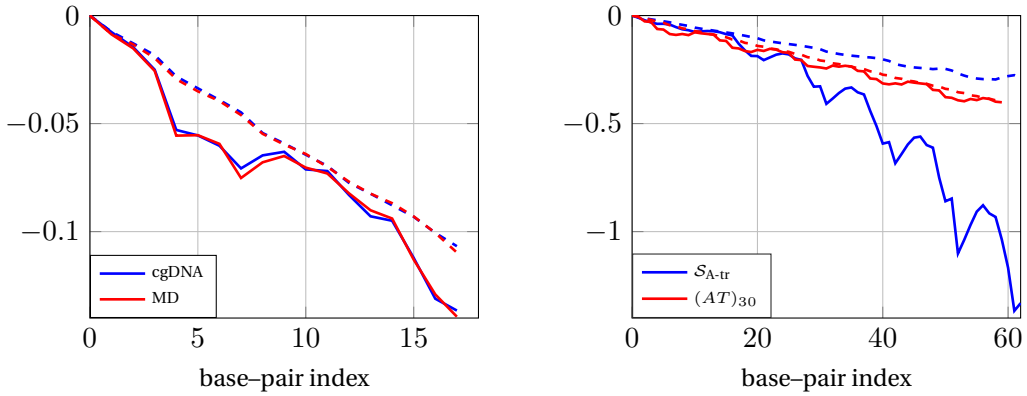



Figure 4.2 – Example of tangent–tangent correlation in *solid* and with shape factorization in *dashed*. Left: comparisons for $\mathcal{S}_4$ between cgDNA and MD. Right: comparison between $(AT)_{30}$ and $\mathcal{S}_{\text{A-tr}}$.

To better illustrate better the unreliability of the apparent persistence length approximation we consider the poly dimer sequence $(AT)_{30}$ whose bichain representation of the ground–state is shown in figure 4.3. As the ground–state is intrinsically straight we observe a very linear decay in the log–ttc, see for instance (4.2) right. Moreover, we obtained the following values for persistence length: $\ell_p = 137$ bp and $\ell_d = 147$ bp. By comparing the values of persistence length obtained for the two sequences, $\mathcal{S}_{\text{A-tr}}$ and $(AT)_{30}$, it is clear that the only comparison that make sense is the one between the values of $\ell_d$, and interestingly enough, the A–tract sequence has a higher value.

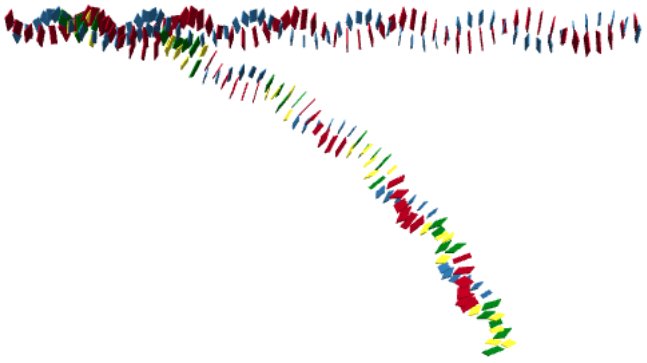


Figure 4.3 – Bichain representation of the ground–state of $\mathcal{S}_{\text{A-tr}} = (A_6CGCGA_6CGGGC)_3$, the bent one, and $(AT)_{30}$, the straight one.

In section 2.2 we also introduced the notion of sequence–averaged persistence lengths. Thus, we randomly generated an ensemble of $10^4$ 220 base–pair long sequences by assigning the same probability to each base. Then we computed the apparent and dynamic persistence lengths for every sequence in the ensemble and plotted the

resulting spectra in figure 4.4. Moreover, we studied also the persistence length of six independent poly dimer sequences because their ground–states are intrinsically straight and thus a direct comparison between both definitions of persistence lengths is possible. In figure 4.4 we show all the results. In blue we have the spectra for $\ell_p$ while in red we have the one for $\ell_d$. The values of $\ell_p$ and $\ell_d$ for each independent poly dimer and the sequence average over the ensemble are reported respectively in italic and bold. The first observation is that the shape of the two spectra are completely different. In particular the spectra for $\ell_p$ is flat and reach very low values ($< 120$) which are related to bent sequences. In contrast the spectra of $\ell_d$ is more peaked with some very large values ($> 190$). The sequence averaged values of both persistence lengths are indicated by *Avg*($= 160$ bp) and **Avg**($= 178$ bp) respectively for $\ell_p$ and $\ell_d$, see table 4.1 for the values of the poly–dimers. But, in both cases it is clear that the sequence–dependence plays a central role in the study of the *rigidity* of DNA, and thus, in the context of sequence–dependent modelling of DNA, both sequence–dependent definition of persistence length could help studying and understanding the mechanical property of DNA.

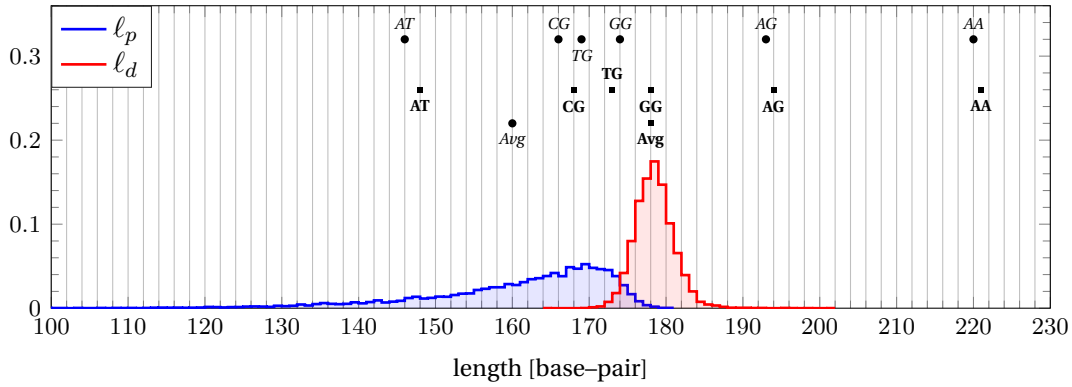



Figure 4.4 – Histograms of apparent (blue) and dynamic (red) persistence lengths computed using the cgDNA model (trained on the ABC data set) over a sequence ensemble of 10K randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence length. The values of the persistence lengths for six independent poly–dimers of length 220 are reported: italic for the apparent and bold for the dynamic. The positions of the values of the apparent persistence length is given by a circle while the positions for the dynamic is given by a square.

| | AA | AG | GG | TG | CG | AT | Avg |
|---|---|---|---|---|---|---|---|
| $\ell_p$ | 220 | 193 | 174 | 169 | 166 | 146 | 160 |
| $\ell_d$ | 221 | 194 | 178 | 173 | 168 | 148 | 178 |

Table 4.1 – Values of $\ell_p$ an d $\ell_d$ for six poly–dimers and the sequence–average. The values are expressed in base–pairs.

# 5 Estimation of cgDNA parameter sets

## 5.1 Kullback-Leibler divergence

Let $p(x)$ and $q(x)$ be two continuous multivariate probability density functions defined on $\Omega \subset \mathbb{R}^N$. The Kullback–Leibler divergence (KLd), or relative entropy, between $p$ and $q$ is [33, 34]

$$D_{KL}(p(x), q(x)) = \int_{\Omega} p(x) \log \frac{p(x)}{q(x)} dx. \tag{5.1}$$

KLd is in general non symmetric, namely $D_{KL}(p(x), q(x)) \neq D_{KL}(q(x), p(x)))$, it vanishes if and only if $p = q$, and it is positive for any two densities $p \neq q$, namely $D_{KL}(p(x), q(x)) > 0$. The fact that (5.1) is not symmetric and, moreover, does not satisfy the triangle inequality, implies that KLd does not define a metric, but only a premetric on the set of probability density functions. KLd is invariant under rescaling meaning that if $X_p, X_q$ are the random variables associated to $p, q$ and $\tilde{X}_p = MX_p$, $\tilde{X}_q = MX_q$ are rescaled random variable respectively associated to $\tilde{p}, \tilde{q}$ we have that

$$D_{KL}(p(x), q(x)) = D_{KL}(\tilde{p}(x), \tilde{q}(x)). \tag{5.2}$$

The latter property has two direct consequences in the context of modelling DNA mechanics. The first is the rescaling factors introduced in the definition of the Cayley transformation (2.15), more precisely, the rescaling of the rotational coordinates by a factor of 5 used in the cgDNA model [55, 19] does not affect the values of (5.1). The second is the linear relation between the coordinates of a sequence and the coordinates read from its complementary shown in (4.2), and thus the invariance of the KLd under change of reading strand.
Another, essential for us, feature of KLd is that is has an explicit algebraic form when both probability density functions $p$ and $q$ are multivariate normal distributions. More precisely if $p(x) \approx \mathcal{N}(\mu_p, C_p)$ and $q(x) \approx \mathcal{N}(\mu_q, C_q)$ the KLd between the two Gaussians

can be written as

$$D_{KL}(p,q) = \frac{1}{2}\left(\text{tr}(K_p^{-1}K_q) + \ln\left(\frac{\det K_p}{\det K_q}\right) - N\right) + \frac{1}{2}(\mu_q - \mu_p)^T K_q(\mu_q - \mu_p), \quad (5.3)$$

with $K_p = C_p^{-1}$ and $K_q = C_q^{-1}$. The previous expression can be separated into two parts, i.e $D_{KL} = D^\dagger + \mathcal{M}$ where

$$D^\dagger(p(x),q(x)) = \frac{1}{2}\left(\text{tr}(K_p^{-1}K_q) + \ln\left(\frac{\det K_p}{\det K_q}\right) - N\right), \quad (5.4)$$

$$\mathcal{M}(p(x),q(x)) = \frac{1}{2}(\mu_p - \mu_q)^T K_q(\mu_p - \mu_q). \quad (5.5)$$

The square root of the second term, $\sqrt{\mathcal{M}}$, is called the Mahalanobis distance [14] it measures the distance between the point $\mu_p$ and the distribution $\mathcal{N}(\mu_q, C_q)$. We recall that (5.3) can be expressed as a function of the covariance matrices of $p$ and $q$, but can simple be written in term of their precision matrices, or in our context in terms of their stiffness matrices.
The non symmetry of the KLd clearly implies two different ways of comparing two probability density function. For example, KLd can be used as an objective function for computing parameters of a model density compared to observed ones. More precisely, if $q$ is considered as an observed pdf one can minimise the KLd in order to find a model $p(\cdot;\theta)$ close to $q$. But the choice of the ordering in the argument of the KLd function leads to two different approaches. In particular, the solution of the following problem

$$\min_{\theta\in\mathcal{C}} D_{KL}(q(x),p(x;\theta)), \text{ where } \mathcal{C} \text{ is the contraint space for the parameter } \theta, \quad (5.6)$$

is equivalent to the maximum log-likelihood of the data whereas the opposite order do not relate to any other known method. Now, instead of considering only one observed density we can consider a family of $N$ distinct pdfs denoted by $\{q_i\}_{i=1}^N$ and again, a single model pdf noted by $p(\cdot;\theta)$. We are now interested in the first order conditions related to the minimisation problem and the two different orderings of the KLd, i.e

$$\min_{\theta\in\mathcal{C}} \sum_{i=1}^{N} D_{KL}(q_i(x),p(x;\theta)) \equiv \min_{\theta\in\mathcal{C}} \mathcal{F}_1(\theta;q), \text{ or} \quad (5.7)$$

$$\min_{\theta\in\mathcal{C}} \sum_{i=1}^{N} D_{KL}(p(x;\theta),q_i(x)) \equiv \min_{\theta\in\mathcal{C}} \mathcal{F}_2(\theta;q), \quad (5.8)$$

For sake of simplicity we will consider now that all the probability functions in the previous problems are Gaussian distributions, thus $q_i \approx \mathcal{N}(\mu_i, K_i)$ and $\theta = (\mu_m, K_m)$. Due to the latter simplification we can use the explicit formulation of the KLd for deriving explicit algebraic formulation of both sums (5.7-5.8) and we can easily compute the

first–order conditions for finding explicit forms for the model parameters $\mu$ and $K$. In detail, for Gaussians

$$\mathcal{F}_1(\theta;q) = \frac{1}{2}\left(K_m : \left(\sum_{i=1}^{N} K_i^{-1}\right) + \sum_{i=1}^{N}(\mu_m-\mu_i)\cdot K_m(\mu_m-\mu_i)\right) + N\ln Z \tag{5.9}$$

$$\mathcal{F}_2(\theta;q) = \frac{1}{2}\left(K_m^{-1} : K^{\Sigma} - N\ln|K_m| + (\mu_m - [K^{\Sigma}]^{-1}\sigma^{\Sigma})\cdot K^{\Sigma}(\mu_m - [K^{\Sigma}]^{-1}\sigma^{\Sigma})\right), \tag{5.10}$$

where $K^{\Sigma} = \sum_{i=1}^{N} K_i$ and $\sigma^{\Sigma} = \sum_{i=1}^{N} K_i\mu_i$. Then we can compute the first order necessary conditions for each $\mathcal{F}_i$, $i = 1,2$ to obtain the estimation of the parameter $\theta = \text{param}(\mu_m, K_m)$, i.e, we compute

$$\frac{\partial \mathcal{F}_i}{\partial K_m} = 0, \quad \frac{\partial \mathcal{F}_i}{\partial \mu_m} = 0,\ i = 1,2. \tag{5.11}$$

We finally obtained the resulting formulas

$$\mu_{m,1} = \frac{1}{N}\sum_{i=1}^{N}\mu_i, \qquad K_{m,1}^{-1} = \frac{1}{N}\left(\sum_{i=1}^{N} K_i^{-1} + \mu_i\otimes\mu_i\right) - \mu_{m,1}\otimes\mu_{m,1} \tag{5.12}$$

$$\mu_{m,2} = [K^{\Sigma}]^{-1}\sigma^{\Sigma}, \qquad K_{m,2} = \frac{1}{N}K^{\Sigma}. \tag{5.13}$$

It is interesting to notice first the difference between the estimation of the mean, in particular the second involves the sigma vectors which have an important role in the cgDNA model but which are not directly observable from an ensemble of configuration snapshots generated from MD simulations. The second difference is in the estimation of the covariance for the maximum likelihood way and the estimation of the stiffness for the other choice of argument in the KLd. Away from the global minimum of $\mathcal{F}_i$ there is no reason to believe that the estimators (5.12) and (5.13) give the same results.

## 5.2 Estimation of parameters

In this section we denote a Gaussian probability density function with mean $\mu \in \mathbb{R}^n$ and stiffness matrix $K \in \mathbb{R}^{n\times n}$ by $\rho(x;\theta)$, where $\theta \in \mathbb{R}^N$ is a vector whose components are all the entries of $\mu$ and $K$, and the notation (2.9) will be used

$$\theta = \text{param}(\mu, K).$$

Let now $Lb := \{\mathcal{S}_i\}_{i=1}^{M}$, be a training of sequences library in which we have computed statistics from molecular dynamic trajectories as presented in section 3.3. Thus, for each $\mathcal{S}_i \in Lb$, $i = 1,\ldots,M$ we have estimated a mean $\mu(\mathcal{S}_i)$ and a stiffness $K(\mathcal{S}_i)$ for which the associated Gaussian distribution will be denoted by $\rho(x;\theta_i)$. We recall that a

cgDNA parameter set is denoted by $\mathcal{P} = \{\sigma^{\alpha}, \sigma^{\alpha\beta}, K^{\alpha}, K^{\alpha\beta}\}_{\alpha\in M,\alpha\beta\in D} \in \mathbb{P}_{\text{tot}}$, where

$$\mathbb{P}_{\text{tot}} = [\mathbb{R}^6]^2 \times [\mathbb{S}^6]^2 \times [\mathbb{R}^{18}]^{10} \times [\mathbb{S}^{18}]^{10}, \tag{5.14}$$

where $\mathbb{S}^N$ is the set of $N \times N$ symmetric matrices. We can refine the parameter space by using the Crick–Watson symmetries presented in (4.2). In fact we can define the subset of $\mathbb{P}_{\text{self}} \subset \mathbb{P}_{\text{tot}}$

$$\mathbb{P}_{\text{self}} = \{\mathcal{P} \in \mathbb{P}_{\text{tot}} \mid \sigma^{\alpha\beta} = E_2\sigma^{\alpha\beta},\ K^{\alpha\beta} = E_2K^{\alpha\beta}E_2,\ \forall\alpha\beta \in D^{'}\}, \tag{5.15}$$

where $D^{'}$ contains only the four palindromic dimers. As the goal of the parameter set $\mathcal{P}$ is to reconstruct the parameters of a Gaussian probability density function, another subspace of $\mathbb{P}_{\text{tot}}$ arise naturally by considering the training library. Given the set of sequences $Lb$ it is rational to consider the following subset

$$\mathbb{P}_{\text{train}} = \{\mathcal{P} \in \mathbb{P}_{\text{tot}} |\ K(\mathcal{P}, S_i) > 0,\ \forall i = 1, \ldots, M\}, \tag{5.16}$$

where $K(\mathcal{P}, S_i)$ is reconstructed using the rule defined in (4.12). Finally we define the parameter space of all admissible cgDNA parameter sets as $\mathbb{P} = \mathbb{P}_{\text{self}} \cap \mathbb{P}_{\text{train}}$. Now, given a family of estimated Gaussian densities $\{\rho(x, \theta_i)\}_{i=1}^{M}$, one for each sequence in $Lb$, we defined the best fit cgDNA parameter set $\mathcal{P}$ as the solution of the following optimization problem

$$\mathcal{P} = \underset{\text{P}\in\mathbb{P}}{\text{argmin}}\ \mathbb{F}(\text{P}; Lb) \tag{5.17}$$

where $\mathbb{F}(\mathcal{P}; Lb) : \mathbb{P} \to \mathbb{R}$ is defined as the sum of Kullback-Leibler divergences over the training library, more precisely

$$\mathbb{F}(\mathcal{P}; Lb) = \sum_{i=1}^{M} D_{KL}(\rho(x, \theta_i(\mathcal{P})), \rho(x, \theta_i)), \tag{5.18}$$

where $\rho(x; \theta_i(\mathcal{P}))$ is the Gaussian probability density function in which parameters are reconstructed using the parameter set $\mathcal{P}$ and the rules (4.12–4.14) , and

$$\theta_i(\mathcal{P}) = \text{param}\left(\mu(\mathcal{P}, \mathcal{S}_i, ), K(\mathcal{P}, \mathcal{S}_i)\right), \quad \forall i = 1, \ldots, M. \tag{5.19}$$

We stress here that in [55, 19, 20] the choice of order in the KLd is as follows: model in first position and data in second position, which corresponds to the case (5.8). In order to solve problem (5.17) the use of numerical methods is necessary. Explicit expressions for the gradient and Hessian matrix can be computed for the function (5.18), and a combination of gradient flow and Newton-Broyden methods can be used to solve (5.17). There were many challenges faced when trying to solve (5.17): the first problem is clearly the large dimension of the unknown vector, which is of dimension 1592. The high number of dimensions implies a large number of operations in matrix–vector and

vector–vector multiplication which slows down the entire computational procedure. Moreover, the numerical evaluation of the Hessian matrix of (5.18) is costly, which motivated the use of a quasi–Newton method, namely the Broyden method. Secondly, find a starting point is not a trivial task, meaning, constructing an initial parameter set $\mathrm{P}_{\mathrm{ini}} \in \mathbb{P}$ is not, in general an easy exercise. Later in chapter 9 of this work we will present a new methodology, which allows computations of an admissible initial guess in a rational and rather simple way. Moreover a new numerical scheme to solve problem (5.17) will be presented.

## 5.3 Positiveness of the best–fit parameter set

The main objective of the cgDNA model is to predict a Gaussian distribution for any arbitrary sequence $\mathcal{S}$. This implies in particular that the reconstructed stiffness matrix $K(\mathcal{S})$ must be both symmetric and positive definite for any arbitrary sequence $\mathcal{S}$. If a parameter set $\mathcal{P}$ satisfies both conditions for any arbitrary sequence, we will refer to it as a *positive definite parameter set*. The symmetry is trivially satisfy by the fact that symmetry has been imposed on each block in the derivation of the best–fit parameter set $\mathcal{P}$. Positivity however is not, in general, easy to impose in the numerics, thus an a posteriori criterion is necessary in order to guarantee the positiveness of any matrix $K(\mathcal{S})$. In [20] the authors managed to find a set of sufficient conditions for $\mathcal{P}$ to be positive definite, which are satisfied on the actual estimated parameter set. In detail, let us define the following two matrices

$$K^{\alpha\beta}_{\frac{1}{2}} := K^{\alpha\beta} + \begin{bmatrix} \frac{1}{2}K^{\alpha} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}K^{\beta} \end{bmatrix} \tag{5.20}$$

$$K^{5'\alpha\beta} := K^{\alpha\beta} + \begin{bmatrix} K^{\alpha} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2}K^{\beta} \end{bmatrix} \tag{5.21}$$

for all $\alpha, \beta \in M$ and $\alpha\beta \in \overline{D}$, where $\overline{D}$ comprises all the 16 dimers. Using Crick–Watson symmetries it is sufficient to consider ten independent dimer dependent matrices for (5.20) and just the sixteen for (5.21) because

$$K^{3'\alpha\beta} = E_2 K^{5'\overline{\beta}\overline{\alpha}} E_2. \tag{5.22}$$

We can now state that a best-fit parameter set $\mathcal{P}$ is positive definite if it satisfies the following conditions

$$K^{\alpha\beta}_{\frac{1}{2}} > 0, \ \forall \alpha\beta \in D, \tag{5.23}$$

$$K^{5'\alpha\beta} > 0, \forall \alpha\beta \in \overline{D}. \tag{5.24}$$

# Part II

# Comparison of cgDNA parameter sets

# 6 Sensitivity of cgDNA parameter sets to training data

An MD simulation requires a large ensemble of input parameters that the user should define. Consequently the cgDNA best-fit parameter set $\mathcal{P}$ (4.9) is directly related to the choices made in the MD protocol used in the training library simulations. In this chapter we study how mechanical properties of DNA predicted by the cgDNA model depend upon the *training sets* constructed in the MD simulations. We define a training set as the following ensemble of MD variables

$$\mathfrak{MD} = (St, Lb, Ff, Io), \tag{6.1}$$

where

$St$ is the simulation time,
$Lb$ is the training library of sequence,
$Ff$ is the force field,
$Io$ is the ion type and concentration.

All the other input variables have been fixed to the standard ABC protocol [37, 52]. The methodology we will follow here for comparing different training sets is based on the simplest well–known *one–factor–at–a–time*, or just *one–at–a–time*, sensitivity analysis. It consists simply in studying how the outcome of a model varies as function of its input variables. In our context, we will change one–at–a–time the component of $\mathfrak{MD}$ and the outcome will be some specific predictions of the related cgDNA parameter set $\mathcal{P}$. In the next section we will introduce the training set we will consider and the chain of comparisons we will study. Moreover we will also introduce the predictions we will take into account in order to be able to compare the different parameter sets. Before going further in the comparison we will briefly discuss how in practice the different training sets are computed.
For each training set one has to solve numerically the high dimensional optimization problem (5.17). We have implemented a parameter continuation method in order

to ensure the finding of a solution and thus in order to automatize the computation of new cgDNA parameter sets. The parameter continuation technique consists in studying the solutions of parameter dependent non–linear systems of the form

$$F(w;\varepsilon) = 0, \tag{6.2}$$

as the parameter $\varepsilon \in \mathbb{R}^n$ varies. Let us introduce the weighted objective function $\mathbb{F}(\mathcal{P}, Lb; \omega) : \mathbb{P} \to \mathbb{R}$ defined by

$$\mathbb{F}(\mathcal{P}, Lb; \omega) = \sum_{i=1}^{M} \omega_i D_{KL}(\rho(x, \theta_i(\mathcal{P})), \rho(x, \theta_i)), \tag{6.3}$$

where $\omega \in R^M$ are the weights, the first argument of the KL divergence is the Gaussian defined by the reconstructed parameter $\theta_i$ for the sequence $\mathcal{S}_i \in Lb$ using $\mathcal{P}$, and the second argument is the observed Gaussian estimations for, again, the sequence $\mathcal{S}_i$. For more detail about the latter notation we refer to the chapter 2 of this work. Given oligomer–based statistics for the libraries $Lb_1$ and $Lb_2$ and given a best–fit parameter set $\mathcal{P}_1$ computed using $Lb_1$ we want to compute the best-fit parameter set $\mathcal{P}_2$ for the data using the parameter continuation technique on the following function

$$\mathbb{F}(\mathcal{P}, Lb; \omega) = \mathbb{F}(\mathcal{P}, Lb_1; \omega_1) + \mathbb{F}(\mathcal{P}, Lb_2; \omega_2), \tag{6.4}$$

where $\omega = (\omega_1, \omega_2)$, $Lb = Lb_1 \cup Lb_2$. The best–fit cgDNA parameter set $\mathcal{P}_2$ will be computed using the algorithm (2). The convergence is ensured as long as the total number of continuation iterations $n_c \in \mathbb{N}$ is large enough. Practically speaking we have used $n_c = 200$ and, in order to speed up the computation we have computed the Hessian matrix only twice during the run of algorithm (2): once at iteration $k = 1$ and once at iteration $k = 100$. For detail about the computation of the hessian matrix of a function of the form (5.17), see section E in appendix.

**Algorithm 2** Parameter continuation of function (6.3)

1: **Initialize:** $\omega_1 = (1, \ldots, 1), \omega_2 = (0, \ldots, 0), \mathcal{P}^{(0)} = \mathcal{P}_1, \varepsilon = 1/n_c, n_c \in \mathbb{N}$
2: **for** $k = 1 : n_c$ **do**
3: **Vary weights:** $\omega_1 = \omega_1 - k\varepsilon, \omega_2 = \omega_2 + k\varepsilon$
4: **Initial Guess:** $\mathcal{P}^{(k-1)}$
5: **Compute using Broyden:** $\mathcal{P}^{(k)} = \underset{P \in \mathbb{P}}{\operatorname{argmin}} \mathbb{F}(\mathcal{P}, Lb; \omega)$
6: **end for**
7: **Finalise:** $\mathcal{P}_2 = \mathcal{P}^{(n_c)}$

## 6.1 Introduction to training data comparison

We consider six different MD training sets: two training sets will have the ABC sequence library as training library while four will have a library called miniABC. The sequence library miniABC was designed by M. Pasi and R. Lavery and, as the ABC one, it contains at least one copy of each of the 136 independent tetra nucleotides. But it is much compact than ABC; it has 13 sequences each of length 18 base pairs and each with $GpC$ dimers at both ends. In the following table we list all the sequences in the miniABC library:

| |
|---|
| GCAACGTGCTATGGAAGC |
| GCAATAAGTACCAGGAGC |
| GCAGAAACAGCTCTGCGC |
| GCAGGCGCAAGACTGAGC |
| GCATTGGGGACACTACGC |
| GCGAACTCAAAGGTTGGC |
| GCGACCGAATGTAATTGC |
| GCGGAGGGCCGGGTGGGC |
| GCGTTAGATTAAAATTGC |
| GCTACGCGGATCGAGAGC |
| GCTGATATACGATGCAGC |
| GCTGGCATGAAGCGACGC |
| GCTTGTGACGGCTAGGGC |

Table 6.1 – The miniABC training library

The details, along with the naming, of the different training sets, are listed hereafter:

| Label | $St$ | $Lb$ | $Ff$ | $Io$ |
|---|---|---|---|---|
| ABC | 50–100$n$s | ABC | bsc0 | K+ |
| $\mu$ABC | 1$\mu$s | ABC | bsc0 | K+ |
| MABC$_0^K$ | 1$\mu$s | miniABC | bsc0 | K+ |
| MABC$_1^K$ | 1$\mu$s | miniABC | bsc1 | K+ |
| MABC$_1^{KNa}$ | 1$\mu$s | miniABC | bsc1 | 50% K+ 50% Na+ |
| MABC$_1^{Na}$ | 1$\mu$s | miniABC | bsc1 | Na+ |

Table 6.2 – Characteristics of the training sets considered in the MD simulations.

In this chapter we will considering three different predictions of the cgDNA model that will be used to compare the different parameter sets. The first prediction is in fact the Gaussian distribution predicted by the parameter set for the sequences in the training

library, and in particular how far these predictions are from the banded observed distributions. Let us recall that the objective function (5.18) is minimized to extract a cgDNA parameter set $\mathcal{P}$. We now define a function that will be used to compare different training sets. Let $D_{KL}$ be the Kullback-Leibler divergence introduced in (5.3) and let us use the same ordering for the arguments as for parameter extraction, see (5.18). We define the following averaged Kullback-Leibler function per degree of freedom for the best–fit parameter set $\mathcal{P}$ and the training library $Lb$

$$\overline{D}(\mathcal{P}, Lb) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n_{\text{dofs}}} D_{KL}\Big(\rho(x,\theta_i(\mathcal{P})),\rho(x,\theta_i)\Big)\right) \tag{6.5}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n_{\text{dofs}}} D^{\dagger}\Big(\rho(x,\theta_i(\mathcal{P})),\rho(x,\theta_i)\Big) + \frac{1}{n_{\text{dofs}}}\mathcal{M}\Big(\rho(x,\theta_i(\mathcal{P})),\rho(x,\theta_i)\Big)\right)$$

$$= \overline{D^{\dagger}}(\mathcal{P}, Lb) + \overline{\mathcal{M}}(\mathcal{P}, Lb). \tag{6.6}$$

where we use the separation of the KLd presented in (5.4), $\theta_i = \text{param}(\mu(\mathcal{S}_i), K(\mathcal{S}_i))$, $\theta_i(\mathcal{P}) = \text{param}(\mu(\mathcal{P},\mathcal{S}_i), K(\mathcal{P},\mathcal{S}_i))$, $\mathcal{S}_i \in \mathcal{T}$, and $n_{\text{dofs}} = 12n_i - 6$ where $n_i$ is the length of $\mathcal{S}_i$. We also recall that $\rho(\cdot,\theta)$ is a Gaussian distribution parametrized by $\theta = \text{param}(\mu, K)$, see (2.9). Using (6.5) we can quantify the quality of the parameter set in predicting Gaussian distributions for sequence in the training library $Lb$ and by using the decomposition (6.6) we can measure the contribution coming from the stiffness and coming from the ground–state.
For example for the ABC training data presented in chapters 3.2 and 4, we have the following values for the total $\overline{D}$, shape $\overline{\mathcal{M}}$ and stiffness $\overline{D^{\dagger}}$ parts of the KLd:

| Protocol | $\overline{D}$ | $\overline{D^{\dagger}}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|
| ABC | $3.00 \cdot 10^{-2}$ | $1.95 \cdot 10^{-2}$ | $1.05 \cdot 10^{-2}$ |

Thus, the main contribution to $\overline{D}$ for the ABC training set comes from $\overline{D^{\dagger}}$.
The other criterion we will focus on are based on the persistence lengths previously presented in section (2.2), and in particular on the apparent and dynamic persistence lengths, respectively (2.43) and (2.47). We will focus our analysis on the spectra of the latter over an ensemble of randomly generated sequences of length 220 base–pair. The sequence ensemble will be fixed throughout the following study. In figure 6.1, left, we show the spectra of $\ell_p$ and $\ell_d$ already presented in section 4.3. In figure 6.1, right, we show the spectra of differences between the reciprocal of apparent and dynamic persistences for each sequence in the ensemble. This difference defines the reciprocal of the static component of the persistence length, see formula 2.45, and is interesting to analyse because it can be used to look for the most bent ground–states in a sequence ensemble. More precisely, for an arbitrary sequence with straight ground–state the difference $\frac{1}{\ell_p} - \frac{1}{\ell_d}$ will be close to zero.

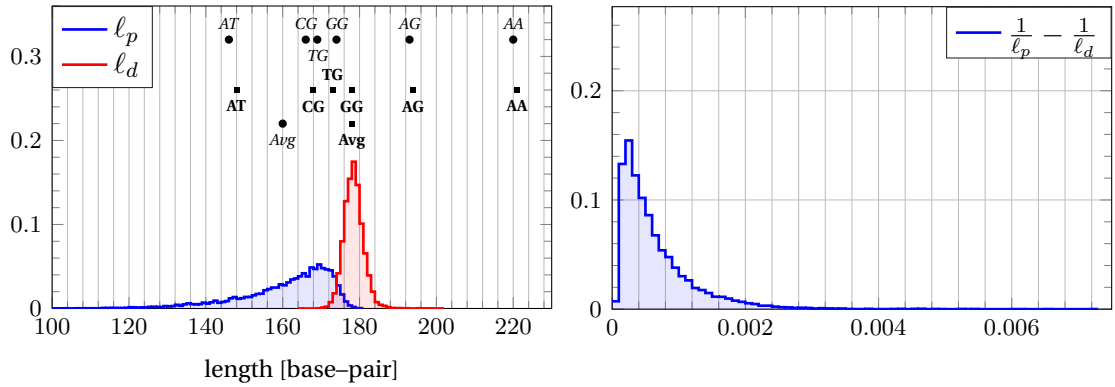



Figure 6.1 – Left: Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA, trained on the ABC data set, over a sequence ensemble of 10K randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence lengths. The values of the persistence lengths for six independent poly–dimers of length 220 are also reported: italic circle for the apparent and bold square for the dynamic. Right: Histogram of differences between reciprocals of apparent and dynamic peristence length computed over the same sequence ensemble used for the histograms shown on the left. This difference is always positive and the magnitude is a indication of how bent the sequence is in an overall sense.

We will take the ABC training set as the starting point for our analysis, in more detail we will consider the following chain of pair wise comparisons:

1. ABC versus $\mu$ABC
2. $\mu$ABC versus MABC$_0^K$
3. MABC$_0^K$ versus MABC$_1^K$
4. MABC$_1^K$ versus MABC$_1^{NaK}$
5. MABC$_1^K$ versus MABC$_1^{Na}$

where in particular we will first study the effect of different simulation durations, then different training libraries, then different force fields, and finally, we will study the effect of different different ion types and concentrations. Again for more detail about the different training sets we refer to table (6.2).

## 6.2 Sensitivity to simulation duration: ABC versus $\mu$ABC

The simulations duration clearly leads to an higher number of trajectory snapshots and consequently to a larger time series of internal coordinates. The latter observations

lead simply to the conclusion that with a longer simulation time the convergence will be better and consequently the overall quality of the data will be better. Thus, we expect to have better estimators for the oligomer–based stationary mean and covariance, hence potentially better fits to the equilibrium distribution values. We can verify the latter by looking at the values for $\overline{D}$, see equation (6.6):

| Protocol | $\overline{D}$ | $\overline{D^{\dagger}}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|
| ABC | $3.00 \cdot 10^{-2}$ | $1.95 \cdot 10^{-2}$ | $1.05 \cdot 10^{-2}$ |
| $\mu$ABC | $2.72 \cdot 10^{-2}$ | $1.77 \cdot 10^{-2}$ | $0.95 \cdot 10^{-2}$ |

The values $\overline{D}$ decreased meaning that the oligomer–based statistic in the training data are better predicted by the model. In figure 6.2 we show the spectra of apparent and dynamic persistence lengths, on the left, and the spectra of the static persistence length on the right. The are no particular differences when compared to the analogous plot shown in figure 6.1, though one can observe that the $\mu$ABC protocol leads to lower values for both sequence averaged persistence lengths:

| Protocol | avg. $\ell_p$ [bp] | avg. $\ell_d$ [bp] |
|---|---|---|
| ABC | 160 | 178 |
| $\mu$ABC | 156 | 173 |

Also by looking at the values of $\ell_p$ and $\ell_d$ (in base–pairs) for the six poly dimer sequences one can notice that for $\mu$ABC data these values tend to be smaller. Thus we can conjecture that simulating up to one microsecond leads to a better exploration of the possible coarse–grained configurations which consequently lead to a slightly softer model.

## 6.3 Sensitivity to training library: $\mu$ABC versus MABC$_0^K$

We continue the chain of comparisons by changing the training library and in particular we will present results for the miniABC training library. We refer to table 6.1 for the list of the sequences. Before discussing the comparison between the two training sets we make a few remarks about the training libraries and especially about the statistics of occurrences of base and dimer sequence sub units contained in both sequence lists. The first remark is that $\mu$ABC contains 39 sequences while miniABC contains just 13, which leads to big differences in the number of instances for different bases and dimers. In figure 6.3 we show the counting of the instances of bases (top row) and dimers (bottom row) for ABC (left column) and miniABC (right column). Beyond the differences in the total numbers of instances, both libraries seems to have equivalent statistics for both bases and dimers. For our purpose it does not matter if, for example, there is a large difference between the number of instances of complementary steps e.g.

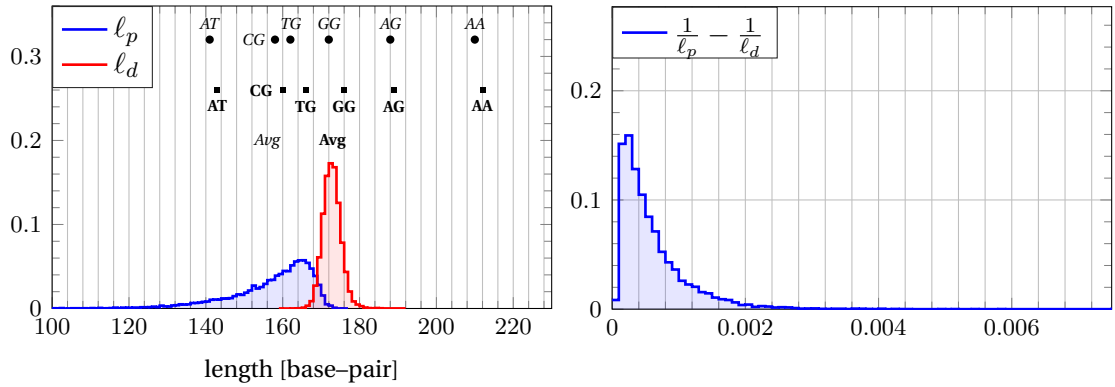



Figure 6.2 – Left: Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA, trained on the $\mu$ABC data set, over a sequence ensemble of 10K randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence lengths. The values of the persistence lengths for six independent poly–dimers of length 220 are also reported: italic circle for the apparent and bold square for the dynamic. Right: Histogram of differences between reciprocals of apparent and dynamic peristence length computed over the same sequence ensemble used for the histograms shown on the left. This difference is always positive and the magnitude is a indication of how bent the sequence is in an overall sense.

$CC$ and $GG$ because in the parameter set computation the instances will be summed by choosing one of the two complementary dimer steps as independent and by using the Crick-Watson symmetries to transform the dependent one. The same reasoning works also for the bases. The only dimers that will be under represented will be the four palindromic ones. In fact one can observe that for miniABC the total number of instances for the palindromic dimer step is around 10 while for the non–palindromic ones it is larger than 15.

By using the cgDNA parameter set computed for the $\mu$ABC we have computed the best-fit parameter set for the protocol MABC$_0^K$ and, as before, we start by looking at the values of the averaged Kullback–Leibler divergence per degree of freedom are reported in the following table:

| Protocol | $\overline{D}$ | $\overline{D^\dagger}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|
| $\mu$ABC | $2.72 \cdot 10^{-2}$ | $1.77 \cdot 10^{-2}$ | $0.95 \cdot 10^{-2}$ |
| MABC$_0^K$ | $2.92 \cdot 10^{-2}$ | $2.01 \cdot 10^{-2}$ | $0.91 \cdot 10^{-2}$ |

The first thing we observe is that we obtained an higher value of $\overline{D}$ for the MABC$_0^K$ protocol. It is quite difficult to interpret this result because the value $\overline{D^\dagger}$ is higher for MABC$_0^K$ while the value $\overline{\mathcal{M}}$ is lower with respect to $\mu$ABC protocol. A deeper investigation on the values of the KLd between reconstructions and oligomer–based statistics for each sequence in the training libraries reveals that for both protocols there

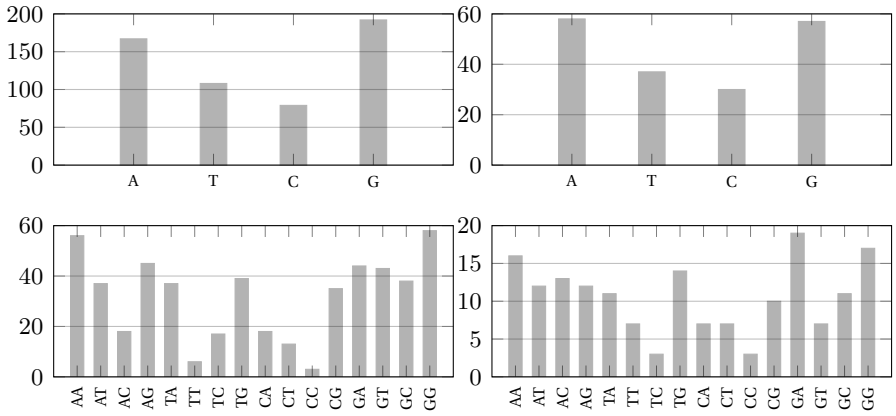



Figure 6.3 – The total number of instances of dimer, and base counting on only one strand for ABC library on the left and muABC library on the right.

are two sequences–(11,28) for $\mu$ABC and (3,8) for MABC$_0^K$–that have significantly larger values for $\overline{D^\dagger}$. The protocol MABC$_0^K$ is indeed penalized by the fact that the average is over a smaller ensemble. If we want really to compare both protocols the figure 6.4 is more pertinent. Again on the left we show the histograms of apparent and dynamic persistence lengths computed using cgDNA model trained on the MABC$_0^K$ data set while on the right we show the histogram of the static component. In the following table we report first the values of the sequence averaged quantities $\ell_p$ and $\ell_d$

| Protocol | avg. $\ell_p$ [bp] | avg. $\ell_d$ [bp] |
|---|---|---|
| $\mu$ABC | 156 | 173 |
| MABC$_0^K$ | 158 | 174 |

By comparing figures 6.2 and 6.4 we notice that MABC$_0^K$ has a slightly higher value for both sequence averaged persistence lengths. By inspecting better also the values of $\ell_p$ and $\ell_d$ for the poly dimer sequences we can observe a difference between the two protocols. In fact for MABC$_0^K$ all the values are higher but for poly A it is nearly unchanged. Thus, for MABC$_0^K$ we have a general trend that leads to a slight increase in the rigidity of the parameter set that can also be seen in the spectra of static persistence length, figure 6.4, where one can observe that the tail on the right–hand side of the histograms get shorter meaning that in the sequence ensemble some sequence will have a ground–state closer to straight, or at least, not as bent.

## 6.4 Sensitivity to force field: MABC$_0^K$ versus MABC$_1^K$

The next comparison is on the force field and in particular on the switch from bsc0 to bsc1. The authors in [24] highlight many different points where the force field bsc1 is actually better then other classical MD force fields. But, in the context of this work, one

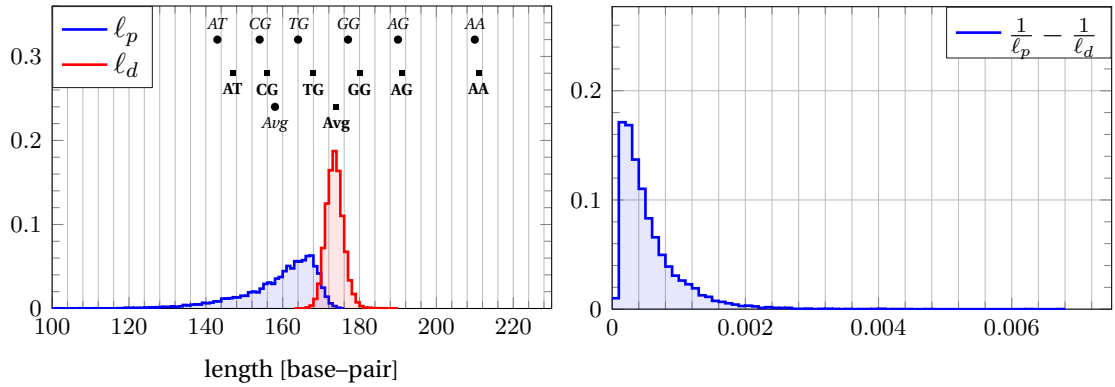



Figure 6.4 – Left: Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA, trained on the MABC$_0^K$ data set, over a sequence ensemble of 10K randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence lengths. The values of the persistence lengths for six independent poly–dimers of length 220 are also reported: italic circle for the apparent and bold square for the dynamic. Right: Histogram of differences between reciprocals of apparent and dynamic peristence length computed over the same sequence ensemble used for the histograms shown on the left. This difference is always positive and the magnitude is a indication of how bent the sequence is in an overall sense.

particular aspect captured our interest: the bsc1 force field provides better stability at the end of the molecule during the simulation, meaning that it suffers less end fraying and broken hydrogen bonds. In section 3.3.1 we introduced the hydrogen bond filtering that we adopted to discard MD snapshots that lead to exceptional outliers in the internal coordinates–especially in the intra–base–pair rotational components–that cannot be included in the estimation of first and second moments. Both protocols have a simulation durations of $1\mu$s which converts into $10^6$ snapshots before HB filtering. The miniABC library has 13 sequence, thus potentially with both protocols we could reach an accumulated time series of length $13 \cdot 10^6$. We computed the percentage of accepted snapshots after HB filtering and we obtained: $71.7$ % for MABC$_0^K$ and even $90.4$ % for MABC$_1^K$. In the following table we show the average KLd values:

| Protocol | $\overline{D}$ | $\overline{D^\dagger}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|
| MABC$_0^K$ | $2.92 \cdot 10^{-2}$ | $2.01 \cdot 10^{-2}$ | $0.91 \cdot 10^{-2}$ |
| MABC$_1^K$ | $2.58 \cdot 10^{-2}$ | $1.59 \cdot 10^{-2}$ | $0.99 \cdot 10^{-2}$ |

The first interesting thing is that sequences 3 and 8 do not have as high values of $\overline{D^\dagger}$ for the bsc1 protocol as was the case for the bsc0 one. The latter could be explained by the higher stability of the bsc1 force field that lead to better statistics for the latter two sequences. Secondly the value of $\overline{\mathcal{M}}$ is higher for protocol MABC$_1^K$ which is again difficult to interpret. We, thus, move to the second step of comparison: sequence–

dependent persistence length analysis. In figure 6.5, left, we show the spectra of $\ell_p$ and $\ell_d$ and we can say straightaway that both histograms looks qualitatively different from their analogues for the bsc0 force field. In detail, we notice a substantial shift to the right of the distribution of $\ell_p$ and for $\ell_d$ the distribution of the values are more peaked and the tail of the right–hand side of the spectra vanishes. The sequence–averaged

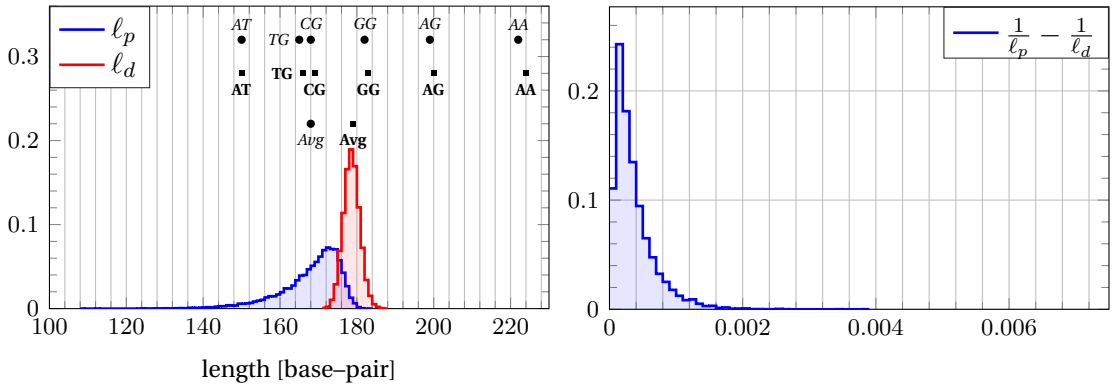



Figure 6.5 – Left: Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA, trained on the MABC$_1^K$ data set, over a sequence ensemble of 10K randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence lengths. The values of the persistence lengths for six independent poly–dimers of length 220 are also reported: italic circle for the apparent and bold square for the dynamic. Right: Histogram of differences between reciprocals of apparent and dynamic peristence length computed over the same sequence ensemble used for the histograms shown on the left. This difference is always positive and the magnitude is a indication of how bent the sequence is in an overall sense.

values of both persistence lengths are:

| Protocol | avg. $\ell_p$ [bp] | avg. $\ell_d$ [bp] |
|---|---|---|
| MABC$_0^K$ | 158 | 174 |
| MABC$_1^K$ | 168 | 179 |

The data actually confirm the intuition that the protocol MABC$_1^K$ lead to an overall more rigid cgDNA parameter set. Also all the poly dimer sequences have a consistent shift toward the right of the graph and, for example, poly A reached a value higher then 220 bp. The left plot in (6.5) shows a change in the shape of the distribution off the reciprocal of the static persistence length with the major feature that the tail on the right-hand side got shorter, indicating again that the protocol MABC$_1^K$ tends to have less bent sequences.

## 6.5 Sensitivity to ions: MABC$_1^K$ versus MABC$_1^{NaK}$ versus MABC$_1^{Na}$

In this section we compare two different ion types against the standard potassium. We recall that the overall ion concentration has been kept the same for all choice of the protocols. In the following table we reported the values of $\overline{D}$ along with its two main contributions:

| Protocol | $\overline{D}$ | $\overline{D^\dagger}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|
| MABC$_1^K$ | $2.58 \cdot 10^{-2}$ | $1.59 \cdot 10^{-2}$ | $0.99 \cdot 10^{-2}$ |
| MABC$_1^{NaK}$ | $2.82 \cdot 10^{-2}$ | $1.73 \cdot 10^{-2}$ | $1.09 \cdot 10^{-2}$ |
| MABC$_1^{Na}$ | $2.88 \cdot 10^{-2}$ | $1.81 \cdot 10^{-2}$ | $1.07 \cdot 10^{-2}$ |

We notice that changing ion type leads to an increase of approximation error compared to the baseline values given by the protocol MABC$_1^K$. In figures 6.6 and 6.7 the sequence–averaged $\ell_p$ and $\ell_d$ are

| Protocol | avg. $\ell_p$ [bp] | avg. $\ell_d$ [bp] |
|---|---|---|
| MABC$_1^K$ | 168 | 179 |
| MABC$_1^{NaK}$ | 168 | 179 |
| MABC$_1^{Na}$ | 169 | 179 |

Thus, we basically have no change in the overall rigidity of the model by changing ion type. There are a few changes in the spectra of the reciprocal of the static persistence length, left plot in figures 6.6 and 6.7. When adding more sodium ions the trend seems to be that the distribution of $\frac{1}{\ell_s}$ get more and more peaked to the left, i.e, more sequences in the ensemble have potentially a straighter ground–state.

## 6.6 Discussion and Conclusions

We summarize the conclusion of the analysis made in the previous sections:

- **Simulation time**: The major change is in the higher reliability of the data because longer time series are closer to converged and thus the oligomer–based equilibrium statistics estimated for the time series will be more accurate. Apart for the latter point no major differences between the two protocols has been identified.
- **Training library**: No important changes have been observed while passing from $\mu$ABC to miniABC libraries for the same MD protocol. Even though both libraries have at least one instance of all the 136 distinct tetranucleotides, they differ by the numbers of instances of distinct bases and dimers, which could be a reason for the small changes between the two associated coarse–grain parameter sets.

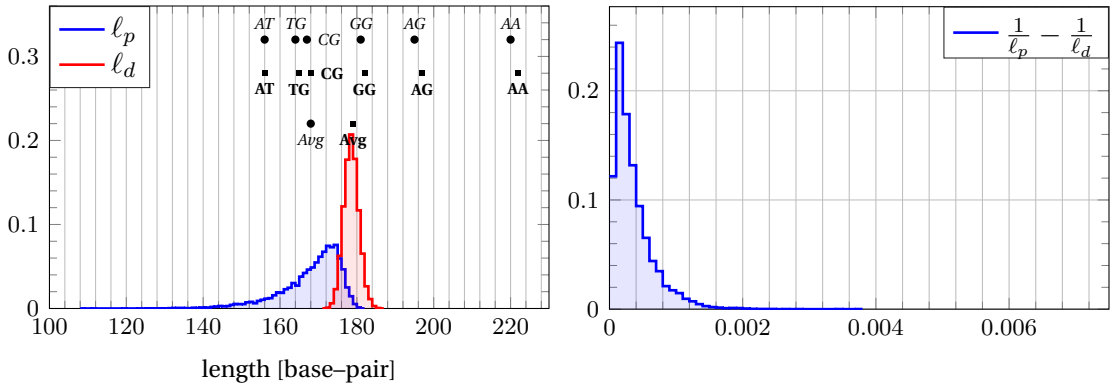



Figure 6.6 – Left: Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA, trained on the MABC$_1^{NaK}$ data set, over a sequence ensemble of 10K randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence lengths. The values of the persistence lengths for six independent poly–dimers of length 220 are also reported: italic circle for the apparent and bold square for the dynamic. Right: Histogram of differences between reciprocals of apparent and dynamic peristence length computed over the same sequence ensemble used for the histograms shown on the left. This difference is always positive and the magnitude is a indication of how bent the sequence is in an overall sense.

- **Force field**: Changing the force field from bsc0 to bsc1 leads to an increase in the sequence averaged persistence length with the major increase being in the sequence averaged $\ell_p$. This is due to the fact that in the ensemble of sequences, the population of sequences with small $\ell_p$ (bent ground state) decreases, i.e, the cgDNA model parameters trained on bsc1 simulations lead to coarse–grained reconstructions that are on average straighter. The increase in $\ell_d$ is less as the dynamic persistence length depends less on the ground state and more (in a non linear way) on the stiffness. Moreover the total number of accepted MD snapshots after HB filtering increased considerably for the protocol MABC$_1^K$ due to the better stability of the force field at the end of the oligomer.

- **Ion type, NaK** (50%–50%): No major changes between protocols MABC$_1^K$ and MABC$_1^{NaK}$ (50/50) have been identified. Both protocols lead to close values of sequence averaged $\ell_p$ and $\ell_d$.

- **Ion type, Na**: No major changes between protocols MABC$_1^K$ and MABC$_1^{Na}$ have been identified. Both protocols lead to close values of sequence averaged $\ell_p$ and $\ell_d$.

As the only major observed change is between the two force fields we make some additional remarks about the changes in the rigidity of the model. For each of the $10^4$ sequences in the ensemble we can compute the differences in $\ell_p$ and $\ell_d$ computed using

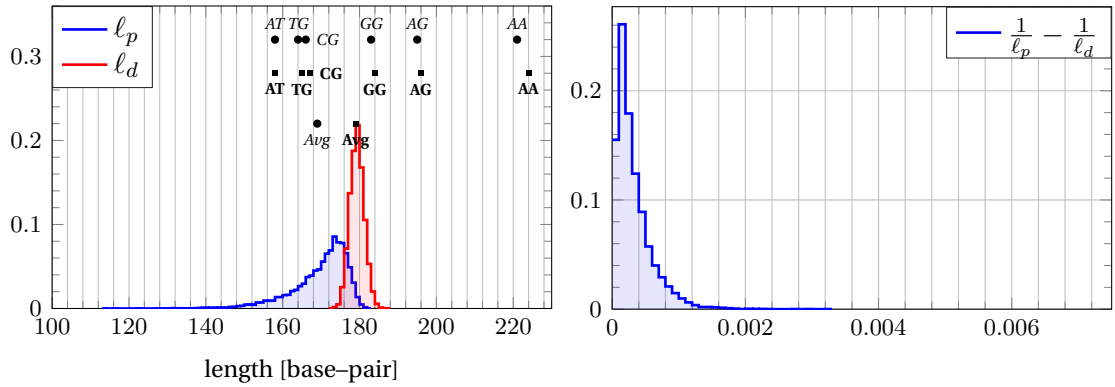



Figure 6.7 – Left: Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA, trained on the MABC$_1^{Na}$ data set, over a sequence ensemble of 10K randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence lengths. The values of the persistence lengths for six independent poly–dimers of length 220 are also reported: italic circle for the apparent and bold square for the dynamic. Right: Histogram of differences between reciprocals of apparent and dynamic peristence length computed over the same sequence ensemble used for the histograms shown on the left. This difference is always positive and the magnitude is a indication of how bent the sequence is in an overall sense.

the two protocols MABC$_0^K$ and MABC$_1^K$. In figure 6.8 we show the histograms of these differences: left for apparent persistence length, and right for dynamic persistence length. The order we chose for the subtraction is "bsc0 values" minus "bsc1 values". For the apparent persistence length we can observe that most of the distribution is located in the negative region of the graph, meaning that most of the $\ell_p$ compute with the bsc1 protocol have higher values compared to the one computed with the bsc0 protocol. The same observation can be done for the $\Delta\ell_d$ spectra where actually the shift towards the negative values is much more prominent. In fact $\ell_d$ has only contributions from the stiffness as the ground–state part has been factorised out. This means that the difference of $\ell_d$ in the right–hand side histogram in figure 6.8 actually says something about the difference in the rigidity between the two force fields. Finally, from these two histograms it is even more clear that the cgDNA model parametrized with the bsc1 protocol is more "rigid" compared to the model parametrized with the protocol MABC$_1^K$. We next consider again the A–tracts sequence $S_{\text{A-tr}} = (A_6CGCGA_6CGGGC)_3$ already mentioned in section 4.3 and compute using the bsc1 cgDNA parameter set its tangent–tangent correlation with and without shape factorization. In figure 6.9 left we plot the ttc (solid line) and its factorized version (dashed line). Moreover in blue we show the result obtained with the cgDNA parameter set trained with the protocol MABC$_1^K$ while in red we show the results obtained with the published bsc0 cgDNA parameter set [55, 19] trained on the original ABC protocol. In figure 6.9 on the right, we also show the rigid–body reconstruction of the ground–state for both protocols.

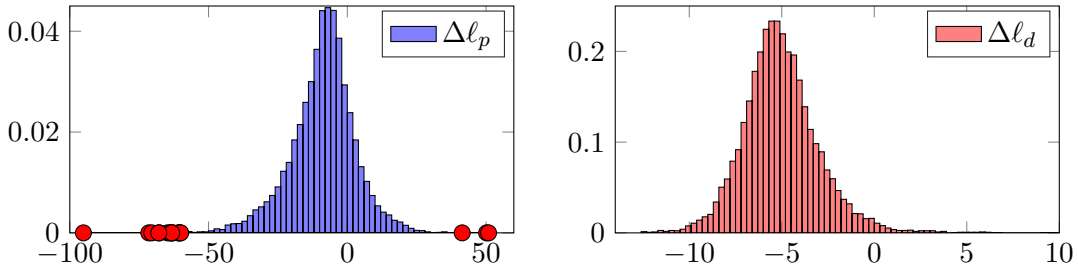



Figure 6.8 – Histograms of differences between persistence lengths computed for the same sequences using bsc0 and bsc1trained cgDNA parameter set. We computed bsc0 predictions minus bsc1 predictions for apparent persistence length, left, and dymanic persistence length right. The fact that the $\Delta\ell_d$ are almost all negative indicates that bsc1 id effectivily stiffer than bsc0.

The more bent molecule corresponds to the ground–state reconstructed from the bsc0 parameter set. In conclusion, in the coarse–grained context, the only major changes

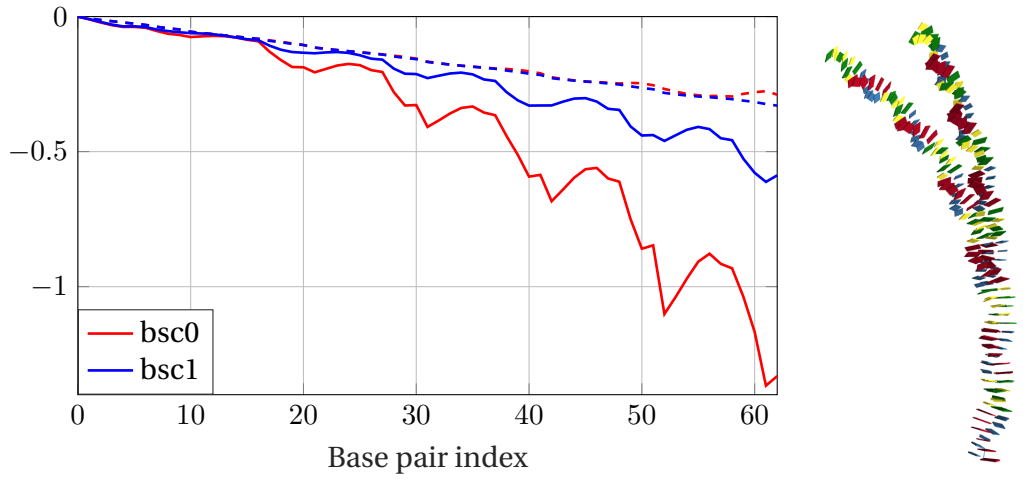



Figure 6.9 – Left: comparison between tangent–tangent correlations (solid) and factorized tangent–tangent (dashed) computed for the sequence $S_{\text{A-tr}} = (A_6CGCGA_6CGGGC)_3$ using cgDNA trained on MABC$_1^K$ (blue) and ABC data (red). Right: 3D visualisation of the ground–state of $S_{\text{A-tr}}$ predicted by the cgDNA model trained on MABC$_1^K$ data set (more straight) and ABC data set (more bent). The 3D figures have been obtained using the web–based viewer for the cgDNA model [13].

in the cgDNA model, and its predictions, is made by the change of force field and in particular by changing from the bsc0 force field [58] to the state of the art force field bsc1 [24].

# 7 A Palindromic training library

In the previous chapter we computed cgDNA parameter sets for different MD protocols but we never discussed in detail the convergence of the MD trajectories. More precisely, we did not discuss the convergence of the first and second (centred) moments estimated for MD time series of internal coordinates. In fact, using the ABC (B.1) or the miniABC (B.2) training library and exhaustive convergence test for the estimators is not possible. The primer goal of this chapter is to introduce a training library comprising of only palindromic sequences which will allow a convergence test based on the Crick–Watson symmetry (4.2), for mean and covariance estimated from MD time series . We first focus on the designing of this palindromic library and in particular we present the algorithm developed to construct it. Then, we study the convergence of the oligomer–based statistics for each sequence of the palindromic library estimated from MD trajectories of $3\mu$s long simulations. Finally we define a new format for the cgDNA parameter set which contains only dimer–based elements and comprises dedicated blocks for each dimer end. Finally we compute a cgDNA parameter set for the new format trained on the $3\mu$s palindromic training data and compare it to a cgDNA parameter set with the old format trained on the same data.

## 7.1 Designing palindromic libraries

The minimal conditions we want to impose on the library $Lb_{\text{palin}}$ are the following:

1. every sequence should be a palindrome,

2. the library should contain at least one instance of all the independent tetramers without counting end dimers,

3. both ends should be $G\text{p}C$ step (for stability against fraying).

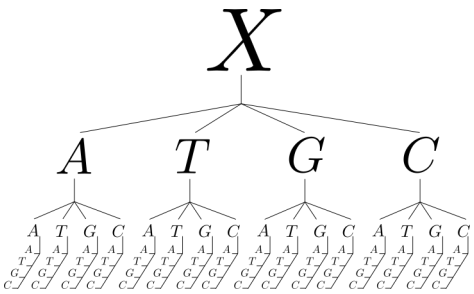
Condition **1** implies that each sequence must have an even total number of base–pairs and that the central dimer must be a palindrome. Moreover the palindromic conditions implies simply that one half of the sequence is in fact the Crick–Watson complement of the other half. Condition **2** implies that in $Lb_{\text{palin}}$ one should be able to count 136 independent tetramers 16 of which are palindromic. Thus, the remark on condition **1** together with the last statement lead to the choice of placing a palindromic tetramer in the middle of a palindromic sequences. The library $Lb_{\text{palin}}$ will thus contains sixteen sequences of the following form:

$$\mathcal{S} \in Lb_{\text{palin}} \Rightarrow \mathcal{S} = GCX_N \dots X_1 P_1 P_2 \overline{P}_2 \overline{P}_1 \overline{X}_1 \dots \overline{X}_N GC. \tag{7.1}$$

Now, to be able to fulfil condition **2** we will need to find $N$ in the above equation, one for all the sequences in $Lb_{\text{palin}}$. A simple computation show that we need at least $N = 8$ which implies that each sequence in the library will be of length 24 and each palindrome will contain a palindromic tetramer and eight non palindromic ones. Clearly, there will room for more than 120 tetramer, more precisely 128, thus 8 tetrameters will appear twice. Condition number **3** comes from the fact that $G\text{p}C$ step have been observed to be the most stable end dimers, meaning that that end step has less probability of broken hydrogen bonds.
Practically to find a library $Lb_{\text{palin}}$ which satisfies the three conditions mentioned earlier, we developed an iterative ad hoc MATLAB algorithm that tries to fill up all the 16 sequences starting from the middle palindromic tetramer. At each iteration it uses a simple yet efficient way for updating the entire library. Before explaining the algorithm we will make a few comments on the methodology and fix some notation. The solution to the problem of finding a library $Lb_{\text{palin}}$ satisfying conditions **1–3** is not unique. Thus in the algorithm it is possible to introduce a random step that will allow us to explore the *space* of admissible libraries. As a palindrome is defined just by one of its halves, the following notation will be used: let $\mathcal{S}$ be a palindromic sequence then

$$\mathcal{S} = (X_N \dots X_1 P_1 P_2)^2 = X_N \dots X_1 P_1 P_2 \overline{P}_2 \overline{P}_1 \overline{X}_1 \dots \overline{X}_N,\ P_i, X_i \in \{A, T, G, C\}. \tag{7.2}$$

In our algorithm we will start just with sixteen sequences of length 2, one for each dimer. At each iteration we will extend all the sequences by one base until we have added eight bases to each of the sixteen initial dimers. We will denote by $\mathcal{S}_i^{(k)} = (X_k^i \dots X_1^i P_1^i P_2^i)^2$ the $i$–th partial palindromic sequence computed after $k$ steps of the algorithm , with $\mathcal{S}_i^0 \in D$, where $D$ is the set of all the possible dimer steps. In order to satisfy condition **2** we have to keep track of the added tetramers, thus we introduce four tree structured graphs, denoted $\mathsf{tree}(X)$, with the following form:



where $X \in \{A, T, G, C\}$. A path in tree$(X)$ defines a tetramer: for example we have

$$X \leftrightarrow A \leftrightarrow C \leftrightarrow G \Leftrightarrow 5'\, GCAX\, 3'.$$

We then give a score, denoted by

$$\text{SCORE}(X \leftrightarrow A \leftrightarrow C \leftrightarrow G) \in \mathbb{N},$$

of zeros to each non–palindromic tetramer in each of the four trees and a score of 9999 to each palindromic tetramer. Whenever the algorithm finds a new tetramer, the score for that tetramer will be increased by one and the score of its complement will be increased by 10 (to avoid selecting it in a later iteration). We finally introduce the notation:

$$\mathbf{T}(\mathcal{S}) = X_1X_2X_3, \text{ for } \mathcal{S} = X_1X_2X_3 \ldots X_N, \tag{7.3}$$

where it just select the first three bases of a sequence $\mathcal{S}$ in the $5'$–$3'$ direction. Equation (7.3) is well defined for sequences with length larger than 3. For dimers we define

$$\mathbf{T}(\alpha\beta) = \alpha\beta\overline{\beta}.$$

We can now introduce the algorithm (3) used to find palindromic libraries. The step at line number 6 in the algorithm does not always find a unique solution, especially during the first iterations. One can introduce a random choice for the base $X_4$ when the number of bases that minimise the score of the current tetramer is bigger than one. By introducing a random step it is clear that each time the algorithm is used it can produce a different outcome because the palindromic library satisfying conditions **1–3** is not unique. Another modification one can make at line 6 is to introduce a check on the next tetramer that will assure that by adding the current base $X_4$ will not arrive at a dead end when considering the leaves of the tree $X_2 \leftrightarrow X_3 \leftrightarrow X_4$, meaning that by adding $X_4$, there exist at least one base $X_{\text{next}}$ with $\text{SCORE}(X_2 \leftrightarrow X_3 \leftrightarrow X_4 \leftrightarrow X_{\text{next}})$=0. We have run algorithm (3) with the mentioned modifications and we found 174 different palindromic libraries satisfying the desired conditions. Based on counting instances of trimers, dimers, and monomers we have chosen the following library:

**Algorithm 3** Palindromic Library

**Given:** $\{\mathcal{S}_i^{(0)}\}_{i=1}^{16}$, initialize the score for each tree$(X)$, $X \in \{A,T,G,C\}$
**Initialize:** $Lb_{\text{palin}}^{tmp} = \{\mathcal{S}_i^0\}_{i=1}^{16}$
**for** $k = 1:8$ **do**
  **for** $i = 1:16$ **do**
    **Define:** $X_3X_2X_1 = \mathbf{T}(\mathcal{S}_i^{(k-1)})$
    **Find:**

$$X_4 = \underset{X\in\{A,T,G,C,\}}{\text{argmin}} \ \text{SCORE}(X_1 \leftrightarrow X_2 \leftrightarrow X_3 \leftrightarrow X)$$

    **Update scores:**

$$\text{SCORE}(X_1 \leftrightarrow X_2 \leftrightarrow X_3 \leftrightarrow X_4) += 1$$
$$\text{SCORE}(\overline{X_4} \leftrightarrow \overline{X_3} \leftrightarrow \overline{X_2} \leftrightarrow \overline{X_1}) += 10$$

    **Update sequence:** $\mathcal{S}_i^{(k)} = X_4\,\mathcal{S}_i^{(k-1)}$
  **end for**
  **Update library:** $Lb_{\text{palin}}^{tmp} = \{\mathcal{S}_i^k\}_{i=1}^{16}$
**end for**
**Finalise library:** $Lb_{\text{palin}} = \{(GC\,\mathcal{S}_i^k)^{\frac{1}{2}}\}_{i=1}^{16}$

| | |
|---|---|
| 1 | GCTTAGTTCAAATTTGAACTAAGC |
| 2 | GCTCTCTGTATTAATACAGAGAGC |
| 3 | GCCCTTGGCGATATCGCCAAGGGC |
| 4 | GCTAAAGCCTTATAAGGCTTTAGC |
| 5 | GCGGTAGAAAACGTTTTCTACCGC |
| 6 | GCCAAGACATTGCAATGTCTTGGC |
| 7 | GCAGATGGTCAGCTGACCATCTGC |
| 8 | GCCTCACCGCTCGAGCGGTGAGGC |
| 9 | GCAGTGGAATCATGATTCCACTGC |
| 10 | GCTTTACTTCGTACGAAGTAAAGC |
| 11 | GCTACCTATGCTAGCATAGGTAGC |
| 12 | GCGCACTGGGGATCCCCAGTGCGC |
| 13 | GCTGAGGAGTCCGGACTCCTCAGC |
| 14 | GCTGCCGTCGGGCCCGACGGCAGC |
| 15 | GCGCACAACACGCGTGTTGTGCGC |
| 16 | GCCTAACCCTGCGCAGGGTTAGGC |

Table 7.1 – The 16 palindromic sequences of the palindromic library.

In figure 7.1 we present the aforementioned counting of instances of monomers, dimers, and trimers. We can observe that in the occurrences for all the trimers, dimers, and monomers no element is over represented. The latter will be useful for the cgDNA parameter estimation in section 7.3. Moreover, by reading the sequences from only

one strand in each oligomer each of all possible 256 tetramers appears at least once. In fact the non–palindromic tetramers appear at least twice while the palindromic tetramer appear only once.

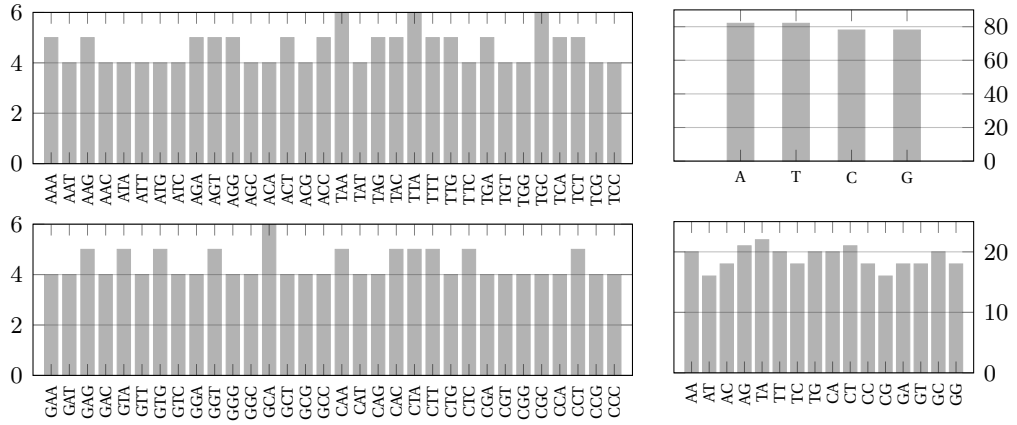



Figure 7.1 – The total number of trimers, dimers, and bases counting on only one strand.

## 7.2 The palindromic training sets

We have performed MD simulations of the palindromic library presented in the previous section, see table 7.1, using the standard ABC protocol, see appendix C, but with the bsc1 force field [24]. We have chosen a simulation time of $3\mu$ for each sequence in the training library plus an additional accumulation of $10\mu$s of simulation, for just three sequences: (1,5,11), see table 7.1. For the additional $10\mu$s trajectories we computed 10 independent parallel simulations using random initial velocities. In the next section we will explain how to use Crick–Watson symmetry to gain insight of the convergence of these MD simulations of the palindromic sequences.

### 7.2.1 Assessing convergence of MD simulations

The physical Crick–Watson symmetries for the ground–state and the covariance matrix of a palindromic sequence $\mathcal{S}$ are:

$$\mu(\mathcal{S}) = E_{2n-1}\mu(\mathcal{S}), \tag{7.4}$$

$$C(\mathcal{S}) = E_{2n-1}C(\mathcal{S})E_{2n-1}. \tag{7.5}$$

When the mean and the covariance, are estimated from an MD time series the latter two conditions are not in general satisfied due to lack of complete convergence of the time series. We can thus introduce the following error functions for the mean and the

covariance to measure deviation from palindromy of both estimators:

$$\text{ERR}(\mu) = \|\mu - E_{2n-1}\mu\|, \tag{7.6}$$

and for the covariance we can introduce the error function

$$\text{ERR}(C) = \|C - E_{2n-1}CE_{2n-1}\|, \tag{7.7}$$

where for the sake of simplicity we have dropped the dependence on the sequence $\mathcal{S}$. For the covariance matrix we can moreover refine the error function by considering only the entries inside the cgDNA stencil (due to the banded reconstruction technique (4.12) ) and, as the covariance matrix is symmetric, we can just consider the diagonal entries plus, for example, the upper triangular part. We introduce then the following notation: Let $C$ be a covariance matrix, we define by $C_{\text{sym}}$ the matrix defined by

$$(C_{\text{sym}})_{ij} = 0, \text{ if } ij \text{ are outside the stencil or if } i < j, \tag{7.8}$$

$$(C_{\text{sym}})_{ij} = (C)_{ij} \text{ otherwise.} \tag{7.9}$$

Moreover, we introduce the notation $H = E_{2n-1}CE_{2n-1}$. We can then redefine the error function for covariance matrices as follow

$$\text{ERR}(C) = \|C_{\text{sym}} - H_{\text{sym}}\|. \tag{7.10}$$

The norm in (7.10) is the Frobenius norm introduce in chapter (2) which is equivalent to the classic 2–norm for vectors. Thus both error functions can be seen as an error function in $\mathbb{R}^n$. This property will be useful for the interpretation of different values of (7.6) and (7.10).Now, suppose that $\mathbf{w}(\mathcal{S}) = \{w_i(\mathcal{S})\}_{i=1}^{M}$ is a time series of internal coordinates for a $n$ base–pair long palindromic sequence. We say that the time series $\mathbf{w}(\mathcal{S})$ converges as function of $M$ when

$$\text{ERR}(\mu) \underset{M\to+\infty}{\longrightarrow} 0, \tag{7.11}$$

$$\text{ERR}(C) \underset{M\to+\infty}{\longrightarrow} 0, \tag{7.12}$$

where $\mu$ and $C$ are the respectively the estimators (3.6) and (3.7) computed from $\mathbf{w}(\mathcal{S})$, and the error functions are respectively defined in (7.6) and (7.10). The first question we want to pose is:

> *Do the Crick–Watson errors (7.6) and (7.10) decrease as the total number of snapshots increases?*

We will answer this question by considering the $10\mu$s trajectories computed for the palindromic sequences (1,5,11). We start by evaluating the error function (7.6) for these three sequences as function of simulation duration. In the following table we show the errors for the mean estimator:

| $\#\mathcal{S}$ | $1\mu$s | $2\mu$s | $3\mu$s | $4\mu$s | $5\mu$s |
|---|---|---|---|---|---|
| 1 | 0.5775 | 0.4596 | 0.3655 | 0.3638 | 0.2425 |
| 5 | 0.7932 | 0.3603 | 0.3232 | 0.3304 | 0.3527 |
| 11 | 0.7140 | 0.3146 | 0.2796 | 0.2233 | 0.2163 |
| | $6\mu$s | $7\mu$s | $8\mu$s | $9\mu$s | $10\mu$s |
| 1 | 0.1957 | 0.1941 | 0.1784 | 0.1976 | 0.1740 |
| 5 | 0.4120 | 0.4009 | 0.3057 | 0.2883 | 0.2204 |
| 11 | 0.2008 | 0.1996 | 0.1990 | 0.1864 | 0.1631 |

Table 7.2 – Palindromic error for the mean estimator for palindromic sequence (1,5,11) as function of simulation duration.

We stress here that the simulation lengths considered are the same for each sequence but the actual number of snapshots used in the computations varies between sequence as it depends upon hydrogen bond or HB filtering. Hereafter we have reported the percentage of trajectories accepted after filtering out the broken HB for the three sequences (1,5,11) for the simulation lengths considered in (7.2):

| $\#\mathcal{S}$ | $1\mu$s | $2\mu$s | $3\mu$s | $4\mu$s | $5\mu$s |
|---|---|---|---|---|---|
| 1 | 0.90 | 0.91 | 0.90 | 0.90 | 0.90 |
| 5 | 0.94 | 0.89 | 0.88 | 0.90 | 0.91 |
| 11 | 0.74 | 0.83 | 0.84 | 0.85 | 0.82 |
| | $6\mu$s | $7\mu$s | $8\mu$s | $9\mu$s | $10\mu$s |
| 1 | 0.91 | 0.90 | 0.90 | 0.90 | 0.90 |
| 5 | 0.89 | 0.89 | 0.89 | 0.90 | 0.90 |
| 11 | 0.84 | 0.85 | 0.85 | 0.86 | 0.86 |

Table 7.3 – Percentage of accepted snaphots per simulation lengths

We recall that the writing rate from the MD simulation is each $1p$s, thus a $1\mu$s simulations should lead to a maximum of one million snapshots before HB filtering.
Returning to table (7.2), we notice that the general trend for each sequence is that both error functions decreases with the increased number of snapshots. We can actually see a strong convergence for the mean estimators. Now that we have observed that effectively the convergence error decreases by increasing the number of snapshots considered a last question arises naturally:

*How big is the obtained error?*

We have defined the error function (7.6) using the classic 2-norm for vectors, also called the *euclidean norm*, thus the convergence errors we have computed can be interpreted

as an average error per degree of freedom by simply assuming that each component contribute the same to the norm, meaning that the average error is just the actual error divided by the square root of the total number of components. For example in table (7.2) for sequence 1 at 10$\mu$s we have computed an error of 0.1740: the associated average per degree of freedom error is 0.0104. Now by considering that the entries in the mean are translations measured in angstroms, and rotations in fifth of radiants we obtain that on averaged the mean estimator is off by approximately 1/100Å and $1/100\frac{\text{rad}}{5}$, which can be considered as small. A further analysis that can be done is to consider separately each helical parameter because for example the rise and the twist tend to have a larger magnitude compared to any of the intra translations and rotations. Thus in percentage the average error per degree of freedom will be smaller for these two quantities. Hence in figure 7.2 we show the difference $|\mu(\mathcal{S}) - E_{2n-1}\mu(\mathcal{S})|$ for sequence 1 at 10$\mu$s where the entries of each helical parameters are plotted separately. We can observe that the helical parameters *buckle*, *twist*, *shift* and *slide* have the highest errors among all the helical parameters and probably not coincidentally the inter variable twist, shift, and slide are the three variable known to sometimes have bimodal distributions. Moreover in figure 7.3 we show the comparison between estimated and

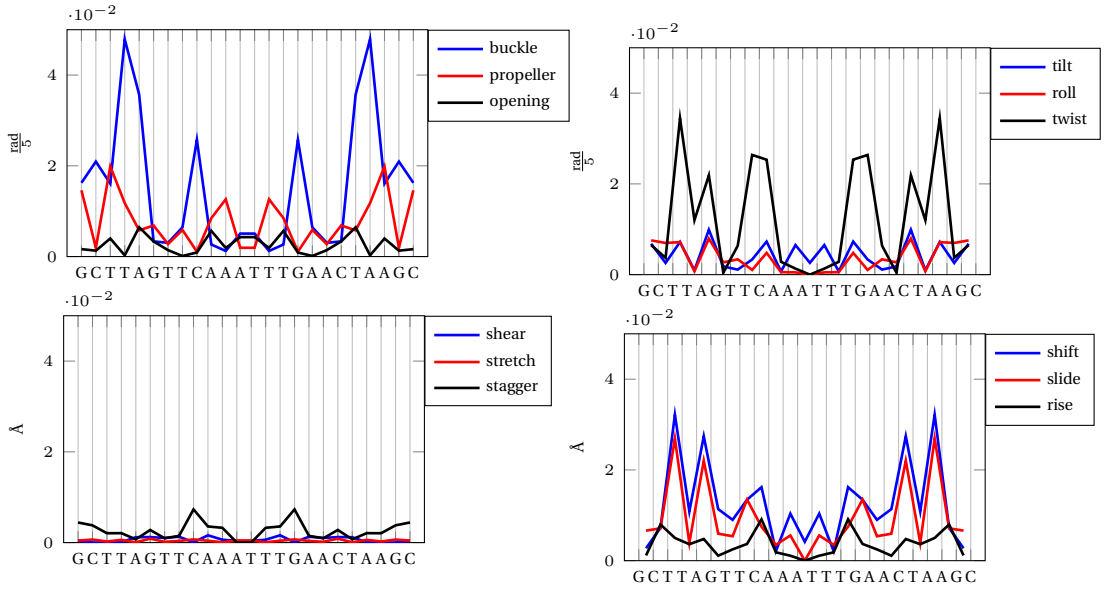



Figure 7.2 – Entry-by-entry palindromic error in the mean estimator at 10$\mu$s for sequence 1.

palindromic symmetrized buckle, shift, twist, and slide, for sequence 1 at 10$\mu$s. We can visually see that the Crick–Watson symmetry after 10 microseconds is achieved to a rather small tolerance for all the degrees of freedom.

We can continue with the analysis of the convergence for the covariance matrix $C$. The result for the palindromic error of the covariance matrix as a function of simulation duration are:

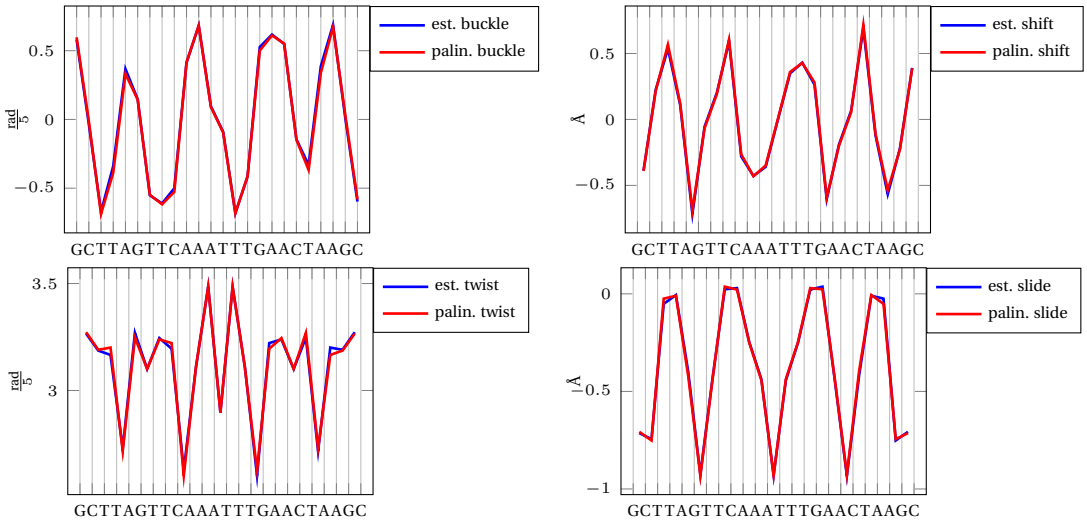



Figure 7.3 – Estimated (est.) and palindromic means (palin.) for four selected helical parameters at 10$\mu$s for sequence 1. Top left: buckle, top right: shift, bottom left: twist, bottom right: slide.

| # $\mathcal{S}$ | 1$\mu$s | 2$\mu$s | 3$\mu$s | 4$\mu$s | 5$\mu$s |
|---|---|---|---|---|---|
| 1 | 0.5496 | 0.4013 | 0.3144 | 0.3303 | 0.2262 |
| 5 | 0.6165 | 0.4204 | 0.2894 | 0.2537 | 0.2573 |
| 11 | 0.7499 | 0.3635 | 0.2923 | 0.2429 | 0.2567 |
| | 6$\mu$s | 7$\mu$s | 8$\mu$s | 9$\mu$s | 10$\mu$s |
| 1 | 0.1790 | 0.1643 | 0.1490 | 0.1564 | 0.1418 |
| 5 | 0.2557 | 0.2440 | 0.2039 | 0.1961 | 0.1598 |
| 11 | 0.2366 | 0.2154 | 0.2036 | 0.1897 | 0.1602 |

Table 7.4 – Palindromic error in the estimator of the covariance as function of simulation length

Thus for the covariance matrix the trend of the error function (7.10) is to decrease when the simulation duration increases. However the interpretation of the errors in table (7.4) using the same reasoning as for the error function of the mean make less sense because the entries of the covariance are correlations between different degrees of freedom and thus do not have a direct physical interpretation. But to judge how big the obtained error is we can introduce a relative covariance error defined by

$$\overline{\mathrm{ERR}}(C) = \frac{\mathrm{ERR}(C)}{\|H_{\mathrm{sym}}\|}, \tag{7.13}$$

where $\mathrm{ERR}(C)$ has been defined in (7.10). For example for sequence 1 we have computed the following relative errors for different simulations lengths:

| $1\mu$s | $2\mu$s | $3\mu$s | $4\mu$s | $5\mu$s |
|---|---|---|---|---|
| 0.0764 | 0.0564 | 0.0440 | 0.0458 | 0.0314 |
| $6\mu$s | $7\mu$s | $8\mu$s | $9\mu$s | $10\mu$s |
| 0.0248 | 0.0228 | 0.0207 | 0.0217 | 0.0197 |

which shows that after 10 microseconds the relative error is slightly less than $2\%$. We want to stress here that already for $1\mu$s the error is lower than $10\%$ which is already acceptable and it drops down to less than $5\%$ in the first $3\mu$s of simulation.
In tables (7.5-7.6) we show respectively the error in the mean and the error in the covariance for each sequence in the palindromic training library at different simulation duration. Moreover in table (7.7) we indicate the percentage of accepted snapshot for each entries in tables (7.5–7.10). We notice that some of the sequences are more converged than others, but finally we can reasonably conclude that both estimators are well converged. Certainly it would be better to extend the simulations of all the palindromic sequences to $10\mu$s, or more, but for the purpose of this work we use the palindromic training data with $3\mu$s simulations length.

| $\#\,\mathcal{S}$ | $100ns$ | $1\mu$s | $2\mu$s | $3\mu$s |
|---|---|---|---|---|
| 1 | 0.8337 | 0.5308 | 0.2104 | 0.2359 |
| 2 | 1.3790 | 0.6458 | 0.3709 | 0.3544 |
| 3 | 1.1501 | 0.5190 | 0.2964 | 0.2364 |
| 4 | 0.8125 | 0.7620 | 0.4284 | 0.4409 |
| 5 | 3.1382 | 0.8064 | 0.4879 | 0.3797 |
| 6 | 3.1841 | 0.6203 | 0.3461 | 0.2837 |
| 7 | 0.8542 | 0.8194 | 0.5507 | 0.4538 |
| 8 | 0.9575 | 0.4286 | 0.3121 | 0.2786 |
| 9 | 2.1321 | 0.3649 | 0.3690 | 0.2635 |
| 10 | 1.2389 | 0.4776 | 0.3570 | 0.5034 |
| 11 | 2.9449 | 1.2054 | 0.6228 | 0.5478 |
| 12 | 1.4374 | 0.7056 | 0.4854 | 0.4680 |
| 13 | 1.0173 | 0.3367 | 0.3293 | 0.2257 |
| 14 | 0.6689 | 0.4699 | 0.2723 | 0.2348 |
| 15 | 1.6356 | 0.8731 | 0.3359 | 0.3278 |
| 16 | 1.1962 | 0.5761 | 0.6560 | 0.3892 |

Table 7.5 – Palindromic error in the estimator of the mean as function of simulation length. In table (7.7) one can find the actual percentage of accepted trajectories for each simulation length and each sequence.

### 7.2.2 Estimation of 1st and 2nd moments using palindromic symmetry

In statistics it common to take advantage of known symmetries in the analysis of time series of data. More precisely it is good practice to compute estimators from a times

| # $\mathcal{S}$ | $100ns$ | $1\mu$s | $2\mu$s | $3\mu$s |
|---|---|---|---|---|
| 1 | 0.7350 | 0.5820 | 0.2928 | 0.2181 |
| 2 | 1.4683 | 0.5765 | 0.3530 | 0.3328 |
| 3 | 0.9808 | 0.3813 | 0.3006 | 0.2560 |
| 4 | 0.8838 | 0.9820 | 0.4589 | 0.4274 |
| 5 | 2.3520 | 0.7814 | 0.5258 | 0.4302 |
| 6 | 1.4567 | 0.5847 | 0.3588 | 0.2689 |
| 7 | 1.0474 | 0.6137 | 0.4657 | 0.4062 |
| 8 | 0.9003 | 0.4028 | 0.2604 | 0.2264 |
| 9 | 1.7451 | 0.3943 | 0.3885 | 0.3129 |
| 10 | 1.3966 | 0.4018 | 0.2478 | 0.3549 |
| 11 | 1.9304 | 0.9960 | 0.5543 | 0.5571 |
| 12 | 1.3977 | 0.6904 | 0.5096 | 0.4084 |
| 13 | 1.0167 | 0.3396 | 0.3234 | 0.2613 |
| 14 | 0.8852 | 0.4616 | 0.3311 | 0.2753 |
| 15 | 1.2996 | 0.8019 | 0.3354 | 0.2977 |
| 16 | 1.3265 | 0.6571 | 0.6327 | 0.3853 |

Table 7.6 – Palindromic error in the estimator of the covariance as a function of simulation duration. In table (7.7) one can find the actual percentage of accepted trajectories for each simulation length and each sequence.

series of data that satisfy the physical symmetries of the underlying object. In our context the interpretation of the mean as a ground–state of the considered sequence makes it sensible to take into account the Crick–Watson symmetry of palindromes. We can actually define two estimators one for the mean and one the covariance that will account for Crick–Watson symmetry and that consequently can enhance the quality of both estimators. Let $\mu$ and $C$ be the standard estimators for, respectively, mean and covariance computed from a time series of a simulated palindrome $\mathcal{S}$ of length $n$. We introduce the following palindromic symmetrised first moment, where for sake of compactness we have dropped the dependence on $\mathcal{S}$:

$$\mu_{\text{palin}} = \frac{1}{2}\left(\mu + E_{2n-1}\mu\right), \tag{7.14}$$

where $E_{2n-1} \in \mathbb{R}^{12n-6\times 12n-6}$ has been introduced in (4.3). In the estimator (7.14) we have basically doubled our time series by considering also the estimation of the mean of the complementary sequence. For the symmetrized covariance matrix we first compute the symmetrized second moment

$$S_{\text{palin}} = \frac{1}{2}\left(S + E_{2n-1}SE_{2n-1}\right), \text{ where } S = C + \mu\mu^T, \tag{7.15}$$

and then we compute the symmetrized centred second moment as

$$C_{\text{palin}} = S_{\text{palin}} - \mu_{\text{palin}}(\mu_{\text{palin}})^T. \tag{7.16}$$

| # $\mathcal{S}$ | 100$ns$ | 1$\mu$s | 2$\mu$s | 3$\mu$s |
|---|---|---|---|---|
| 1 | 0.84 | 0.79 | 0.85 | 0.87 |
| 2 | 0.91 | 0.92 | 0.90 | 0.90 |
| 3 | 0.81 | 0.84 | 0.88 | 0.90 |
| 4 | 0.91 | 0.92 | 0.92 | 0.92 |
| 5 | 0.62 | 0.91 | 0.92 | 0.93 |
| 6 | 0.63 | 0.90 | 0.90 | 0.79 |
| 7 | 0.96 | 0.93 | 0.90 | 0.91 |
| 8 | 0.94 | 0.94 | 0.93 | 0.93 |
| 9 | 0.95 | 0.94 | 0.88 | 0.90 |
| 10 | 0.94 | 0.92 | 0.93 | 0.93 |
| 11 | 0.89 | 0.89 | 0.90 | 0.91 |
| 12 | 0.85 | 0.92 | 0.93 | 0.93 |
| 13 | 0.88 | 0.92 | 0.78 | 0.82 |
| 14 | 0.94 | 0.94 | 0.93 | 0.94 |
| 15 | 0.95 | 0.90 | 0.92 | 0.93 |
| 16 | 0.92 | 0.93 | 0.93 | 0.93 |

Table 7.7 – The actual percentage of accepted snapshots per simulation for each sequence in the training library.

From $C_{\text{palin}}(\mathcal{S})$ we then compute a banded stiffness matrix $K_{\text{palin}}$ using the algorithm (1).

## 7.3 Palindromic cgDNA parameter set

In section (7.2.1) we have introduced a notion of convergence for MD simulations of palindromic sequences based on the classical estimators and we gave also an interpretation of the values of the convergence error that help in understanding how big the error actually is. For palindromic error we would like to conclude that after three microseconds all the 16 sequences are rather well converged even if after having analysed the three sequences with 10$\mu$s long trajectories it would clearly be better to extend the simulation lengths for every single palindrome. We also would like to stress here that the original cgDNA parameter set has been computed on the ABC training set that contains sequences simulated for a duration from 50 to 100 $n$s, thus a 3$\mu$s training set is already an extreme enhancement in the quality of the data. Thus, for the purpose of this work we will consider the 3$\mu$s palindromic data set along with the palindromic symmetric oligomer–based statistics and in the next section we will compute and show prediction of the best-fit cgDNA parameter fit. For the sake of simplicity we drop the notation

$$\mu_{\text{palin}}(\mathcal{S}) \quad , C_{\text{palin}}(\mathcal{S}) \quad \text{and } K_{\text{palin}}(\mathcal{S})$$

to identify the palindromic symmetric estimation of respectively the mean and the covariance for palindromic sequence $\mathcal{S}$ and use the classic notation $\mu(\mathcal{S})$ and $C(\mathcal{S})$. Thus, in what follows all oligomer–based statistics should always be interpreted as palindromically symmetrized.
Before deriving the best–fit parameter set we introduce a new format for the cgDNA parameter set that we will refer to as the *dimer–based* parameter set. We no longer consider the monomer–based stiffness metric blocks and sigma vectors, but add dedicated independent blocks for each 5' end dimers. In detail the dimer–based cgDNA parameter set has the following format:

$$\mathcal{P} = \left\{\sigma^{5'\alpha\beta}, \sigma^{\alpha\beta}, K^{5'\alpha\beta}, K^{\alpha\beta}\right\}_{\alpha\beta\in D, 5'\alpha\beta\in\overline{D}}, \tag{7.17}$$

and the reconstruction scheme for an arbitrary DNA sequence $\mathcal{S} = X_1X_2,\ldots,X_n$ reads simply

$$K(\mathcal{P},\mathcal{S}) = P_d^T K_d P_d, \tag{7.18}$$
$$\sigma(\mathcal{P},\mathcal{S}) = P_d^T \sigma_d, \tag{7.19}$$
$$\mu(\mathcal{P},\mathcal{S}) = K(\mathcal{P},\mathcal{S})^{-1}\sigma(\mathcal{P},\mathcal{S}), \tag{7.20}$$
$$\tag{7.21}$$

where

$$K_d = \text{diag}(K^{5'X_1X_2},\ldots,K^{X_iX_{i+1}},\ldots,K^{X_{n-1}X_n3'}),$$
$$\sigma_d = (\sigma^{5'X_1X_2},\ldots,\sigma^{X_iX_{i+1}},\ldots,\sigma^{X_{n-1}X_n3'}),$$

with $P_d$ the matrix defined in (4.15). Moreover the $3'$ end parameter set elements are computed from the corresponding $5'$ ones using the Crick–Watson symmetry.
Clearly we have increased considerably the number of independent unknowns in the parameter set $\mathcal{P}$. The reasons behind this choice can be summarized in two points: 1) it is rational to have that each local energy (monomer–based and dimer–based) has a positive definite stiffness matrix. But that is in fact not the case for the current cgDNA parameter set where the set of conditions (5.20) and (5.21) must be satisfied in order to prove its positiveness. The conditions themselves suggested a way for transforming the old cgDNA parameter set format into a dimer–based model one. 2) Dimers at the ends have been observed to have quite different statistics compared to the same dimers in the interior. Thus allowing a specific end element in the parameter set can certainly enhance the quality of the predictions, especially for non $G$p$C$ ends that have in general less instances, and thus less data.
We can now compute two best-fit cgDNA parameter: one using the parameter set format (4.9) and the other using the parameter set (7.17). For sake of simplicity we will refer to the cgDNA parameter set (4.9) as the *old* one while the cgDNA parameter set (7.17) will be called the *new* one. We recall that the cgDNA parameter set is the

solution of the optimization problem (5.17), where the objective function is defined by (5.18). The differences between the two computations will just be in the reconstruction scheme and in the total number of unknowns. The algorithm (2) implements a parameter continuation approach for computing a cgDNA parameter set starting from an already computed one. We have used this approach to compute the best-fit parameter set for the old format, denoted by $\mathcal{P}_{\text{old}}$. For the new format a less direct computation has been used to construct an admissible initial guess. Once the initial guess has been computed a combination of gradient descent and Broyden method have been adopted in order to compute the first cgDNA parameter for the new format, denoted by $\mathcal{P}_{\text{new}}$. All the latter computations were carried by O. Gonzalez. On the other hand the new format was actually proposed in light of the cgDNA+ model implemented in chapter 9 where the end blocks are actually of a different dimension from interior blocks.
With both parameter sets $\mathcal{P}_{\text{old}}$ and $\mathcal{P}_{\text{new}}$ in hand we start by first comparing the prediction of the ground–state shape vectors at the ends of oligomers for non $G\text{p}C$ dimer ends. For example in figure 7.4 we show the first four bases of the sequence $\mathcal{S} = AAGCAAACTAGC$ which was used for deriving the end block $5'AA$. We can see that in general the approximation of the helical parameters at the end dimer $5'AA$ is better for the new format of the parameter set. The same results can be observed in any of the other fifteen non $G\text{p}C$ ends. By following the same comparison technique used in chapter 6, we can compute the averaged Kullback-Leibler divergence per degree of freedom (6.5) for $\mathcal{P}_{\text{old}}$ and $\mathcal{P}_{\text{new}}$ and separately the palindromic sequence library, denoted PALIN, and the ends sequence library, denoted ENDS:

| Format | Library | $\overline{D}$ | $\overline{D^{\dagger}}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|---|
| New | PALIN | $2.09 \cdot 10^{-2}$ | $1.31 \cdot 10^{-2}$ | $0.78 \cdot 10^{-2}$ |
| New | ENDS | $2.25 \cdot 10^{-2}$ | $1.34 \cdot 10^{-2}$ | $0.91 \cdot 10^{-2}$ |
| Old | PALIN | $2.29 \cdot 10^{-2}$ | $1.43 \cdot 10^{-2}$ | $0.86 \cdot 10^{-2}$ |
| Old | ENDS | $3.00 \cdot 10^{-2}$ | $1.79 \cdot 10^{-2}$ | $1.20 \cdot 10^{-2}$ |

We can first observe that for the end data we have a better approximation for the new format which confirms the improved choice of having dedicated parameter set elements for the end dimers. Secondly we observe that even for the palindromic sequences the new format of the parameter set is better at predicting the data. We would like to point the reader to the tables presented in chapter 6 and observe that the average per degree of freedom values of the Kullback-Leibler for the PALIN $\mathcal{P}_{\text{new}}$ is the lowest obtained value between all the other parameter sets and libraries.
In figure 7.5 we show an example of predictions for sequence number 1 of PALIN using the parameter set $\mathcal{P}_{\text{new}}$. As the sequence is a palindrome, and as cgDNA reconstruction and the MD mean estimator defined by (7.14) both satisfy perfectly the palindromic symmetries we show the helical parameters for just the half of the sequence. For $\mathcal{P}_{\text{new}}$ we also compute the spectra of apparent and dynamic persistence length but over an ensemble of one million randomly generated DNA sequences each of length

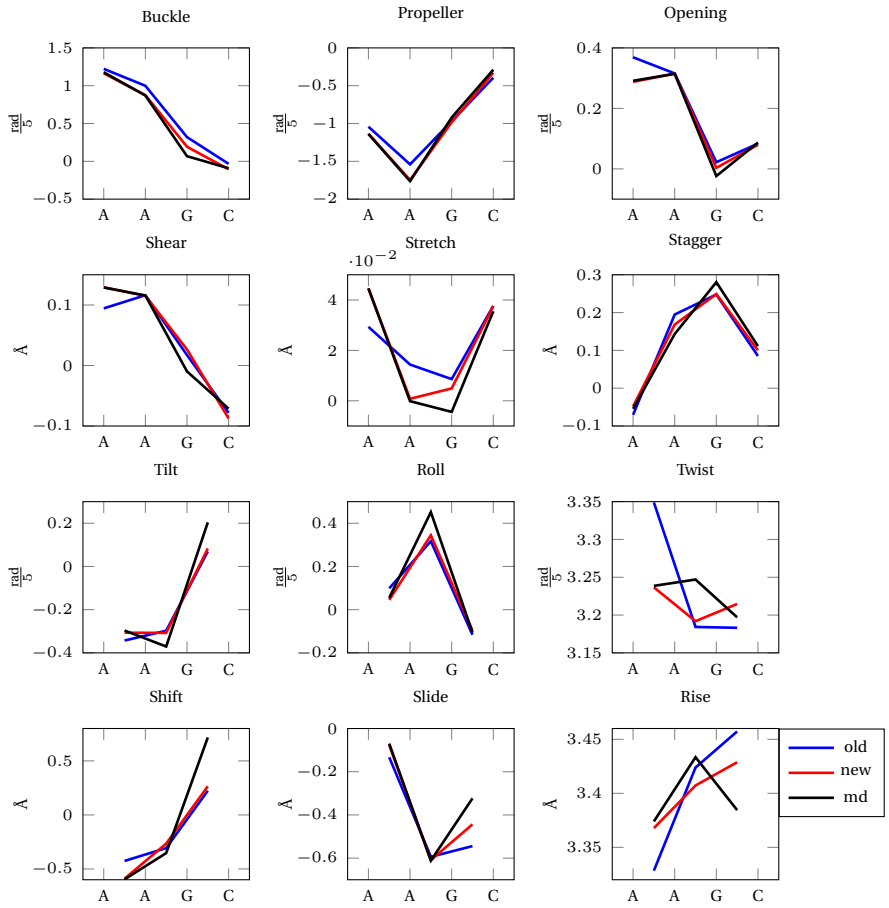



Figure 7.4 – Comparisons of internal coordinates of MD ground–state (black) and cgDNA predictions with the old parameter set format (blue) and the new format (red).

220. In figure 7.6 we show the two histograms and the values of $\ell_p$ and $\ell_d$ for the six independent poly dimers. The first observation is that the overall features of the histograms are in general the same as the spectra showed in chapter 6, meaning that for the apparent persistence lengths the histogram is more wide with a long tail spreading to the low values region of $\ell_p$, while the histogram for $\ell_d$ is more peaked around the mean, with a thin tail spreading in the direction of high values.

### 7.3.1 Positiveness of the new format

In the previous section we have presented a new format of the cgDNA parameter set which in particular is composed of only dimer–based units blocks for the stiffness matrices and vectors for the weighted shapes. This particular format implies that there is not an unique solution to the optimization problem (5.17). More precisely, the linear transformation that maps a parameter set $\mathcal{P}$ and a training library $Lb$ to their respectively reconstruction, is not injective. This concept will be better explored in the

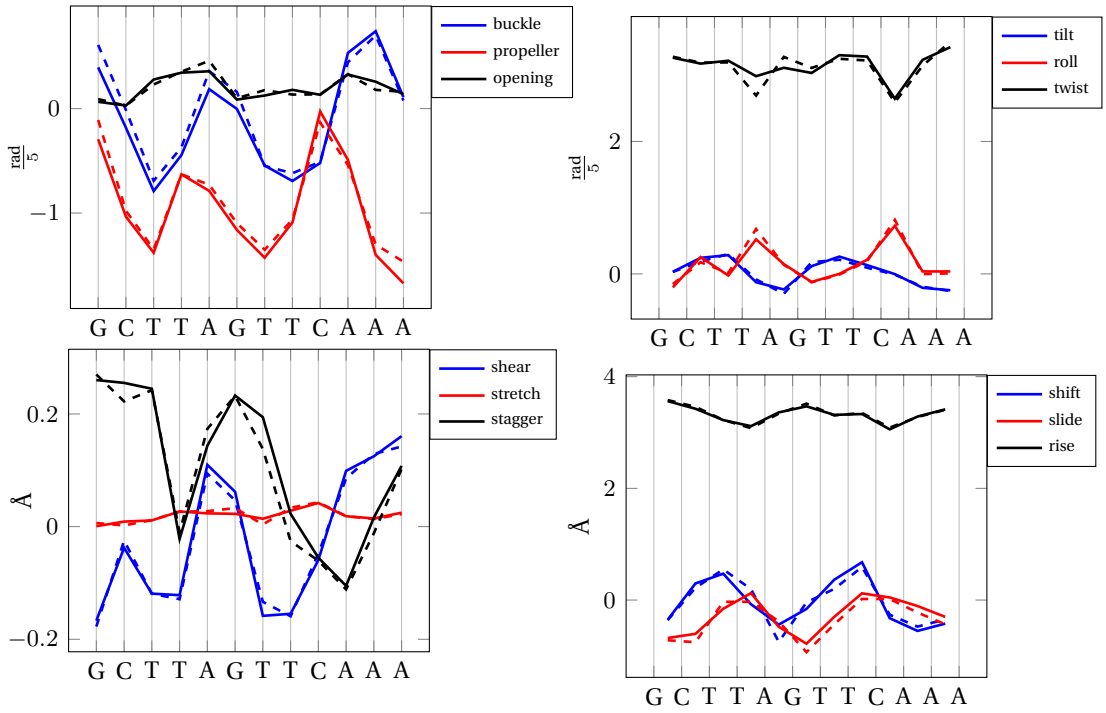



Figure 7.5 – Comparison between cgDNA (*solid*) and MD (*dashed*) ground–states for sequence 1 in the palindromic library.

context of the cgDNA+ model in chapter 9. In this section we want just to make the following point: with parameter set format (7.17) the sufficient conditions (5.20) and (5.21) become simply that each dimer–based stiffness block must be positive definite. For the best-fit palindromic cgDNA parameter the latter condition is fully satisfied and thus we can refer to $\mathcal{P}$ trained on the palindromic data set as a positive definite cgDNA parameter set.
In chapter 9 we will discuss the case when the stiffness parameter set blocks are not positive definite and thus the sufficient conditions are not satisfied. We will then show how to take advantage of the non–uniqueness of the parameter set in order to be able to recover the positiveness.

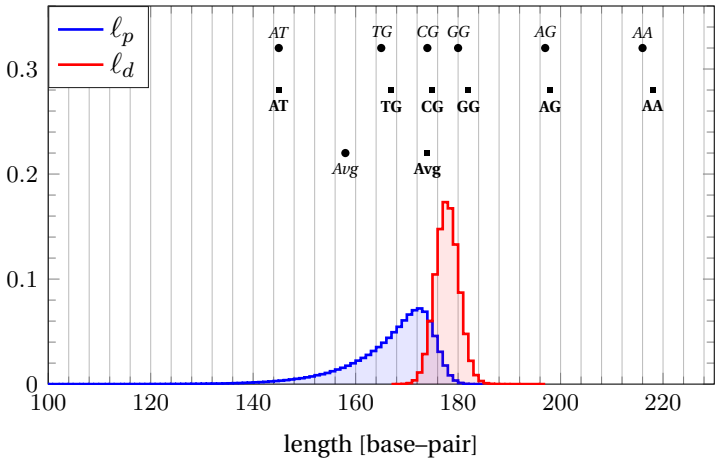



Figure 7.6 – Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA, trained on the Palindromic data set, over a sequence ensemble of 1 million randomly generated sequences of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for the apparent and bold font for the dynamic persistence length. The values of the persistence lengths for six independent poly–dimers of length 220 are reported: italic for the apparent and bold for the dynamic. The positions of the values of the apparent persistence length is given by a circle while the positions for the dynamic is given by a square.

# Part III

# cgDNA+

# 8 Coarse–grain configuration variables for double stranded DNA

In this chapter, we coarse–grain double–stranded DNA by considering both strands as single unit chains composed by a repeating pattern of a composite unit comprising a phosphate group rigid–body and a nucleic rigid–body. We introduce two representations of coarse–grain double–stranded DNA: two *interacting strands* and the *tetrachain*. The two coarse–grained models are in analogy with, respectively, the double chain and bichain description presented in [21]. The main difference between the bichain and the tetrachain is in the composition of a base–pair level which, for the tetrachain, is formed by four rigid–bodies and not two as in the bichain model. Consequently, the definition of the microstructure will also be different. For the tetrachain model, we define then an internal elastic energy describing nearest–neighbour interactions and we compute its coordinate free first variation. From the first variation, we then compute the formulas for the total external loads acting on a single phosphate group, again, in a coordinate free framework. Finally, we define the internal coordinates for the tetrachain representation by parametrising the inter– and intra–base–pair relative displacement using the Cayley transformation for the rotation part and by writing the translational part in the mid frames. The microstructure of the tetrachain is composed of two base–to–phosphate group rigid–body transformations which are parametrised using, again, the Cayley transformation for the rotation part and the translational part written in the base frame. Thanks to the choice of internal coordinates we can write explicit formulas for the total external loads acting on a single phosphate group.

## 8.1 Double stranded DNA configurations with explicit backbone treatment

We will consider a double–stranded DNA (dsDNA) molecule as the interaction of two rigid body chains representing the anti–parallel strands. Each rigid body chain is formed by the repetition of a phosphate group followed by a nucleic base in the $5' \to 3'$ direction, as shown in figure 8.1. Let us consider a DNA fragment with sequence

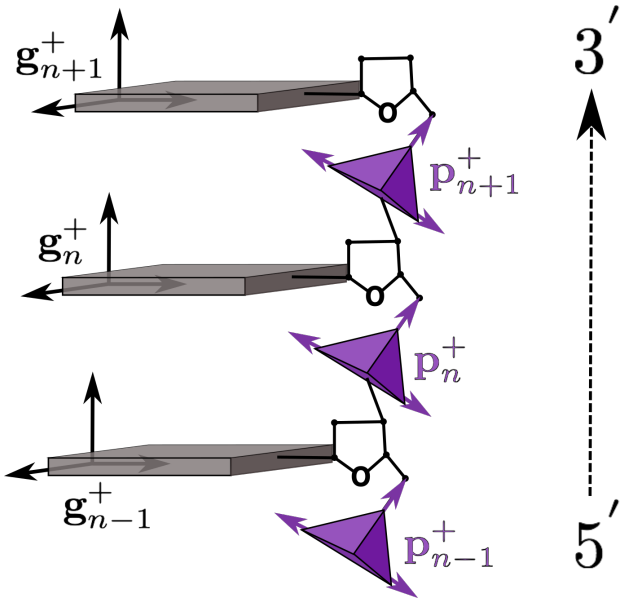



Figure 8.1 – Schematic representation of single stranded DNA $\mathcal{S}$ fragment composed by three consecutive nucleotides. In the figure the sugar ring is shown but is only treated implicitly in the model.

$S = X_1, X_2, \ldots, X_N$, $X_i \in \{A, T, G, C\}$. Mathematically, we will define by $\mathcal{S} \in SE(3)^{2N}$ the *reading strand* and by $\overline{\mathcal{S}} \in SE(3)^{2N}$ the *complementary strand*. By using , $\mathbf{g}^+_{\in} SE(3)$ for the nucleic bases and $\mathbf{p}^+_{\in} SE(3)$, for the phosphate group we formally have that the two strands can be written as

$$\mathcal{S}^+ = (\mathbf{p}^+, \mathbf{g}^+) = (\mathbf{p}^+_1, \mathbf{g}^+_1, \mathbf{p}^+_2, \mathbf{g}^+_2 \ldots, \mathbf{p}^+_N, \mathbf{g}^+_N) \in SE(3)^{2N}, \tag{8.1}$$

and

$$\overline{\mathcal{S}}^+ = (\overline{\mathbf{p}}^+, \overline{\mathbf{g}}^+) = (\overline{\mathbf{p}}^+_1 \overline{\mathbf{g}}^+_1, \overline{\mathbf{p}}^+_2, \overline{\mathbf{g}}^+_2, \ldots, \overline{\mathbf{p}}^+_N, \mathbf{g}^+_N) \in SE(3)^{2N}, \tag{8.2}$$

where for sake of compactness, each rigid–body associated to a base, on both strands, is in fact related to the considered sequence as the bases have different atomistic composition depending on their base type. For example for the reading strand we should have written

$$\mathcal{S}^+ = (\mathbf{p}^+, \mathbf{g}^+) = (\mathbf{p}^+_1, \mathbf{g}^+_1(X_1), \mathbf{p}^+_2, \mathbf{g}^+_2(X_2), \ldots, \mathbf{p}^+_N, \mathbf{g}^+_N(X_N)) \in SE(3)^{2N}. \tag{8.3}$$

On the other hand, all the phosphate groups have the same chemical structure thus they have no explicit dependence on the sequence. For what follows we have decided to drop the dependence on the sequence for all the base rigid–bodies in the notation. We can introduce now the two *backbone rigid body displacements* $a^\zeta_n$, $a^\alpha_n \in SE(3)$, along the reading strand, describing the rigid body motion from $\mathbf{g}^+_n$ to $\mathbf{g}^+_{n+1}$ through

the phosphate group $\mathbf{p}_n^+$ as

$$\mathbf{g}_{n+1}^+ = \mathbf{g}_n^+ a_n^{+\zeta} a_n^{+\alpha},$$

where

$$a_n^{+\zeta} = [\mathbf{g}_n^+]^{-1}\mathbf{p}_{n+1}^+,$$
$$a_n^{+\alpha} = [\mathbf{p}_{n+1}^+]^{-1}\mathbf{g}_{n+1}^+.$$

With the latter notation we have directly the following relation to the double chain inter rigid body displacement $a_n^+ = a_n^{+\zeta} a_n^{+\alpha}$ introduced in [21] For the complementary strand the notation will be the same but with an over line. The two strands interact at the level of the bases. Thus we introduce an *intra–base–pair* rigid–body displacement. As both strands are oriented in the $5' \to 3'$ direction we apply a transformation to every single base frame on the complementary strand to avoid rotations close to $\pi$ in the intra–base–pair rigid–body displacement. We denote the intra rigid-body displacement by $b_n \in SE(3)$ defined as

$$b_n = [\overline{\mathbf{g}}_{N-n+1}^+ t]^{-1}\mathbf{g}_n^+ = t[\overline{\mathbf{g}}_{N-n+1}^+]^{-1}\mathbf{g}_n^+,$$

where $t \in O(3) \times \mathbb{R}^{3\times 3}$ is the re-framing transformation

$$t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{8.4}$$

which satisfies $t = t^{-1}$. For sake of simplicity we introduce the frame

$$\mathbf{g}_n^- = \overline{\mathbf{g}}_{\sigma(n)}^+,$$

and the corresponding intra rigid–body transformation

$$b_n = t[\mathbf{g}_n^-]^{-1}\mathbf{g}_n^+, \tag{8.5}$$

where $\sigma(n) = N - n + 1$. We can also introduce an equivalent transformation for the phosphate frames on the complementary strand which simply reorder them in the $3' \to 5'$ direction, namely:

$$\mathbf{p}_n^- = \overline{\mathbf{p}}_{\sigma(n)}^+.$$

We finally obtain that a dsDNA can be coarse grained using two interacting single chains $(\mathcal{S}^-, \mathcal{S}^+)$, where $\mathcal{S}^\pm = (\mathbf{g}_1^\pm, \mathbf{p}_1^\pm, \ldots, \mathbf{p}_{N-1}^\pm, \mathbf{g}_N^\pm)$ with direct interactions at base–pair level given by (8.5). The coarse grain model $(\mathcal{S}^-, \mathcal{S}^+)$ will be called *two interacting strands*. In figure (8.2) we show a schematic representation of a two interacting strands configuration. The two interacting strand point of view, is an extension of the double-chain representation of the dsDNA where the phosphate group were not explicitly included in the model. Thus, we can continue the analogy, by introducing an equiva-

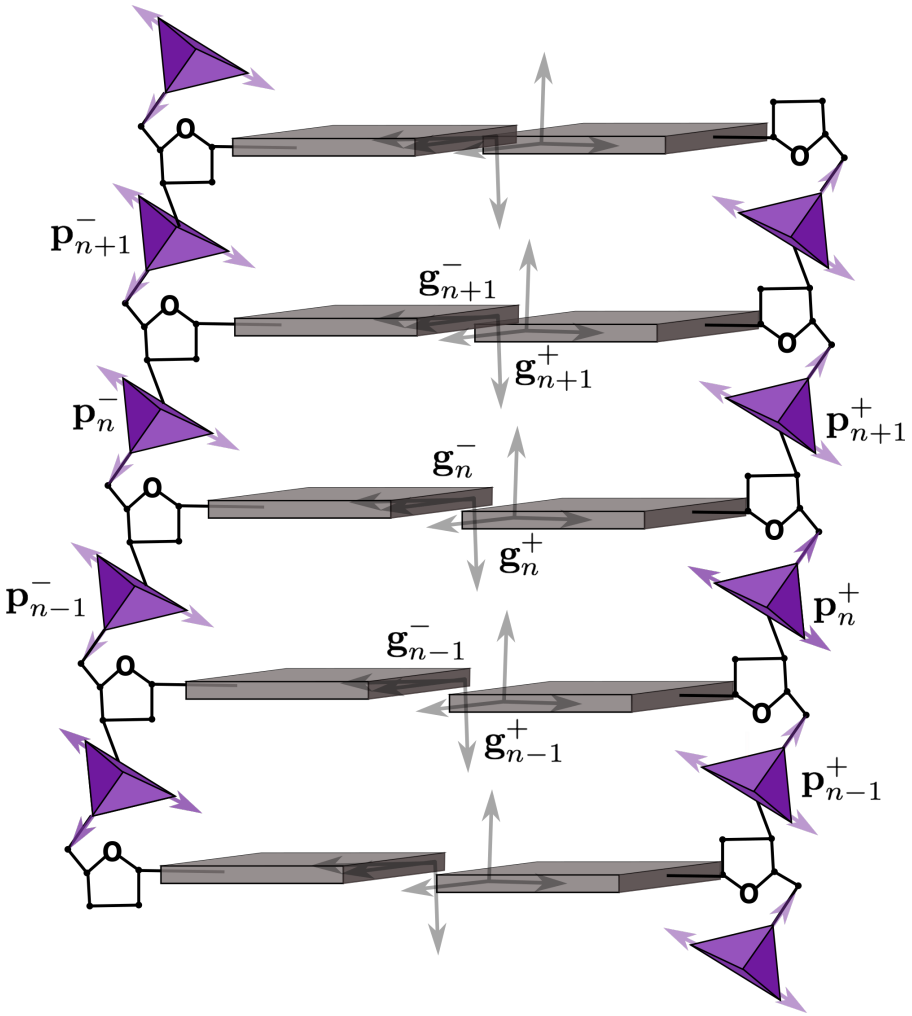



Figure 8.2 – Schematic representation of two the interacting strands representation of double stranded DNA. The sugar ring is shown but is not explicitlty modelled.

lent point of view to the bichain model, see for instance section (2.4), by introducing a *macrostructure* and a *microstructure*. The macro structure will be identified with the ensemble of base–pair rigid–bodies $\mathbf{g} \in SE(3)^N$ defined as

$$\mathbf{g}_n = \mathbf{g}_n^- t\mathcal{P}_n^H = \mathbf{g}^+ \mathcal{P}_n^{-H},$$

where $\mathcal{P}_n^H \in SE(3)$ is defined by

$$\mathcal{P}^H = \begin{bmatrix} ([R_n^-]^T R_n^+)^{\frac{1}{2}} & \frac{1}{2}[R_n^-]^T (r_n^+ - r_n^-) \\ 0 & 1 \end{bmatrix}.$$

The micro structure associated to the two interacting strands configuration contains the intra–base–pair rigid body displacement introduced in (2.4), just as for the bichain model, but now additionally with two base–to–phosphate group rigid body displacements. As a nucleic acid base is covalently bonded to two adjacent phosphate groups, we need to consider two possible different definitions for the microstructure.

### 8.1.1 Microstructure with explicit treatment of the phosphate group

In order to integrate the phosphate groups into the definition of the micro structure we have to include two base–to–phosphate group rigid body displacements per base-pair level. We will use the notion of *base–pair level* to refer to a unit composed of four rigid bodies: two complementary bases and two phosphate groups, one on the reading strand and the other on the complementary strand. Is clear from the schematic representation of dsDNA 8.1 that each base is in fact covalently bonded to two phosphate groups, a $3'$ one and a $5'$ one. The first possible choice is to consider two units formed by a base – sugar ring – $3'$–phosphate group, formally composed by the following rigid–body frames:

$$(\mathbf{p}_{n-1}^-, \mathbf{g}_n^-, \mathbf{g}_n^+, \mathbf{p}_{n+1}^+) \in SE(3)^4, \tag{8.6}$$

while the second option is to consider the base–pair level as the set of two nucleotide, base – sugar ring – $5'$–phosphate group. With this approach the $n$th base–pair level is formed by the following rigid-body frames:

$$(\mathbf{p}_n^-, \mathbf{g}_n^-, \mathbf{g}_n^+, \mathbf{p}_n^+) \in SE(3)^4. \tag{8.7}$$

For simplicity of the Crick–Watson symmetry transformation (discussed later) it is important that in both cases the definition of the units has precisely the same chemical structure on both strands. We define now the *local micro structure configuration* as the following ensemble of three rigid–body, motions associated to the chain of four

rigid–bodies forming the base-pair:

$$\mathcal{M}_n = (\mathcal{B}_n^{\mp}, \mathcal{P}_n, \mathcal{B}_n^{\pm}) \in SE(3)^3, \tag{8.8}$$

where $\mathcal{P}_n$ is the standard intra-base-pair displacement, and $\mathcal{B}_n^{\pm}$ represents the $n$th base to phosphate group rigid–body motions, according to the strand, $+$ or $-$, and to the definition of base–pair level, I or II. In detail:

$$\text{for I} \quad : \quad \mathcal{B}_n^{+} = [\mathbf{g}_n^{+}]^{-1}\mathbf{p}_{n+1}^{+},\ \mathcal{B}_n^{-} = [\mathbf{g}_n^{-}]^{-1}\mathbf{p}_{n-1}^{-}, \tag{8.9}$$

$$\text{for II} \quad : \quad \mathcal{B}_n^{\pm} = [\mathbf{g}_n^{\pm}]^{-1}\mathbf{p}_{n}^{\pm}, \tag{8.10}$$

Without lost of generality we will assume that the base–pair level is always formed by four rigid bodies, which implies that the composition of the strand should be adapted according to the base–pair level type. Thus we can introduce *tetrachain* configurations $(\mathbf{g}, \mathcal{M})$ for double stranded DNA where

$$\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_N) \in SE(3)^N, \text{ and } \mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_N) \in SE(3)^N. \tag{8.11}$$

In general, independent of the composition of the base-pairs level, we will have an invertible mapping between two interacting strand configurations and tetrachain configurations, i.e,

$$(\mathcal{S}^{+}, \mathcal{S}^{-}) \leftrightarrow (\mathbf{g}, \mathcal{M}).$$

## 8.2 Internal energy for tetrachains

We introduce now the notion of internal energy for tetrachain configurations. The first assumption on our energy is that the interaction are physical local, and more precisely, we will consider only *nearest neighbour interactions*. This assumption of only local interactions, and thus local contribution to energies, has been already used for previous models of DNA such as the cgDNA model [55, 19] and the rigid–base–pair model [51, 78]. In both coarse grain models the total energy of the molecule is expressed as a sum of local energies defined at the level of single *junction* (two base-pairs level). In the previous paragraph we extended the notion of base–pair level from two complementary bases to two complementary bases plus two phosphate groups. In fact, this new notion of base–pair level can be seen as an ensemble of two units, where each unit is composed of a base and an adjacent phosphate group. With the latter point of view a junction, in a double strand configuration is composed by four units, and the nearest neighbour assumption will lead to the following statement: *each unit has five nearest neighbours* just as in the cgDNA rigid–base model, see chapter 4.

Mathematically the internal energy $E : SE(3)^{4N} \longrightarrow \mathbb{R}$ takes the following form

$$\mathbf{E}(\mathbf{g}, \mathcal{M}) = \sum_{n=1}^{N-1} w_n(\mathcal{M}_n, a_n, \mathcal{M}_{n+1}), \tag{8.12}$$

where $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_N)$, $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_N)$, $\mathcal{M}_n$ are defined in (8.8), and $a_n = \mathbf{g}_n^{-1}\mathbf{g}_{n+1}$. In equation (8.12) the local contributions $w_n$ to the energy **E** are given in general form but in the context of DNA modelling these contributions will be chosen to be quadratic in the internal coordinates and, moreover, with local sequence–dependence.

### 8.2.1 Equilibrium configurations and variational principle

We compute the first variation of the internal elastic energy (8.12) with respect to the macro structure **g** and the micro structure $\mathcal{M}$:

$$D_l E(\delta\mathbf{g}, \delta\mathcal{M}) = D_l E \delta\mathbf{g} + D_l E \delta\mathcal{M}. \tag{8.13}$$

For sake of compactness we will use the short notation $w_n \equiv w_n(\mathcal{M}_n, a_n, \mathcal{M}_{n+1})$. The expression for the first quantity has already been computed in (2.74) and (2.76), thus we only focus on the first variation of the energy with respect to the microstructure:

$$\begin{aligned} D_l E \delta\mathcal{M} = \sum_{n=1}^{N} & \mathcal{T}^* \left( \partial_{\mathcal{B}_n^+} (w_n + w_{n-1}) [\mathcal{B}_n^+]^T \right) \cdot \Theta_n^+ \\ & + \mathcal{T}^* (\partial_{\mathcal{P}_n} \left( w_n + w_{n-1} \right) \mathcal{P}_n^T) \cdot \Theta_n^{\mathcal{P}} \\ & + \mathcal{T}^* (\partial_{\mathcal{B}_n^-} \left( w_n + w_{n-1} \right) [\mathcal{B}_n^-]^T) \cdot \Theta_n^- \\ & = \sum_{n=1}^{N} \mu_n^+ \cdot \Theta_n^+ + \mu_n^{\mathcal{P}} \cdot \Theta_n^{\mathcal{P}} + \mu_n^- \cdot \Theta_n^-, \end{aligned} \tag{8.14}$$

where

$$\begin{aligned} \mu_n^{\pm} &= \mathcal{T}^* \left( \partial_{\mathcal{B}_n^{\pm}} (w_n + w_{n-1}) [\mathcal{B}_n^{\pm}]^T \right), \\ \mu_n^{\mathcal{P}} &= \mathcal{T}^* \left( \partial_{\mathcal{P}_n} (w_n + w_{n-1}) \mathcal{P}_n^T \right), \end{aligned} \tag{8.15}$$

and, by convention, $w_0^{\mathcal{M}} = w_N^{\mathcal{M}} = 0$. We want to emphasize here that the expressions (8.15) are completely coordinate free, in the sense that no explicit parametrisation of the relative rigid–body transformation has been introduced yet. This implies a high level of generality. On the other hand for the computation of (8.14) the use of exponential coordinates has been adopted for the rigid–body absolute coordinates representing the bases and the phosphate groups.

## 8.3 Total force and torque acting on a single phosphate group

The derivation of a close form expression of the total external load acting on a single phosphate group is obtained as the simplification of the first variation (8.14) under the following constraints:
$\forall n = 1, \dots, N$

$$\mathbf{g}_n \text{ is fixed } \Rightarrow \delta\mathbf{g}_n = \mathcal{T}\theta_n\mathbf{g}_n = 0 \Rightarrow \theta_n = 0,$$
$$\mathbf{g}_n^{\pm} \text{ is fixed } \Rightarrow \delta\mathbf{g}_n^{\pm} = \mathcal{T}\theta_n^{\pm}\mathbf{g}_n^{\pm} = 0 \Rightarrow \theta_n^{\pm} = 0.$$

The latter constraints implies that each intra–base–pair rigid body displacement is also fixed which, moreover, implies that
$\forall n = 1, \dots, N$

$$\mathcal{P}_n \text{ is fixed } \Rightarrow \delta\mathcal{P}_n = \mathcal{T}\Theta_n\mathcal{P}_n = 0 \Rightarrow \Theta_n = 0,$$

If now one wants to consider the total external load acting on the $m$–th phosphate group on the Watson strand, and thus one wants to compute the variation of the energy (8.12) with respect to the rigid–body $\mathbf{p}_m^+$, two additional constraints should be added:
$\forall n = 1, \dots, N$

$$\mathbf{p}_n^- \text{ is fixed } \Rightarrow \delta\mathbf{p}_n^- = \mathcal{T}\vartheta_n^-\mathbf{p}_n^- = 0 \Rightarrow \vartheta_n^- = 0,$$

and
$\forall n \neq m$

$$\mathbf{p}_n^+ \text{ is fixed } \Rightarrow \delta\mathbf{p}_n^+ = \mathcal{T}\vartheta_n^+\mathbf{p}_n^+ = 0 \Rightarrow \vartheta_n^+ = 0.$$

Again by combining the constraints on the bases and on the phosphate groups we can conclude that,
$\forall n = 1, \dots, N$

$$\mathcal{B}_n^- \text{ is fixed } \Rightarrow \delta\mathcal{B}_n^- = \mathcal{T}\Theta_n^-\mathcal{B}_n^- = 0 \Rightarrow \Theta_n^- = 0, \tag{8.16}$$

and
$\forall n \neq m$

$$\mathcal{B}_n^+ \text{ is fixed } \Rightarrow \delta\mathcal{B}_n^+ = \mathcal{T}\Theta_n^+\mathcal{B}_n^+ = 0 \Rightarrow \Theta_n^+ = 0. \tag{8.17}$$

Thus expression (8.14), under the latter constraint, becomes

$$(D_l E)\vartheta_m^+ = \mathrm{Ad}_{\mathbf{g}_m^+}^{-T}\mu_m^+ \cdot \vartheta_m^+. \tag{8.18}$$

By using the relation between left and right infinitesimal variations

$$\vartheta_m^+ = \mathrm{Ad}_{\mathbf{p}_m^+}\varphi_m^+, \tag{8.19}$$

we obtain that the variation, on the right of the energy (8.12) with respect to the $m$–th phosphate group on the Watson strand $\mathbf{p}_m^+$ can be written as:

$$(D_r E)\varphi_m^+ = \mathrm{Ad}^{-T}_{\mathcal{B}_m^+}\mu_m^+ \cdot \varphi_m^+. \tag{8.20}$$

Now we can define the total external load acting on $\mathbf{p}_m^+$ by

$$\lambda_m^+ = -\mathrm{Ad}^{-T}_{\mathcal{B}_m^+}\mathcal{T}^*\left(\partial_{\mathcal{B}_m^+}(w_m + w_{m-1})[\mathcal{B}_m^+]^T\right) \in \mathbb{R}^6. \tag{8.21}$$

with

$$\lambda_m^+ = \begin{bmatrix} c_m^+ + [r_m^p]^+ \times f_m^+ \\ f_m^+ \end{bmatrix}, \tag{8.22}$$

where $c_m^+$ is the external couple, around the phosphate position $[r_m^p]^+$, and $f_m^+$ is the total external force acting on $\mathbf{p}_m^+ = ([R_m^p]^+, [r_m^p]^+)$.

## 8.4 Internal coordinates for tetrachain configurations

In practice the internal energy will be given as a function of internal coordinates and not as a function of rigid body internal displacements. In this paragraph we define the internal coordinates associated to tetrachain configurations. First we recall the notation of the internal coordinates for the intra– and inter– base–pair of bichain configurations:

$$y_n = (\eta_n, \mathrm{w}_n) \leftrightarrow \mathcal{P}(y_n), \equiv \mathcal{P}_n$$
$$x_n = (u_n, v_n) \leftrightarrow a(x_n) \equiv a_n.$$

For the base-to-phosphate group coordinates we recall that two definitions are possible, and, in particular, a base–pair level is composed by three internal coordinates. For both definitions we have that, in internal coordinates, the micro structure at each base–pair level is

$$m_n = (z_n^\pm, y_n, z_n^\mp), \tag{8.23}$$

where

$$z_n^\pm = (\eta_n^\pm, \mathrm{w}_n^\pm) \leftrightarrow \mathcal{B}_n^\pm(z_n^\pm) \equiv \mathcal{B}_n^\pm. \tag{8.24}$$

The parametrisation of the base–to–phosphate rigid–body displacement has a different definition than the one chosen in the inter– and intra– rigid body transformations. In fact, we have decided to simplify the definition by writing the base–to–phosphate translation directly in the base frame. We recall that for the cgDNA internal coordinates the translational parts are written in the mid–frame: i.e. the junction frame for the inters and the base–pair frame for the intras. The motivation behind the introduction of the mid–frames in the definition of the translation rely primarily on the fact that such coordinates lead to a simple linear change of reading strand transformation, see for instance chapter 4. For the phosphate degrees of freedom the choice of writing the translation in the base frame will also lead to a simple change of reading strand transformation that will be detailed in the next section (8.4.1).
In general, let $\mathbf{g} = (R, r) \in SE(3)$ be a base frame and let $\mathbf{p} = (R^p, r^p) \in SE(3)$ be a phosphate group frame. The base–to–phosphate rigid–body transformation is defined by

$$\mathcal{B} = \mathbf{g}^{-1}\mathbf{p} = \begin{bmatrix} R^T R^p & R^T(r^p - r) \\ 0 & 1 \end{bmatrix}.$$

Let $z = (\eta, \mathrm{w}) \in \mathbb{R}^6$ such that $\mathcal{B} \equiv \mathcal{B}(z)$, or,

$$\eta = cay_\alpha^{-1}(R^T R^p), \tag{8.25}$$
$$\mathrm{w} = R^T(r^p - r), \tag{8.26}$$

where $cay_\alpha^{-1} : SO(3) \to \mathbb{R}^3$ is the inverse Cayley transformation defined by

$$cay_\alpha^{-1}(R) = \frac{2\alpha}{1 + \mathrm{trace}(R)}\mathrm{vec}(R - R^T), \tag{8.27}$$

with $\mathrm{vec}(S) = (S_{32}, S_{13}, S_{21})$ and $\alpha = 5$ in our context.
Finally the internal coordinates for tetrachain configurations can be concatenated in the following way, independent of the definition of base–pair level:

$$\omega = (m_1, x_1, m_2, x_2, \ldots, x_{n-1}, m_n) \in \mathbb{R}^{24N-18}. \tag{8.28}$$

The internal energy (8.12) can then be written in term of the internal coordinates $\omega$ associated to the coarse–grained representation $(\mathbf{g}, \mathcal{M})$. In fact we have simply the following relationship

$$\mathbf{E}(\mathbf{g}, \mathcal{M}) = \sum_{n=1}^{N-1} w_n(\mathcal{M}_n, a_n, \mathcal{M}_{n+1}) = \sum_{n=1}^{N-1} w_n^\omega(m_n, x_n, m_{n+1}) = \mathbf{E}(\omega), \tag{8.29}$$

with the additional property that

$$w_n(\mathcal{M}_n, a_n, \mathcal{M}_{n+1}) = w_n^\omega(m_n, x_n, m_{n+1}),\ \forall n = 1, \ldots, N-1. \tag{8.30}$$

The term-by-term equivalence (8.30) is important for example in order to derive a tractable expression for (8.15). More precisely, we will derive an analytical expression for the micro structure internal loads that will depend on the geometric configuration, the ground–state, and the stiffness matrix. The key to derive such expression is the linearisation of the expansion of the element of the algebra $se(3)$ in term of the internal coordinates. More precisely, let $\mathcal{B} \in SE(3)$ and consider the following expansion of $\mathcal{B}$ around $\overline{\mathcal{B}}$

$$\mathcal{B} = (I_4 + \mathcal{T}\Theta)\,\overline{\mathcal{B}} + o(|\Theta|). \tag{8.31}$$

Consider now that $\mathcal{B} \equiv \mathcal{B}(z)$ and $\overline{B} \equiv \overline{\mathcal{B}}(\overline{z})$, with $z, \overline{z} \in \mathbb{R}^6$, thus we have that $\Theta \equiv \Theta(z)$. We then linearise the latter equation, meaning that there exist $\mathbb{L}_{\overline{z}} \in \mathbb{R}^{6\times 6}$ such that

$$\Theta(z) = \mathbb{L}_{\overline{z}}(z - \overline{z}) + o(|z - \overline{z}|). \tag{8.32}$$

For the cgDNA coordinates, the matrix $\mathbb{L}$ was computed in [21] and is reported in (2.87). For the base–to–phosphate internal coordinates presented in this section, the definition of the matrix $\mathbb{L}_z$ is straightforward because the rotational and the translational parts of the coordinates are decoupled. The linear mapping $\mathbb{L}_z$, for $z = (\eta, \mathrm{w}) \in \mathbb{R}^6$, is simply defined by

$$\mathbb{L}_z = \begin{bmatrix} \mathbb{P}_1(\eta) & 0 \\ 0 & I \end{bmatrix} \in \mathbb{R}^{6\times 6}, \text{ with } \mathbb{P}_1(\eta) = \frac{4\alpha^2}{4\alpha^2 + \left(\frac{|\eta|}{2}\right)^2}\left(I + \frac{1}{2\alpha}[\eta\times]\right). \tag{8.33}$$

Finally we compute the first variation on the left for the left–hand side term in (8.30) and the partial derivative of the right–hand side term of the same equation with respect, for example, to $\mathcal{B}_n^+ \equiv \mathcal{B}^+(z_n^+)$:

$$\begin{aligned} &D_l(w_n)\Theta_n^+ = \partial_{\mathcal{B}_n^+} w_n : \mathcal{T}\Theta_n^+\mathcal{B}_n^+ = \partial_{\mathcal{B}_n^+} w_n [\mathcal{B}_n^+]^T : \mathcal{T}\Theta_n^+ = \mathcal{T}^*\left(\partial_{\mathcal{B}_n^+} w_n [\mathcal{B}_n^+]^T\right)\cdot\Theta_n^+, \\ &D(w_n^\omega)\delta z_n^+ = \partial_{z_n^+} w_n^\omega \cdot \delta z_n^+. \end{aligned} \tag{8.34}$$

We use now the linearisation assumption (8.32)

$$\Theta_n^+ = \mathbb{L}_{z_n^+}\delta z_n^+,$$

and replace it in (8.34), to finally obtain

$$\mathbb{L}_{z_n^+}^T \mathcal{T}^*\left(\partial_{\mathcal{B}_n^+} w_n [\mathcal{B}_n^+]^T\right)\cdot\delta z_n^+ = \partial_{z_n^+} w_n^\omega \cdot \delta z_n^+,$$

and

$$\mathcal{T}^*\left(\partial_{\mathcal{B}_n^+} w_n [\mathcal{B}_n^+]^T\right) = \mathbb{L}_{z_n^+}^{-T}\partial_{z_n^+} w_n^\omega. \tag{8.35}$$

The coordinate version of (8.15) reads

$$\begin{aligned}\mu_n^{\pm} &= \mathbb{L}_{z_n^{\pm}}^{-T}\partial_{z_n^{\pm}}\left(w_n^{\omega}+w_{n-1}^{\omega}\right),\\ \mu_n^{\mathcal{P}} &= \mathbb{L}_{y_n}^{-T}\partial_{y_n}\left(w_n^{\omega}+w_{n-1}^{\omega}\right),\end{aligned} \tag{8.36}$$

where $\mathbb{L}_{z_n^{\pm}}$ is defined in (8.33) while $\mathbb{L}_{y_n}$ is defined in (2.87). Consequently we have also derived explicit formulas for the total external loads acting on a single phosphate by using (8.36) in (8.18-8.20). We can also rewrite the total external load acting on a single phosphate as

$$\lambda_n^{\pm} = -\mathrm{Ad}_{\mathcal{B}_n^{\pm}}^{-T}\mathbb{L}_{z_n^{\pm}}^{-T}\partial_{z_n^{\pm}}\left(w_n^{\omega}+w_{n-1}^{\omega}\right)\in\mathbb{R}^6, \tag{8.37}$$

$\forall n=1,\dots,N$.

### 8.4.1 Change of reading strand transformation

In chapter 4 we showed the relation between internal coordinates of a sequence $S$ and its complementary $\overline{S}$ in the context of bichain configurations. We now just discuss the analogous relation for the case of tetrachain configurations. In particular the intra– and the inter– base–pair coordinates still satisfy the linear relations given by the matrix $E=\mathrm{diag}(-1,1,1,-1,1,1)$, see for instance (4.2). Formally, for tetrachain configurations of DNA, the change of reading strand reads:

$$\overline{\mathcal{S}}^{+}\text{ is the reading strand }\;\Rightarrow\;\overline{\mathbf{g}}_n^{-}=\mathbf{g}_{\sigma(n)}^{+}\;\text{ and }\;\overline{\mathbf{p}}_n^{-}=\mathbf{p}_{\sigma(n)}^{+},\;\forall n=1,\dots,N, \tag{8.38}$$

where $\sigma(n)=N-n+1$ and $(\overline{\mathcal{S}}^{+},\overline{\mathcal{S}}^{-})$ denote the double stranded configuration of the DNA molecule with sequence $\overline{S}$. This implies that, for example in the case of base–pair level of type I, see for instance (8.9):

$$\overline{\mathcal{B}}_n^{\pm}=[\overline{\mathbf{g}}_n^{\pm}]^{-1}\overline{\mathbf{p}}_n^{\pm}=[\mathbf{g}_{\sigma(n)}^{\mp}]^{-1}\mathbf{p}_{\sigma(n)}^{\mp}=\mathcal{B}_{\sigma(n)}^{\mp},\;\forall n=1,\dots,N. \tag{8.39}$$

which implies the following relation of the internal coordinates:

$$\overline{z}_n^{\pm}=z_{\sigma(n)}^{\mp},\;\forall n=1,\dots,N. \tag{8.40}$$

In the case of the base-pair level type II, see 8.10, the change of reading strand relations are the same as for type I, and a similar computation, to that leading to (8.39), can be used to obtain the relations (8.40). Finally, the change of reading strand transformation for tetrachain internal coordinates is given by:

$$\begin{aligned}&\text{when }(\mathcal{S}^{+},\mathcal{S}^{-})\mapsto(\overline{\mathcal{S}}^{+},\overline{\mathcal{S}}^{-})\\ &\qquad(x,m)\mapsto(z_N^{\pm},\mathbb{E}_2y_N,z_N^{\mp},\mathbb{E}_2x_N,z_{N-1}^{\pm},\dots,z_1^{\pm},\mathbb{E}_2y_1,z_1^{\mp})\end{aligned} \tag{8.41}$$

# 9 A sequence–dependent coarse–grain model of B-DNA with explicit treatment of the phosphate groups

The cgDNA+ model is a sequence–dependent coarse–grain model of double stranded B-form DNA with explicit treatment of bases and phosphate groups. The parameters of the model will be trained on the Palindromic data set. The main goal of the cgDNA+ model is to predict the sequence–dependent ground–state and flexibility of double–stranded B–form DNA. The cgDNA+ model is a natural extension of the cgDNA model which relies on the same assumptions presented in section 4.1 with the only addition assumption on the rigidity of the atoms forming the phosphate group. In this chapter we present the entire process that goes from extensive molecular dynamics simulations of palindromic sequences to the estimation of the parameter set of the cgDNA+ model. Given a parameter set $\mathcal{P}$ and an arbitrary DNA sequence $\mathcal{S}$, cgDNA+ predicts a Gaussian equilibrium probability density function in the configuration space by reconstructing the ground–state $\mu \equiv \mu(\mathcal{P},\mathcal{S}) \in \mathbb{R}^M$ and the stiffness matrix $K \equiv K(\mathcal{P},\mathcal{S}) \in \mathbb{R}^{M\times M}$:

$$\rho(w;\mathcal{S},\mathcal{P}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}(w-\mu)\cdot K(w-\mu)\right\}, \tag{9.1}$$

where $M = 24N - 18$, with $N$ the length in base–pairs of the sequence $\mathcal{S}$. Any configuration $w \in \mathbb{R}^{24N-18}$ can be divided into $N-1$ inter–base–pair internal coordinates, $N$ intra–base–pair internal coordinates, and $2(N-1)$ base–to–phosphate internal coordinates. As a single internal is represented by a six dimensional vector the total number of components in the configuration $w$ are $24N-18$. The inter– and intra–base–pairs coordinates have been already introduced in chapter 2 while the base–to–phosphate internal coordinates have been introduced in chapter 8 and further discussed in the next section.

From the MD trajectories we extract time series for the two interacting strand configuration, denoted by $(\mathcal{S}^+,\mathcal{S}^-)$, where each strand is coarse–grained at the level of base and phosphate groups. The only atom group that is not explicitly treated in the model is the sugar ring. For this purpose, we generalize the definitions (2.51) and (2.52). Let $\mathbf{p}(\mathcal{S}) \in \mathbb{R}^{3L}$ be the vector containing all the Cartesian coordinates of the atoms of the DNA molecule with sequence $\mathcal{S}$ of length N. The main modelling decision underlying

the cgDNA+ model is to coarse–grain the bases and the phosphate group, therefore we can redefine (2.51) as follow

$$\mathbf{r}(\mathcal{S}) \to (\mathbf{p}_{\mathcal{S}^+}, \mathbf{p}_{\overline{\mathcal{S}}^+}) = (\mathbf{p}^{X_1}, \mathbf{p}^{b_1}, \ldots, \mathbf{p}^{b_{N-1}}, \mathbf{p}^{X_N}, \mathbf{p}^{\overline{X}_1}, \mathbf{p}^{\bar{b}_1}, \ldots, \mathbf{p}^{\bar{b}_{N-1}}, \mathbf{p}^{\overline{X}_N}) \in \mathbb{R}^{3\ell}, \quad (9.2)$$

where $\mathbf{p}^{X_i} = (p_1^{X_i}, \ldots, p_{n_i}^{X_i}) \in \mathbb{R}^{3n_i}$ is the atom group considered for the base $X_i$ of $\mathcal{S}$ and $\mathbf{p}^{b_i} = (p_1^{b_i}, \ldots, p_{m_i}^{b_i}) \in \mathbb{R}^{3m_i}$ is the atom group for the $i$–th phosphate group. Straight away we can introduce a simplification because all the phosphate groups are composed of the same group of atoms independent of the sequence. Thus we can set $\mathbf{p}^{b_i} = \mathbf{p}^{b}\ \forall i = 1, \ldots, N-1$. The atoms on the complementary bases are distinguished by the over line notation. We recall that both strands are read in the $5' \to 3'$ direction and thus the numbering of the atoms groups in $(\mathbf{p}_{\mathcal{S}^+}, \mathbf{p}_{\overline{\mathcal{S}}^+})$ follow the same rule. Now we can generalize (2.52) in the following way

$$\mathfrak{R}^{\cdot\cdot}(\mathbf{p}_{\mathcal{S}^+}, \mathbf{p}_{\overline{\mathcal{S}}^+}) = (\mathcal{S}^+, \overline{\mathcal{S}}^+) \in SE(3)^{4N},$$

where $\mathcal{S}^+$ is the coarse–grain representation of the reading strand of $\mathcal{S}$ while $\overline{\mathcal{S}}^+$ is the representation of the complementary strand. We recall that the fitting procedure is described in section 2.3 and the detail about ideal atoms and ideal frames for both base and phosphate groups can be founded in section 1.3 and in appendix A. For convenience we can actually use transformation introduced in section 8.1 that maps $(\mathcal{S}^+, \overline{\mathcal{S}}^+) \to (\mathcal{S}^+, \mathcal{S}^-)$ to redefine $\mathfrak{R}^{\cdot\cdot}$ as follows:

$$\mathfrak{R}^{\cdot\cdot}(\mathbf{p}_{\mathcal{S}^+}, \mathbf{p}_{\overline{\mathcal{S}}^+}) = (\mathcal{S}^+, \mathcal{S}^-) \in SE(3)^{4N}.$$

## 9.1 On the base–to–phosphate degrees of freedom

### 9.1.1 Two possible definitions of internal coordinates

In section (8.4) we introduced the general invertible relationship between internal coordinates and tetrachain configurations and we have explicitly written the two possible ways of defining the coordinates of the microstructures of tetrachain configurations. In figure 9.1 we show a schematic representation of the two possible definitions of the base–to–phosphate degree of freedom. We recall hereafter the general notation for internal coordinates of tetrachain configurations of double stranded DNA of $N$ base–pairs:

$$\omega = (m_1, x_1, m_2, x_2, \ldots, m_N) \in \mathbb{R}^{24N-18},$$

where $x_n \in R^6$ parametrise the inter–base–pair displacement, $m_n$ is the coordinates of the microstructure defined by

$$m_n = (z^{\mp}, y_n, z^{\pm}) \in R^{18}, \quad (9.3)$$

$y_n \in \mathbb{R}^6$ parametrise the intra–base–pair displacement, and $z_n^{\pm} \in \mathbb{R}^6$ parametrise the base–to-phosphate displacements. We recall that the two possible internal coordinates, denoted by I and II in section 8.1.1 are respectively related to the base–to–3′ phosphate coordinates and base–to–5′ phosphate. In order to choose which one of the two internal

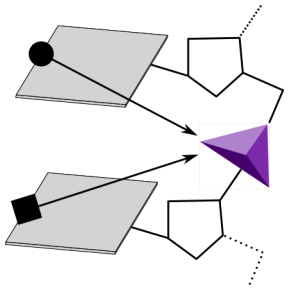


Figure 9.1 – Schematic representation of the two possible base–to–phosphate relative displacements. One way is base–to–3' phosphate (square), called *version 1*, and the second is the base–to–5' phosphate (circle), called *version 2*.

coordinate definitions to use for the cgDNA+ model we will compute oligomer–based statistics based on the palindromic data set. In particular we will extract time series of both internal coordinates from the MD trajectories. Thus, the first simple thing one can compare between the two version is the one dimensional histogram of all the components of the coordinates. For the seek of compactness we show just one example. We have chosen the sequence

$$\mathcal{S}_{11} = GCTACCTATGCTAGCATAGGTAGC, \tag{9.4}$$

of the Palindromic Library and in particular the phosphate located in the 16–th junction $A\text{p}T$. We recall that for *version 1* the internal coordinates parametrise the relative displacement between the base $A$ and the phosphate group in the $A\text{p}T$ junction while for *version 2* the relative displacement is between the base $T$ and the phosphate unit. In figures (9.2) we show the one dimensional histograms for each component of the internal coordinates mentioned above. In the first row we can observe the histograms for *version 1* while the ones for *version 2* are in the second row. In the first column one can find the rotational components while in the second columns the translational one. We can notice that the internal coordinates vary substantially between the two versions. In particular the translational components for *version 2* are way more non-Gaussian compared to *version 1*. It is clear here that a direct comparison between the two versions is not possible but, by looking at the histograms in (9.2), we can gain some insight about the difference in parametrising a relative displacement between two rigid bodies that are separated by a different number of covalent bonds, see figure (9.1) for details about the chemical structure. After a more careful analysis of all the phosphate coordinates in all the junctions of all the 16 sequences we can conclude that *version 2* of the internal coordinates tend to have one–dimensional histograms of base–to–phosphate degree of freedom that deviate more from Gaussianity, and in many

cases, for the translational part, a clear bi–modality is present. This outcome could

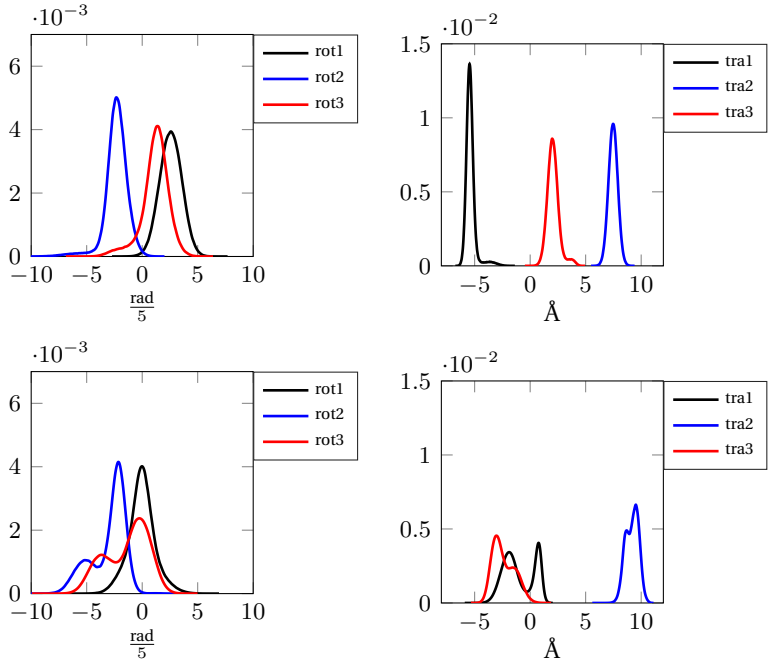



Figure 9.2 – For sequence $\mathcal{S}_{11}$ in the palindromic library we show the one dimensional histogram of both versions of the base–to–phosphate internal coordinate for the phosphate in the 16th junction: $ApT$. The first row is *version 1* and the second *version 2*. In the first column the rotational components, in the second the translational ones.

have been guest beforehand as the extra covalent bond that is parametrised by the version 2 is related to the BI–BII state already introduced in section 1.1. By recalling, the transitions between the states BI and BII can be related to the formation of a sequence– and context–dependent hydrogen bond between a phosphate group and its 5' base. In the context of our coarse–grained model we do not have a direct relation between the percentage of occupancy of BI and BII and internal coordinates, but it is not really surprising that the version 2 of coordinates leads to bi–modal behaviour of, in particular, the translational components.

Moreover, in figure (9.3) we show also the observed stiffness matrix for sequence $\mathcal{S}_{11}$ and, in particular, its sparsity pattern. One can notice that the major difference between the two stiffness matrices is in the magnitude of the entries. In fact for version 2 the entries seems to have a bigger magnitude compared to the same entries in the version 1 stiffness. We can also remark that in both stiffness there are 6 times 6 dimensional blocks with almost zeros entries. We will come back to this remark later on in this chapter, but at the moment the important thing is that both choices of internal coordinates bring to stiffness with *holes* in the same place. On top of the comment about the differences between the two definition of internal coordinates we would like to focus the attention of the reader to the particular sparsity pattern of

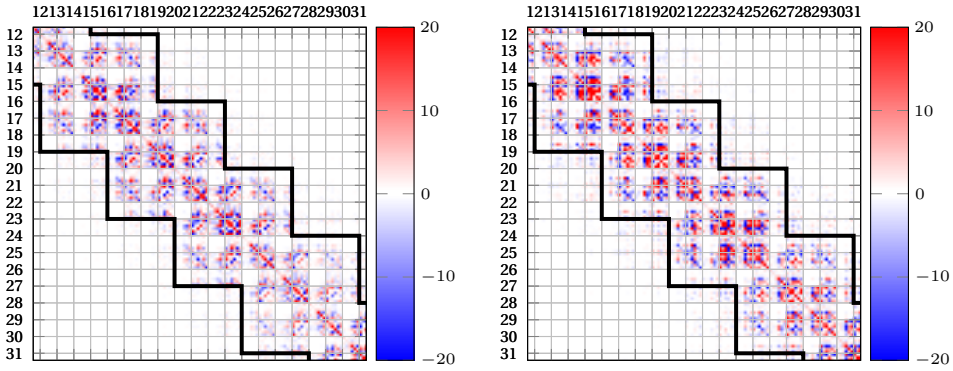



Figure 9.3 – Example of the sparsity pattern of observed stiffness matrices obtained from times series of both internal coordinates: left *version 1*, right *version 2*. The sequence is $\mathcal{S}_{11}$ from the palindromic training library.

both stiffness matrices. In particular, compared to the one showed in figure (3.2) for the cgDNA internal coordinates, the locality assumption is even more clear when the phosphate degrees of freedom are explicitly treated. A more detailed description of the sparsity pattern will be presented later on in this chapter.
Now that we have observed the qualitative differences between the two definitions of internal coordinates in reproducing some observable of the system we can introduce the following quantitative procedure. We will first compute the parameters, mean and banded stiffness, of the banded Gaussian for all sequences in the palindromic data set and then we will use the Kullback–Leibler divergence, with the maximum likelihood order for the input arguments, for computing the *approximation error* between banded Gaussian and observed Gaussian for both definitions of internal coordinates. First, the mean of the banded stiffness is the mean of the observed Gaussian, so that in the KLd the Mahalanobis part (5.4) will be zero. Thus, we can focus our attention only on the stiffness part and in particular we can use the following sequence averaged per degrees of freedom definition of the Kullback–Leibler divergence:

$$\overline{D^\dagger}(\theta_{\text{obs}}, \theta_{\text{band}}) = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{n_{\text{dofs}}} D^\dagger(\rho(x, \theta_{(\text{obs},i)}), \rho(x, \theta_{(\text{band},i)})). \tag{9.5}$$

For helping the reader we have labelled by *band* the parameter of the banded Gaussian and by *obs* the parameter of the observed, raw, Gaussian. The algebraic expression of $D^\dagger$ for two Gaussian can be found in (5.4).
Before going any further we recall that the marginal of a Gaussian distribution is another Gaussian distribution. For example, let $\rho(\omega)$ be a multivariate Gaussian distribution with mean $\mu = (x, y, z) \in \mathbb{R}^{n+m+p}$ with $x, y, z$ vectors of dimension $n, m, p$, and covariance matrix $C \in \mathbb{R}^{(n+m+p)\times(n+m+p)}$. The marginal of $\rho(\omega)$ over the components $y$ is the Gaussian defined by the mean $\mu_y = (x, z) \in \mathbb{R}^{n+p}$ and the covariance $C_y \in \mathbb{R}^{(n+p)\times(n+p)}$ defined by extracting the $x$–$x$, $x$–$z$ and $z$–$z$ blocks of $C$.

The methodology for quantifying the difference between the two versions involved also two different level of marginalisations of the Gaussian distributions $\rho(x, \theta_{(\text{obs},i)})$ and $\rho(x, \theta_{(\text{band},i)})$. The first marginalisation is over the base–to–phosphate degrees of freedom and will be denoted using the super script $m$. The second marginalisation is over the base–to–phosphate degree of freedom and the intra–base–pair coordinates and will be denoted using the super script $m^2$. The reason behind the decision of comparing also the marginals of the original distribution is to test how much information, about the inter– and intra–base-pair degree of freedom, is carried by the banded stiffness in both cases. In table 9.1 we show the value of (9.5) for the three proposed comparisons. We can observe that *version 2* of coordinates leads to a smaller value

| Version | $\overline{D^\dagger}(\theta_{\text{obs}}, \theta_{\text{band}})$ | $\overline{D^\dagger}(\theta^m_{\text{obs}}, \theta^m_{\text{band}})$ | $\overline{D^\dagger}(\theta^{m^2}_{\text{obs}}, \theta^{m^2}_{\text{band}})$ |
|---|---|---|---|
| 1 | $7.9 \cdot 10^{-3}$ | $2.9 \cdot 10^{-3}$ | $2.7 \cdot 10^{-3}$ |
| 2 | $6.4 \cdot 10^{-3}$ | $2.0 \cdot 10^{-3}$ | $2.0 \cdot 10^{-3}$ |

Table 9.1 – Values of $\overline{D^\dagger}$, defined in (9.5) for distributions with different degrees of freedom. The values in the first column quantify the error in the banded approximation of the observed Gaussian for both cgDNA+ internal coordinates. The Gaussian denoted by the super script $m$, second column, are the marginalisation over the phosphate components and the Gaussian denoted by the super script $m^2$, third column, are the further marginalisation over the intra–base–pair coordinates.

of (9.5) for all of the three proposed computations. As the first value in table (9.1) is smaller for *version 2* we can conclude that the observed stiffness matrix for that particular choice of coordinates is in fact closer to the banded assumption. By marginalising the first and the second time we actually quantify the information that the banded approximation is carrying. Again from table (9.1) we can conclude that version 2 is actually a better choice in this sense.
After the considerations made in this section we chose *version 2* as the definition of the internal coordinates for the cgDNA+ model. The motivation underlying the these decision are summarized as:

1. *Version 2* lead to one dimensional marginals for the phosphate degrees of freedom that are noticeably more non–Gaussian compared to *version 1*, but the latter phenomena can be explained by the BI/BII transition. For future development of the cgDNA+ model the *version 2* seems to be the natural choice of internal coordinates for the study of these transitions.

2. It is rational to chose the internal coordinates that minimize the banded approximation error as in any case for the purpose of this work we will deal only with the Gaussian model and the banded approximation represent the data that is used for the estimation of the cgDNA+ parameter set. Thus, the closer the banded distributions are to the observed data, the more accurate the predictions of the cgDNA+ model should be.

Finally the cgDNA+ internal coordinates we will consider read

$$\omega = (m_1, x_1, m_2, x_2, \ldots, m_N) \in \mathbb{R}^{24N-18}. \tag{9.6}$$

for a $N$ base–pair long arbitrary sequence, where the microstructure is defined by $m_n = (z_n^+, y_n, z_n^-)$. The phosphate degrees of freedom are defined as follow

$$z_n^\pm = \mathcal{C}\left(\mathcal{B}_n^\pm\right) = \mathcal{C}\left([\mathbf{g}_n^\pm]^{-1}\mathbf{p}_n^\pm\right) = (\eta_n^{\pm p}, \mathrm{w}_n^{\pm p}), \tag{9.7}$$

where, $\mathcal{C}$ is defined by

$$\mathcal{C}(\mathcal{B}_n^+) = \left(cay([R_n^+]^T[R_n^p]^+), [R_n^+]^T([r_n^p]^+ - r_n^+)\right) \in \mathbb{R}^6. \tag{9.8}$$

with $\mathbf{g}_n^+ = (R_n^+, r_n^+) \in SE(3)$, $\mathbf{p}_n^+ = ([R^p]_n^+, [r_n^p]^+) \in SE(3)$ and

$$\mathcal{B}_n^+ = ([R_n^+]^T[R_n^p]^+, [R_n^+]^T([r_n^p]^+ - r_n^+)).$$

The Crick–Watson symmetry for the internal coordinates (9.6) has already been introduced in section (8.4.1), here we just recall that the phosphate coordinates of the ground–state of a palindromic sequence satisfy

$$z_n^+ = z_{\sigma(n)}^-, \tag{9.9}$$

where $\sigma(n) = N - n + 1$, $N$ being the total number of base–pair of the sequence.
From the palindromic data set we can extract oligomer–based mean of the internal coordinates (9.6) and compute the sequence averaged mean of the phosphate degrees of freedom. In the following table we report the values:

| Coord. type | Seq. Avg. MD |
|---|---|
| $\eta_1^p$ | 0.9534 $\frac{\mathrm{rad}}{5}$ |
| $\eta_2^p$ | -3.3880 $\frac{\mathrm{rad}}{5}$ |
| $\eta_3^p$ | -1.7000 $\frac{\mathrm{rad}}{5}$ |
| $\mathrm{w}_1^p$ | -0.5302 Å |
| $\mathrm{w}_2^p$ | 9.5188 Å |
| $\mathrm{w}_3^p$ | -1.7211 Å |

Table 9.2 – Sequence–averaged values of the base–to-phosphate components computed from the palindromic 3$\mu$s long data set. We drop the $\pm$ notation because we also averaged Crick and Watson degrees of freedom as all the sequences in the training library are palindromic.

### 9.1.2 On the cgDNA+ sparsity pattern

As anticipated in the previous section, the cgDNA+ degrees of freedom lead to some interesting properties that we will describe in this section. The first property is actually inherited from the cgDNA degrees of freedom namely the topological tree structure of the internal coordinates as in scheme (9.4). It basically shows the connectivity between rigid–bodies defined by the internal coordinates and it defines also the composition of a base–pair level. We recall here that Crick and Watson phosphate groups that are in the same base–pair level are not in the same junction. The latter comment seems to introduce an asymmetry, but in fact the choice to integrate the phosphate group through a relative displacement from a base actually preserves the natural orientation of both strands. The most interesting property of the internal coordinates is the sparsity

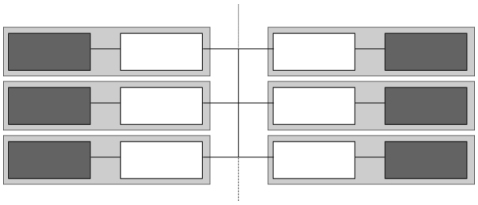


Figure 9.4 – Tree structure of the cgDNA+ internal coordinates. The white blocks represent the bases while the dark grey represent the phosphate groups. The light gray blocks containing a base and a phosphate group represent a base–phosphate unit.

pattern of the observed stiffness matrix computed as the inverse of the covariance matrix. In figure (9.5) we show a dinucleotide in the matrix shown on the right–hand side of figure 9.3. By only considering the entries inside the stencil defined by the black lines, we can start by presenting the overlap highlighted on the left–hand side part of figure 9.5 by the green box: its dimension is 18 times 18 and the degrees of freedom involved are Crick base–to–phosphate, intra–base–pair, and Watson base–to–phosphate forming the a base–pair level. It is clearly a natural extension of the cgDNA model in the sense that the overlap is actually composed by the coupling related to the degrees of freedom defining the base–pair level, or equivalently, the micro structure:

$$m_n = (z_n^+, y_n, z_n^-).$$

The cyan highlights the inter–base–pair block which lays in the middle of the 42 times 42 block. Let now fix the attention on the yellow block. This block is related to the base-to–Crick phosphate group degrees of freedom. On the same line of the matrix we show in red all the phosphate blocks in the stencil that are coupled with the yellow base–to–Crick phosphate group mentioned above. Being on the same line of the matrix means that the entries quantify the coupling between the related degrees of freedom. From left to right each odd red blocks is the coupling between Crick phosphate in the downstream and up stream base–pair level, while the even ones are the Watson phosphate also in the downstream and up stream base–pair level. It is interesting to

notice that the Crick–Watson physical coupling is really close to be zero, meaning that Crick phosphate group does not see any of the Watson phosphate from the elastic interaction point of view. Finally, the big orange box contains the interaction between three base–pair level that can be interpreted as follow: each base–phosphate unit on the Watson or Crick strand, interacts with its five nearest neighbour units. For example if one consider again the schematic representation of the internal coordinates in figure 9.4 the latter comment can be visually explained by fixing, for example, the middle right light gray box. Its five nearest neighbour are just the other light gray boxes in the scheme. We recall that a light gray box is in fact a unit composed by a base and a phosphate group attached to it. The main properties of the interactions between units is that there is very weak physical coupling between phosphate groups on different strands.

The last point we want to present is shown in figure 9.6 where we compare the banded

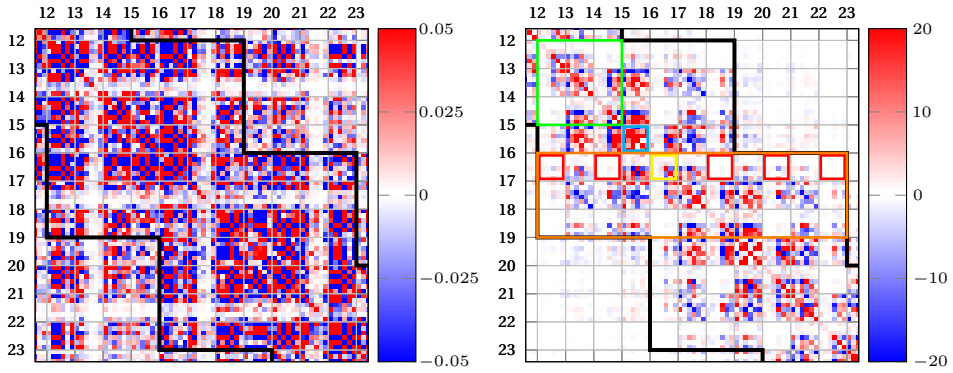



Figure 9.5 – Left, details of the covariance matrix estimated from time series of internal coordinates (9.6) for sequence $\mathcal{S}_{11}$, see sequence list (7.1). Right, we show the same detail, but for its inverse, the stiffness matrix. Some parts of the stiffness matrix are highlighted as explained in the text.

stiffness estimation and the observed one. One can remark that the *holes* corresponding to the Crick phosphate– Watson phosphate degrees of freedom interactions are still present in the banded stiffness. We recall that in algorithm 1 used for the computation of the best fit banded stiffness is not possible to impose constraints on the values of the entries inside the desired stencil. Hence, the fact that the banded stiffness preserves the physical properties observed in the data could indicate that the origin of this weak coupling can be explained by the entries of the covariance matrix just inside the stencil.

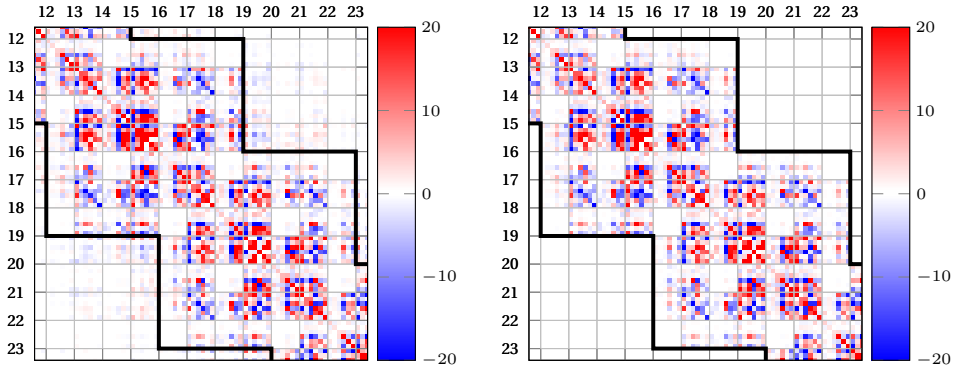



Figure 9.6 – Left: observed stiffness matrices Right: banded stiffness best estimate. The sequence is $\mathcal{S}_{11}$ from the palindromic training library.

## 9.2 Convergence of the phosphate degrees of freedom

In this section we will generalize the definition of convergence error, in the sense of palindromic symmetry, of the first and centred second moment estimator given by

$$\text{ERR}(\mu) = \|\mu - E_N^+ \mu\|, \tag{9.10}$$

$$\text{ERR}(C) = \|C - E_N^+ C E_N^+\|, \tag{9.11}$$

for MD time series of internal coordinates (9.6). The matrix $E_N^+$ represent the linear map of the change of reading strand transformation and its definition has not yet been explicitly given, it will be discussed in detail in section 9.3. Before proceeding with the convergence study of the Palindromic sequences in the context of cgDNA+ internal coordinates, we first consider sequence $\mathcal{S}_{11}$ and focus attention on the phosphate in the 16–th junction $ApT$ on the reading strand which in fact is the coordinates from base $T$ to the $5'$ phosphate. We recall that for this particular sequence we have simulated $10\mu$s which are a total of 10 million snapshots before hydrogen–bond filtering. For more detail about HB filtering see section 3.3.1. In figure (9.7) we show, in solid lines, the histograms over all the accepted snapshots for the base–to–phosphate internal coordinates in the $ApT$ junction. In dashed line we plot its palindromic complement which actually is the base–to–phosphate degrees of freedom on the Crick strand on the 8–th base–pair level, see equation (9.9). From the histograms we can observe that both rotational and translational parts of the coordinates are extremely well converged even if most of the coordinates present a clear bi-modal behaviour. We now compute the palindromic error for the mean (9.10 ) for sequences (1,5,11) for which we have up to $10\mu$s of trajectories. In the following table we report the results:

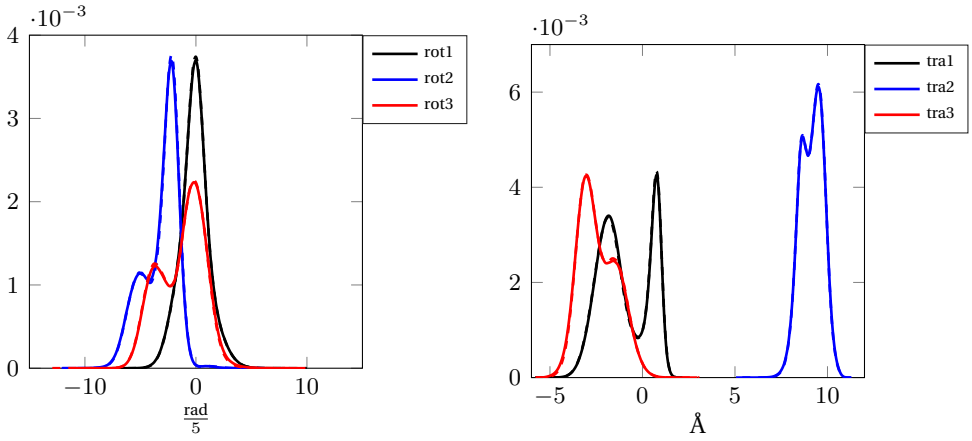



Figure 9.7 – One dimensional histograms of the base–to–phosphate coordinates in the 16–th junctions on Watson strand (solid line) compared to its palindromic symmetric degree of freedoms on the Crick strand (dashed line) for sequence $\mathcal{S}_{11}$ over $10\mu$s long simulation. The pairs of curves are virtually indistinguishable.

| # $\mathcal{S}$ | $1\mu$s | $2\mu$s | $3\mu$s | $4\mu$s | $5\mu$s |
|---|---|---|---|---|---|
| 1 | 2.3753 | 1.7361 | 1.3174 | 1.3920 | 0.9332 |
| 5 | 2.8186 | 1.8046 | 1.1890 | 1.1418 | 1.2626 |
| 11 | 2.7406 | 1.1944 | 1.1404 | 0.8462 | 0.8567 |
| | $6\mu$s | $7\mu$s | $8\mu$s | $9\mu$s | $10\mu$s |
| 1 | 0.7830 | 0.7486 | 0.6460 | 0.6675 | 0.5994 |
| 5 | 1.4387 | 1.3878 | 1.0316 | 0.9691 | 0.7266 |
| 11 | 0.7917 | 0.7589 | 0.7510 | 0.6982 | 0.6111 |

Table 9.3 – Palindromic error in the estimator of the mean as function of simulation length for sequence (1,5,11) of the Palindromic Library.

The interpretation of the latter values can be done using the same methodology proposed in section 7.2.1. But, it is important to take into account two things: 1) is clear that the palindromic error is bigger for cgDNA+ coordinates because in the error function we basically add more non–zero terms that correspond to the phosphate degrees of freedom. 2) From table (9.2) we noticed that the sequence–averaged values for the phosphates are bigger in magnitude then the intra– or inter–base–pair ones, thus if we consider the average error per degree of freedom then, we have to take into account the differences in magnitude. For example, for sequence $\mathcal{S}_{11}$ the error after $10\mu$s is $0.7266$ which implies an averaged per degree of freedom error of $0.0308$. For a better understanding of the convergence of the mean we extract just the phosphate components, and compute the palindromic error just for these degree of freedoms. In the following table we show the findings:

| # $\mathcal{S}$ | 1$\mu$s | 2$\mu$s | 3$\mu$s | 4$\mu$s | 5$\mu$s |
|---|---|---|---|---|---|
| 1 | 2.3040 | 1.6742 | 1.2657 | 1.3437 | 0.9012 |
| 5 | 2.7047 | 1.7683 | 1.1442 | 1.0929 | 1.2124 |
| 11 | 2.6459 | 1.1522 | 1.1056 | 0.8162 | 0.8289 |
| | 6$\mu$s | 7$\mu$s | 8$\mu$s | 9$\mu$s | 10$\mu$s |
| 1 | 0.7582 | 0.7231 | 0.6209 | 0.6376 | 0.5736 |
| 5 | 1.3784 | 1.3286 | 0.9853 | 0.9252 | 0.6924 |
| 11 | 0.7658 | 0.7322 | 0.7241 | 0.6728 | 0.5890 |

Table 9.4 – Palindromic error in the phopshate components of the mean estimator as function of simulation length

We can now better understand which degrees of freedom contribute the most to the error by simply computing the square of any entry in table (9.3) and then verifying that it is in fact the sum of the squares for the same entries in tables (7.2) and (9.4). For example in (9.3) we can read 1.3174 $(1.3174^2 = 1.7355)$ for $\mathcal{S}_1$ at 3$\mu$s, and from tables (7.2) and (9.4) we read that the contributions from the cgDNA degrees of freedom is $0.3655$ $(0.3655^2 = 0.1336)$ which the contribution from the base–to–phosphate coordinates is $1.2657$ $(1.2657^2 = 1.6019)$. Thus, we obtain that in the mean estimator of $\mathcal{S}_1$ at 3$\mu$s the major contribution to the palindromic error, for cgDNA+ internal coordinates, comes from the phosphate degrees of freedom. In figure 9.8, we show the comparison between Crick and Watson degrees of freedom for sequence $\mathcal{S}_1$ computed at 10$\mu$s. In solid we show the phosphate coordinates for the Crick strand in reverse order to match the Watson phosphate coordinates showed in dashed line. We can observe an astonishing match between the two curves.
In conclusion, the cgDNA+ mean estimator converges, in the sense of palindromic symmetry, with a reasonably high speed even if the error coming from the phosphate degree of freedom tend to be larger compared to the cgDNA part.
Next, we study the convergence of palindromic error in the estimator of the covariance

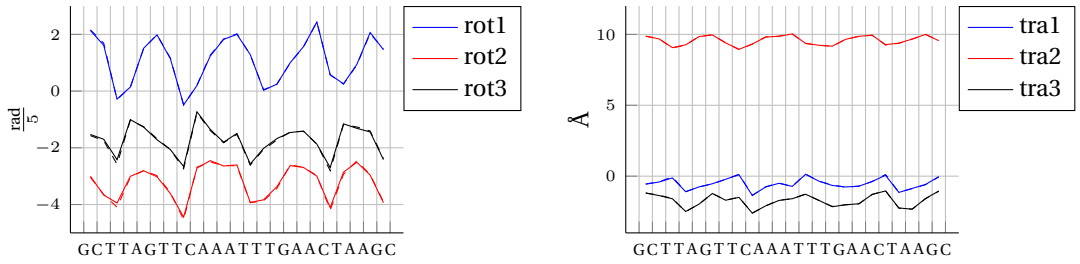



Figure 9.8 – Comparison of Crick and Watson phopshate degrees of freedom for $\mathcal{S}_1$ computed at 10$\mu$s. In solid we show the Crick phosphate coordinates in reverse order, and in dashed the Watson phosphate degrees of freedom.

matrix. In the following table we report the palindromic error (9.11):

| # $\mathcal{S}$ | 1$\mu$s | 2$\mu$s | 3$\mu$s | 4$\mu$s | 5$\mu$s |
|---|---|---|---|---|---|
| 1 | 12.7157 | 9.1412 | 6.7662 | 7.6078 | 5.2133 |
| 5 | 11.8900 | 8.3603 | 5.5511 | 4.7657 | 5.3002 |
| 11 | 14.1163 | 6.8554 | 5.9344 | 4.7001 | 4.7712 |
| | 6$\mu$s | 7$\mu$s | 8$\mu$s | 9$\mu$s | 10$\mu$s |
| 1 | 4.2848 | 3.7680 | 3.3660 | 3.2651 | 3.1135 |
| 5 | 5.2586 | 5.0496 | 4.2272 | 4.3219 | 3.4921 |
| 11 | 4.4921 | 3.9930 | 3.4361 | 3.2679 | 2.9514 |

Table 9.5 – Palindromic error in the estimator of the covariance as function of simulation length

We recall that we only consider the entries of the covariance that are inside the cgDNA+ stencil and, because of the symmetry of the covariance matrix, we consider only the diagonal entries plus the upper (or lower) triangular part. The first observation of the results reported in table (9.5) is that the error for the cgDNA+ internal coordinates is much bigger than the one obtained for cgDNA. Again, it is normal that the error increases because, in the covariance matrix, there are the additional blocks corresponding to phosphate–phosphate and phosphate–cgDNA correlations. We can again consider the sub–blocks corresponding to phosphate degrees of freedom, meaning that we can compute the marginal covariance over the cgDNA components. Once we have computed the marginal covariance we consider only the phosphate–phosphate blocks that are inside the stencil. We compute then the following errors:

| # $\mathcal{S}$ | 1$\mu$s | 2$\mu$s | 3$\mu$s | 4$\mu$s | 5$\mu$s |
|---|---|---|---|---|---|
| 1 | 12.2666 | 8.8037 | 6.5136 | 7.3374 | 5.0284 |
| 5 | 11.5300 | 8.1239 | 5.3805 | 4.5998 | 5.1328 |
| 11 | 13.5723 | 6.6229 | 5.7468 | 4.5436 | 4.6014 |
| | 6$\mu$s | 7$\mu$s | 8$\mu$s | 9$\mu$s | 10$\mu$s |
| 1 | 4.1397 | 3.6408 | 3.2511 | 3.1492 | 3.0045 |
| 5 | 5.0939 | 4.8932 | 4.0977 | 4.1819 | 3.3824 |
| 11 | 4.3381 | 3.8465 | 3.3017 | 3.1371 | 2.8363 |

Table 9.6 – Palindromic error (9.11) in the phopshate-phosphate covariance sub–blocks as function of simulation duration.

Here it is important to mention that in the process of marginalisation, the phosphate–cgDNA blocks are lost. We can retrieve the palindromic error for the phosphate–cgDNA blocks by subtracting from the square of the entries in table (9.5) the square of the corresponding values in tables (7.4) and (9.6). By doing this we observe that the major contribution to the palindromic error for the covariance matrix comes from the

phosphate–phosphate blocks, which leads to the conclusion that phosphate blocks tend to converge slower compared to the cgDNA–cgDNA and the phosphate–cgDNA ones. Finally in the tables (D.1) and (D.2) of the Appendix we report the palindromic error for the palindromic data set. We can conclude that for the phosphate coordinates $3\mu$s simulation are insufficient and it would be better to extend all the 16 simulations to 10 microsecond or more. Unfortunately at the moment a data set containing $10\mu$s long simulations of each sequence in the Palindromic library is not available. But, for the purpose of these work we will again consider the $3\mu$s Palindromic data set and work with it. In fact, we defined the Palindromic data set for the cgDNA+ internal coordinates by computing palindromic symmetric mean and covariance estimators of each palindromic sequence as well as the banded stiffness approximation of each symmetrised second centred moments.

### 9.2.1 Convergence of oligomer–based Gaussian

In the previous section we studied the convergence of first and second (centred) moments separately. In this section we will use the Kullback–Leibler divergence per degree of freedom between observed oligomer–based Gaussian and its palindromic symmetric Gaussian to assess the convergence of the MD simulation. In detail, for a palindromic sequence $\mathcal{S}$ of the palindromic training library we estimate the oligomer–based Gaussian $\rho(w;\mathcal{S})$ which parameters are the mean $\mu(\mathcal{S})$ estimated using the standard estimator and the banded stiffness matrix, $K(\mathcal{S})$, computed from $C(\mathcal{S})$, the standard estimator for the covariance. The palindromic symmetric Gaussian is denote by $\tilde{\rho}(w;\mathcal{S})$ and its parameter are $\tilde{\mu}(\mathcal{S}) = E_N^+\mu(\mathcal{S})$ and $\tilde{K}(\mathcal{S}) = E_N^+K(\mathcal{S})E_N^+$. The Kullback–Leibler divergence per degree of freedom is defined by

$$\widehat{D}\left(\rho(w;\mathcal{S}),\tilde{\rho}(w,Sw)\right) = \frac{1}{n_{dofs}}D_{KL}\left(\rho(w;\mathcal{S}),\tilde{\rho}(w,Sw)\right) \tag{9.12}$$

$$= \frac{1}{n_{dofs}}D^{\dagger}\left(\rho(w;\mathcal{S}),\tilde{\rho}(w,Sw)\right) + \frac{1}{n_{dofs}}\mathcal{M}\left(\rho(w;\mathcal{S}),\tilde{\rho}(w,Sw)\right) \tag{9.13}$$

$$= \widehat{D^{\dagger}}\left(\rho(w;\mathcal{S}),\tilde{\rho}(w,Sw)\right) + \widehat{\mathcal{M}}\left(\rho(w;\mathcal{S}),\tilde{\rho}(w,Sw)\right), \tag{9.14}$$

where $D_{KL}(\cdot,\cdot)$, $D^{\dagger}(\cdot,\cdot)$, and $\mathcal{M}(\cdot,\cdot)$ where defined in 5.3 and 5.4. In table 9.7 we report the value of the Kullback–Leibler divergence per degree of freedom (as function of simulation duration) for each sequence in the palindromic sequence library. We observed that the stiffness part $\widehat{D^{\dagger}}$ contributes the most to the value of $\widehat{D}$ for all palindromic sequences. Consequently, also in the sequence average values (last row of table 9.7) we notice that. Finally we can conclude that the stiffness part converge slower, with respect to (9.14), than the shape part. This is consistent with the conclusions of the previous section 9.2. But, using the Kullback–Leibler per degree of freedom as error function we can compare the result of the sequence average values at $3\mu$s with the

| | $1\mu$s | | | $2\mu$s | | | $3\mu$s | | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{S}$ | $\widehat{D}$ | $\widehat{D^\dagger}$ | $\widehat{\mathcal{M}}$ | $\widehat{D}$ | $\widehat{D^\dagger}$ | $\widehat{\mathcal{M}}$ | $\widehat{D}$ | $\widehat{D^\dagger}$ | $\widehat{\mathcal{M}}$ |
| 1 | 0.0689 | 0.0643 | 0.0046 | 0.0149 | 0.0144 | 0.0005 | 0.0098 | 0.0094 | 0.0004 |
| 2 | 0.0616 | 0.0553 | 0.0063 | 0.0297 | 0.0276 | 0.0021 | 0.0205 | 0.0190 | 0.0014 |
| 3 | 0.0357 | 0.0328 | 0.0030 | 0.0108 | 0.0103 | 0.0005 | 0.0082 | 0.0079 | 0.0003 |
| 4 | 0.0465 | 0.0418 | 0.0047 | 0.0193 | 0.0180 | 0.0013 | 0.0169 | 0.0154 | 0.0015 |
| 5 | 0.0764 | 0.0652 | 0.0112 | 0.0227 | 0.0206 | 0.0021 | 0.0117 | 0.0106 | 0.0010 |
| 6 | 0.0574 | 0.0529 | 0.0044 | 0.0223 | 0.0211 | 0.0012 | 0.0136 | 0.0128 | 0.0007 |
| 7 | 0.0587 | 0.0528 | 0.0058 | 0.0237 | 0.0218 | 0.0019 | 0.0141 | 0.0129 | 0.0012 |
| 8 | 0.0264 | 0.0253 | 0.0011 | 0.0135 | 0.0129 | 0.0006 | 0.0066 | 0.0063 | 0.0003 |
| 9 | 0.0341 | 0.0326 | 0.0015 | 0.0161 | 0.0153 | 0.0008 | 0.0123 | 0.0117 | 0.0006 |
| 10 | 0.0370 | 0.0348 | 0.0022 | 0.0147 | 0.0140 | 0.0007 | 0.0156 | 0.0147 | 0.0009 |
| 11 | 0.1232 | 0.1021 | 0.0210 | 0.0450 | 0.0409 | 0.0042 | 0.0247 | 0.0223 | 0.0025 |
| 12 | 0.0516 | 0.0474 | 0.0041 | 0.0378 | 0.0342 | 0.0036 | 0.0223 | 0.0206 | 0.0018 |
| 13 | 0.0244 | 0.0234 | 0.0010 | 0.0206 | 0.0196 | 0.0010 | 0.0135 | 0.0129 | 0.0006 |
| 14 | 0.0200 | 0.0182 | 0.0019 | 0.0106 | 0.0100 | 0.0006 | 0.0127 | 0.0123 | 0.0004 |
| 15 | 0.0733 | 0.0658 | 0.0075 | 0.0168 | 0.0158 | 0.0010 | 0.0132 | 0.0124 | 0.0008 |
| 16 | 0.0540 | 0.0505 | 0.0035 | 0.0273 | 0.0251 | 0.0023 | 0.0134 | 0.0123 | 0.0011 |
| Avg | 0.0531 | 0.0478 | 0.0052 | 0.0216 | 0.0201 | 0.0015 | 0.0143 | 0.0133 | 0.0010 |

Table 9.7 – Palindromic convergence error computed using the Kullback–Leibler divergence per degree of freedom between observed banded Gaussian and its palindromic symmetric Gaussian as function of simulation duration. The Avg values are obtained by averging over all the 16 palidromic sequences. We can observe that at $3\mu$s the average values of the KLdalmost the half of the values of the KLd per degree of freedom obtained in chapters 6 and 7. The main contribution to the error comes from the stiffness part of the KLd per degree of freedom.

values obtained in chapters 6 and 7. We actually remark that the palindromic error is almost half of the total error between model and data. This suggest again that the simulations of the palindromic library should be extended. In table 9.8 we show the palindromic error at $10\mu$s for the sequences $(1, 5, 11)$ of the palindromic training library computed using (9.14). We observe that the palindromic errors for the three $10\mu$s long MD simulations decrease substantially. In conclusion the Kullback–Leibler divergence per degree of freedom (9.14) allows to get a more insight understanding of the convergence of the oligomer–based Gaussian and allows to compare the palindromic error with the modelling error.

## 9.3 The cgDNA+ parameter set

The parameter set format for the cgDNA+ model is a natural extension of the cgDNA one, but with the particular property that the end sigma vector and stiffness matrices will be of different dimension than the interior ones. In detail the cgDNA+ parameter

| | $10\mu$s | | |
|---|---|---|---|
| $\mathcal{S}$ | $\widehat{D}$ | $\widehat{D^\dagger}$ | $\widehat{\mathcal{M}}$ |
| 1 | 0.0034 | 0.0033 | 0.0001 |
| 5 | 0.0039 | 0.0037 | 0.0002 |
| 11 | 0.0026 | 0.0025 | 0.0001 |

Table 9.8 – Palindromic convergence computed using the Kullback–Leibler per degree of freedom (9.14) computed from $10\mu$s long MD simulations.

set is defined by

$$\mathcal{P} = \left\{\sigma^{5'\alpha\beta}, \sigma^{\alpha\beta}, K^{5'\alpha\beta}, K^{\alpha\beta}\right\}_{\alpha\beta\in D', 5'\alpha\beta\in D} \subset \mathbb{P}_{\text{tot}}, \tag{9.15}$$

where $\mathbb{P}_{\text{tot}} = [\mathbb{R}^{36}]^{16} \times [\mathbb{R}^{42}]^{10} \times [\mathbb{S}^{36}]^{16} \times [\mathbb{S}^{42}]^{10}$, and $\mathbb{S}^N$ is the set of $N \times N$ symmetric matrices. The end sigma vectors are of dimension 36 while the interior ones are of dimension 42. Equivalently, the stiffness end blocks are of dimension $36 \times 36$ while the interior ones are of dimension $42 \times 42$. The difference in dimension between interior and end blocks is due to the fact that in the MD simulations the first phosphate group on both strands is absent. Consequently, the first and last base–pair levels are composed of only an intra–base–pair degree of freedom and a single base–to–phosphate set of internal coordinates.
In the following we will mainly focus on the computation of the cgDNA+ parameter set (9.15). It involves, in particular, the problem of computing an initial guess to initialise for the optimization problem. But first, we need to introduce the cgDNA+ reconstruction rules for the mean and the stiffness matrix and the parameter extraction problem.
Let $\mathcal{P}$ be a cgDNA+ parameter set of the form (9.15) and let $\mathcal{S}$ be a $N$ base–pair long DNA sequence. We can define the reconstruction rule for the stiffness matrix $K(\mathcal{P},\mathcal{S})$ and the weighted shape vector $\sigma(\mathcal{P},\mathcal{S})$ in the following way:

$$K(\mathcal{P},\mathcal{S}) = P_d^T K_d P_d, \tag{9.16}$$
$$\sigma(\mathcal{P},\mathcal{S}) = P_d^T \sigma_d, \tag{9.17}$$
$$\mu(\mathcal{P},\mathcal{S}) = K(\mathcal{P},\mathcal{S})^{-1}\sigma(\mathcal{P},\mathcal{S}), \tag{9.18}$$

where

$$K_d = \text{diag}(K^{5'X_1X_2}, \ldots, K^{X_iX_{i+1}}, \ldots, K^{X_{n-1}X_n3'}),$$
$$\sigma_d = (\sigma^{5'X_1X_2}, \ldots, \sigma^{X_iX_{i+1}}, \ldots, \sigma^{X_{n-1}X_n3'}),$$

and the matrix $P_d \in \mathbb{R}^{42N-12\times 24N-18}$ reads

$$P_d = \begin{bmatrix} I_{18} & & & & & & \cdots & \\ & I_{18} & & & & & & \\ & I_{18} & & & & & & \\ & & I_6 & & & & & \\ & & & I_{18} & & & & \\ & & & I_{18} & & & & \\ & & & & I_6 & & & \\ \vdots & & & & & & \ddots & \vdots \\ & & & & & & & I_{18} \end{bmatrix}, \tag{9.19}$$

where we use the notation $I_n$ for the $n$–dimensional identity matrix. We recall that the $3'$ end blocks for both stiffness and weighted shape can be computed using Crick–Watson symmetry. More precisely,

$$K^{\overline{\alpha\beta}3'} = E^{5'} K^{5'\alpha\beta} E^{5'}, \tag{9.20}$$

where $E^{5'}$ is defined by

$$E^{5'} = \begin{bmatrix} & & & & & E \\ & & & & I_6 & \\ & & & E & & \\ & & I_6 & & & \\ & E & & & & \\ I_6 & & & & & \end{bmatrix}, \tag{9.21}$$

with $E = \mathrm{diag}(-1,1,1,-1,1,1) \in \mathbb{R}^6$, and $E^{5'} \in \mathbb{R}^{36\times 36}$ satisfy $[E^{5'}]^{-1} = [E^{5'}]^T = E^{3'} \in \mathbb{R}^{36\times 36}$ with

$$E^{3'} = \begin{bmatrix} & & & & & I_6 \\ & & & & E & \\ & & & I_6 & & \\ & & E & & & \\ & I_6 & & & & \\ E & & & & & \end{bmatrix}. \tag{9.22}$$

For the interior blocks the Crick–Watson symmetry rule is given by

$$K^{\overline{\alpha\beta}} = E^{\mathrm{int}} K^{\alpha\beta} E^{\mathrm{int}}, \tag{9.23}$$

with

$$E^{\text{int}} = \begin{bmatrix} & & & & & & I_6 \\ & & & & & E & \\ & & & & I_6 & & \\ & & & E & & & \\ & & I_6 & & & & \\ & E & & & & & \\ I_6 & & & & & & \end{bmatrix}, \tag{9.24}$$

which satisfy $E^{\text{int}} = [E^{\text{int}}]^T = [E^{\text{int}}]^{-1}$. Finally we have that the complementary sequence $\overline{\mathcal{S}}$ of $\mathcal{S}$ must satisfy

$$\begin{aligned} \mu(\mathcal{P},\mathcal{S}) &= E_N^+ w(\mathcal{P},\overline{\mathcal{S}}), \\ K(\mathcal{P},\mathcal{S}) &= E_N^+ K(\mathcal{P},\overline{\mathcal{S}}) E_N^+, \end{aligned}$$

where

$$E_N^+ = \begin{bmatrix} & & & & E^{5'} \\ & & & E^{\text{int}} & \\ & & E^{\text{int}} & & \\ & \iddots & & & \\ E^{3'} & & & & \end{bmatrix}. \tag{9.25}$$

The best–fit cgDNA+ parameter set $\mathcal{P}$ is defined as

$$\mathcal{P} = \underset{P\in\mathbb{P}}{\text{argmin}}\, \mathbb{F}(P; Lb), \tag{9.26}$$

with

$$\mathbb{F}(P; Lb) = \sum_{i=1}^{M} D_{KL}(\rho(x,\theta_i(P)), \rho(x,\theta_i)), \tag{9.27}$$

where $Lb$ is the training library containing $M$ sequences, $\rho(x,\theta_i)$ is the observed banded Gaussian for sequence $\mathcal{S}_i$ and $\rho(x,\theta_i(P))$ is the predicted Gaussian for sequence $\mathcal{S}_i$ and parameter set $P$ where $\theta_i(\mathcal{P}) = \text{param}(\sigma(\mathcal{P},\mathcal{S}_i), K(\mathcal{P},\mathcal{S}_i))$, see for instance (2.9), where the weighted shape $\sigma$ and the stiffness matrix $K$ have been reconstructed using the rules (9.16–9.17). We define the set of admissible cgDNA+ parameter sets by $\mathbb{P} = \mathbb{P}_{\text{self}} \cap \mathbb{P}_{\text{train}} \subset \mathbb{P}_{\text{tot}}$. We recall that $\mathbb{P}_{\text{self}}$ is the subset of parameter sets which interior sigma vectors and stiffness matrices for palindromic dimers satisfy the Crick–Watson symmetry (9.23). The subset $\mathbb{P}_{\text{train}}$, instead, contains only parameter sets $P$ that, for all $\mathcal{S} \in Lb$, reconstruct a positive definite stiffness matrix $K(P,\mathcal{S})$.

For the numerical resolution of problem (9.27) we need an initial parameter set $\mathcal{P}_{ini} \in \mathbb{P}$. For this purpose we introduce the Fisher information matrix and its relationship with Kullback–Leibler divergence.

### 9.3.1 Fisher information matrix

An interesting and useful feature of the Kullback–Leibler divergence is its relationships with the Fisher information [28]. We recall first that, under some regularity conditioned, the Fisher information is the second centred moment of $F(x,\theta) = \log p(x;\theta)$ conditional to the parameter $\theta \in \mathbb{R}$, namely

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}F(x,\theta)\middle|\theta\right], \tag{9.28}$$

where $p(x;\theta)$ is a probability density function conditioned on $\theta$. The definition (9.28) can be extended, again under some regularity conditions, to parameters $\theta \in \mathbb{R}^N$ with $N > 1$ which leads to the Fisher information matrix defined, entry–by-entry, by

$$[\mathcal{I}(\theta)]_{ij} = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}F(x,\theta)\middle|\theta\right]. \tag{9.29}$$

Now consider $p(x;\hat{\theta})$, a probability density function parametrized by $\hat{\theta} \in \mathbb{R}^N$, and let $\theta = \hat{\theta} + \delta\theta \in \mathbb{R}^N$ with $\delta\theta \ll 1$. The KLd between $p(x;\hat{\theta})$ and $p(x;\theta)$ is

$$D_{KL}(p(x;\hat{\theta}), p(x;\theta)) = \int_\Omega p(x;\hat{\theta}) \log\frac{p(x;\hat{\theta})}{p(x;\theta)}dx. \tag{9.30}$$

A direct computation shows that a first relation between the Fisher information (9.28) of $p(x,\hat{\theta})$ and the KLd (9.30) is

$$\mathcal{I}(\theta) = -\int_\Omega p(x;\hat{\theta})\frac{\partial^2}{\partial\hat{\theta}^2}\log p(x;\hat{\theta})dx = \frac{\partial^2}{\partial\theta^2}D_{KL}(p(x;\hat{\theta}), p(x;\theta))\bigg|_{\theta=\hat{\theta}}. \tag{9.31}$$

We now expand (9.30) at $\theta = \hat{\theta}$:

$$\begin{aligned} D_{KL}(p(x,\hat{\theta}), p(x,\theta)) &= D_{KL}(p(x,\hat{\theta}), p(x,\hat{\theta})) + \frac{\partial}{\partial\theta}D_{KL}(p(x,\hat{\theta}), p(x,\theta))\bigg|_{\theta=\hat{\theta}} \cdot \delta\theta \\ &+ \frac{1}{2}\delta\theta \cdot \frac{\partial^2}{\partial\theta^2}D_{KL}(p(x,\hat{\theta}), p(x,\theta))\bigg|_{\theta=\hat{\theta}}\delta\theta + \mathcal{O}(|\delta\theta|^3). \end{aligned} \tag{9.32}$$

As the KLd vanishes at $\theta = \hat{\theta}$ and this point is also its global minima, the first and second term in (9.32) vanish. By using equation (9.31) we can write the second relation

between KLd and the Fisher information matrix:

$$D_{KL}(p(x,\hat{\theta}),p(x,\hat{\theta}+\delta\theta)) = \frac{1}{2}\delta\theta\cdot\mathcal{I}(\hat{\theta})\delta\theta + \mathcal{O}(|\delta\theta^3|). \tag{9.33}$$

Finally we obtain that, for a parametric probability density function $p(\cdot,\hat{\theta})$, its Fisher information matrix defines a quadratic approximation of the Kullback-Leibler divergence between $p(\cdot,\hat{\theta})$ and any other probability density function parametrised a neighbour of $\hat{\theta}$. As a last remark we point out that equation (9.33) holds for both orderings of the argument in the KLd, meaning that we have also

$$D_{KL}(p(x,\hat{\theta}+\delta\hat{\theta}),p(x,\hat{\theta})) = \frac{1}{2}\delta\theta\cdot\mathcal{I}(\hat{\theta})\delta\theta + \mathcal{O}(|\delta\theta^3|). \tag{9.34}$$

The (standard) theory mentioned above can be applied whenever two parametric probability density functions are close in parametric space. In particular, in the context of DNA modelling, one can use (9.31) or (9.34) to compute the Kullback–Leibler divergence between two banded Gaussians parametrised respectively by $\theta_m = (\sigma_m, K_m) \in \mathbb{R}^N$ and $\theta_d = (\sigma_d, K_d) \in \mathbb{R}^N$. But in practice the computation of the Fisher Information matrix as the second derivative of the Kullback–Leilber divergence (9.31) is not trivial and deserves to be better explained. For sake of completeness let us rewrite the Kullback-Leibler divergence between two banded Gaussians denoted by $\rho_m = \rho(x;\theta_m)$ and $\rho_d = \rho(x;\theta_d)$:

$$\begin{aligned} D_{KL}(\theta_m,\theta_d) =& \frac{1}{2}\left(\text{trace}(K_m^{-1}K_d) + \ln\left(\frac{\det K_m}{\det K_d}\right) - N\right) \\ &+ \frac{1}{2}(\sigma_m - K_m\mu_d)K_m^{-1}K_dK_m^{-1}(\sigma_m - K_m\mu_d) \\ &= D_{KL}(\sigma_m,K_m,\sigma_d,K_d), \end{aligned} \tag{9.35}$$

where we have explicitly written the second term, namely the Mahalanobis distance, as a function of $\sigma_m$. The important point we want to the describe hereafter is the relation between variations of (9.35) with respect to $\sigma_m$ and $K_m$, and its differentiation with respect to $\theta_m$. The first directional derivative of (9.35) with respect to $\sigma_m$ and $K_m$, in the direction $\mathbf{d} = \text{param}(\lambda,\Lambda) \in \mathbb{R}^{N+N^2}$ is

$$\nabla_{\mathbf{d}} D_{KL} := \partial_K D_{KL} : \Lambda + \partial_\sigma D_{KL}\cdot\lambda, \tag{9.36}$$

where, for the sake of compactness, we drop the argument of the KLd. More details about the expression (9.36) can be found in appendix E. We define the first derivative of (9.35) with respect to $\theta_m$ by

$$\text{Grad}D_{KL} := \text{param}(\partial_\sigma D_{KL}, \partial_K D_{KL}), \tag{9.37}$$

where the bijective operator param$(\cdot,\cdot)$ is defined in (2.9) and is the same operator that maps $\theta_m$ to the couple $(\sigma_m, K_m)$. For the second derivative with respect to $\theta_m$ we

can proceed in an equivalent way by defining first the second directional derivatives of the Kullback–Leibler divergence with respect to $\sigma_m$ and $K_m$ in the directions $\mathbf{d} = \text{param}(\lambda, \Lambda) \in \mathbb{R}^{N+N\times N}$ and $\mathbf{d}' = \text{param}(\lambda', \Lambda') \in \mathbb{R}^{N+N^2}$

$$\begin{aligned}\nabla_{\mathbf{d}'}\nabla_{\mathbf{d}} D_{KL} = &\partial_K \text{trace}\left(\Lambda^T \partial_K D_{KL}\right) : \Lambda' \\ &+ \partial_\sigma \text{trace}\left(\Lambda^T \partial_K D_{KL}\right) \cdot \lambda' \\ &+ \partial_K \text{trace}\left(\lambda^T \partial_\sigma D_{KL}\right) : \Lambda' \\ &+ \partial_\sigma \text{trace}\left(\lambda^T \partial_\sigma D_{KL}\right) \cdot \lambda'. \end{aligned} \tag{9.38}$$

Again the explicit algebraic expression for all the terms in (9.38) can be found in appendix E. We rearrange the right–hand side of (9.38) in order to obtain an expression of the form

$$\nabla_{\mathbf{d}'}\nabla_{\mathbf{d}} D_{KL} = \mathcal{H}_K(\rho_m, \rho_d); \mathbf{d}') : \Lambda + \mathcal{H}_\sigma(\rho_m, \rho_d); \mathbf{d}') : \lambda. \tag{9.39}$$

Finally we can define the Hessian matrix of the KLd with respect to the parameter $\theta_m$ columns–wise by

$$[\text{Hess} D_{KL}]_{(\cdot,j)} = \text{param}(\mathcal{H}_\sigma(\rho_m, \rho_d); \mathbf{e}_j), \mathcal{H}_K(\rho_m, \rho_d); \mathbf{e}_j)), \tag{9.40}$$

where $\mathbf{e}_j \in \mathbb{R}^{N+N^2}$ is the $j$–th element of the standard basis of $\mathbb{R}^{N+N^2}$.
Finally, the relation between the Hessian matrix defined column–wise in (9.40) and the Fisher Information matrix (9.29) is simply given, column–wise, by

$$[\mathcal{I}(\theta_d)]_{(\cdot,j)} = \left[\left.\frac{\partial^2 D_{KL}(\theta_m, \theta_d)}{\partial \theta_m^2}\right|_{\theta_m=\theta_d}\right]_{(\cdot,j)}. \tag{9.41}$$

In the next section we will show how from the standard theory of this paragraph we can derive an approximation of the Kullback–Leibler divergence when one banded Gaussian is in fact reconstructed from a cgDNA+ parameter set and a sequence, which the other probability density function is a banded Gaussian estimated from a MD time series of internal coordinates of cgDNA+ degrees of freedom for the same sequence.

### 9.3.2 Computation of an admissible initial cgDNA+ parameter set

Let $\mathcal{P}$ be a cgDNA+ parameter set in the format (9.15) and let $\mathcal{S}$ be a $N$ base–pair long DNA sequence. We assume moreover that $\sigma(\mathcal{S}) \in \mathbb{R}^{24N-18}$ and $K(\mathcal{S}) \in \mathbb{R}^{(24N-18)\times(24N-18)}$ are respectively the weighted–shape and the banded stiffness matrix estimated from a MD time series of cgDNA+ internal coordinates for the sequence $\mathcal{S}$. We denote by $\sigma(\mathcal{P}, \mathcal{S}) \in \mathbb{R}^{24N-18}$ and $K(\mathcal{P}, \mathcal{S}) \in \mathbb{R}^{(24N-18)\times(24N-18)}$ the weighted shape and stiffness matrix reconstructed by a cgDNA+ parameter set $\mathcal{P}$. For sake of simplicity we will use the notation $\theta_d = (\sigma(\mathcal{S}), K(\mathcal{S}))$ and $\theta_m = (\sigma(\mathcal{P}, \mathcal{S}), K(\mathcal{P}, \mathcal{S}))$ to denote respectively the

parameters for the observed banded Gaussian and the parameters of the Gaussian reconstructed by the cgDNA+ parameter set $\mathcal{P}$. The goal of this section is to derive an approximate expression for the Kullback–Leibler divergence

$$D_{KL}(\rho(x,\theta_m(\mathcal{P},\mathcal{S})),\rho(x,\theta_d(\mathcal{S})), \tag{9.42}$$

between $\rho(x,\theta_d)$ and $\rho(x,\theta_m)$. Supplementary detail about the following computation can be found in appendix E.
The main goal of the next paragraph is to derive an approximation of the Kullback–Leibler divergence not in terms of perturbation of banded Gaussian but instead as a function of a cgDNA+ parameter set. For doing that we first start by recalling the reconstruction map

$$\mathcal{R}(\mathcal{P},\mathcal{S}) = (\sigma(\mathcal{P},\mathcal{S}),K(\mathcal{P},\mathcal{S})), \tag{9.43}$$

given by the rules presented in (9.17) and (9.16). An important point to remark here is that the mapping (9.43) is not invertible because of the overlapping structure of the cgDNA+ stiffness matrices. Next, we introduce a bijective mapping between the parameter set $\mathcal{P}$ in the format (9.15) and its vectorial form

$$P = \text{vec}(\mathcal{P}) \in \mathbb{R}^L, \tag{9.44}$$

where $L$ is the total number of independent entries in $\mathcal{P}$. The interesting point here is that we can generalize the reconstruction rule (9.43) for $P$ and $\theta_m(\mathcal{P})$ in the following way

$$\mathcal{R}_{\text{vec}}(P,\mathcal{S}) = R(\mathcal{S})P = \theta(\mathcal{P},\mathcal{S}), \tag{9.45}$$

where the matrix $R \in \mathbb{R}^{N+N^2\times L}$ is called the *parameter reconstruction matrix*. It maps the element of the parameter set to each entry of $\sigma$ and $K$ according to the sequence. The main consequence of (9.45) is that it shows the linear relationship between the parameter set $P$ and the Gaussian parameter $\theta_m(\mathcal{P})$. Thus, we can easily define the gradient and the Hessian matrix of the KLd divergence with respect to the parameter set $P$:

$$\text{Grad}_P D_{KL} := R(\mathcal{S})^T \text{Grad} D_{KL}, \tag{9.46}$$
$$\text{Hess}_P D_{KL} := R(\mathcal{S})^T \text{Hess} D_{KL} R(\mathcal{S}). \tag{9.47}$$

From a computational point of view we never compute the above expression because of the complexity of defining the matrix $R(\mathcal{S})$ explicitly. In appendix E we present a more computational efficient way that we have implemented and used in practise.
We now approximate the Kullback–Leibler divergence (9.42) using

$$D_{KL}(\theta_m,\theta_d) \approx \frac{1}{2}\theta_m\cdot\mathcal{I}(\theta_d)\theta_m - \theta_m\cdot\mathcal{I}(\theta_d)\theta_d + \frac{1}{2}\theta_d\cdot\mathcal{I}(\theta_d)\theta_d,$$

where $\theta_m = \text{param}(\sigma(\mathcal{P},\mathcal{S}), K(\mathcal{P},\mathcal{S}))$ and $\theta_d = \text{param}(\sigma(\mathcal{S}), K(\mathcal{S}))$. We can now use the reconstruction rule (9.45) to get the following approximation

$$D_{KL}(\theta_m,\theta_d) \approx \frac{1}{2}P\cdot R(\mathcal{S})^T\mathcal{I}(\theta_d)R(\mathcal{S})P - P\cdot R(\mathcal{S})^T\mathcal{I}(\theta_d)\theta_d + \frac{1}{2}\theta_d\cdot\mathcal{I}(\theta_d)\theta_d. \quad (9.48)$$

The linear change of variable obtained in the above expression gives a direct relationship between the Fisher information matrix computed for the parameter $\theta$ and the Fisher information matrix computed for the cgDNA+ parameter set:

$$\mathcal{I}_{(\mathcal{P},\mathcal{S})}(\theta_d) = R(\mathcal{S})^T\mathcal{I}(\theta_d)R(\mathcal{S}). \quad (9.49)$$

Finally we obtain the following approximation of the Kullback–Leibler divergence (9.42)

$$D_{KL}(\rho(x,\theta_m),\rho(x,\theta_d)) \approx \frac{1}{2}P\cdot\mathcal{I}_{(\mathcal{P},\mathcal{S})}(\theta_d)P - P\cdot R(\mathcal{S})^T\mathcal{I}(\theta_d)\theta_d + \frac{1}{2}\theta_d\cdot\mathcal{I}(\theta_d).\theta_d. \quad (9.50)$$

We can now take advantage of the approximation (9.50) for computing a candidate initial guess. Let us consider the palindromic sequence library $Lb_{\text{palin}}$ and its corresponding oligomer–based statistics $\left\{(\sigma(\mathcal{S}_i), K(\mathcal{S}_i)|\mathcal{S}_i \in Lb_{\text{palin}}\right\}$. Then, for a given cgDNA+ parameter set $\mathcal{P}$ consider the following functional

$$\sum_{i=1}^{16} D_{KL}(\rho(x,\theta(\mathcal{P},\mathcal{S}_i)),\rho(x,\theta(\mathcal{S}_i)), \quad (9.51)$$

and in particular, by using (9.50) on each element of the summation, we obtain the approximation of (9.42)

$$\begin{aligned}\sum_{i=1}^{16} D_{KL}(\rho(x,\theta(\mathcal{P},\mathcal{S}_i)),\rho(x,\theta(\mathcal{S}_i)) &= \sum_{i=1}^{16}\frac{1}{2}P\cdot\mathcal{I}_{(\mathcal{P},\mathcal{S}_i)}(\theta_d)P - P\cdot R(\mathcal{S}_i)^T\mathcal{I}(\theta_d)\theta_d + C\\ &= \frac{1}{2}P\cdot\mathcal{I}_{(\mathcal{P},Lb_{\text{palin}})}(\theta_d)P - P\cdot B + C. \quad (9.52)\end{aligned}$$

where

$$\begin{aligned}\mathcal{I}_{(\mathcal{P},Lb_{\text{palin}})} &= \sum_{i=1}^{16}\mathcal{I}_{(\mathcal{P},\mathcal{S}_i)}(\theta_d),\\ B &= \sum_{i=1}^{16} R(\mathcal{S}_i)^T\mathcal{I}(\theta_d)\theta_d,\\ C &= \frac{1}{2}\sum_{i=1}^{16}\theta_i\cdot\mathcal{I}(\theta_i)\theta_i.\end{aligned}$$

The candidate initial guess can then be computed by minimizing (9.52) and, consequently, by solving the following least square system, or Fisher system,

$$\mathcal{I}_{(\mathcal{P},Lb_{\text{palin}})}P = B. \tag{9.53}$$

In equation (9.53) the matrix $\mathcal{I}_{(\mathcal{P},Lb_{\text{palin}})}$ is not invertible because of the non injectivity of the reconstruction mapping for the cgDNA+ model. Thus, the matrix $\mathcal{I}_{(\mathcal{P},Lb_{\text{palin}})}$ has as many zero eigenvalues as the dimension of the null–space generated by the parameter set format and the reconstruction scheme. In section (9.4) we will better present this null–space, but for the moment we stress that, due to the zero eigenvalues of $\mathcal{I}_{(\mathcal{P},Lb_{\text{palin}})}$, the solution $P$ in (9.53) is computed using the Moore–Penrose pseudo–inverse. Finally, we denote

$$\widehat{P} = \mathcal{I}^{\dagger}_{(\mathcal{P},Lb_{\text{palin}})}B,$$

a *candidate* initial guess were $A^{\dagger}$ is the psuedo–inverse of A. Before claiming that $\widehat{P}$ is an actual initial guess have to verify that it belongs to the space of admissible parameters set $\mathbb{P}$. Unfortunately there is no known proof that ensures that for an arbitrary training library $Lb$ the data $\{(\sigma(\mathcal{S}_i), K(\mathcal{S}_i)|\mathcal{S}_i \in Lb)\}_{i=1}^{M}$ leads to a solution of (9.53) that always lies in $\mathbb{P}$. Thus for any set of data one should verify that $\widehat{P} \to \widehat{\mathcal{P}} \in \mathbb{P}$. If the latter condition is satisfied we will called $\widehat{P}$ an initial guess, denoted $P_{ini} \to \mathcal{P}_{ini}$, and it will be used to solve (9.27) numerically.
Practically, we have compute the initial guess by solving (9.53) for the Palindromic data set and after checking that the parameter set reconstructs positive definite stiffness matrices for each palindromic sequence, we have computed the averaged Kullback-Leibler divergence per degrees of freedom (6.6) in order to be able to compare the accuracy of $\mathcal{P}_{ini}$ to test the results obtained in chapters 6 and 7. We present the findings in table 9.9.

| $\mathcal{P}_{ini}$ | $\overline{D}$ | $\overline{D}^{\dagger}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|
| PALIN | $5.59{\cdot}10^{-2}$ | $5.44{\cdot}10^{-2}$ | $0.15{\cdot}10^{-2}$ |

Table 9.9 – Average Kullback-Leibler divergence per degrees of freedom (6.6) computed for the initial guess parameter set $\mathcal{P}_{ini}$ and the Palindromic data set.

We notice that the stiffness part of KLd is the major contributor to the overall value of the error, while the shape part as a surprisingly low value. In conclusion we can state that $\mathcal{P}_{ini}$ is a good point from where to start the numerics' but still too poor to avoiding the actual resolution of (9.27) by minimisation. In the next section we will focus on a numerical scheme for computing best-fit cgDNA+ parameter sets.

### 9.3.3 Fisher informed gradient

In this paragraph we will expose the methodology used for deriving the first best–fit cgDNA+ parameter set. The scope of this section is not to analyse the proposed method but to present the numerical scheme adapted to resolution of problem (9.27) and to explain its efficacy.
Once the initial guess $\mathcal{P}_{ini}$ is computed we can start the minimization of (9.27). In this section we will present a different approach to numerics than the one proposed at the end of chapter 5 or with the algorithm (2). In particular in 5.2 we presented a scheme for the equivalent cgDNA problem (9.27) that is divided into two major steps: a gradient descent stage and a Quasi-Newton method stage using the Broyden method. The first step aims at decreasing the residue (norm of the gradient computed at the current iteration) until a certain threshold. This first step can last for multiple days/weeks especially because the computation is very sensitive on the step size. The second stage aims at decreasing the residue down until convergence is achieved (residue smaller then $10^{-11}$), which can take a few days. Practically speaking one of the most expensive steps in the Broyden algorithm is to compute the Hessian matrix. It could take up to 9–12 hours to get the Hessian matrix on a regular laptop, and up to one hour to compute the same matrix on a CPU server with 12 cores. The choice of using a Broyden method was actually motivated by this high cost, in term of time, for computing the Hessian matrix, which made infeasible the use of a classic Newton method. Historically, the first cgDNA parameter set was computed using these two–stage procedures described above by starting from an admissible initial parameter set constructed from the data. In particular the computation of the initial guess presented in paragraph (9.3.2) was not known at that time which made the construction of the initial parameter set very hard and delicate. The computations were very intense and lasted for multiple months: from the computation of the initial guess to the convergence of the Broyden algorithm. Once the first parameter set was computed, all the following ones were computed using the parameter continuation procedure presented in chapter (6), summarized by the algorithm(2).
For computing the very first cgDNA+ parameter set we changed completely the procedure thanks to the Fisher information matrix (9.49). For the first cgDNA+ parameter set the gradient descent–Broyden approach could easily take months to converge even if the initial guess is given by (9.53) as the dimension of the unknown is significantly larger. We thus implement a different approach that uses the matrix $\mathcal{I}^{\dagger}_{(\mathcal{P},Lb_{\text{palin}})}$ as pre–conditioner in the gradient descent step, defining, at step $k$, the following updating scheme

$$P^{(k+1)} = P^{(k)} - \alpha \mathcal{I}^{\dagger}_{(\mathcal{P},Lb_{\text{palin}})} \text{Grad}_P \mathbb{F}(P^{(k)}; Lb), \tag{9.54}$$

where Grad$\mathbb{F}$ is the gradient of sum of KLs in (9.27) computed using the gradient of a single Kullback–Leibler divergence (9.46) and $\alpha \in ]0\ 1]$ is the step size. In appendix

E one can find the detailed computation of the gradient of (9.27) with respect to the parameter set. The updating scheme (9.54) leads to a preconditioned–gradient flow method that we call, *Fisher–informed* gradient flow. The Fisher–informed gradient started from the initial guess (9.53) converges very fast which means, in our context, that in less then one hour we managed to compute the best–fit cgDNA+ parameter set trained on the Palindromic data set. The computational time refers to the computation run on a CPU server with 24 cores and MATLAB multi–threaded computations and initially taking into account only ten independent interior dimers and only $G\mathrm{p}C$ end dimers. In the next paragraph we better explain this point. Even if we are far from being able to compute in a fast way a cgDNA+ parameter set on an everyday laptop, the Fisher–informed gradient method improved substantially the entire parameter extraction procedure and, combined with the computation of the initial guess of section (9.3.2), it leads to the conclusion that the parameter continuation in the context of cgDNA+ parameter set is not needed. Meaning that for any new data set it is faster to compute the initial guess, by solving the Fisher system, (9.53) and by running the Fisher–Informed gradient method (9.54).
A possible explanation of the efficacy of the Fisher–informed method is that the Fisher information matrix evaluated at the data brings the additional information of the curvature of the Gaussian parameters space at the point given by the parameters of the banded Gaussians. The inverse of the Fisher information matrix then plays the role of rescaling the gradients by taking into account the geometry of the space around the solution. For completeness of the discussion we would like to address the attention of the reader to the fact that the Fisher information matrix as defined in (9.29) defines a metric on the Riemann manifold whose points are parametric probability measures, thus it induces an inner product. Finally, the Fisher information metric is also related to the Hessian of the KLd with respect to the parameter of the Gaussian. The Hessian we are considering is actually the second derivative with respect to the cgDNA+ parameter set thus, the latter theory cannot be directly applied but, nevertheless, the matrix used in the updating scheme (9.54) helps substantially the convergence of the gradient flow. This if is for surely related to the fact that the Hessian matrix used in the updating scheme changes the inner product and, consequently, the geometry of the problem.
The total number of independent entries in the cgDNA+ parameter set (9.15) is, 20,682 which means that this is the size of the optimization problem (9.27). To avoid working in such a big space we adopted the following technique:

1. We consider only the Palindromic library, meaning only $G\mathrm{p}C$ end blocks.
2. We add each end block independently using the corresponding sequence of the end library B.4, keeping fixed the values of the interior blocks.

The dimension of the system in 1) is 10152 while the dimension of the system in 2) is only 702. The main reason underlying the two points presented above is, of course,

to decrease the dimension of the system. Also because the end sequences are not palindromic so that a proper convergence test is not feasible. Consequently, we trust less the estimated mean and covariance for the end data set. On the other end many applications of the cgDNA model involve localised sequence marginals, so that an accurate parameter set of end blocks is not so crucial. In conclusion the ends data should be simulated for longer time in order to enhance the quality of the resulting oligomer–based statistics. In the next section we will discuss the positiveness of the best–fit cgDNA+ parameter set.

## 9.4 Proving positiveness of the best-fit parameter set

Using the palindromic, oligomer–based statistics we have numerically solved problem (9.27) starting from the admissible initial guess computed by (9.53), see table (9.9) for the accuracy of the initial guess. We started first by computing the ten interior dimer elements and only $G\text{p}C$ end dimers. Now, the next step of the cgDNA+ parameter estimation procedure is to prove that the obtained best–fit parameter set is actually positive definite, in the sense that, the stiffness matrix predicted for an arbitrary sequence (with $G\text{p}C$ ends) is positive definite. We need to recall that uniqueness of the best-fit parameter is not satisfied because of the non injectivity of the following linear map

$$\mathcal{R} : (\mathrm{P}, Lb) \to \{\sigma(\mathrm{P}, \mathcal{S}_i), K(\mathrm{P}, \mathcal{S}_i)\}_{i=1}^{M}, \tag{9.55}$$

which represents the reconstruction rule (9.17,9.18,9.16) for the sequence library $Lb$. More precisely, let $\mathrm{P} \in \mathbb{P}$ be an arbitrary parameter set (9.15) and define $\Gamma^A, \Gamma^G \in \mathbb{R}^{18\times 18}$ and $\gamma^A, \gamma^G \in \mathbb{R}^{18}$ respectively two non zero matrices and two non zero vectors. Define then $\Gamma^{\overline{\alpha}} = E^{\Gamma}\Gamma^{\alpha}E^{\Gamma}$ and $\gamma^{\overline{\alpha}} = E^{\Gamma}\gamma^{\alpha}$ with $\alpha = \{A, G\}$, and $E$ defined by

$$E^{\Gamma} = \begin{bmatrix} & & I_6 \\ & E & \\ I_6 & & \end{bmatrix}, \; E = diag(-1, 1, 1, -1, 1, 1).$$

We can now define a new parameter set $\mathrm{P}^{'}$ whose elements are, for the stiffness part,

$$K_{\Gamma}^{\alpha\beta} = K^{\alpha\beta} + \mathrm{diag}(-\Gamma^{\alpha}, \mathbf{0}_6, \Gamma^{\beta}), \; K_{\Gamma}^{5'\alpha\beta} = K^{5'\alpha\beta} + \mathrm{diag}(\mathbf{0}_{18}, \Gamma_{\alpha}), \tag{9.56}$$

where $K^{\alpha\beta}, K^{5'\alpha\beta} \in \mathrm{P}$. With a similar construction for the weighted shape parameters $\sigma^{\alpha\beta}$ and $\sigma^{5'\alpha\beta}$ using $\gamma^X$, $X = A, T, G, C$ we obtain a parameter set that will satisfy

$$\mathcal{R}(\mathrm{P}, Lb) = \mathcal{R}(\mathrm{P}^{'}, Lb), \text{ for any arbitrary set of sequences } Lb. \tag{9.57}$$

We can take advantage of the non trivial null space in the parameter set for proving that the best-fit parameter set $\mathcal{P}$ is actually reconstructing a positive definite matrix for a

arbitrary sequence. We have already presented in section 7.3.1, that the stiffness block elements in the cgDNA parameter set (7.17) are all symmetric and positive definite, but as a matter of fact, the stiffness blocks of the best-fit parameter set are not positive definite. We can therefore use the freedom given by the null space to find the two matrices $\Gamma^A$ and $\Gamma^G$ and use the equations (9.56) to find a positive definite cgDNA+ parameters. set. We present a technique we have designed that searches in the null space for the $\Gamma^\alpha$ elements.
Let $K \in \mathbb{R}^{n\times n}$ be a matrix, with eigenvectors and eigenvalues $\{x_i, \lambda_i\}_{i=1}^n$, and further assume that the eigenvalues are all distinct. Now introduce a small perturbation matrix $\delta K$ and the perturbed matrix $K' = K + \delta K$. We can study how much $\{x_i, \lambda_i\}_{i=1}^n$ will change with respect to the perturbation $\delta K$. In fact, it is expected that if $\delta K$ is small enough, the perturbation in the eigenvectors and eigenvalues of $K$ will also be small. More precisely, we expect that

$$\lambda_i' = \lambda_i + \delta\lambda_i, \tag{9.58}$$
$$x_i' = x_i + \delta x_i, \tag{9.59}$$

for all $i = 1, \ldots, n$. In the particular case when $K$ is symmetric and positive definite, explicit formulas for $\{\delta x_i, \delta\lambda_i\}_{i=1}^n$ can be derived, see for instance [73]. We can now ask the following questions: can we find two matrices $\Gamma^A, \Gamma^G \in \mathbb{R}^{18\times 18}$ and use equation (9.56) to find a positive defined best-fit parameter set $\mathcal{P}$. How can we construct such matrices?
A priori the first question is not answerable because there is no guarantee that a solution of (9.27) does reconstruct positive definite matrices for any arbitrary sequence, but an extensive study of the positiveness can be done to be gain insight on the problem. It is sufficient that positiveness is not satisfied for a single sequence to break any hope to prove the positiveness of $\mathcal{P}$. Practically, we have used the best–fit cgDNA+ parameter set to reconstruct stiffness matrices for millions of sequences (with $G\text{p}C$) and not a single indefinite matrix has been found. To prove then that the parameter set is actually positive definite we used the following approach to construct the elements of the null space.

Given a best-fit parameter set $\mathcal{P} = \{\sigma^{\alpha\beta}, \sigma^{5'\alpha\beta}, K^{\alpha\beta}, K^{5'\alpha\beta}\}_{\alpha\beta\in D}$,
find $\Gamma^A, \Gamma^G \in \mathcal{C}$, where
$$\mathcal{C} = \left\{\Gamma^A, \Gamma^G \in \mathbb{R}^{6\times 6} | K_\Gamma^{\alpha\beta} > 0,\ \forall\alpha\beta,\ K_\Gamma^{5'\alpha\beta} > 0,\ \forall 5'\alpha\beta\right\}, \tag{9.60}$$

and $K_\Gamma^{\alpha\beta}$, $K_\Gamma^{5'\alpha\beta}$, are defined in (9.56). For doing that, one can use a standard algorithm to minimize the following constrained nonlinear problem

$$\min_{\Gamma^A,\Gamma^G\in\mathcal{C}} F(\Gamma^A,\Gamma^G;\mathcal{P}), \tag{9.61}$$

with $F : \ker(\mathbb{P}) \to \mathbb{R}$, with the specific choice of $F$ as the zero functions. The latter choice for the objective function implies that one just wants to find an arbitrary element of the null space of $\mathbb{P}$, meaning an admissible point of $\mathcal{C}$. As the construction of an admissible point is not trivial, one must start from a random initial guess for $\Gamma^A$ and $\Gamma^G$ and need to iterate by adding the current solution to the previous initial guess. More precisely, the element in the null space after $k+1$ iterations will be computed as

$$[\Gamma^\alpha]^{k+1} = [\Gamma^\alpha]^k + [\Gamma^\alpha_{\text{sol}}]^k, \text{ for } \alpha \in \{\alpha\beta\}, \tag{9.62}$$

where $[\Gamma^\alpha_{\text{sol}}]^k$ is the solution after $k$ iterations obtained using the initial guess $[\Gamma^\alpha]^k$. We noticed that, with this strategy, at each iteration i) the total number of negative eigenvalues ii) the sum of the absolute value of the negative eigenvalues iii) and the Frobenius norm of $\Gamma^\alpha_{\text{sol}}$, decrease in an oscillatory fashion. This means that the procedure is actually lifting up the negative eigenvalues. Unfortunately, the convergence of the latter procedure is quite slow and another procedure should be used. Starting from a solution computed after some iteration of the previous strategy, one can use the following non trivial objective function

$$F(\Gamma^A,\Gamma^G;\mathcal{P}) = \sum_{\alpha\beta\in D} \|K_\Gamma^{\alpha\beta}\|_{\text{F}} + \sum_{5'\alpha\beta\in D} \|K_\Gamma^{5'\alpha\beta}\|_{\text{F}}, \tag{9.63}$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm, and continue to compute the element in the null space in an iterative way. With this non trivial objective function, we found a solution to (9.60) and thus we can prove a positive best-fit parameter set $\mathcal{P}$ trained on the Palindromic data set.

For the missing end dimers blocks we first reconstructed a large number of sequence which present all the possible combination of ends dimers. Unfortunately, we concluded that any sequence ending with the dimer blocks $G\text{p}A$ or $T\text{p}G$ is reconstructed with an indefinite stiffness matrix. On the other hand, we used the method present above to find a positive definite cgDNA+ parameter set for ten interior dimer steps and fourteen ends dimer steps different than $G\text{p}A$ or $T\text{p}G$. We conjectured that indefiniteness of the two end dimer blocks is related to a lack of convergence of the corresponding statistics, but unfortunately we do not have more data to prove this statement. In any case, we already mentioned that reliability of the statistics estimated for the end data set cannot be verified compared to the palindromic data set for which we could study the convergence of both, mean and covariance, estimators. But, as future development, the end data set needs to be extended in simulation duration in order to ensure that the parameter estimation procedure proposed in the two point

steps method presented at the end of section 9.3.3 is efficient and reliable. In the following section we will consider the best–fit parameter set $\mathcal{P}$ as the parameter estimated only from the Palindromic data set.

## 9.5 Assessing the best–fit cgDNA+ parameter set

We can check accuracy of the best–fit cgDNA+ parameter set trained on the palindromic data set. We first start by looking at the average Kullback-Leibler function per degree of freedom introduced in (6.6):

| cgDNA+ | $\overline{D}$ | $\overline{D}^{\dagger}$ | $\overline{\mathcal{M}}$ |
|---|---|---|---|
| PALIN | $2.40{\cdot}10^{-2}$ | $2.29{\cdot}10^{-2}$ | $0.12{\cdot}10^{-2}$ |

It is interesting to see how well approximated the ground–state are, in the sense of the Mahalanobis distance. In fact, no single comparison done in chapters 6 and 7 with the cgDNA model has shapes predicted with such a great precision. This implies that most of the cgDNA+ approximation error is in the stiffness, but it is still in the range of errors we have obtained in section 7.3. This suggest that the whole process of extracting a coarse–grained model of DNA from MD trajectories presented in this work is stable and accurate even for a very large set of parameters such as for the cgDNA+ model. In the next paragraph we present the comparison between data and reconstruction of ground–states and tangent–tangent correlations.

### 9.5.1 Ground–state

The first comparison we make is between ground–states of palindromic sequences in the training library and the corresponding cgDNA+ predictions. For detailed comparison we have selected three sequences: $\mathcal{S}_1$, $\mathcal{S}_5$, and $\mathcal{S}_{11}$. For the selected sequences, in figure 9.9 we show the comparisons of the phosphate degrees of freedom observed from MD simulations and predicted by the model. As all the sequences are palindromes we can just show the base–to–phosphate coordinates for one of the strand, and in figure 9.9 we have selected the reading strand. We notice a really good agreement between predictions and observation for all the selected sequences. For the inter– and intra–base–pair internal coordinates we decided to show the differences (in absolute value) between the predictions of the cgDNA parameter set derived in chapter 7 and cgDNA+ for the same sequences. In figure 9.10 in solid lines we show the errors for cgDNA+ predictions while in dashed lines we show the error for the cgDNA ones. In general for each rigid–base–pair degree of freedom the cgDNA+ prediction are closer to the data. In figure 9.11 we show the same errors for the intra–base–pair internal coordinates and again in general the cgDNA+ model reduces the error between obser-

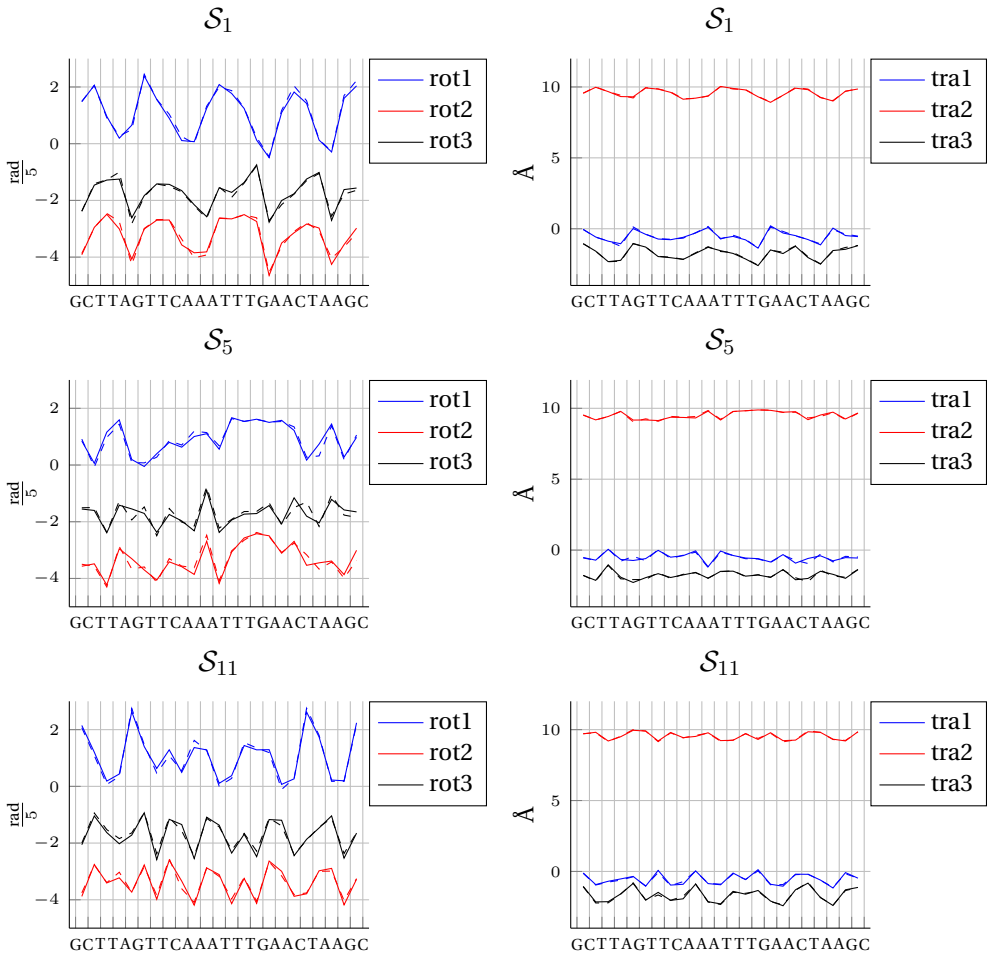



Figure 9.9 – Comparison of base–to–phosphate degrees of freedom, on the reading strand, between cgDNA+ predictions (solid line) and MD observation (dashed line) for the sequences (1,5,11) of the Palindromic Library. In the first column we show the rotational coordinates, while in the second the translations.

vation and predictions.

In conclusion, the ground–states predicted by the cgDNA+ model are in excellent agreement with the MD observation and in general the cgDNA+ model enhances the quality of the predictions of the inter– and intra–base–pair components, in addition to predicting base–to–phosphate degrees of freedom.

### 9.5.2 Tangent–tangent correlation

In this paragraph we consider the sequences $(1, 7, 10, 14)$ and we compare the tangent–tangent correlation computed using the training data and using the cgDNA+ predicted Gaussians. First we recall that for a $N$ base–pair long sequence $\mathcal{S}$ the tangent–tangent

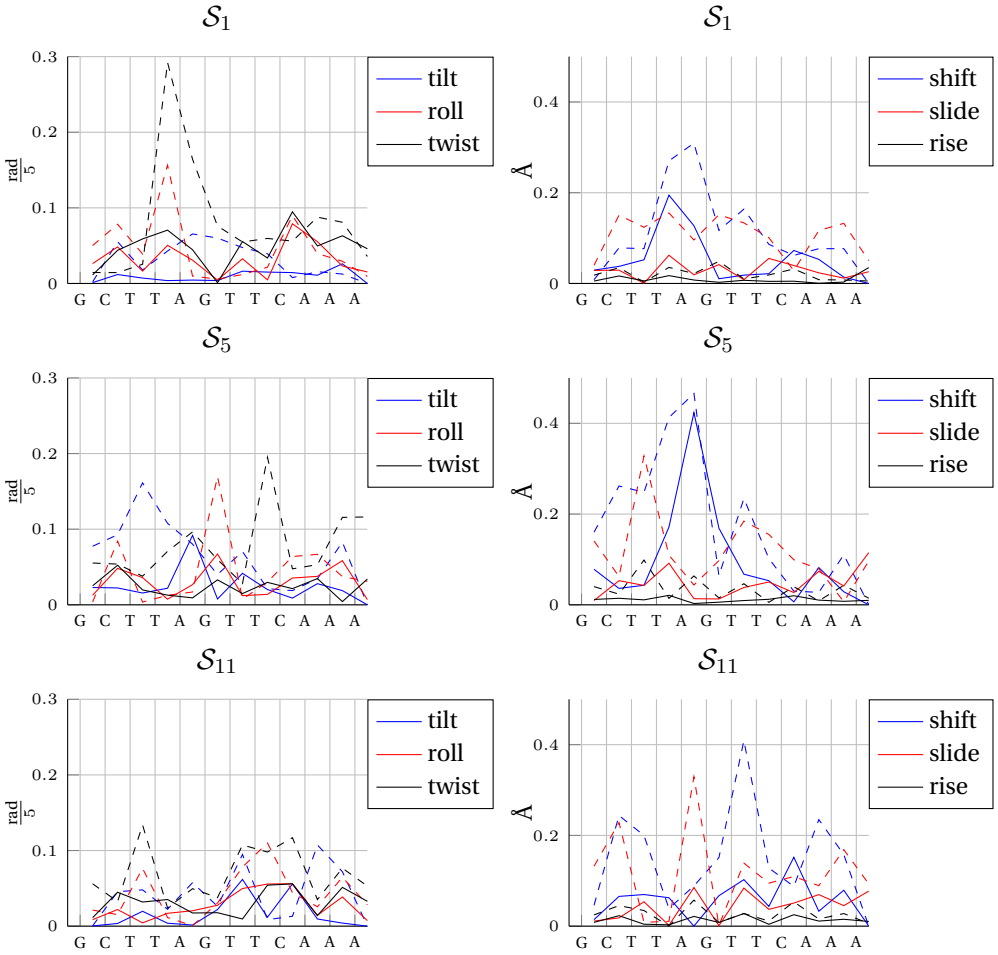



Figure 9.10 – Absolute error between model predicted inter–base–pair degree of freddom and MD obsrvation. In solid, we show the error obtained by the cgDNA+ model and in dashed the error obtained by the cgDNA model. The sequences considered are (1,5,11) in the Palindromic Library. If two plots were super imposed they would be indistinguishable, which is why we chose to plot errors.

correlation function reads:

$$\langle \mathbf{t}_n \cdot \mathbf{t}_0 \rangle = \int_{\mathbb{R}^{12N-6}} (\mathbf{t}_n(w) \cdot \mathbf{t}_0) \rho(w;\mathcal{S}) dw \tag{9.64}$$

where $\mathbf{t}_0$ is the tangent vector of a fixed base–pair (usually taken away from an end) and $\rho(w;\mathcal{S})$ is a probability density function in the configuration space of $\mathcal{S}$. Since an analytical expression of (9.64) is not available for a general $\rho(w;\mathcal{S})$, one can only approximate it. Therefore, we estimate (9.64) using the Monte Carlo method where the configuration ensemble to use for the evaluation of the tangent–tangent function will be generated in two different ways:

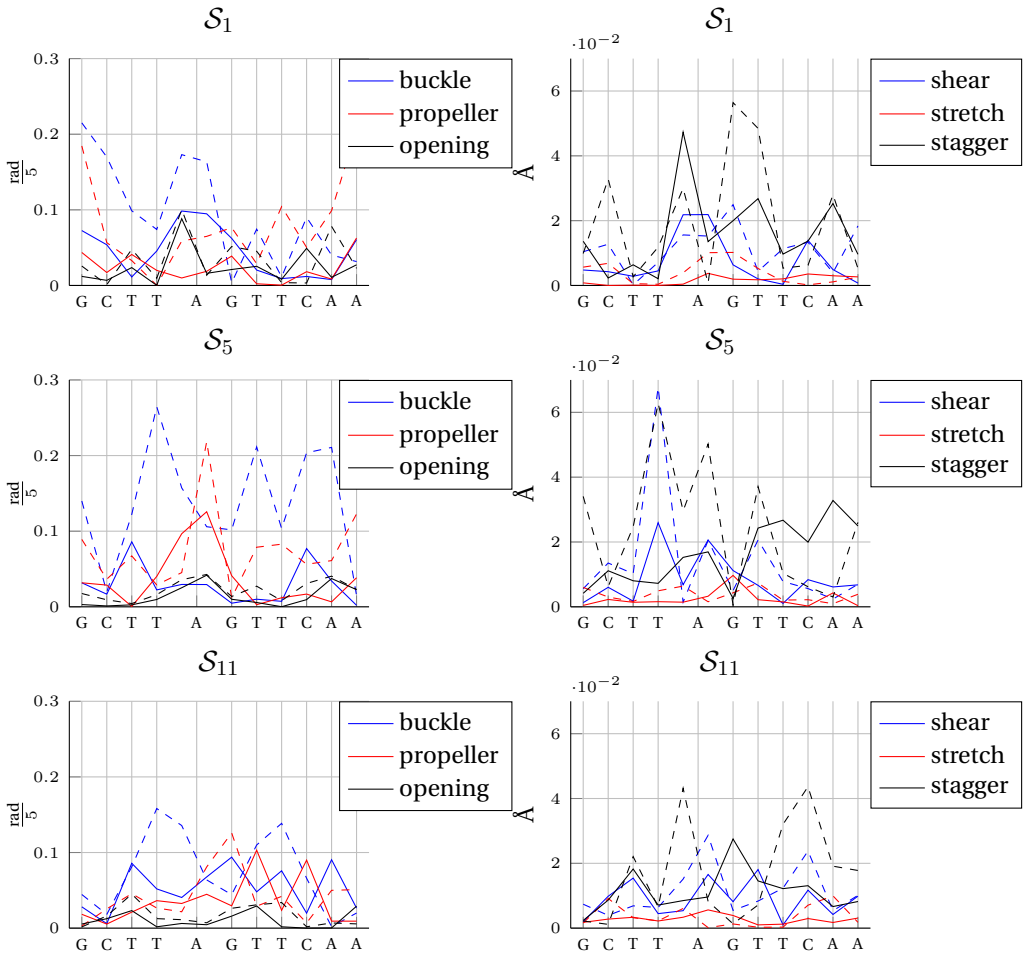



Figure 9.11 – Absolute error between model predicted intra–base–pair degree of freedom and MD obsrvation. In solid, we show the error obtained by the cgDNA+ model and in line the error obtained by the cgDNA model. The sequences considered are (1,5,11) in the Palindromic Library

1. from the (filtered) MD time series of trajectories of internal coordinates,
2. from direct sampling of the observed Gaussian.

In figure 9.12 we label by *MD* the tangent–tangent computed from the ensemble generated by the MD trajectories after hydrogen–bond filtering (see for detail section 3.3.1) while with the label *MC* the ttc computed using the Monte Carlo method and direct sampling of the observed Gaussian. We can observe that both ttc curves gives the same result. By consequence we can conclude that the tangent–tangent correlation is not influenced by the non–Gaussian behaviour of configurations in the ensemble generated from the filtered MD trajectories as its value is totally defined by the first and second (centred) moments. Moreover, we computed the ttc using the *unfiltered* trajectories and noticed that the HB filtering does not influence the ttc computation.

We decided to omit this curve in figure 9.12 because it is strictly the same as the others two. From now on we will use the label *MD* for the ttc computed with method 2) mentioned above.

In the next comparisons we will consider only direct sampling method from the

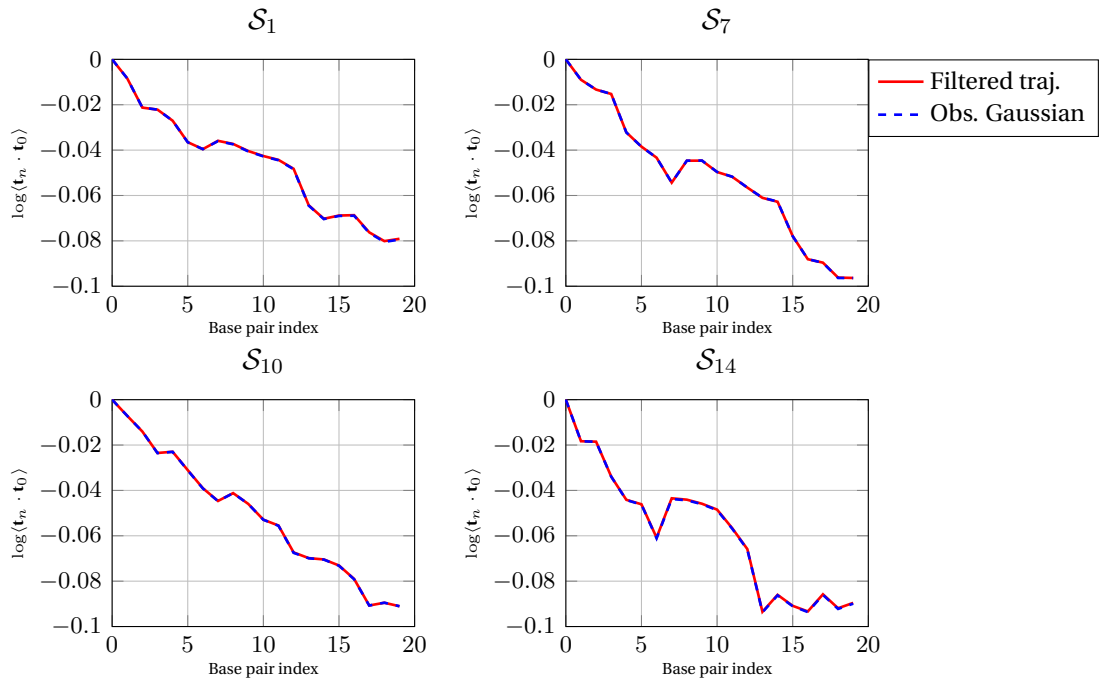



Figure 9.12 – Comparison of tangent–tangent correlation functions computed from an ensemble of filtered MD trajectories (solid) and computed from an ensemble generated by direct sampling of the oligomer–based Gaussian of the filtered MD trajectories. The sequence considered are (1,7,10,14) of the Palindromic Library.

following different Gaussians defined by different sparsity pattern and different degrees of freedom. We will consider:

1. the observed Gaussian as in figure 9.12 (*MD*),
2. the truncated approximation of the observed one, with cgDNA+ degree of freedom (*NN*+),
3. the truncated approximation for cgDNA degree of freedom (*NN*),
4. the truncated version for rigid–base–pair degree of freedom (*rbp*).

One of the key points is that the integral (9.64) depends only on the inter–base–pair internal coordinates and thus in practice to estimate the tangent–tangent correlation function one just needs a distribution on these degree of freedom. In figure 9.13 we show in black the ttc's for the MD distribution defined on the cgDNA+ internal

coordinates, but this curve is the same for any other choice of degrees of freedom as long as it contains the inter–base–pair ones and the covariance is the observed one. After these observation we can properly compare the consequence of nearest–neighbour interactions assumption on three different level of coarse–graining. In figure 9.13 we show: in green the rigid–base-pair model, in red the rigid–base model and in blue the rigid–base and rigid–phosphate model. It was already observed in section 9.1 that by adding more detail to the coarse–grain model the more accurate the assumption of locality is. But here we further show how accurate the banded Gaussians are in predicting a non–trivial observable from the MD data, and again the more detailed coarse–grain model performs better. We can finally compare the

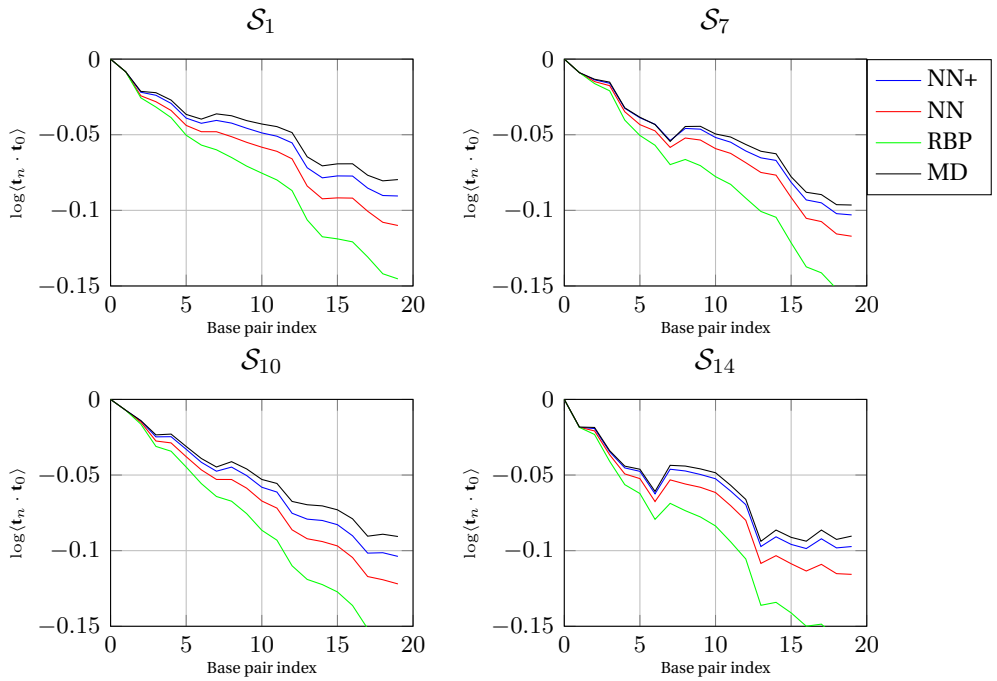



Figure 9.13 – Computation of the tangent–tangent correlation function for different banded matrices estimated from MD simulations. In black we show the ttc for the observed distributions, in blue we show the ttc for the banded approximation for rigid–base rigid–phosphate model, in red the one for rigid–base model, and in green the one for the rigid–base–pair model. The banded approximation for the three coarse–grain levels represent the corresponding nearest–neighbour interactions assumption. The sequence considered are (1,7,10,14) of the Palindromic Library. We conclude that a more detailed model combined with the nearest–neighbour assumption lead to a better appriximation of the ttc observed from MD.

performance of the best–fit cgDNA+ parameter set on predicting tangent–tangent correlation function. In figure 9.14 in blue we plot the cgDNA+ prediction while in black we show the MD one, and in red the banded oligomer–based Gaussian. We can see that the prediction of the model stays in the range of the truncation error and

moreover, it implies that it can predicts really well the rigidity, in the sense of apparent persistence length, of the data. For sake of completeness in figure (9.15) we show also

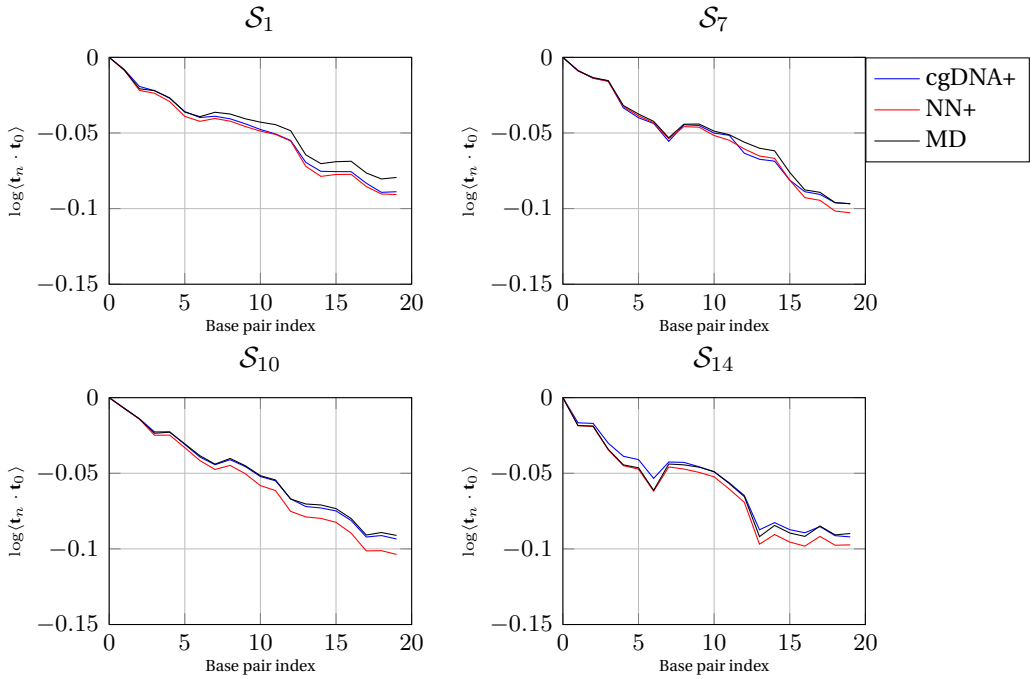



Figure 9.14 – Tangent–tangent correlation functions computed using the observed Gaussian (black), its banded approximation Gaussian (red), and the predicted cgDNA+ Gaussian (blue). The sequences considered are (1,7,10,14) of the Palindromic Library. We conclude that the error between ttc predicted by cgDNA+ and ttc observed from MD is in the range of error of the banded approximation.

the the factorised version of the tangent–tangent correlation where one again observe that the predictions of the cgDNA+ model are still remarkably good.

## 9.6 Beyond nearest neighbour interactions

In this section we discuss a possible further development of the cgDNA+ model and in particular we discuss a possible extension of the band of the stiffness matrix by introducing the interaction between an inter–base–pair coordinates and its two neighbours. For sake of simplicity in this section we will refer to the cgDNA+ stencil as the *nearest–neighbour* (NN) pattern while the extended stencil will be called *next–to–nearest–neighbour* (NNN). In figure 9.16 on the top left panel we show the NN stencil, in blue, and the NNN sparsity pattern, in red. Moreover we highlighted in dark green the overlapping blocks while in we highlighted red the inter–base–pair one. The choice of discussing the NNN sparsity pattern comes simply from the observation made on the raw stiffness matrices for the palindromic training set. In figure 9.16 we show

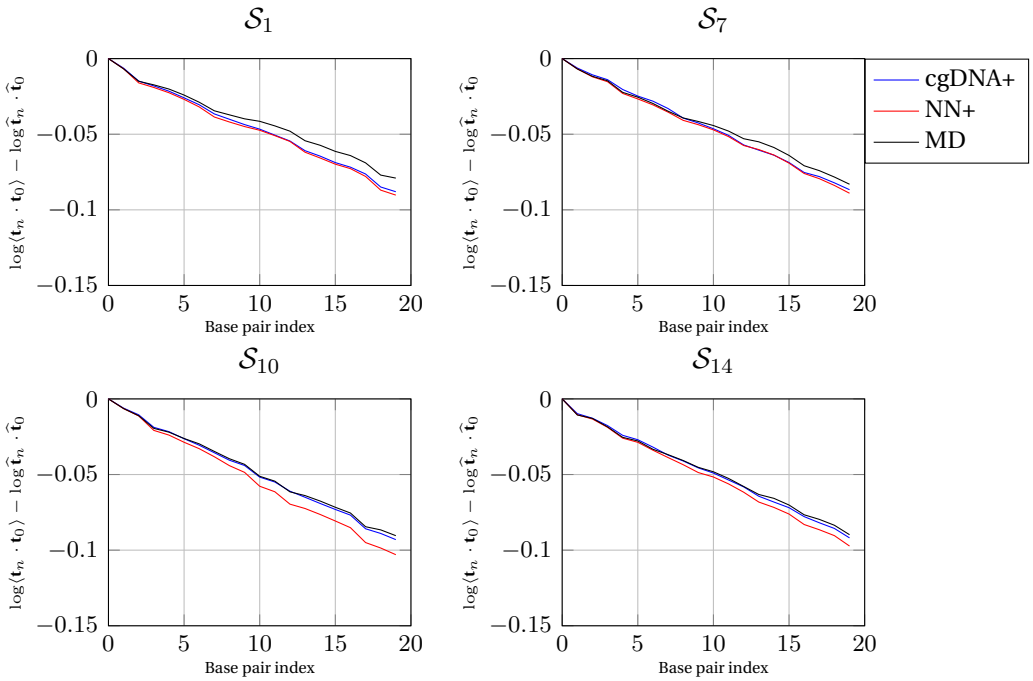



Figure 9.15 – Factorised tangent–tangent correlation function computed using the observed Gaussian (black), its banded approximation Gaussian (red), and the predicted cgDNA+ Gaussian (blue). The sequence considered are (1,7,10,14) of the Palindromic Library.

three different examples of observed stiffness matrices for the palindromic sequences $(7, 12, 16)$, chosen arbitrarily, where one can see the extra $6 \times 6$ blocks that characterise the NNN sparsity pattern. In the next paragraph we apply a model selection criterion to study the NNN model on the oligomer–based statistics

In [1, 2] a model selection theory based on the Kullback–Leibler divergence and the maximum likelihood was developed. The aim of a model selection procedure is, given some data, to select, from an ensemble of statistical models, the most suitable one. The rational behind the choice is based upon the context of the theory introduced by Akaike: suppose that $q$ is a probability density function and that $\widehat{\theta}$ is the maximum likelihood estimator given some realisation of $q$, denoted by $\mathbf{w} = \{w_i\}_{i=1}^{L}$. The Akaike information criterion is given by

$$\text{AIC}(\widehat{\theta}) = 2n_{\text{param}} - 2\log(\mathcal{L}(\widehat{\theta}), \mathbf{w}), \tag{9.65}$$

where $n_{\text{param}}$ is the total number of parameters to be estimated in $\widehat{\theta}$, and $\mathcal{L}(\cdot)$ is the maximum–likelihood function. Given an ensemble of candidate statistical models the best one has the minimal value of the AIC within the ensemble. The main goal of the Akaike information criterion is not only to find the model that best fits the data, to also avoid over fitting thanks to the penalty term $n_{\text{param}}$. But, it is important

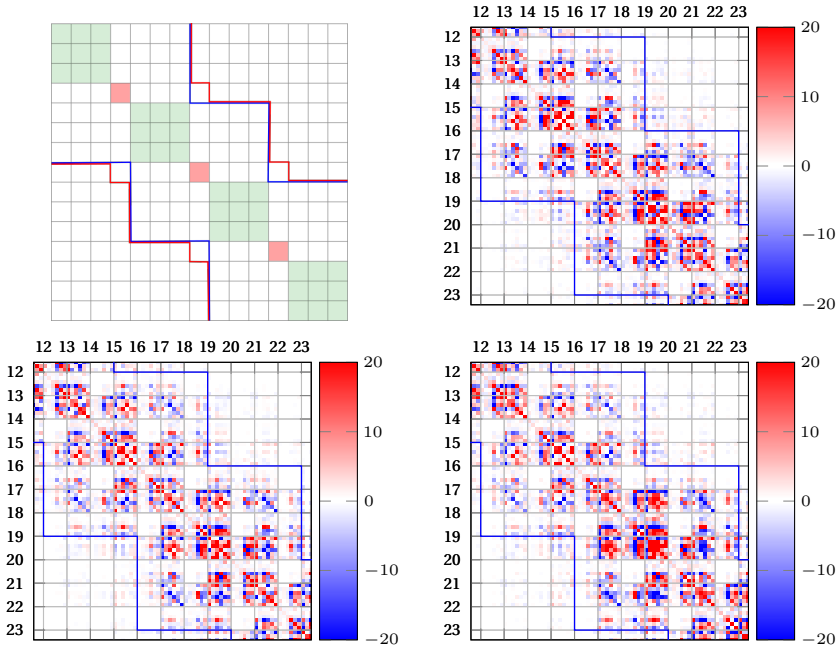



Figure 9.16 – In the top left panel we show two different stencils: in blue the nearest–neighbour stencil used in the cgDNA+ model and in red the next–to–nearest–neighbour stencil which includes an extra 6 times 6 blocks related to the interactions between a inter–base–pair degree of freedom and its adjacent one. In dark green we show the overlaps related to the micro structure and in red the inter–inter interactions. In the other three panels we show three examples of observed stiffness matrices for the palindromic sequences $(7, 12, 16)$. The extra 6 times 6 blocks outside the blue stencil is clearly visible in the example, as being the largest addition contribution.

to mention that (9.65) is derived from minimization of Kullback-Leibler divergence, which in this context is also called *Kulback–Leibler information*, between an unknown pdf and different approximations to it: Let $q$ be the unknown pdf and let $p(x;\theta)$ an approximate model, pdf parametrised by a parameter $\theta \in \mathbb{R}^N$. We recall that the KLd between $p$ and $q$ reads:

$$D_{KL}(q,p) = \int_\Omega q(x) \log\left(\frac{p(x;\theta)}{q(x)}\right) dx. \tag{9.66}$$

In this context the above KLd divergence can be interpreted as a measure of discrepancy between model and data. The best approximation of (9.66) within a statistical model ensemble $\{p_i(x;\theta)\}_{i=1}^L$ should be the one that minimise it. In section 3.4.1 of [29] the authors show the relation between (9.66) and (9.65) by assuming that both unknown and model distributions are not tractable in the sense that an analytic expression of (9.66) is not derivable. In our context the unknown pdf is actually the observed (true) Gaussian distribution, and all the considered model are themselves Gaussian densities functions. Therefore, we can evaluate (9.66) using its explicit algebraic form. Clearly

for Gaussian or parametric distributions in general, the KLd (9.66) is minimized when the parameters of both distribution match.
To apply the KLd for the model selection in DNA context we can estimate from MD trajectories oligomer–based mean and covariance for the rigid–base–pair degrees of freedom, meaning that we consider only inter–base–pair internal coordinates. We then call the obtained Gaussian $\rho(\cdot;\theta_{\text{obs}})$ where $\theta_{\text{obs}} \in \mathbb{R}^{n_{\text{tot}}}$. We now propose the following models for approximating the observed Gaussian all which have model mean coinciding with the oligomer–based observed one. Thus, we consider only different approximations for the covariance matrix based on different coarse–grain models and different assumptions on the local interactions. The models we have selected are:

$m_1$) Diagonal,

$m_2$) $3 \times 3$ block diagonal,

$m_3$) $6 \times 6$ block diagonal,

$m_4$) cgDNA marginal,

$m_5$) cgDNA+ marginal,

$m_6$) NNN marginal.

The last three approximations are the marginalisation, down to inter–base-pair internal coordinates, of the banded stiffness matrix estimated from time series of cgDNA internal coordinates ($m_4$), the NN banded stiffness matrix estimated form time series of cgDNA+ internal coordinates ($m_5$), and the NNN banded stiffness matrix estimated from time series of cgDNA+ internal coordinates ($m_6$). From the Palindromic data set we have computed the KLd (9.66) between $\rho(\cdot,\theta_{\text{obs}})$ and the Gaussian parametrised by one of the above approximations of the observed stiffness matrix. In the following table we have reported the results for three palindromic training library sequences, randomly chosen. In the last two rows of table (9.10) we have also report the averaged value over all the palindromic sequences and the total number of estimated parameters for each of the selected approximation.

| | $m_1$ | $m_2$ | $m_3$ | $m_4$ | $m_5$ | $m_6$ |
|---|---|---|---|---|---|---|
| $\mathcal{S}_1$ | 24.8836 | 20.7152 | 12.9734 | 5.8152 | 0.2893 | 0.1473 |
| $\mathcal{S}_7$ | 25.6658 | 21.4481 | 13.4720 | 6.2784 | 0.2203 | 0.1175 |
| $\mathcal{S}_{14}$ | 29.5059 | 25.1930 | 15.4379 | 8.0900 | 0.2254 | 0.1157 |
| avg | 26.5893 | 22.5815 | 14.2088 | 6.9528 | 0.2430 | 0.1348 |
| $n_{\text{param}}$ | 138 | 276 | 483 | 3057 | 16857 | 17649 |

Table 9.10 – In the first three rows we report the value of the KLd between the observed Gaussian, for the corresponding palindromic sequence, and different Gaussian approximatons. The details about the model $m_n$ can be found in the text. In the fourth row we reported the value of the average KLd over all the sequences in the Palindromic Library. In the last row we report the total number of parameters that have been estimated for each model.

We can observe that the value of the KLd decreases as a function of the complexity of the model. Next step is to relate the decrease reported in table 9.10 with the increase in the total number of estimated parameters. For sake of simplicity we will now consider only the averaged values reported in the last row of table 9.10 and we will define the increase of accuracy from model $m_i$ to model $m_{i+1}$, $i = 1, \ldots, 5$, by the ratio, denoted $r_1$, between the values of the averaged KLd for $m_{i+1}$ and the corresponding one for $m_i$. We will then compare the obtained results with the ratio, denoted $r_2$, between the total number of estimated parameters for model $m_i$ divided by the number of parameters in model $m_{i+1}$. In the following table we report the findings:

| | $m_1 \to m_2$ | $m_2 \to m_3$ | $m_3 \to m_4$ | $m_4 \to m_5$ | $m_5 \to m_6$ |
|---|---|---|---|---|---|
| $r_1$ | 1.1775 | 1.5893 | 2.0436 | 28.6098 | 1.8031 |
| $r_2$ | 2.1562 | 1.7500 | 6.3292 | 5.5142 | 1.0470 |

Table 9.11 – In the first row we reported he ratios $\frac{m_i}{m_{i+1}}$ which quantify the increase in accuracy of the $m_{i+1}$ model compared to the accuracy of the $m_i$ model, or they quantify the factor of decrease in the Kullback–Leibler divergence between model and data. In the second row the ratio between the number of estimated parameters for model $m_i$ and the number of estimated parameters $m_{i+1}$ which quantify the factor of augmentation of complexity of the model.

It is interesting to notice that the even if in the model $m_5$ we increase the number of parameters by a factor $5.5$ the decrease in the KLd is impressive.
The primary goal of this section is to discuss the NNN stencil and a possible parameter set which would take into account an extra coupling between adjacency of inter–base–pair coordinates. From the oligomer–based study it is totally rational to consider the NNN, stencil because of the simple fact that a little increase in the number of parameters divides the KLd by almost a factor of two. To better understand the improvement

in $m_5 \rightarrow m_6$ one can take as comparison the entries in table 9.11 corresponding to $m_3 \rightarrow m_4$ where the number of parameters increased by a factor of $6$ to get a decrease of a factor $2$.
The discussion in this section is just at the oligomer–based level because a parameter set reproducing the NNN sparsity pattern is not available, and the derivation of such a model is beyond the scope of this work. Nevertheless we provide some remarks about the feasibility of the derivation of such a model. The first step is about assumptions to be make on the sequence dependence of the extra $6 \times 6$ blocks, because the latter can influence the overall feasibility of the parameter estimation. For example assuming a trimer dependence could be problematic especially for the blocks that are toward the ends. From a feasibility point of view a suitable assumption would be to treat the extra blocks as base–dependent which will lead to just four additional $6 \times 6$ blocks in the current cgDNA+ parameter set. It seems easy, but there is still the problem of proving that any reconstructed stiffness matrix with the NNN stencil will be positive definite for any arbitrary sequence. The latter problem is closely related to the main assumption about how these extra blocks contribute to the total elastic energy. In fact a possible way of proceeding would be to consider $12 \times 12$ blocks of the form

$$K^{YXZ} = \begin{bmatrix} \delta K^{YX}_{(x,x)} & K^X \\ [K^X]^T & \delta K^{XZ}_{(x,x)} \end{bmatrix}, \tag{9.67}$$

where $K^X$ is the extra $6 \times 6$ blocks for the base $X$, and $\delta K^{YX}_{(x,x)}$ is a $6 \times 6$ matrix extracted from the inter block $(x,x)$ of $K^{YX}$. For the positive definiteness of the parameter set with the extra blocks the condition to be satisfied is the existence of the $\delta K$ blocks in (9.67) for any triplet $YXZ$ such that $K^{YXZ} > 0$. For the moment this problem has not been explored in any detail.
In conclusion in this section we have shown that at the oligomer–based level the cgDNA+ stencil carries a large amount of information when marginalised down to rigid–base–pair degrees of freedom, which implies a good approximation of the data. Moreover, the NNN stencil improves even more the accuracy of approximation of the data, which could be a good reason to dedicate more attention to the NNN sparsity pattern and a possible extension of the cgDNA+ parameter set. The choice of focusing the analysis only on the rigid–base–pair internal coordinates is motivated by the fact that for many applications it is important to have an accurate prediction of the macro structure, see for example the prediction of the tangent–tangent correlation and the related computation of the persistence lengths.

# Part IV

# Applications of cgDNA+

# Preliminary remarks

In the following chapters, we present four different applications of the cgDNA+ model. In chapter 11 we study sequence–dependent persistence length as done in [18, 45]. This application is not specific to the presence of explicit phosphate groups, but it has the purpose of showing that the overall rigidity of the best–fit cgDNA+ parameter set, trained on the palindromic training data, and of comparing it to the analysis done in section 7.3 for the cgDNA model. In contrast, the applications presented in chapters 11,12, and 13 are original and are specific to the cgDNA+ model. The main motivation is to present how to take advantage of the phosphate group rigid body configurations to study sequence–dependent mechanical properties of B–form DNA. We use the mathematical tools introduced in previous chapters in the applications. The primary objective is more methodological than scientific in the sense that we focus attention more toward the techniques rather than the scientific conclusions. We believe that a more in-depth knowledge of chemical and biological aspects are needed in order to better understand and interpret the outcomes. Moreover, we stress that the cgDNA+ model is a sequence-dependent coarse–grain model train on MD simulations. In this work we have made the conscious decision of considering only AMBER molecular dynamics simulations with the bsc1 force field for the training library. Therefore, any sequence–dependent mechanical property studied using the cgDNA+ model should only be interpreted as the coarse–grain consequence of the considered MD protocol. In particular, the cgDNA family of coarse–grain models represent a precise and detailed paradigm of mathematical modelling of double-stranded DNA which (must) assume as a starting point some specific MD simulation training library.

# 10 Study of sequence–dependent persistence lengths using cgDNA+

In this chapter we will use the cgDNA+ model to study the sequence–dependent *rigidity* of B–form DNA by computing the apparent and dynamic persistence lengths. We start by recalling the main concept related to both definitions of persistence length in the context of the cgDNA+ model, see for more detail chapter 2.
Consider an $N$ base-pair long sequence $\mathcal{S}$. Using the cgDNA+ model we can predict its ground–state $\mu(\mathcal{P},\mathcal{S}) \in \mathbb{R}^{24N-18}$ and its stiffness matrix $K(\mathcal{P},\mathcal{S}) \in \mathbb{R}^{(24N-18)\times(24N-18)}$ by using the reconstruction rules described in (9.16-9.18). We recall that the internal coordinates can be divided into inter–base–pair internal coordinates, $x = (x_1,\ldots,x_{N-1}) \in \mathbb{R}^{6(N-1)}$, and micro structure internal coordinates, $m = (m_1,\ldots,m_N) \in \mathbb{R}^{18N-12}$, in the following manner

$$\mu(\mathcal{P},\mathcal{S}) = (x,m) = (m_1,x_1,m_2,x_2,\ldots,x_{N-1},m_N) \in \mathbb{R}^{24N-18}, \tag{10.1}$$

where the micro structure $m_n \in R^{18}$ contains two base–to–phosphate degrees of freedom and intra–base–pair relative coordinates for $n = 2,\ldots,N-1$, while the first and last micro structure internal coordinates contain only only one base–to–phosphate relative coordinates and an intra–base–pair degree of freedom. Moreover the inter–base–pair internal coordinates are related to the macro structure of the DNA molecule represented by a chain of rigid bodies denoted by $\mathbf{g}(x) = (\mathbf{g}_1,\ldots,\mathbf{g}_N) \in SE(3)^N$. The relation between $x$ and $\mathbf{g}$ is given by the following recursion relation

$$g_n = g_1 \prod_{k=1}^{n-1} a(x_k), \tag{10.2}$$

where the rigid body transformation $a(x_k)$ is given by

$$a(x_k) = \begin{bmatrix} Q(u_k) & Q(u_k)^{\frac{1}{2}} v_k \\ 0 & 1 \end{bmatrix}, \tag{10.3}$$

with $x_k = (u_k, v_k) \in \mathbb{R}^6$ and

$$Q(u) = cay_\alpha(u) = I + \frac{4\alpha^2}{4\alpha^2 + |u|^2}\left(\frac{1}{\alpha}[u\times] + \frac{1}{2\alpha^2}[u\times]^2\right), \tag{10.4}$$

with $\alpha = 5$. This is the parametrisation of the rotation group used in [55, 19] where the factor 5 is used in order to introduce a better scaling between rotation and translation stiffnesses. There is a freedom in choosing the first rigid body absolute coordinates $\mathbf{g}_1$ and, in general, we chose the identity matrix $I \in \mathbb{R}^{4\times4}$.
In order to compute persistence lengths the key quantity to evaluate is the following function of the inter–base–pair internal coordinates:

$$F(x;i) = \left(R_i^T(x)R_0(x)\right)_{(3,3)} = \mathbf{t}_n(x)\cdot\mathbf{t}_0(x), \tag{10.5}$$

where $R_0(x) \in SO(3)$ is the orientation of the reference, fixed, rigid body $\mathbf{g}_0$, $R_i(x) \in SO(3)$ is the orientation of the $i$–th rigid body $\mathbf{g}_i \in SE(3)$ along the chain $\mathbf{g}$ following the reference rigid body $\mathbf{g}_0$. In (10.5) the $(\cdot,\cdot)$ notation selects one entry of the 3 by 3 matrix, which, in this application, is the $(3,3)$ entry. The $(3,3)$ entry can also be obtained as the inner product between the third column of $R_0$ and the third column of $R_i$ denoted respectively by $\mathbf{t}_0$ and $\mathbf{t}_i$. The reference rigid–body orientation $R_0$ can be chosen to be any rigid body along the chain. In practice for a DNA fragment of length $N > 10$ the reference rigid body is chosen to be at least the third rigid body ($\mathbf{g}_4$) to avoid any significant end effect.
For the sequence $\mathcal{S}$ we can now introduce the tangent–tangent correlation as the expectation of (10.5) with respect to an underlying distribution which, in our context, will be the cgDNA+ Gaussian probability density function $\rho(w;\mu,K)$ predicted by $\mathcal{P}$ for the sequence $\mathcal{S}$

$$\langle F(x;i)\rangle = \int_{\mathbb{R}^{24N-18}} \mathbf{t}_i(x)\cdot\mathbf{t}_0(x)\rho(w;\mu,K)dw, \text{ with } w = (x,m) \in \mathbb{R}^{24N-18}. \tag{10.6}$$

The integral (10.6) can be approximated numerically using the Monte Carlo method which consist in generating an ensemble of configuration $\mathbf{w} = (\mathbf{x},\mathbf{m}) = \left\{(x^{(k)}, m^{(k)})\right\}_{k=1}^{M}$ directly for the Gaussian $\rho(w;\mu,K)$ and then computing the average

$$\overline{F}(\mathbf{x},i) = \frac{1}{M}\sum_{k=1}^{M} F(x^{(k)};i). \tag{10.7}$$

Clearly the value of the average (10.7) depends upon the number of generated configuration $M$. In [18, 45] it has been shown that (10.7) converge well for values of $M > 10^5$ of configurations sampled by predicted cgDNA density functions. For more detail about the Monte Carlo method and the computation of (10.6) we refer to [18]. For the following computations we used $M = 10^6$. The apparent persistence length is then computed as the negative reciprocal of the slope of the linear fit, passing through zero, to the

observation $\big(\{i, \log \overline{F}(\mathbf{x}, i)\} | i = 0, \ldots, N\big)$. The dynamic persistence length is obtained similarly but the linear fit is made to the data $\big(\{i, \log \overline{F}(\mathbf{x}, i) - \log \widehat{\mathbf{t}}_i \cdot \widehat{\mathbf{t}}_0\} | i = 0, \ldots, N\big)$, where $\widehat{\mathbf{t}}_i \cdot \widehat{\mathbf{t}}_0$ is computed on the ground–state $\mu(\mathcal{P}, \mathcal{S})$.
We consider now the following 24 base-pair long palindromic sequence

$$\mathcal{S}_{17} = GCGACTCATAGGCCTATGAGTCGC.$$

We use the same MD protocol as in the palindromic data set for generating 3 microsecond trajectories for $\mathcal{S}_{17}$. After hydrogen bond filtering, see for instance section 3.3.1, we obtained the percentage of accepted snap shots and palindromic errors reported in table 10.1. We can observe that the palindromic errors for $\mathcal{S}_{17}$ are in the same range as obtained for the palindromic training set, see for instance tables D.1-D.2-9.7. In figure 10.1 we compare the cgDNA+ ground–states reconstructed (solid line) for $\mathcal{S}_{17}$ and the mean configuration observed from MD (dashed line). We can remark that the cgDNA+ prediction is in excellent agreement with the observation.
We compute now the tangent–tangent correlation (10.6) for the cgDNA+ predicted

| | acc. snap. | $err(\mu)$ | $err(C)$ | $\widehat{D}$ |
|---|---|---|---|---|
| $\mathcal{S}_{17}$ | 72.44 % | 1.2520 | 6.3703 | 0.0163 |

Table 10.1 – Palindromic errors for the sequence $\mathcal{S}_{17}$.

Gaussian and the MD observed Gaussian of figure 10.2. We can again observe that the cgDNA+ model predicts with good agreement both factorised and non–factorised tangent–tangent correlations functions. For sake of completeness in figure 10.2 we also show the prediction obtained with the cgDNA model trained on the palindromic data set. The cgDNA+ model performs significantly better then cgDNA.

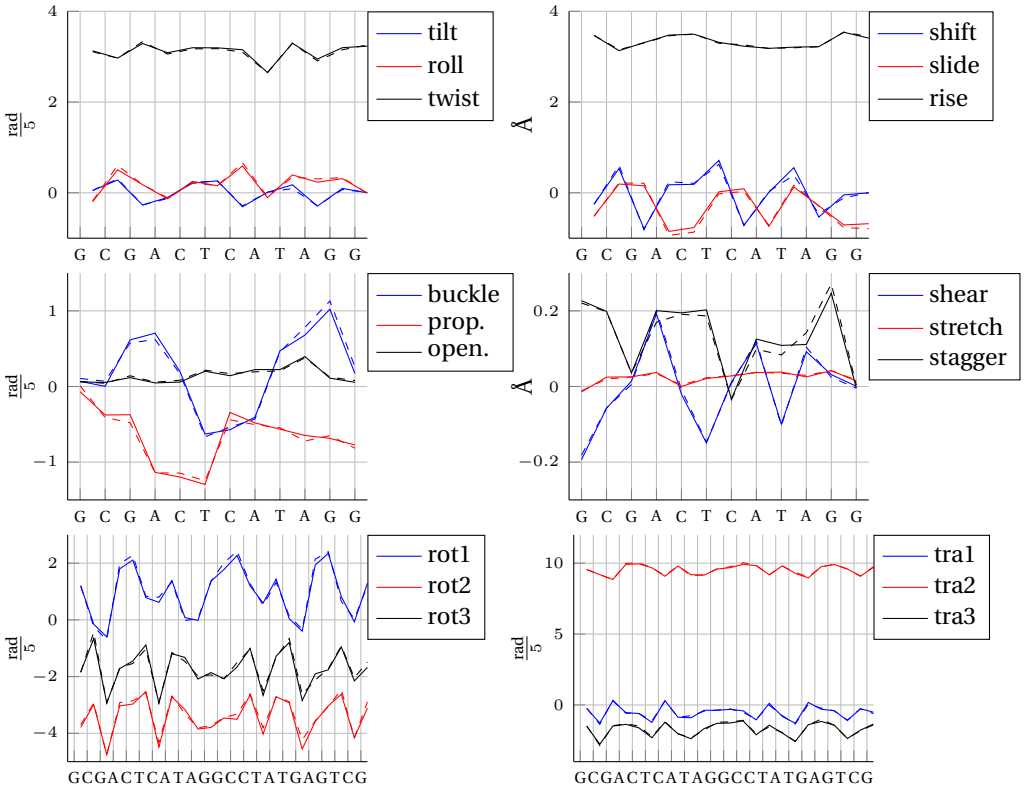



Figure 10.1 – Comparison between the ground–state components predicted by cgDNA+ (solid) and observed from MD simulation (dashed) grouped as inters, intras, and base–to–phosphate degrees of freedom from top to bottom with rotations on left and translations on right. The sequence considered is a 24 bp palindrome which is not part of the palindromic training data set simulated for 3$\mu$s. Due to the palindromic symmetry of the internal coordinates we show just the half of the sequence for the inter and intras component and just the base–to–phosphate degrees of freedom of the reading strand. The MD palindromic error is totally negligible.

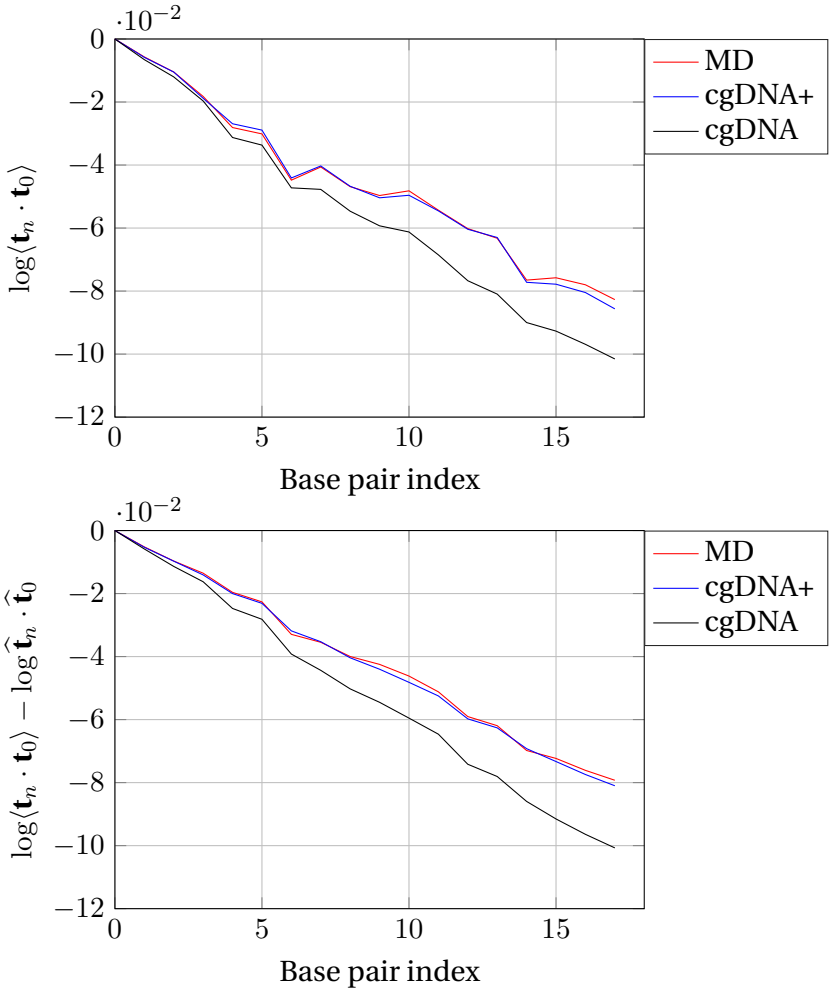



Figure 10.2 – Comparison between tangent–tangent correlation predicted by Monte Carlo simulation on cgDNA (black), cgDNA+ (blue), and computed using the Gaussian estimated from MD time series (red). The sequence considered is a 24 bp palindrome simulated for 3$\mu$s which is not part of the palindromic training data set. The first three base–pairs have been dropped to avoid end effects.

We also randomly generated an ensemble of 1 million sequences of length 220 base–pairs, by using equal probabilities for each nucleic bases $\{A, T, G, C\}$, and computed apparent and dynamic persistence lengths. In figure 10.3 we show the obtained histograms. The first remark is that, compared to the histograms obtained with cgDNA trained on the palindromic data set, see for instance figure 7.6, both spectra have shifted considerably toward the right implying that the sequence average apparent and dynamic persistence lengths increase. In the following table we report the value of the sequence–averaged apparent and dynamic persistence lengths obtained for cgDNA and cgDNA+ trained on the palindromic data set:

| Model | $\bar{\ell}_p$ [bp] | $\bar{\ell}_d$ [bp] |
|---|---|---|
| cgDNA | *158* | **174** |
| cgDNA+ | *204* | **217** |

Table 10.2 – Values of sequence–averaged apparent and dynamic persistence lengths (in base–pairs) predicted by cgDNA and cgDNA+. Both model parameter set were trained on the same palindromic data library.

Moreover in table 10.3 we compare the value of apparent and dynamic persistence lengths computed for six poly dimer sequences of length 220 base–pairs using the cgDNA and cgDNA+ models. Again one can observe that the values predicted for cgDNA+ are substantially higher compared to the one obtained for cgDNA. Between all the comparisons done in chapters 6-7 among all the considered sequences, the $A\mathrm{p}A$ poly dimer is consistently the sequence with the highest dynamic persistence length while poly $A\mathrm{p}T$ is the sequence with the lowest value of dynamic persistence length. The latter statements stay true also for the cgDNA+ model.

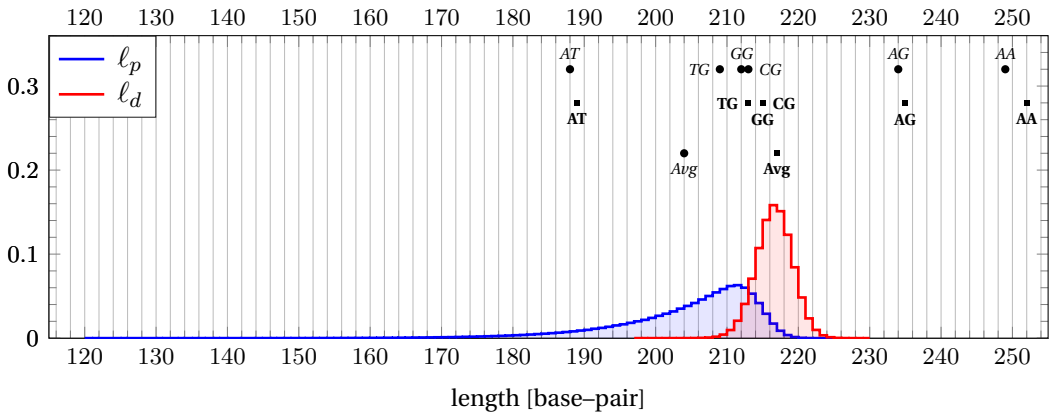



Figure 10.3 – Histograms of apparent (blue) and dynamic (red) persistence lengths computed using cgDNA+, trained on the Palindromic data set, over an ensemble of 1 million randomly generated sequences each of length 220 base pairs. We report the averaged values (Avg) of both spectras: italic font for apparent and bold font for dynamic persistence length. The values of the persistence lengths for six independent poly–dimer sequences of length 220 are alsos reported: again italic for apparent and bold for dynamic. The values of apparent persistence length is given by a circle, while dynamic is given by a square.

| Model | AA | AG | GG | TG | CG | AT |
|---|---|---|---|---|---|---|
| cgDNA | *216*/**218** | *197*/**198** | *180*/**182** | *165*/**167** | *174*/**175** | *145*/**145** |
| cgDNA+ | *249*/**252** | *234*/**235** | *212*/**215** | *209*/**213** | *213*/**215** | *188*/**189** |

Table 10.3 – Values of apparent (italic) and dynamic (bold) persistence length (in base–pairs) for six poly dimer sequences as predicted by cgDNA and cgDNA+. Both models were trained on the same palindromic data library.

In conclusion we studied the rigidity of B–form DNA, in terms of apparent and dynamic persistence lengths, using the cgDNA+ model. We concluded that the cgDNA+ model is more rigid (in the sense of apparent and dynamic persistence lengths) compared to the cgDNA model, when both models are trained on the same MD trajectories. We have also shown how accurately the cgDNA+ can predict the tangent–tangent correlation function, with and without shape factorization, for a sequence not included in the MD training library. The latter experiment suggests that the apparent and dynamic persistence lengths predicted by cgDNA+ are of the right order and, even if the conclusion is just on one sequence, it suggests that we can trust the computation of the persistence length spectra over an ensemble of 1 million sequences. Moreover, in chapter 9 we remarked that the hydrogen bond filtering, see section 3.3.1, has no impact on the overall process leading to a cgDNA+ parameter set.

# 11 Crystal structure packing forces

One of the major advantages of having the explicit rigid–body configuration of the phosphate groups is the possibility of evaluating total external loads acting on them. Because some protein-DNA binding interactions implies the formation of hydrogen bonds between phosphate groups of the DNA and the protein, the cgDNA+ model could in fact help in quantifying the magnitude of such interactions. In particular it could help in identifying which phosphate groups in a DNA–protein complex are actually bonded. Some prior work [7, 6] studied exactly the aforementioned problem, but the coarse–grain model exploited was a rigid–base–pair model [51, 35, 78] and thus only total external forces acting on the, non–physical, base–pair frames could be computed and studied. In this chapter we will present how to practically compute the total external loads acting on the phosphate groups. The following analysis can also be done at the level of nucleic bases, but we have decided to focus our attention just on the phosphate groups. Appendix F provides the explicit formulas for computing the total external loads acting on a single base, as well as all the calculations leading to it. Instead of trying to replicate the analysis of [7, 6] we want to show how to use the cgDNA+ model to compute crystal packing forces in PDB structures of naked DNA. Practically speaking we will select some crystal structure of naked DNA [70], and reconstruct the cgDNA+ ground–state and stiffness matrix for the considered sequences. The packing forces for a particular PDB structure will then be a consequence of the deviation of the shape of the PDB structure with respect to its cgDNA+ ground–state weighted by its stiffness matrix. In more detail, given a PDB structure of naked DNA we will use Curves+ [38] to extract frames for each base and each phosphate group on both strands in order to obtain a coarse–grained representation of the molecule as introduced in section 8.1. From the coarse–grained representation we then compute the cgDNA+ internal coordinates of the PDB structure, see for instance the definition (9.6). The initial step of the analysis is the comparison between the PDB internal coordinates and the cgDNA+ reconstructed ground–state, and the second step consists in the computation of the packing forces of the PBD structure.
The PDB crystal structure of, for example, a naked linear fragment of DNA is obtained

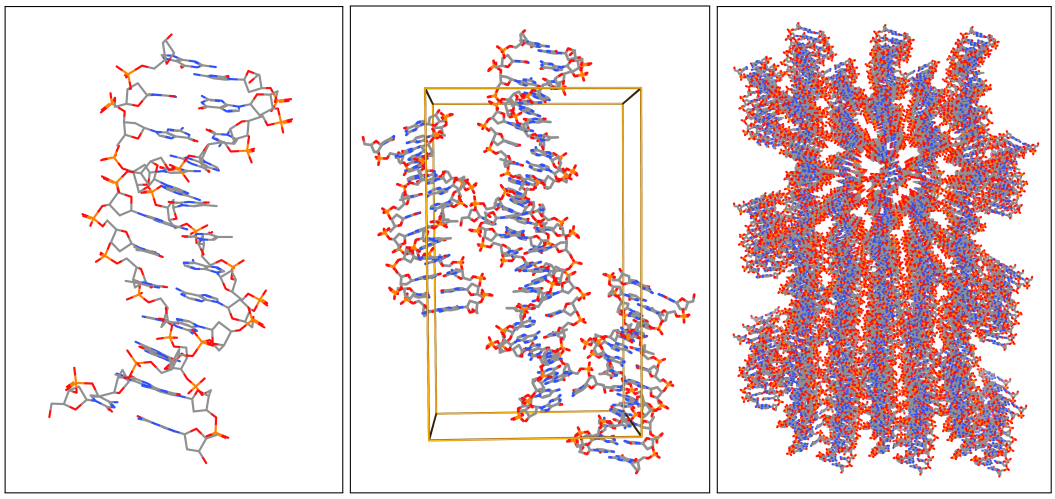


Figure 11.1 – Left: representative (mean) molecule for the *1bna* structure. Center: unit cell of the *1bna* crystal structure. Right: Partial view of the crystal.

by the X-ray diffraction method which, from a crystal formed by multiple copies of the DNA molecule, see for instance figure 11.1 and its caption, extracts the Cartesian coordinates of all the atoms of the molecule.

Before presenting the PDB structures chosen for the analysis we recall the expression for the total external load on phosphate groups in the specific case of the cgDNA+ model. Let $\mathcal{S}$ be a $N$ base–pair long DNA sequence and let $\mu(\mathcal{P},\mathcal{S})$, $K(\mathcal{P},\mathcal{S})$ be respectively its ground–state and its stiffness matrix. The related elastic energy is defined as

$$\mathbf{E}(\omega) = \frac{1}{2}(\omega - \mu(\mathcal{P},\mathcal{S})) \cdot K(\mathcal{P},\mathcal{S})(\omega - \mu(\mathcal{P},\mathcal{S})). \tag{11.1}$$

Let now $\omega = (x,m) = (m_1, x_1, m_2, \ldots, x_{N-1}, m_N) \in \mathbb{R}^{24N-18}$ be a configuration of $\mathcal{S}$ different from the ground–state. We recall that $m_i = (z_i^+, y_i, z_i^-) \in \mathbb{R}^{18}$ is the $i$–th base–pair level defining the microstructure. The total external load acting on the $m$–th phosphate group on the Watson strand for the configuration $\omega$ reads:

$$\lambda_m^+ = -\mathrm{Ad}_{\mathcal{B}_m^+}^{-T} \mathbb{L}_{z_m^+}^{-T} \partial_{z_m^+} \mathbf{E}(\omega) \tag{11.2}$$

where the components of $\lambda_m^+$ are written in the rigid–body of the corresponding phosphate group, the $m$–th base–to–phosphate coordinates is denoted by $z_m^+$,

$$z_m^+ = (\eta_m^+, \mathrm{w}_m^+) \in \mathbb{R}^6,$$
$$\mathcal{B}_m^+ = \mathcal{C}(z_m^+) = (cay_\alpha(\eta_m^+), \mathrm{w}_m^+) \in SE(3).$$

The scaled Cayley transform is defined in (2.15), in our context $\alpha = 5$ and, the matrix $\mathbb{L}_{z_m^+}$ has already been introduced in section 8.4

$$\mathbb{L}_{z_m^+} = \begin{bmatrix} \mathbb{P}_1(\eta_m^+) & 0 \\ 0 & I \end{bmatrix}, \text{ with } \mathbb{P}_1(\eta_m^+) = \frac{4\alpha^2}{4\alpha^2 + \left(\frac{|\eta_m^+|}{2}\right)^2} \left( I + \frac{1}{2\alpha}[\eta_m^+ \times] \right), \text{ with } \alpha = 5.$$

Finally the adjoint operator matrix for an element $\mathbf{g} = (R, r)$ is defined by

$$\mathrm{Ad}_{\mathbf{g}} = \begin{bmatrix} R & 0 \\ [r\times]R & R \end{bmatrix},$$

with inverse

$$\mathrm{Ad}_{\mathbf{g}}^{-1} = \mathrm{Ad}_{\mathbf{g}^{-1}} \begin{bmatrix} R^T & 0 \\ -R^T[r\times] & R^T \end{bmatrix}.$$

The two PDB crystal structures considered are for the Drew–Dickerson dodecamer: *1bna* [68] (1.9 Å res.) and *4c64* [40] (1.32 Å res.). We extracted the internal coordinates from both crystal structures and compared them to the internal coordinates reconstructed by cgDNA+. In figures 11.2 and 11.3 we compare, respectively, the phosphate degrees of freedom and the rigid–base degrees of freedom. In solid line, we show the cgDNA+ coordinates, in dashed line the one from *1bna*, and in dash–dotted the one from *4c64*. We can observe that there are many discrepancies between cgDNA+ and the crystal structure and, moreover, we can notice that between the two PDB structures there are also some important differences.

| PDB | $\mathbf{E}(\omega)$ | err($\omega$) |
|---|---|---|
| *4c64* | $4.79 \cdot 10^3$ | 25.1467 |
| *1bna* | $4.72 \cdot 10^3$ | 23.7398 |

Table 11.1 – Second column: value of the cgDNA+ energy $\mathbf{E}(\omega)$ evaluated on the internal coordinates of the PDB structures *1bna* and *4c64*. Third column: palindromic error err($\mu$) of the PDB structures.

We should also point out that the Drew–Dickerson dodecamer is a palindromic sequence, but the internal coordinates extracted from the two PDB structures do not satisfy the palindromic symmetry. In table 11.1 we report the cgDNA+ energy evaluated on the internal coordinates of the PDB structures and the palindromic error. We notice that the energy values and the palindromic error of both crystal structures are of the same order and, moreover, the palindromic error is significantly larger than the palindromic error of the mean value of the internal coordinates computed on 3$\mu$s of MD trajectories. We want to focus the attention of the reader on the fact that the

value of the energy depends upon the parameter set $\mathcal{P}$ and hence depends upon the data used to train the model. However, the palindromic error reported in table 11.1 is model independent because the palindromic property is a physical property of double stranded DNA. Therefore in the PDB structure, both DNA fragments are *frozen* in a position imposed by the crystal packing structure, see for instance figure 11.1 to see how the structure *1bna* is packed in the crystal.

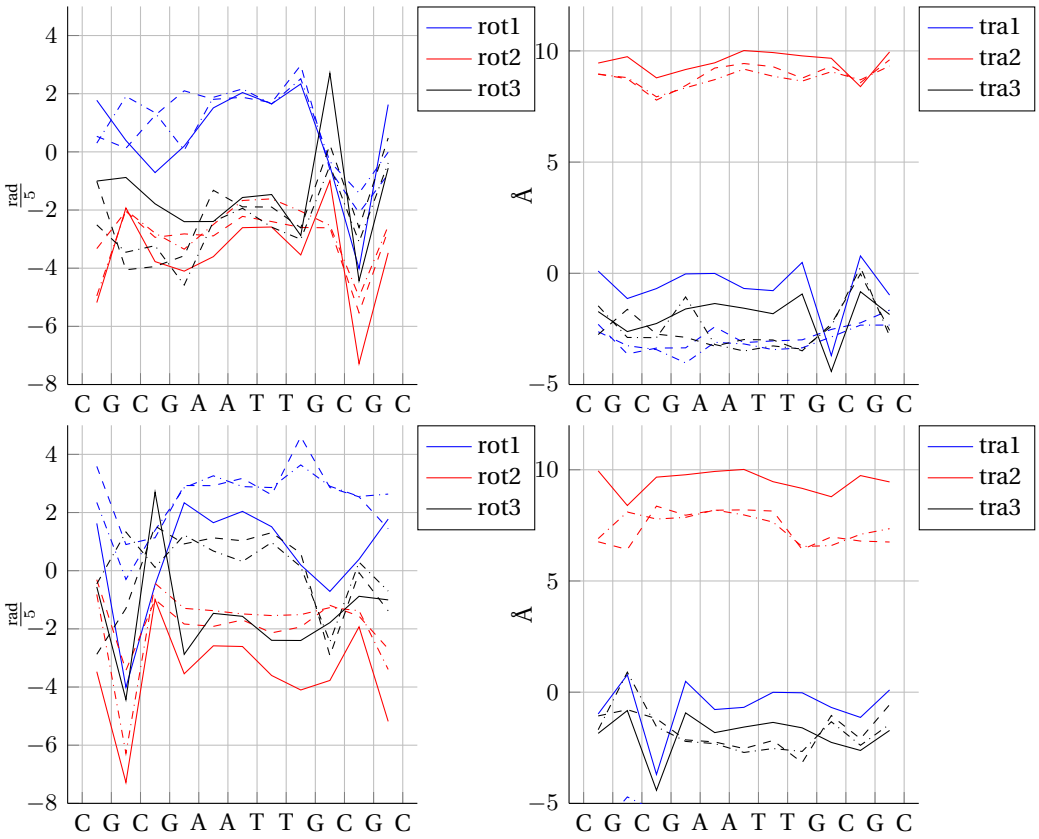



Figure 11.2 – Comparison of base–to–phosphate degrees of freedom for the Drew–Dickerson dodecamer in its ground–state as reconstructed by cgDNA+ (solid), extracted from the *1bna* (dashed), and *4c64* (dash-dotted) PDB structures.

In the context of the cgDNA+ model the crystal structures considered are not in a state of minimal energy, thus some external packing forces constrain the DNA fragments in a configuration that, moreover, need not satisfy the palindromic symmetries that are perfectly satisfied by the cgDNA+ reconstructed ground–state for $\mathcal{S}_{dd}$. Using formula (11.2) and its equivalent for the complementary strand we can compute the total external load acting on each individual phosphate group. In figures 11.4 and 11.5 we report the values of the components of the total external couple (left column) and total external force (right column) acting on the phosphate groups on the Watson strand (first row) and the Crick strand (second row). In the third row, we report the norm of the total

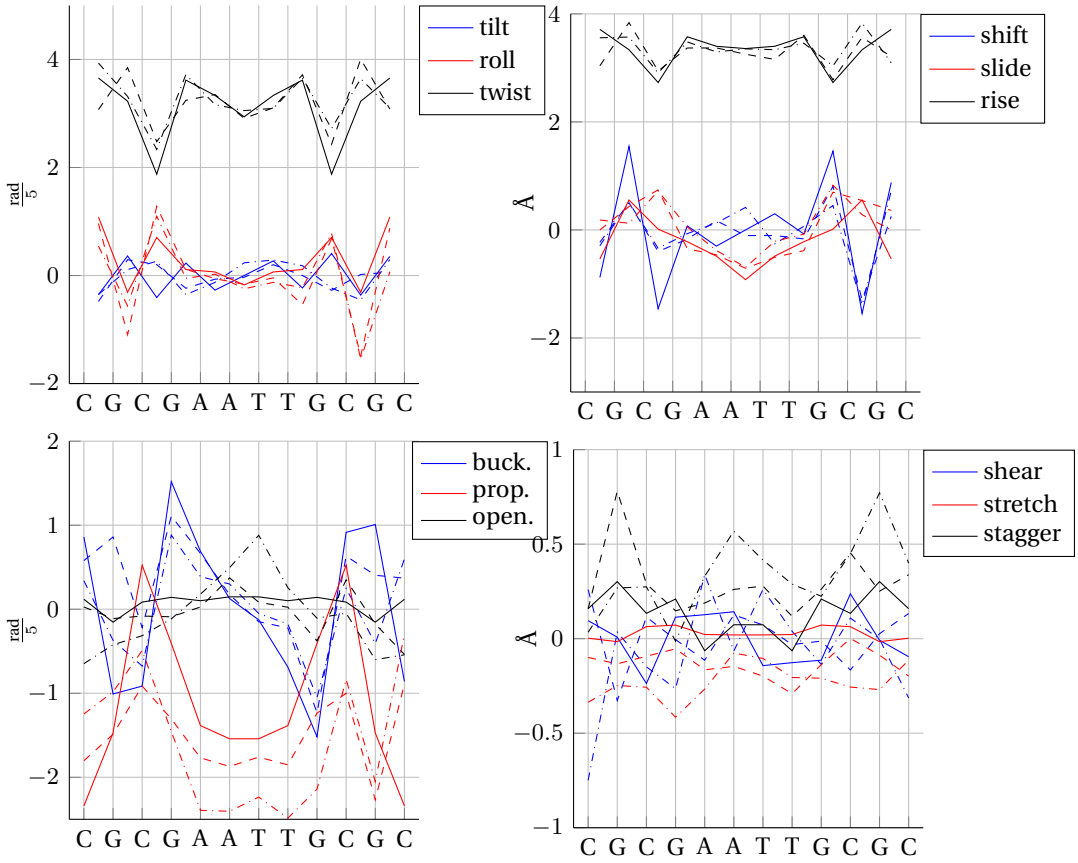



Figure 11.3 – Comparison of inter and intra degrees of freedom for the Drew–Dickerson dodecamer in its ground–state as recontructed by cgDNA+ (solid), extracted from the *1bna* (dashed), and *4c64* (dash-dotted) PDB structures.

couple and force. The interpretation of figures 11.4 and 11.5 is not straightforward, but we describe the findings. A first remark is that the crystal environment acts on the DNA fragment in a non–symmetrical way, as the signals obtained for the Watson strand are not palindromically symmetric to the signals computed for the Crick strand. That was already clear from the fact that the internal coordinates extracted from the crystal structure do not satisfy the palindromic symmetry. Another straightforward conclusion for comparing figures 11.4 and 11.5 is that the crystal packing loads in the structure *1bna* act on the DNA fragment in a different way to the ones in *1c64* as there is no clear similarity between the signals. In figure 11.6 the total external torque and force for *1bna* as a vector with origin at the corresponding phosphate group position, and the colour encodes magnitude, darker the higher the magnitude. Again an interpretation is not straightforward, especially for the couple components.

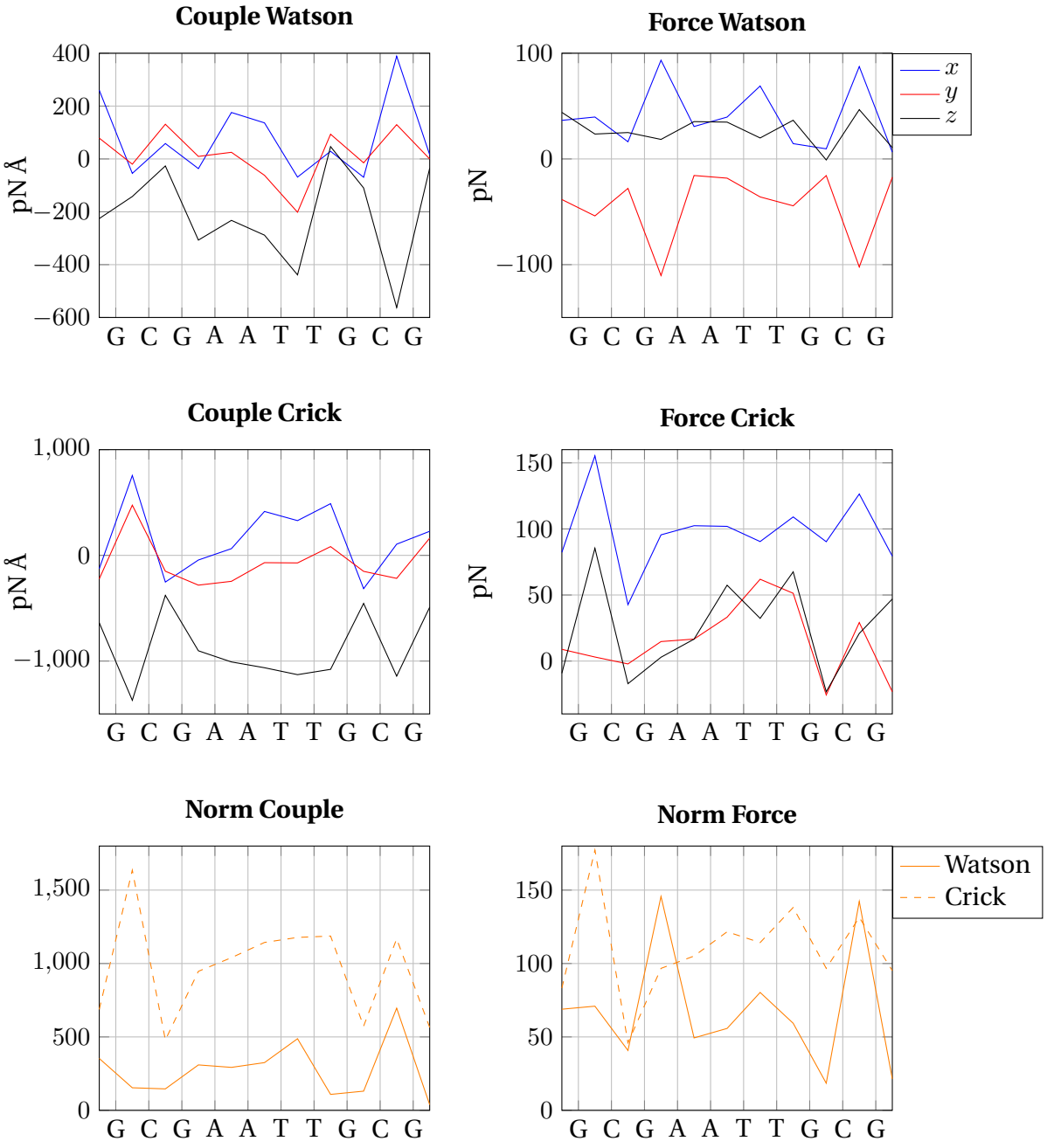



Figure 11.4 – Total external couples and force acting on each phosphate group in each strand computed for the *1bna* crystal structure.

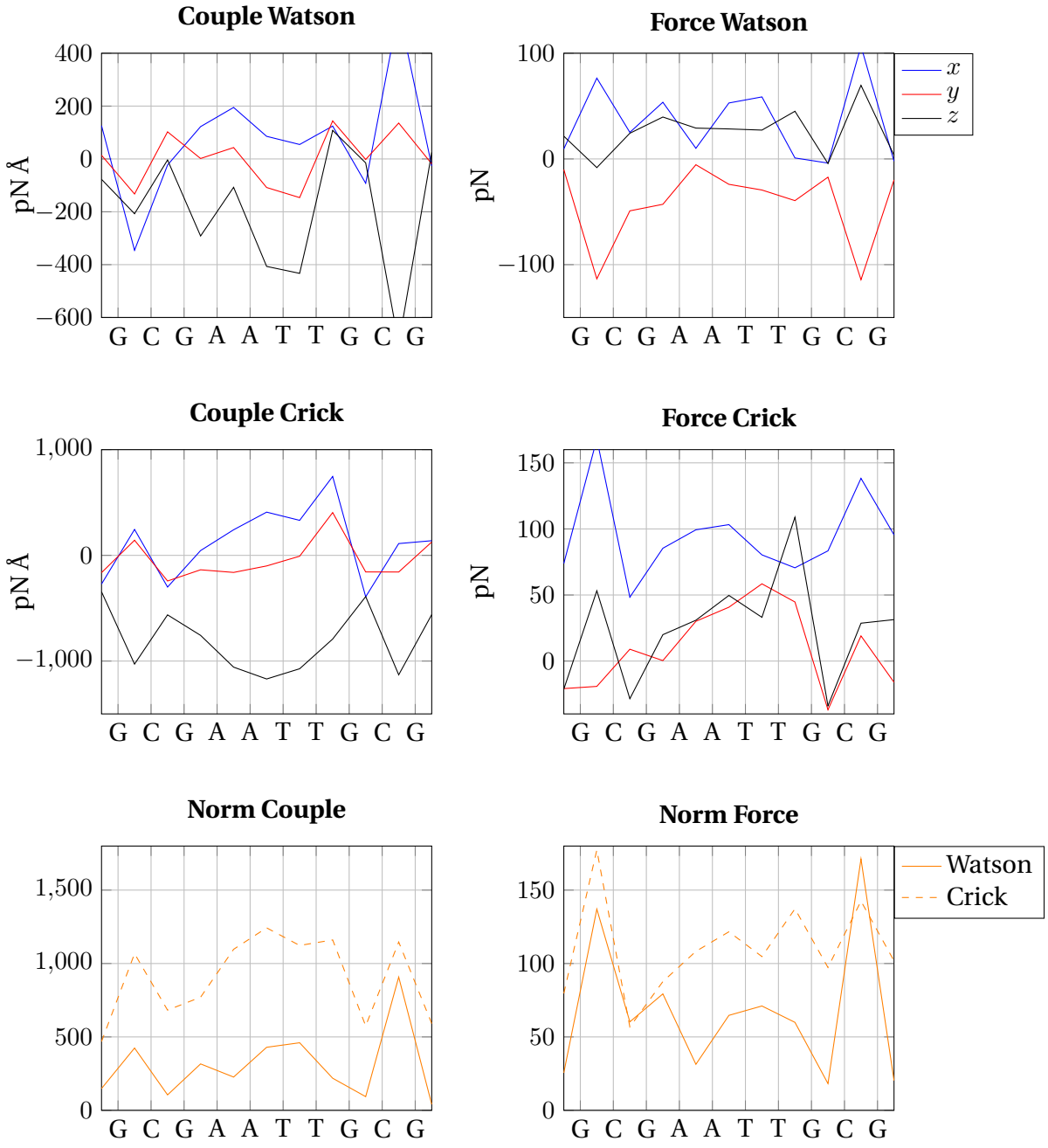



Figure 11.5 – Total external couples and forces acting on each phosphate group in each strand computed for the *4c64* crystal structure.

In conclusion, the cgDNA+ model can be used to compute the packing forces acting on a naked DNA crystal structure in order to study more precisely how the crystal act on the molecule. The interpretation of the obtained signals is not trivial as, we believe, a deeper and more robust knowledge of the X-ray diffraction method and X-ray structure analysis is needed. The example here are only provided as a proof of

concept. The only point we want to discuss in more detail is the magnitude of such external loads. In particular the magnitude seems elevated. In chapter 10 we have shown how well cgDNA+ can predict the tangent–tangent correlation for a sequence which was not part of the training library so we can reasonably claim that the cgDNA+ predicted energy is on the right scale, and thus the values obtained for the external loads are consistent. Thus, the reason behind the high values of the external load rely on the high rigidity of the cgDNA+ model, and thus on the rigidity inherited from the MD force field, and the fact that the DNA fragment in the PDB structure is kept, by the crystal, in a configuration which is very far from its (MD) ground–state. However, the values of the external load could be analysed in a better way by trying to smooth out large load values, using the uncertainty related to the position of each atom of the PDB structure. The Cartesian coordinates of the crystal structure are the mean values of a distribution with a standard deviation given by the *r–factor*. The r-factor represents, consequently, the level of uncertainty of the PDB structure. Thus, a sensible and interesting computation would be the minimisation or relaxation of the cgDNA+ energy in a subspace of configurations defined by the uncertainty on the coordinates of the atoms, meaning that the embedded atoms relate to the solution of such a problem would lie within that uncertainty. Unfortunately, such a method has not been implemented yet, and is still an active topic of research. The main purpose of this section was just to illustrate how we can use the cgDNA+ predicted ground–state and stiffness to study the external forces as extra tools for comparing two configurations of the same sequence. The two examples we have treated are packing forces in PDB crystal structures of naked DNA. As mentioned at the beginning of this chapter, we could also apply our methodology to the study of PDB structures of protein–DNA complexes. We extend this type of application to the study of external loads on averaged protein–DNA configurations computed from large scale MD simulations. With this approach, one could gain some significant insight on the DNA–protein binding dynamics for the external forces without having to deal with packing forces.

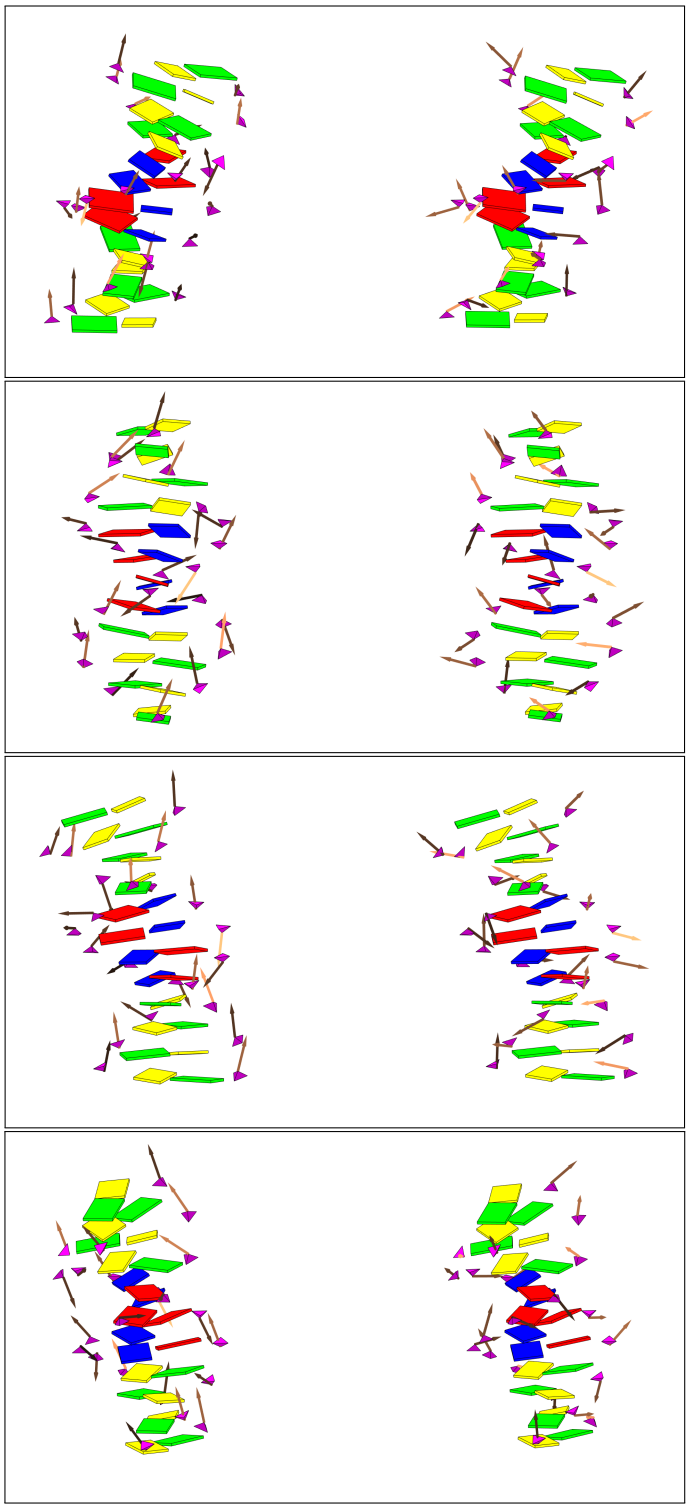


Figure 11.6 – 3D visualisation of the total torques (first column) and forces (second column) acting on each phosphate group from four different points of view. The vectors are coloured as a function of their magnitude. Higher the norm, darker the arrow.

# 12 Fine–graining of cgDNA+ ground–state backbone configurations

## 12.1 Computation of sugar configurations and sugar puckering modes

The cgDNA+ model can predict the ground–state configuration of the coarse–grain internal variable for any arbitrary sequence. From any ground–state given in internal coordinates one can then reconstruct the absolute positions and orientations of each base rigid–body and each phosphate group rigid body. From the absolute coordinates, one can then re–embed idealised atoms localised for all bases and phosphate groups to get an atomistic representation of any arbitrary sequence. However, the DNA structure will not be fully represented because the sugar group is not explicitly considered in the cgDNA+ model. We show here that, by knowing the position of each base and each phosphate group we can compute the configuration of the sugar rings. This, in this chapter, we will present a simple computation that will allow the reconstruction of the entire atomistic configuration of any arbitrary sequence starting from the atomistic configuration for the bases and the phosphate groups as predicted by cgDNA+.
Formally, the strategy we adopt to retrieve the positions of each sugar ring atoms $\mathbf{r}^s$ of a given arbitrary sequence is to minimise the force field potential given by the following function

$$U(\mathbf{r}^s;\mathbf{c}) = U_b(\mathbf{r}^s;\mathbf{c}) + U_{nb}(\mathbf{r}^s,\mathbf{c}), \tag{12.1}$$

under the assumption that some atoms $\mathbf{c}$ are fixed. The potential $U_b$ accounts for the energies coming from covalently bonded interactions while $U_{nb}$is the potential which contributions come from the non covalently bonded interactions. In section (3.1) one can find all the detail about $U_b$ and $U_{nb}$ for any given set of atom positions. In order to compute the position and configuration of the sugar ring along both strands we need to identify the group of atoms for which we will define the potential (12.1),

meaning that we need to introduce the sugar ring atoms position $\mathbf{r}^s$ and we also need to identify which atom positions $\mathbf{c}$ will be considered as fixed. In figure 12.1 we show a schematic representation of an Adenine nucleic acid base, a sugar ring, and two phosphate groups. We highlighted with a red circle all the atoms that are fixed in positions that can be predicted using the cgDNA+ model. Hence in the scheme the non–highlighted atoms are the ones that we should compute. The sugar ring atomistic composition is sequence–independent, which implies that we need to compute the Cartesian coordinates of the following atoms:

$$O'_4,\ C'_2,\ C'_3,\ C'_4,\ C'_5. \tag{12.2}$$

The next step is to identify which fixed atoms should be considered in order to cover all possible covalently and non–covalently bonded interactions of the system. Again, by examining the scheme in figure 12.1 one can convince oneself that all the red dotted atoms present in the scheme are enough in order to express all the interactions of the system. In fact the only fixed atoms of the bases we will consider are: $C'_1,\ N_9,\ C_4,\ C_8$ for purines and $C'_1,\ N_1,\ C_2,\ C_6$ for pyrimidines. On the other hand, we will consider all the atoms of both phosphate groups attached to a sugar ring.

We can now define the problem in a more precise mathematical way. Let $\mathcal{S}$ be

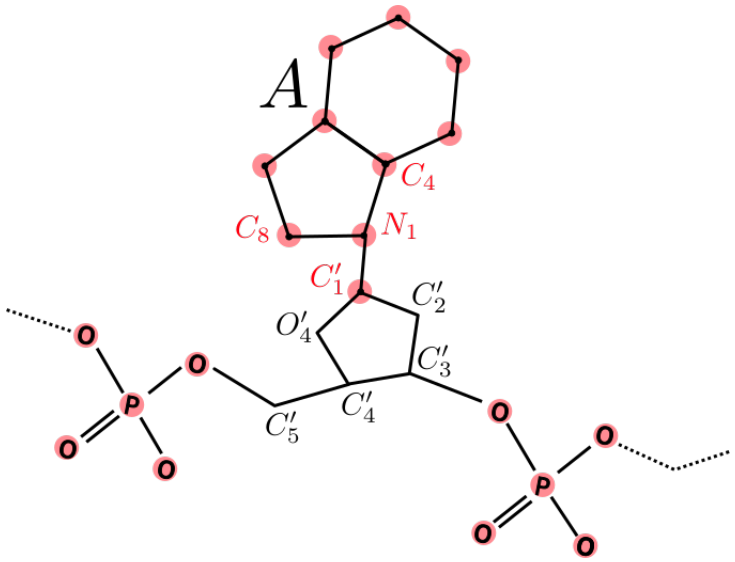



Figure 12.1 – Schematic representation of a base and the backbone composed of two phosphate groups and a sugar ring.

a $N$ base–pair long DNA sequence and $\mu(\mathcal{P},\mathcal{S}) \in \mathbb{R}^{24N-18}$ be the cgDNA+ reconstructed ground–state using a parameter set $\mathcal{P}$ and the sequence $\mathcal{S}$. Let now define by $(\mathcal{S}^+,\mathcal{S}^-) \in SE(3)^{4N-2}$ the configuration containing the absolute positions and orientations of each base and phosphate group rigid body reconstructed from the

ground–state $\mu(\mathcal{P},\mathcal{S})$ where

$$\mathcal{S}^{\pm} = (\mathbf{g}_1^{\pm}, \mathbf{p}_2^{\pm}, \mathbf{g}_2^{\pm} \ldots, \mathbf{p}_{n-1}^{\pm}, \mathbf{g}_n^{\pm}) \in SE(3)^{2N-1}. \tag{12.3}$$

We now define the ideal atoms for each base type $X \in \{A,T,G,C\}$ by $\mathbf{a}^X = (a_1^X, \ldots, a_{n_X}^X) \in \mathbb{R}^{3n_X}$ where $n_X \in \mathbb{N}$ and $n_X$ could vary between base types. The ideal atoms for the phosphate groups are sequence independent and are denoted by $\mathbf{p} = \{a_1^p, \ldots, a_{n_p}^p\} \in \mathbb{R}^{3n_p}$. The details about the atom names and atomic ideal coordinates that are considered for each base type and the phosphate group are reported in appendix A.
As said before we want to compute each sugar ring absolute positions and orientations by taking advantages of the Cartesian coordinates of the embedded ideal atoms of the bases and the phosphate group. This goal can be achieved by computing each sugar ring separately. Thus, for now on, we will focus only on the $m$-th sugar ring along the reading strand. Thus, we will only consider the following three rigid bodies: $(\mathbf{p}_m^+, \mathbf{g}_m^+, \mathbf{p}_{m+1}^+) \in SE(3)^3$. Clearly, the rigid body $\mathbf{g}_m^+$ is related to a specific base type, but for the sake of simplicity we will treat it as general, and thus we will consider an arbitrary base type $X$. Finally, we have to define the embedded atoms of the rigid–body positions and orientations which are computed as the affine transformation of the ideal atoms:

$$c_j^X = \mathbf{g}_m^+ * a_j^X := R_m^+ a_j^X + r_m^+, \forall j = 1, \ldots, n_X, \tag{12.4}$$

$$c_j^p = \mathbf{p}_m^+ * a_j^p := B_m^+ a_j^p + b_m^+, \forall j = 1, \ldots, n_p \tag{12.5}$$

$$c_{j+n_p}^p = \mathbf{p}_{m+1}^+ * a_j^p := B_{m+1}^+ a_j^p + b_{m+1}^+, \forall j = 1, \ldots, n_p \tag{12.6}$$

Finally, given an initial configuration of the sugar ring

$$\mathbf{r}_{ini}^s = (r_1^s, \ldots, r_5^s),$$

and the coordinates of the fixed atoms of the adjacent base and phosphates

$$\mathbf{c}^X = (c_1^X, \ldots, c_{n_X}^X),$$
$$\mathbf{c}^p = (c_1^p, \ldots, c_{n_p}^p),$$

we solve

$$\min_{\mathbf{r}^s \in \mathbb{R}^{15}} U(\mathbf{r}^s; \mathbf{c}^X, \mathbf{c}^p), \tag{12.7}$$

where $U(\cdot; \mathbf{c}^X, \mathbf{c}^p)$ is the force field potential defined in (12.1) with the Cartesian coordinates of the atoms of the base and the phosphate groups being fixed. Thus the only unknowns in the system are the Cartesian coordinates of the sugar ring atoms (12.2). The first example we show in figure 12.2 has been computed from the cgDNA+ reconstructed ground–state of the palindromic sequence $\mathcal{S}_1$. From a pure visualisation point of view, the resulting full atomistic representation of sequence $\mathcal{S}_1$ looks astonishingly

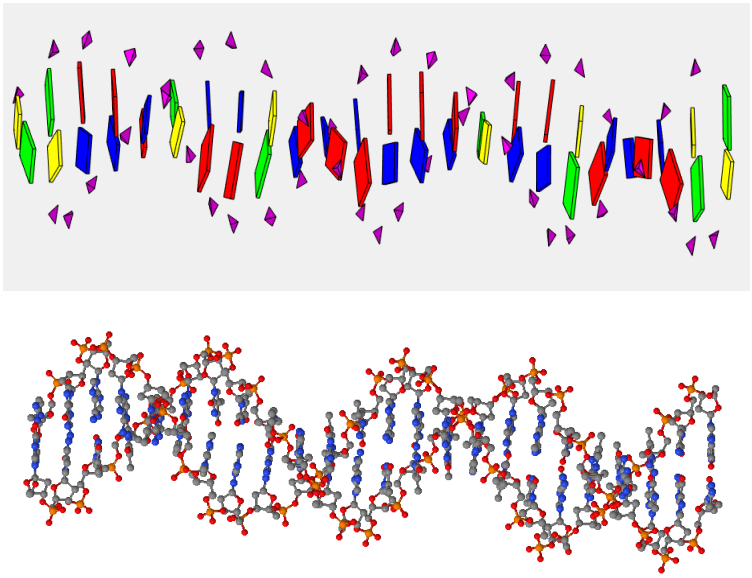


Figure 12.2 – In the first row, we show the double interacting strand representation of the coarse–grain ground–state predicted by cgDNA+ for the sequence $\mathcal{S}_1$ of the palindromic library. In the second row we show the atomistic representation where the base and the phosphate group atoms are just the embeddings of the corresponding idealise atoms and the sugar rings were computed by solving the problem (12.7)

like actual DNA. A closer look at the atomistic structure reveals that no sugar ring is planar. In fact, a five atoms covalently bonded cannot have a minimum energy configuration that is planar [3].

In chapter 11 we have compared the coarse–grain shape extracted from the PDB structures *1bna* and *4c64* of the Drew–Dickerson dodecamer $\mathcal{S}_{dd} = CGCGAATTCGCG$. Now we can also compare the sugar ring configurations between the two X–ray structures and the sugar configurations obtained by the cgDNA+ prediction and solving the optimization problem (12.7).

Before moving to the comparison, we briefly recall a few concepts about sugar rings already introduced in section 1.2. A five–member ring system composed of a oxygen atom and four carbon atoms singly covalently bonded is in general not planar [3]. In fact a sugar ring puckers in one of two main forms called envelope or twist. The envelope form is characterised by four atoms lying in the same plane, with the last one out of the plane of about 0.5Å. While in the twisted form two adjacent atoms have opposite positions to the plane defined by the other three. In figure 1.2 (section 1.2) we showed an example for both forms of sugar ring puckering. The classification of all the sugar ring modes can be done using the so–called *pseudorotation cycle*, see for instance figure 1.3 (section 1.2), and in particular, the value of the phase of the pseudorotation spans every possible conformation of the sugar ring. We recall that in the context of DNA the sugar ring configurations have preferred puckering modes

due to the potential energy of the whole DNA structure. Let us recall the endocyclic sugar torsion angles $v_0, \ldots, v_4$ whose definitions are reported in the table 12.1. The two possible definitions of the pseudorotation phase angle $P$ are

$$\tan P = \frac{\sum_{i=1}^{5} -\theta_i \sin(\frac{4}{5}\pi(i-1))}{\sum_{i=1}^{5} \theta_i \cos(\frac{4}{5}\pi(i-1))}, \tag{12.8}$$

$$\tan P = \frac{(v_4 + v_1) - (v_3 + v_0)}{2v_2\left(\sin\left(\frac{1}{5}\pi\right) + \sin\left(\frac{2}{5}\pi\right)\right)}. \tag{12.9}$$

It should be mentioned that the two different definitions for the pseudorotation phase angle gives slightly different result. Definition (12.9) appears in [77] while (12.8) can be founded in [3]. Definition (12.9) is implemented in the 3DNA software [41] while (12.8) is implemented in the Curves+ software [38]. In this work we have used Curves+ for fitting both bases and phosphate groups to atomic coordinates and thus, for consistency, we will also use Curves+ for the computation of the pseudorotation phase angles for both PDB structures *1bna* and *4c64*. In table (12.1) we report the definition of the dihedral angles in (12.8) and (12.9).

Table 12.2 provides results for the sugar rings on the reading strand. We first observe

| Name | Definition |
|---|---|
| $v_0$ or $\theta_4$ | $C'_4 - O'_4 - C'_1 - C'_2$ |
| $v_1$ or $\theta_5$ | $O'_4 - C'_1 - C'_2 - C'_3$ |
| $v_2$ or $\theta_1$ | $C'_1 - C'_2 - C'_3 - C'_4$ |
| $v_3$ or $\theta_2$ | $C'_2 - C'_3 - C'_4 - O'_4$ |
| $v_4$ or $\theta_3$ | $C'_3 - C'_4 - O'_4 - C'_1$ |

Table 12.1 – Name and definition of the endocyclic sugar torsion angles.

that between the two different PDB structures there is no actual agreement in terms of values of the pseudorotation phase angles in all but one sugar ring. One of the main reason could lie in the difference between the resolution for which the structure has been derived: $1.9$Å for *1bna* and $1.32$ Å for *4c64*. For the cgDNA+ predicted sugar rings puckering we can observe that again there is no agreement with the PDB structures data, but as a consistency test, we can observe that the $C'_2$–endo mode is the most represented which is in line with the statement that the latter is the preferred puckering mode for sugar rings in nucleotides.

| Sugar nbr | $P$ | puck. mode | Sugar nbr | $P$ | puck. mode |
|---|---|---|---|---|---|
| | 161.6 | $C_2'$–endo | | 101.5 | $O_4'$–endo |
| 1 | 169.0 | $C_2'$–endo | 7 | 119.8 | $C_1'$–exo |
| | 122.4 | $C_1'$–exo | | 143.3 | $C_1'$–exo |
| | 139.8 | $C_1'$–exo | | 115.9 | $C_1'$–exo |
| 2 | 159.7 | $C_2'$–endo | 8 | 115.9 | $C_1'$–exo |
| | 148.1 | $C_2'$–endo | | 132.6 | $C_1'$–exo |
| | 92.8 | $O_4'$–endo | | 140.7 | $C_1'$–exo |
| 3 | 52.2 | $C_4'$–exo | 9 | 146.6 | $C_2'$–endo |
| | 147.5 | $C_2'$–endo | | 181.0 | $C_3'$–exo |
| | 166.6 | $C_2'$–endo | | 146.5 | $C_2'$–endo |
| 4 | 165.5 | $C_2'$–endo | 10 | 143.0 | $C_1'$–exo |
| | 124.3 | $C_1'$–exo | | 175.4 | $C_2'$–endo |
| | 128.8 | $C_1'$–exo | | 147.7 | $C_2'$–endo |
| 5 | 151.3 | $C_2'$–endo | 11 | 166.2 | $C_2'$–endo |
| | 108.5 | $C_1'$–exo | | 161.5 | $C_2'$–endo |
| | 127.3 | $C_1'$–exo | | 114.1 | $C_1'$–exo |
| 6 | 131.7 | $C_1'$–exo | 12 | 38.2 | $O_4'$–endo |
| | 145.0 | $C_2'$–endo | | 180.0 | $C_2'$–endo |

Table 12.2 – For each sugar group on the reading strand we report the pseudorotation phase angle $P$ and the corresponding puckering modes computed from three different set of sugar ring data (from top to bottom): *1bna*, *4c64*, and cgDNA+ reconstruction.

## 12.2 A sequence context study of BI–BII backbone conformations

In the previous section, we have shown how to compute the sugar rings on a given cgDNA+ predicted ground–state. In this section, we take advantage of this fine graining property of cgDNA+ reconstructions to study how some torsional angles behave as a function of the sequence context. In particular we are interested in the torsional angles $\varepsilon$ and $\zeta$ that are defined respectively by the backbone atoms $C_4'$–$C_3'$–$O_3'$–$P$ and $C_3'$–$O_3'$–$P$–$O_5'$(upstream). We recall that the torsional angles are the dihedral angles defined by the related four atoms. The two angles $\varepsilon$ and $\zeta$ are important quantities to compute the so–called BI–BII backbone conformations. In figure 12.3 we show an example of both states.

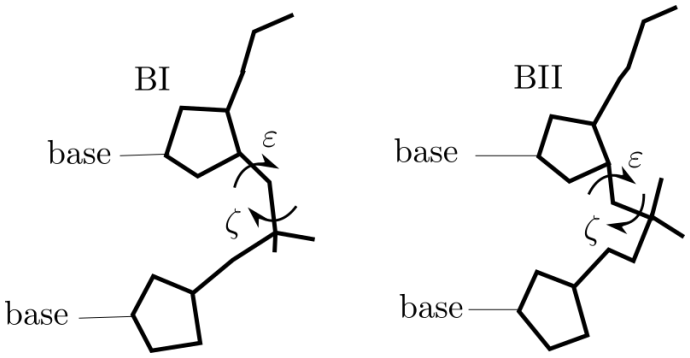



Figure 12.3 – Example of BI and BII states.

The characterisation of the BI–BII state in terms of the two torsional angles $\varepsilon$ and $\zeta$ is quite simple, namely:

$$\varepsilon - \zeta < 0^\circ \rightarrow \text{ BI state ,}$$
$$\varepsilon - \zeta > 0^\circ \rightarrow \text{ BII state .}$$

Next we consider all the possible hexamer sub–sequences, completed to 18mers by adding on both sides four random bases and $G$p$C$ ends. We then reconstruct all the cgDNA+ ground states and compute only the Watson sugar ring for the central dimer. Once the sugar ring atoms are available we compute the two backbone angles $\varepsilon$ and $\zeta$ for each sequence. We recall that the base and phosphate atoms are the embeddings of idealised atom coordinates reported in the tables A.2 and A.3 in appendix A. We then compute the difference $\varepsilon - \zeta$, and identified the related backbone states. We start by first show the percentage (over sequence) of BI–BII states divided into 16 different dimers steps in figure 12.4.

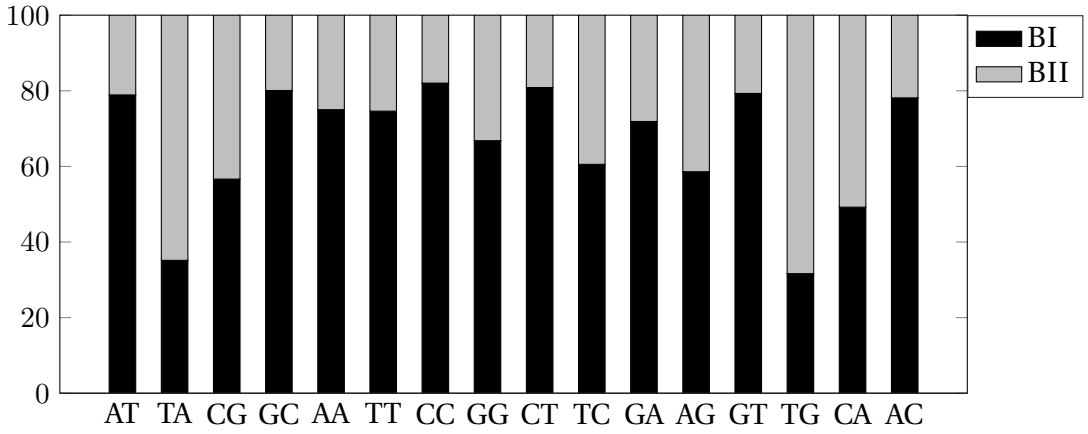



Figure 12.4 – Percentage of BI–BII states computed for for each centred dimer step in alle the possible hexamer contexts.

It is interesting to note that the BI states (darker grey) are the most represented states between all the dimers. We continue the analysis by re–arranging the computed backbone states using the purine–pyrimidine (R–Y) alphabet for the central dimers and the flanking bases. In figure 12.5 we show two four–by–four matrices whose entries are classified as follow: the columns correspond to a specific central dimer in the R–Y alphabet while the row is the flanking bases also expressed in the R–Y alphabet. On the left–hand side we show the result for the BI state while on the right–hand side we visualise the corresponding data for the BII state. By looking at the results from the purine–pyrimidine perspective, one can remark some combinations of central dimers and flanking bases that privilege one specific state. For example, the YR central dimer with R..Y flanking bases is BI state specific while for the same central dimer, but with Y..R flanking sequence the tendency is to be in the BII configuration. The only two combinations of the central dimer that are mostly BII have YR central dimers. It is also interesting to notice that the upstream flanking base seems to be more relevant for the characterisation of the percentage of BI or BII configuration.

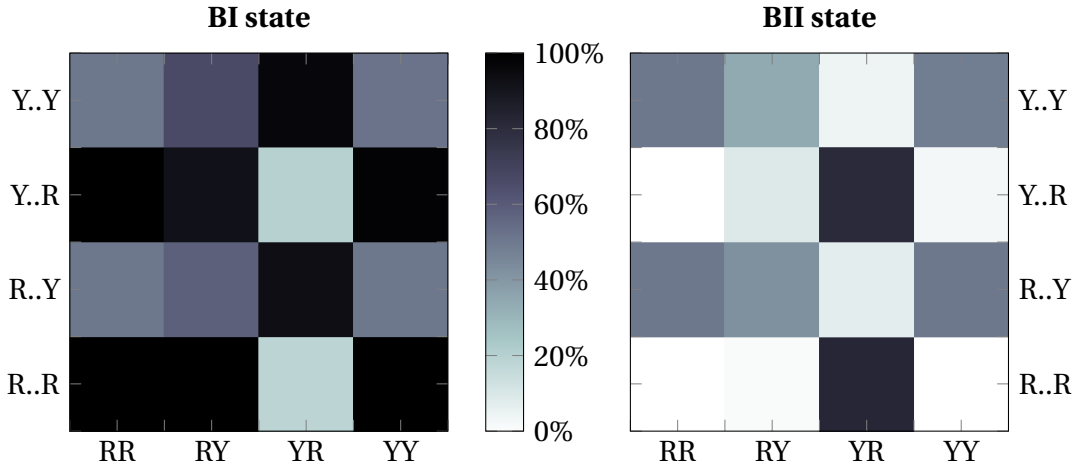



Figure 12.5 – Percentage of BI–BII states in the purine–pyrimidine (R–Y) alphabet for the central dimers and the flanking bases. On the left we show the BI state and on the right the BII state.

In conclusion, thanks to sugar ring reconstruction we study BI–BII states as a function of the dimer and the flanking sequence. The results reveal an overall preference for the BI state, but for the YR central dimer and upstream R bases, the BII configuration is preferred.

# 13 Groove widths prediction

In this chapter, we will show how the cgDNA+ model can be used for studying the groove widths in DNA ground–state and, in particular, how to measure the width of the minor and major grooves on the ground–state of any arbitrary sequence. We first define the minor and major grooves, and then we discuss two approaches to groove width computations using the cgDNA+ model.

The chemical structure of double–stranded B–form DNA is a double helix structure in which the distance between backbone varies along the chain consistently and forms two different grooves: the major groove which is the side of the helix in which the backbones are far apart, and the minor groove which, reversely, is the side of the helix where the strands are closer. In figure 13.1, left, we show a double–stranded B–form DNA, and we indicate the two grooves. One can notice that the major and the minor grooves form two distinct areas that twist around the DNA oligomer. The minor and major grooves are believed to play an important role in DNA–protein binding readout but a complete understanding of this relationship is not yet available [9, 30, 47, 57, 62, 63, 69]. What has been observed is that in DNA–protein complexes the minor and major groove behave oppositely. In fact, in complexes, the minor groove width tend to vary while the major groove width tend to stay closer to the value of the major groove of unbounded DNA [47, 57, 62, 63].

In the context of cgDNA+, one can observe the grooves, for any arbitrary sequence, by using the same method proposed in the Curves+ article [38] which we now summarise. Let $\mathcal{S}$ be a $N$ base–pairs long sequence, and $\mu(\mathcal{P},\mathcal{S})$ its cgDNA+ predicted ground–state. Recall that there exists an explicit relationship between the internal coordinates $\mu$ and the absolute coordinates of the double–stranded representation denoted $(\mathcal{S}^+,\mathcal{S}^-)$. More precisely, we transform the internal coordinates

$$\mu = (m_1, x_1, m_2, \ldots, x_{N-1}, m_n),$$

where $x_n \in \mathbb{R}^6$ are the inter, and $m_n = (z_n^+, y_n, z_n^-) \in \mathbb{R}^{18}$ are the relative coordinates for the microstructure, into the tetra–chain representation $(\mathbf{g},\mathcal{M})$, where

$\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_N) \in SE(3)^N$ is the macrostructure, $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_N)$ are the microstructure, and $\mathcal{M}_n = (\mathcal{B}_n^+, \mathcal{P}_n, \mathcal{B}_n^-) \in SE(3)^3$ is the $n$th base–pair level. The following transformations rules are used:

$$\begin{aligned}
\mathbf{g}_n &= \prod_{k=1}^{n-1} (cay_\alpha(u_n), cay_\alpha(\eta_n)^{\frac{1}{2}} v_n) \in SE(3), \\
\mathcal{P}_n &\equiv \mathcal{P}_n(y_n) = (cay_\alpha(\eta_n), cay_\alpha(\eta_n)^{\frac{1}{2}} \mathbf{w}_n) \in SE(3) \\
\mathcal{B}_n^\pm &= (cay_\alpha(\eta_n^\pm), \mathbf{w}_n^\pm) \in SE(3)
\end{aligned}$$

where $\alpha = 5$, and the scaled Cayley transformation $cay_\alpha$ has been defined in (2.15). We then compute the absolute coordinates of each phosphate group in the following manner

$$\mathbf{p}_n^\pm = \mathbf{g}_n \mathcal{P}_n^{\pm H} B_n^\pm = ([R_n^p]^\pm, [r_n^p]^\pm) \in SE(3),$$

with

$$\mathcal{P}_n^{\pm H} = \begin{bmatrix} cay_\alpha(\eta_n)^{\pm\frac{1}{2}} & \pm\frac{1}{2}\mathbf{w}_n \\ 0 & 1 \end{bmatrix}$$

being the half rigid body motion between the base–pair frames and the base frame. It is important to understand that for all $n = 1, \ldots, N$, $\mathcal{P}_n^{\pm\frac{1}{2}} \neq \mathcal{P}_n^{\pm H}$. The latter statement has some important consequences in, for example, the computation of the total external forces acting on every single base, see for instance the computations in the appendix F.

In summary, for a $N$ base–pair long sequence $\mathcal{S}$ we reconstruct its cgDNA+ ground–states expressed in internal coordinates $\mu(\mathcal{P}, \mathcal{S})$. From the internal coordinates we compute the absolute coordinates of the phosphate groups given by $\mathbf{p}_n^\pm = ([R_n^p]^\pm, [r_n^p]^\pm) \in SE(3)$ for all $n = 1, \ldots, N$. For groove width analysis we just need the positions of the phosphate groups and thus we introduce the notation for the phosphate group origins on both strands: $\mathcal{O}^\pm = ([r_1^p]^\pm, \ldots, [r_{N-1}^p]^\pm) \in \mathbb{R}^{3(N-1)}$. Now by following the methodology proposed in [38] we compute the cubic spline interpolation of both sets of points: $\mathcal{O}^+$ and $\mathcal{O}^-$. In figure 13.1, right, we show the cubic spline interpolation obtained for the cgDNA+ predicted ground–state of the Drew–Dickerson dodecamer. The next step is to select an equal number of equidistributed points on both splines and compute all the pairwise distances between the two set of points. We compute a distance matrix, denoted by $D_m \in \mathbb{R}^{m\times m}$ whose entries $(i, j)$ are the Euclidean distance between the $i$–th point on the reading strand spline and the $j$–th point on the complementary one. Finally the distance matrix can be visualize as the surface generated by the points $(i, j, [D_m]_{(i,j)}) \in \mathbb{R}^3$.

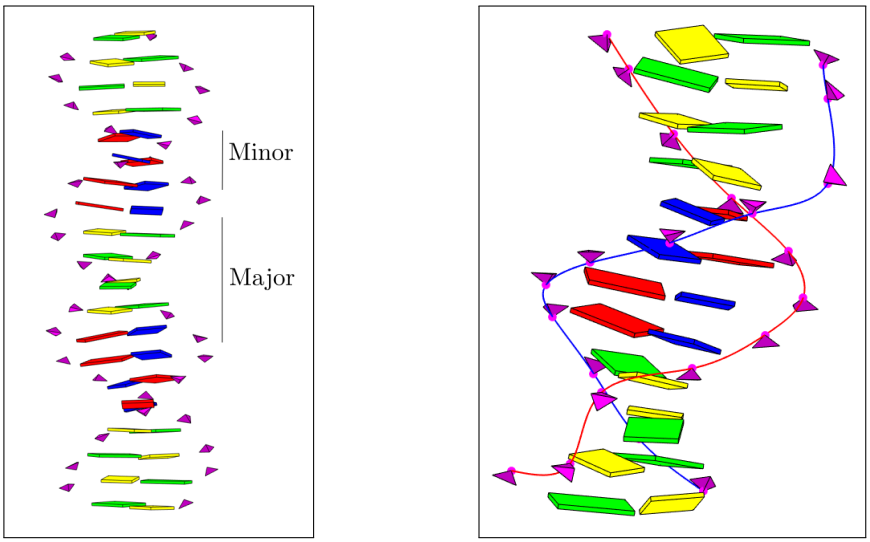



Figure 13.1 – Three dimensional view of the ground–state for the Drew-Dickerson dodecamer with cubic spline interpolaton of the phosphate group positions. The blue spline interpolates the phosphate positions on the reading strand while the red spline interpolates the complementary one. The magenta dots locate each phosphate positions.

Now consider two sequence–specific sequences: $S_1 = CGCGAATTCGCG$ and $S_2 = TATAGGCCTATA$. We study their groove widths by using the method mentioned above. In figures 13.2 and 13.3 we respectively show the outcome of the groove analysis for sequence $\mathcal{S}_1$ and $\mathcal{S}_2$. The contour lines indicate the distance between two points on different backbones. One can observe first that the two plots are qualitatively quite different. Secondly, we can remark that each plot has two local minima which are located at the contour line levels 10 and 16 (Å) for $\mathcal{S}_1$ and levels 11 and 18 (Å) for $\mathcal{S}_2$. We can interpret these two numbers as the minor and major groove width of these two sequences. Thus using cgDNA+ we can easily have the information of the minor and major groove width by first interpolating the phosphate group positions with a cubic spline, one spline for each backbone, secondly, we compute the pairwise distances between points equidistributed along both splines. Finally, we visualise the surface generated by the obtained distance matrix which we can infer the values of both grooves width.

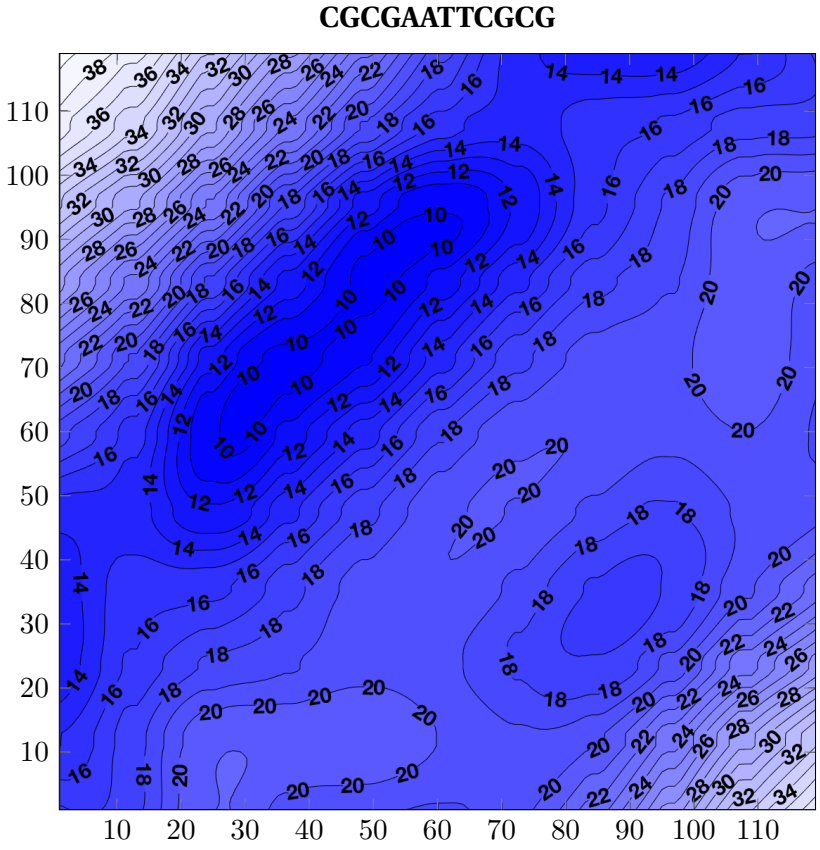



Figure 13.2 – Visualization of grooves width for $\mathcal{S}_1 = CGCGAATTCGCG$ sequence. The contour lines indicate the value of the distance between a point on a strand and a point on the complementary one. We recall that the strand are approximate as a cubic spline passing through the origins of each phosphate group on that strand.

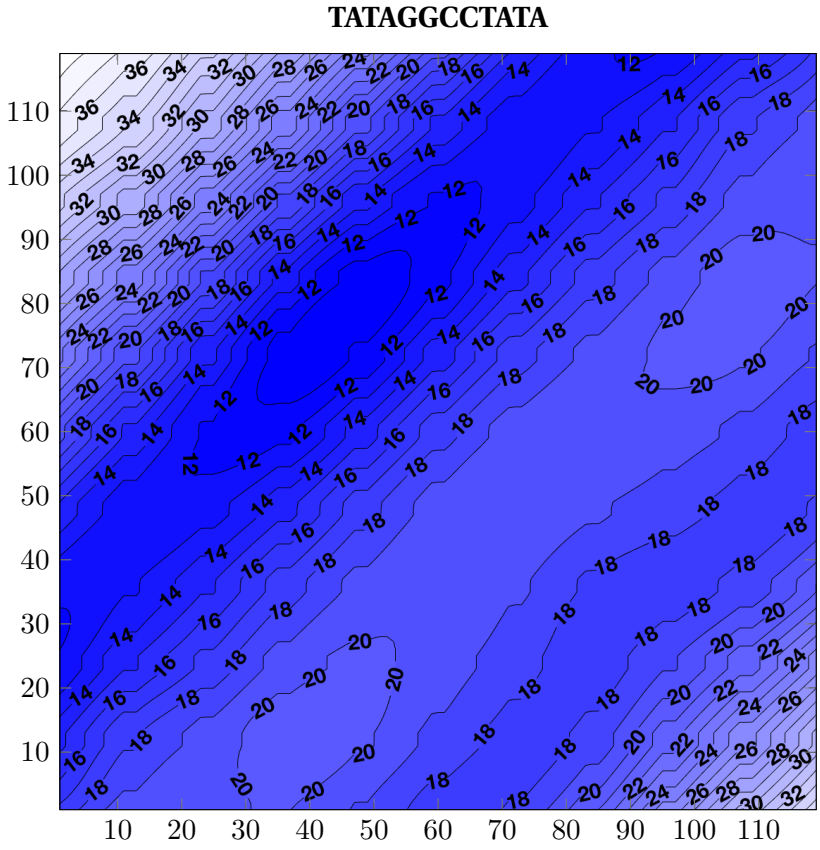



Figure 13.3 – Visualization of grooves width for $\mathcal{S}_2 = TATAGGCCTATA$ sequence. The contour lines indicate the value of the distance between a point on a strand and a point on the complementary one. We recall that the strand are approximate as a cubic spline passing through the origins of each phosphate group on that strand.

We can simplify the computation of minor and major groove by considering only the phosphate group positions on both strand and all the associated pairwise distances. With this approach, the computations are faster allowing the study of many more sequences. In fact in what follows we want to gain some insight on the sequence dependence of the minor and major groove using the cgDNA+ model. More precisely, we want to study the minor and major groove by fixing a dimer step and by changing the surrounding sequence. We thus consider all the 1,048,576 decamers (10 base–pair long sequences) which we then divided by selecting the sequences sharing the same central dimer. We thus divided the decamer into 16 sets of 65,536 sequence. Each decamer is then completed by adding four random bases on each side, and GC ends to obtain the 16 sets of 65536 sequences of length 22 base–pairs. Using cgDNA+, we have then reconstructed all the ground–states and computed all the phosphate group positions. Then we started from the phosphate group in the central dimer and computed the pairwise distances between the phosphate groups on the opposite strand. Naturally, we can divide the phosphate on the other strand into two groups: upstream phosphate

and downstream phosphate. The upstream phosphates are all the phosphates in the 3'–5' direction on the complementary strand while the downstream phosphates are the one in the 5'–3' direction, again, on the complementary strand. Thus, the pairwise distances between the central phosphate on the reading strand and the phosphate on the other strand are divided into upstream distances and downstream distances. From each group, we select the minimum to get the *discrete* major and minor groove width. The computations where carried on a standard notebook and took around 1 hour. We show the result in figure 13.4. It is interesting to see how the minor groove varies in values: between 10–14 Å for $G\mathrm{p}A$, and 10–12 Å for $A\mathrm{p}T$. Also, the shape of the histograms vary considerably: quite peaked for $T\mathrm{p}G$ and bimodal for $T\mathrm{p}T$. On the other hand, the major groove tends to be quite peaked and locate around 18 Å, even if in some case the values range between 16 and 18 Å, see for instance the $G\mathrm{p}T$ case. We can now study the results by collecting the data in another way. Namely, we can select the dimer and the flanking bases to see how the histograms change. In figure 13.5 we show the result for four selected central dimers $C\mathrm{p}T$, $T\mathrm{p}A$, $G\mathrm{p}C$, and $A\mathrm{p}G$. We can remark that the shape of the histograms for the minor groove change substantially revealing multiple behaviours. A careful reader will remark that the chosen central dimers representatives of the four possible choices in the purine–pyrimidine alphabet. We thus rearranged the data using the purine–pyrimidine (R–Y) alphabet for both the central dimers and the flanking bases to get the figure 13.6. We can notice a few interesting things. The first is the change in the shape of the major groove histograms which became considerably less peaked and in some cases a bimodality shows up. Secondly, for the minor groove, the histogram shapes vary again from a clear bimodality to a single peak. The high variability of both grooves leads to the conclusion that characterizing them using the R–Y alphabet is not sufficient in order to understand sequence dependency. Some possible strategies would be to extend the classification to the flanking dimer or to considered separately the sequence dependence of the minor and major groove in order to better explore the hidden structure. It would also be interesting to study the correlation between the grooves and the BI–BII states that may play an important role [46, 16, 72].

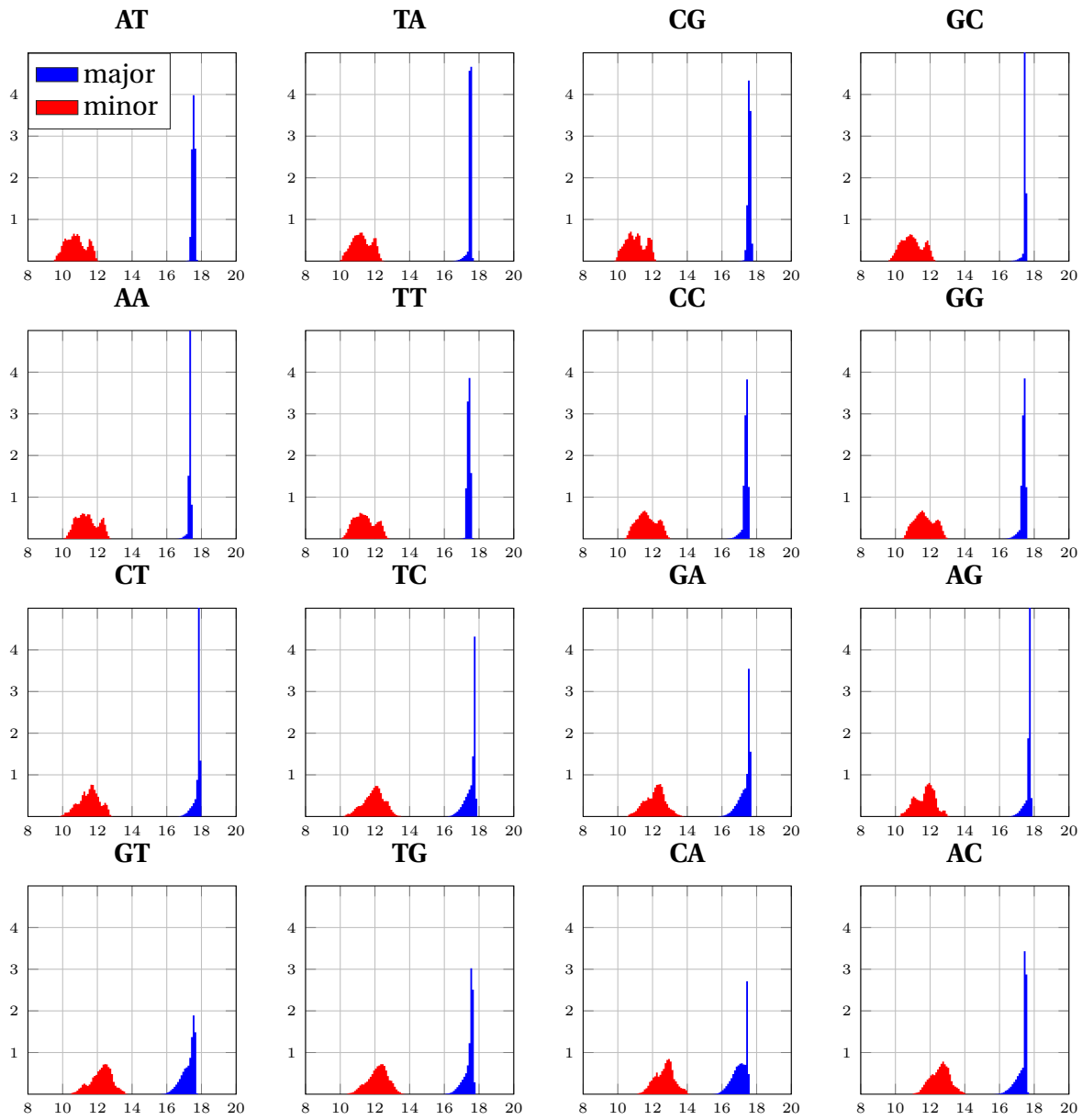



Figure 13.4 – Histogram of minor (red) and major (blue) groove width computed for each central dimers in all the possible tetramer context.

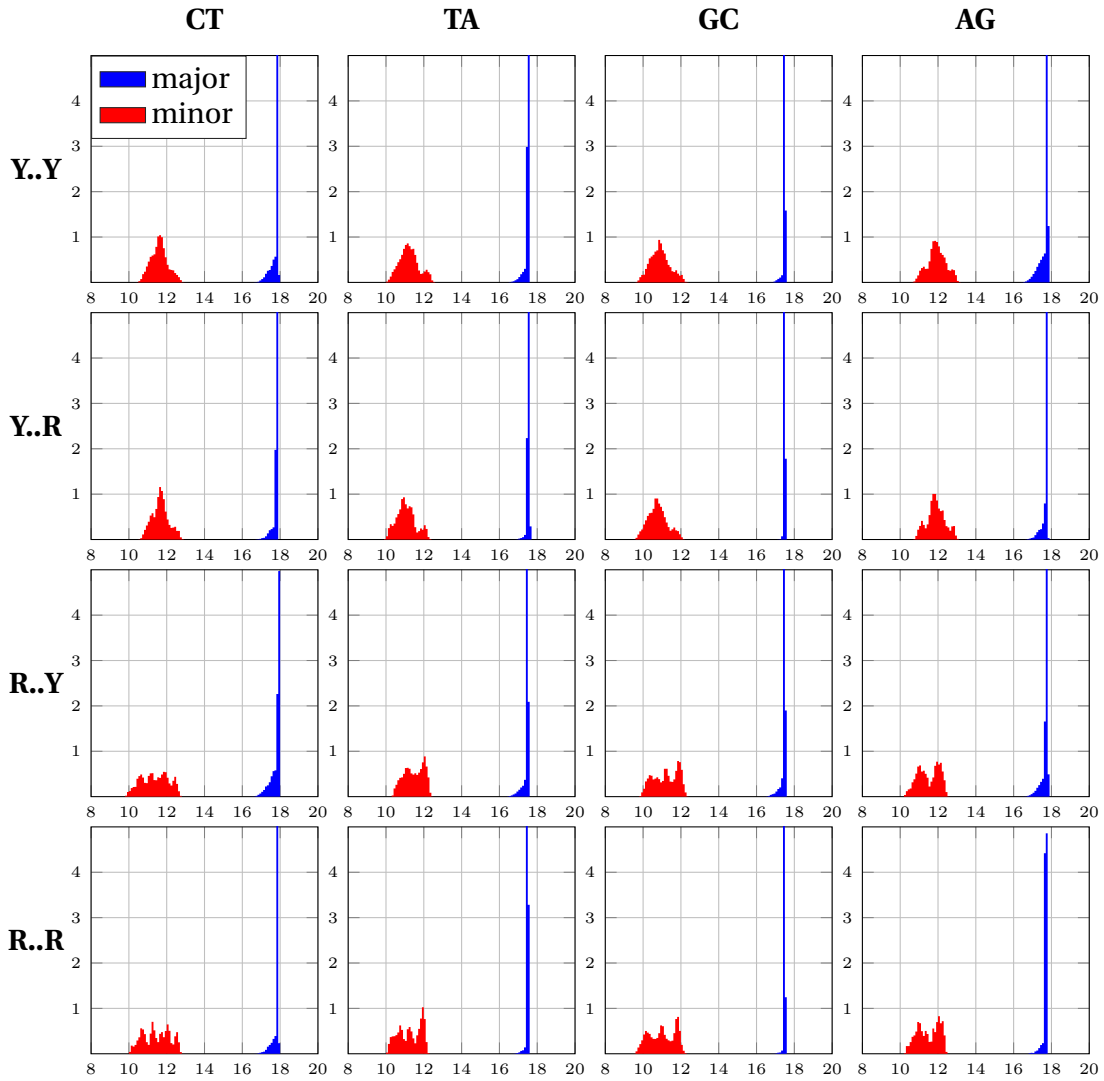



Figure 13.5 – Histogram of minor (red) and major (blue) groove width computed for each the central dimers CT, TA, GC, and AG classified by the flanking bases in the purine pyrimidine alphabet (R–Y).

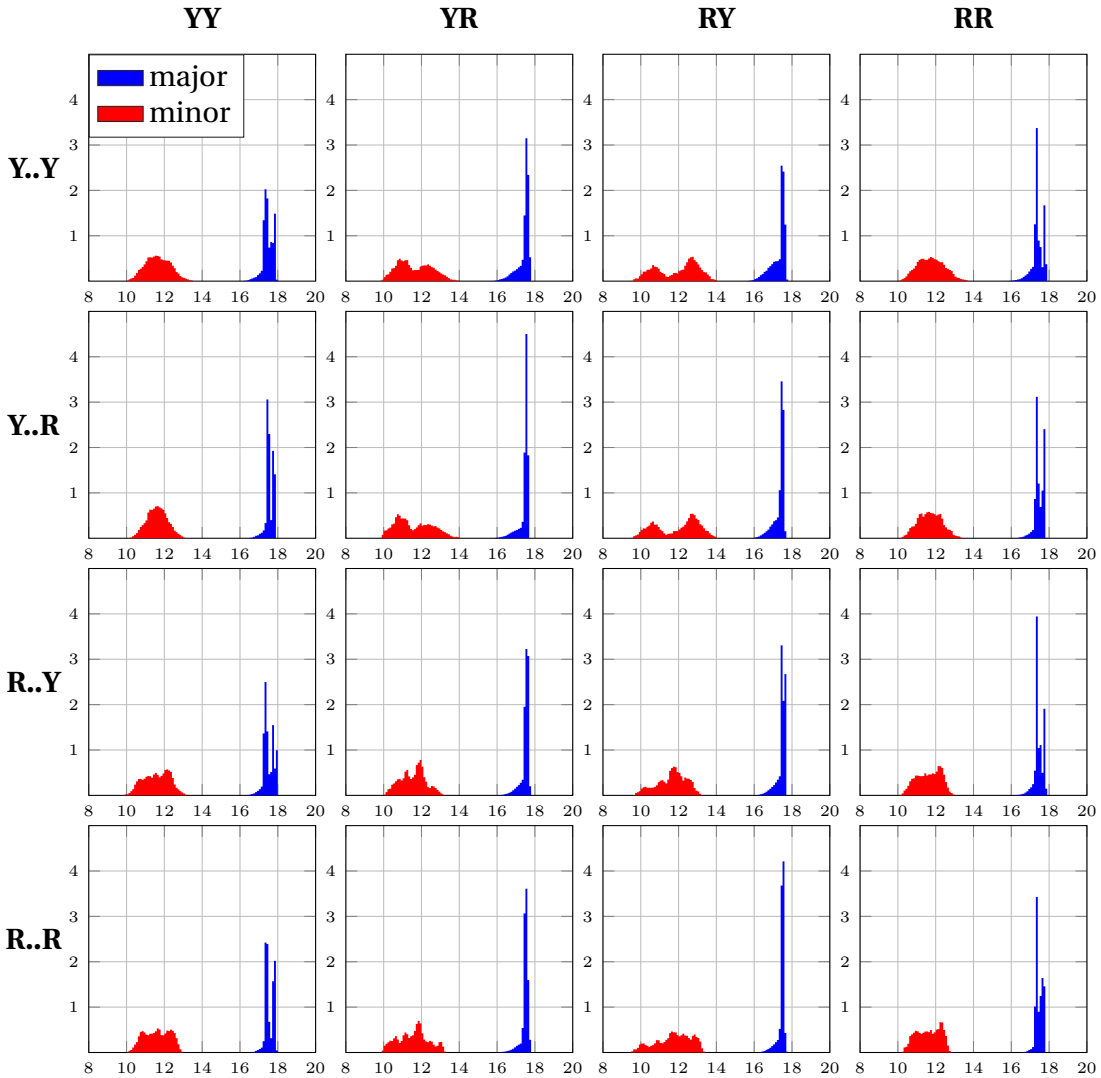



Figure 13.6 – Histogram of minor (red) and major (blue) groove width computed for each the central dimers in the purine–prymidine (R–Y) alphabet classified by the flanking bases also in the purine pyrimidine alphabet.

In conclusion, we have presented another example of how the cgDNA+ model can be used, and in particular, we have showed how to compute the groove widths of B–form DNA by using the phosphate group absolute positions as predicted by cgDNA+. For a given arbitrary sequence one can approximate each backbone as a cubic spline passing through the phosphate group position and can then determine the minor and major groove width by visualising the pairwise distances between an equal number of equidistributed points along both splines. Another method we have presented is easier and faster because it does not approximate the backbones but considers only the pairwise distances between the phosphate group position. We have presented the study of the grooves by fixing a phosphate of a central dimer and by considering

all the possible tetramer context which in particular implies the reconstruction of all the decamer sequences. The findings show that the minor groove varies considerably as function of the sequence while the major groove has the tendency of varying less. However, a natural extension of this study would be to cluster the distribution dshowed in figure 13.6 the the number of the index of the phosphate group associated to the minor and major groove widths. In fact, we conjecture that the distributions will naturally split into at least two distinct histograms.

# Future development of the applications

We have proposed four applications of the cgDNA+ model one for each of the following topics: sequence–dependent persistence length of DNA, packing forces in crystal structures, sequence–dependent sugar ring puckering mode, sequence–dependent BI–BII backbone conformations analysis, and sequence–dependent major and minor groove width computations. We want to discuss possible future developments of the latter three applications.

Firstly, we want to stress that there exist analogous explicit formulas for computing the total external forces acting on a single base, see for instance appendix F. The evaluation of external forces can be used as an extra tool to compare MD and crystal structures through a coarse–grain model. The most natural future development is the use of the cgDNA+ model to study external loads acting on double–stranded DNA computed on averaged configurations extracted from MD simulations of DNA–protein complexes. The principal advantages of working with data generated from MD simulations would be the absence of packing forces.

The study of BI–BII backbone conformations can be extended to an ensemble of configurations drawn from a cgDNA+ predicted Gaussian and thus for the same sequence. More precisely, the BI–BII states occupancy can be studied within the sequence in order to understand if the BI–BII configuration computed on the ground–states is the most common state among its possible configurations.

Again, another next step is to study the distribution of groove widths for single sequences using the cgDNA+ predicted Gaussian. The idea consists in computing, using Monte Carlo, the expectation of major and minor groove width in order to understand if the groove widths computed on ground–states are averaged values, or if the distributions of minor and major grooves have interesting sequence–dependent properties.

The relation between the three applications mentioned above is the understanding of the role of DNA sequence in DNA–protein binding. The computation of the external forces can be used to study specific cases, while the other two methods can be used for in extensive study of sequence–dependent structural properties of DNA. In particular, it would be of great interest to study the relationship between groove widths and BI–BII backbone conformations [49] in the cgDNA+ context.

In conclusion, we have presented illustrative examples of how the cgDNA+ model

might be used as a tool for the analyse of many mechanical and structural properties of the DNA that were not feasible with previous models, and especially in better understanding the mechanics of DNA–protein binding. In particular the computational efficiency of the cgDNA+ model allows behaviour over large ensemble of sequences to be considered

# Conclusions and Outlook

The original contributions of this work are presented in Parts II, III, and IV.

Part II describes enhancements to the cgDNA model [55, 19].
In chapter 6 we used the parameter continuation algorithm 2 to compute cgDNA parameter sets for a variety of MD protocols 6.2. The main conclusion of the sensitivity study is that the force field has the most impact on the parameters of the cgDNA model, and consequently on its predictions, compared to differences in training libraries and ion type. We showed that the prediction of apparent and dynamic persistence lengths for the bsc1 parameter set tend to be significantly higher compared to those with the bsc0 force field, see for instance figures 6.4–6.5. The reciprocal of the static persistence length computed using (2.45) shows a clear changes in rigidity of the bsc1 based cgDNA parameters, right–hand side of figures 6.4–6.5. Moreover, we used the sequence–averaged per–degree of freedom Kullback–Leibler divergence (avgKLd) (6.6) to assess the quality of the best–fit parameter set in reproducing the data and to allow a direct comparison between MD protocols.
In chapter 7 we designed a new sequence library, see table 7.1, using the algorithm 3 which comprises only palindromic sequences, and contains at least one instances of all the possible non–palindromic tetramers. We took advantage of palindromic symmetry to compute the convergence error for the first (7.6) and second centred moments (7.10) estimated from MD simulation, see tables 7.5–7.6. We conclude that the $3\mu$s simulation duration is enough to achieve a negligible palindromic error in both estimators and therefore hopefully an overall negligible error.
We compared the new format dimer–based cgDNA parameter set (7.17) and showed how it improves the quality of the reconstruction, especially at the ends, see figure 7.4. The excellent performance of the new format of the cgDNA parameter set does imply the loss of uniqueness of the best–fit parameter set, but in an interesting way. The main consequence is related to positiveness of the parameter set in the sense of having the property of reconstructing a positive–definite stiffness matrix for any arbitrary sequence. A sufficient condition for this property is that the parameter set is such that each dimer–dependent stiffness block is positive–definite, see section 7.3.1.

Part III is dedicated to the mathematical background of the cgDNA+ model along with the computation of the first best–fit cgDNA+ parameter set, i.e, including an explicit description of the phosphate groups.
In chapter 8 we defined tetrachain configurations of double–stranded DNA by introducing the base–pair level, (8.6) or (8.7), and the microstructure (8.8). Then, for the nearest–neighbour internal energy (8.12) we computed its coordinate free, first variation (8.14) and its coordinate free variation with respect to a single phosphate group rigid–body (8.18). In section 8.4 we defined the internal coordinates for tetrachains by parametrising the base–to–phosphate relative rigid–body motion. We introduced the linearisation matrix (8.33) specific to the base–to–phosphate degree of freedoms, and we gave the explicit coordinate dependent formula for the total external forces acting on a single phosphate group 8.37 in any configuration.
In chapter 9 we introduced the cgDNA+ model. The internal coordinates of the cgDNA+ model are the tetrachain internal coordinates with the modelling choice of parametrising the relative displacement between a base and its $5'$ phosphate group, see (9.6). For the chosen internal coordinates we studied firstly the convergence of mean and covariance estimated from the palindromic MD training library, in particular in tables 9.4 and 9.6, and secondly we compute the convergence error using the Kullback–Leibler divergence between observed banded Gaussian estimated from MD time series and its palindromic symmetric Gaussian, see table 9.7. We showed that the base–to–phosphate entries of the estimators converge slowly compared to the inter– and intra–base–pair degree of freedom and that the main contribution to palindromic error comes from the lack of convergence of the covariance. We then introduce the cgDNA+ parameter set format (9.15) which generalises the cgDNA one. For computing a best–fit parameter set we first approximated the Kullback–Leibler divergence between Gaussians predicted by cgDNA+ and the equivalent Gaussian statistics observed from MD, as a quadratic function (9.52) using the fact that Fisher information is the second derivative of the relative entropy. From the approximation (9.52) we computed an initial parameter set by solving the linear system (9.53). We then used the Fisher–informed gradient numerical scheme (9.54) to compute the best–fit parameter set trained on the palindromic data, which involves the estimation of 10K parameters. The proposed numerical scheme has proven to be efficient and reliable when computing a parameter set using the palindromic data set. Using the method presented in section (9.4) we proved that the parameter set is positive–definite. We then used the end data set to complete the parameter set by solving the system individually for each end block by fixing all the other elements of the parameter set. We observed that the dimer blocks $T\mathrm{p}G$ and $G\mathrm{p}A$ always reconstruct indefinite stiffness matrices. Our conjectured explanation is that the MD simulations of the corresponding sequences are too short in simulation duration. Further simulations are being pursued to verify thus conjecture. Finally, we observed that the comparison between components of the base–to–phosphate internal coordinates predicted by cgDNA+ and observations are astonishingly good, see figure 9.9. In particular, in figure (9.14), we compared predictions

of tangent–tangent correlation computed with the cgDNA+ predicted Gaussian and observed from MD and the agreement is surprisingly good. And this is independent of whether or not the MD time series id filtered for snapshots with broken hydrogen bonds.

Part IV is dedicated to some first illustrative applications of the cgDNA+ model.
In chapter 10, we showed how well the cgDNA+ model can predict the tangent–tangent correlation of a sequence which is not included in the palindromic library (figure 10.3) and we found that the sequence–averaged apparent persistence length is 204 while the dynamic is 217. Both values are significantly higher than the ones obtained for cgDNA trained on the same MD data set but, we recall that the cgDNA+ model predicts tangent–tangent correlations that are closer to those obtained directly from MD. In particular the fact that we obtain persistence lengths that are rather high when compared to the consensus experimental values is not a potential criticism of the cgDNA+ model. Rather it is a criticism of the MD potentials used to train the cgDNA+ parameter set. In fact, the cgDNA+ model accurately reflects implications of the modelling MD simulation protocol, and the computational efficiency of cgDNA+ allows much longer sequence ensembles to be considered that would not be possible with direct MD simulation. This is an important goal of the cgDNA+ model.
In chapter 11 we computed packing forces as external load applied to the phosphate groups for two different PDB crystal structures of the Drew–Dickerson dodecamer sequence. In figures 11.4–11.4, we observed high values of forces and couples exerted by the crystal on the DNA.
In chapter 12 we show how to fine–grain the ground–states predicted by the cgDNA+ model by solving the small dimensional problem (12.7) to obtain all heavy atom configurations. We then, computed the sugar puckering modes on the fine–grained Drew–Dickerson ground–state predicted by cgDNA+ and observed that the modes are different from the ones extracted from the PDB structure of the same sequence, see table 12.2. Thanks to the fine–grained configurations of the ground–state we could study the sequence–dependence of the BI–BII backbone conformation. The main conclusion is that BI–BII states have strong sequence context behaviour and that BI is the most represented states, see figure 12.5.
In chapter 13 we computed the major and minor groove width on cgDNA ground–states using the method in [38], see figures 13.2 and 13.3. We observed a strong sequence–dependence. Then we studied how groove widths behave in different sequence–context. In figures 13.4,13.5, and 13.6, we observed that the minor groove width depends more on the sequence compared to the major groove.

## Future development

For completeness the first future development is related to the $T$p$G$ and $G$p$A$ ends blocks, which currently lead to indefinite reconstructions. We need to simulate the related training sequences for longer simulation duration. In addition, we could also simulate their complementary sequence in order to be able to test their convergence in the sense of Crick–Watson symmetry. We note however that many potential applications of the cgDNA+ model involve sequence dependence far from the ends of DNA sequence. In such applications the pertinent object is a sequence localised marginal probability density function of cgDNA+. Such marginals are easy to compute and far from the ends have little dependence on the actual end sequences. In that framework it is not so important that some end sequence parameter blocks are indefinite, we jut assume other end sequences.
We can divide the other future development in two, inter related, groups: 1) generalisation of the model basic assumptions, 2) extension of the applications.

### Generalisation of the basic assumptions

The first extension of an assumption underlying the cgDNA+ model has already been presented in section 9.6. It consists in extending the stiffness stencil to include the inter–inter interactions of two consecutive junctions. Based on the oligomer–based analysis summarised in tables 9.10–9.11, we concluded that the bigger stencil enhances the quality of the banded approximation with a comparably small number of additional parameters. We want to highlight here that a careful study of the sequence–dependence of the extra blocks should be done in order to define the corresponding parameter set elements.
The second development is on the extension of the alphabet by including the epigenetic base modifications of methylation, hydroxymethylation, etc, in order to be able to study their mechanical properties. It has been observed that epigenetic modifications have an impact on the local property of the double helix, see for examples [66, 48, 56]. The first step in this direction is to construct the appropriately enhanced palindromic library of MD simulations and this step is in progress of the time of submission of this thesis.

### Extension of the applications

In this work we have presented three illustrative applications of the cgDNA+ model and we have discussed their potential connections to the study of DNA–protein binding, with a focus in understanding the role of the DNA sequence. One particularly interesting application is the study of nucleosomes core particle and associated nucleosome positions sequences. A classic approach is to consider that the center line

of the base–pairs of the DNA molecule that wrap around the histone octamer is given by an ideal helix, see for example [75]. Then, given a sequence, the problem is to compute the energy of its configuration under the constraints that the base–pair lie on the helix. We propose two possible extension that could be done using cgDNA+. The first is to constrain only the positions of the phosphate groups that bind to the histone octamer. The second is to describe a set of external forces acting on the phosphate group binding sites. Both proposed method are feasible using the cgDNA+ model but are more challenging from a mathematical point of view, especially the second.
Quantitative and detailed comparison with experimental data remains, as always, a formidable challenge. In the context of modelling DNA–protein interactions one important issue to be addressed is that the cgDNA+ model predicts a range of deformations that can informally be described as ranging from very soft to very stiff modes. As any experimental observations inevitably will contain noise, it seems to be important to develop effective techniques to relax the data within the experimental error in order to decrease the high energy. How to achieve this in a realistic way remain for the moment an elusive objective, that must be reached before quantitative comparison with known DNA–protein structures can be achieved.

# Part V

# Appendices

# A Ideal atom definitions

The Tsukuba convention consists in a set of Cartesian coordinates and atom types that form an idealised nucleic acid base. In table A.1 we report the tsukuba convention as reported in [50]. In table A.2 we report the actual atom coordinates and atom types that are used in the Curves+ software [38] that we have used to fit the fully atomistic molecular dynamics snapshots. The reader can remark that there is a difference in the number of atoms considered for each base type and a difference in their Cartesian coordinates. Moreover in table A.3 we report the reference configuration of the phosphate group which have been also used in the fitting procedure in a modified version of Curves+.

Purine

| Adenine | | | | Guanine | | | |
|---|---|---|---|---|---|---|---|
| $C_1'$ | -2.479 | 5.346 | 0.000 | $C_1'$ | -2.477 | 5.399 | 0.000 |
| $N_9$ | -1.291 | 4.498 | 0.000 | $N_9$ | -1.289 | 4.551 | 0.000 |
| $C_8$ | 0.024 | 4.897 | 0.000 | $C_8$ | 0.023 | 4.962 | 0.000 |
| $N_7$ | 0.877 | 3.902 | 0.000 | $N_7$ | 0.870 | 3.969 | 0.000 |
| $C_5$ | 0.071 | 2.771 | 0.000 | $C_5$ | 0.071 | 2.833 | 0.000 |
| $C_6$ | 0.369 | 1.398 | 0.000 | $C_6$ | 0.424 | 1.460 | 0.000 |
| $N_6$ | 1.611 | 0.909 | 0.000 | $O_6$ | 1.554 | 0.955 | 0.000 |
| $N_1$ | -0.668 | 0.532 | 0.000 | $N_1$ | -0.700 | 0.641 | 0.000 |
| $C_2$ | -1.912 | 1.023 | 0.000 | $C_2$ | -1.999 | 1.087 | 0.000 |
| $N_3$ | -2.320 | 2.290 | 0.000 | $N_2$ | -2.949 | 0.139 | 0.001 |
| $C_4$ | -1.267 | 3.124 | 0.000 | $N_3$ | -2.342 | 2.364 | 0.001 |
| | | | | $C_4$ | -1.265 | 3.177 | 0.000 |

Pyrimidine

| Thymine | | | | Cytosine | | | |
|---|---|---|---|---|---|---|---|
| $C_1'$ | -2.481 | 5.354 | 0.000 | $C_1'$ | -2.477 | 5.402 | 0.000 |
| $N_1$ | -1.284 | 4.500 | 0.000 | $N_1$ | -1.285 | 4.542 | 0.000 |
| $C_2$ | -1.462 | 3.135 | 0.000 | $C_2$ | -1.472 | 3.158 | 0.000 |
| $O_2$ | -2.562 | 2.608 | 0.000 | $O_2$ | -2.628 | 2.709 | 0.001 |
| $N_3$ | -0.298 | 2.407 | 0.000 | $N_3$ | -0.391 | 2.344 | 0.000 |
| $C_4$ | 0.994 | 2.897 | 0.000 | $C_4$ | 0.837 | 2.868 | 0.000 |
| $O_4$ | 1.944 | 2.119 | 0.000 | $N_4$ | 1.875 | 2.027 | 0.001 |
| $C_5$ | 1.106 | 4.338 | 0.000 | $C_5$ | 1.056 | 4.275 | 0.000 |
| $C_{5M}$ | 2.466 | 4.961 | 0.001 | $C_6$ | -0.023 | 5.068 | 0.000 |
| $C_6$ | -0.024 | 5.057 | 0.000 | | | | |

Table A.1 – Tsukuba convention: Atoms type and Cartesian coordinates for the nucleic bases $A, T, G$, and $C$.

Purine

| Adenine | | | | Guanine | | | |
|---|---|---|---|---|---|---|---|
| $C_1'$ | 1.57340 | -2.41044 | -0.12190 | $C_1'$ | 1.58195 | -2.39594 | -0.12320 |
| $N_9$ | 0.37786 | 1.52945 | -0.00531 | $N_9$ | 0.37714 | -1.52785 | -0.00520 |
| $C_8$ | -0.94428 | -1.87750 | 0.15888 | $C_8$ | -0.94239 | -1.88689 | 0.15880 |
| $N_7$ | -1.75109 | -0.86494 | 0.22638 | $N_7$ | -1.76708 | -0.87080 | 0.22880 |
| $C_5$ | -0.91754 | 0.23706 | 0.10110 | $C_5$ | -0.93480 | 0.24071 | 0.10280 |
| $C_6$ | -1.16848 | 1.61858 | 0.09523 | $C_6$ | -1.25149 | 1.62390 | 0.10480 |
| $N_1$ | -0.11123 | 2.43959 | -0.04964 | $N_1$ | -0.10165 | 2.41153 | -0.05020 |
| $C_2$ | 1.10298 | 1.90641 | -0.17768 | $C_2$ | 1.18367 | 1.92662 | -0.18720 |
| $N_3$ | 1.45731 | 0.64079 | -0.18694 | $N_3$ | 1.47916 | 0.62851 | -0.18920 |
| $C_4$ | 0.38110 | -0.16006 | -0.04011 | $C_4$ | 0.37551 | -0.14975 | -0.04020 |

Pyrimidine

| Thymine | | | | Cytosine | | | |
|---|---|---|---|---|---|---|---|
| $C_1'$ | 1.95363 | -1.45659 | -0.19073 | $C_1'$ | 1.94866 | -1.45161 | -0.19029 |
| $N_1$ | 0.75787 | -0.57587 | -0.07419 | $N_1$ | 0.74385 | -0.58352 | -0.07229 |
| $C_2$ | 0.97215 | 0.78019 | -0.13405 | $C_2$ | 0.93020 | 0.79520 | -0.12929 |
| $N_3$ | -0.15841 | 1.56515 | -0.02137 | $N_3$ | -0.15547 | 1.60399 | -0.02329 |
| $C_4$ | -1.45350 | 1.11665 | 0.14102 | $C_4$ | -1.37969 | 1.08626 | 0.13371 |
| $C_5$ | -1.57773 | -0.32139 | 0.19302 | $C_5$ | -1.59254 | -0.32937 | 0.19471 |
| $C_6$ | -0.49400 | -1.10811 | 0.08633 | $C_6$ | -0.49500 | -1.12092 | 0.08671 |

Table A.2 – Curves+ convention: Atoms type and Cartesian coordinates for the nucleic bases $A, T, G$, and $C$.

| Phosphate group | | | |
|---|---|---|---|
| $P$ | 0.000 | 0.000 | 0.000 |
| $O_3'$ | 1.518 | 0.000 | -0.537 |
| $O_5'$ | -0.759 | -1.315 | -0.537 |
| $OP_1'$ | -0.698 | 1.208 | -0.493 |
| $OP_2'$ | 0.000 | 0.000 | 1.480 |

Table A.3 – Convention for the phosphate group ideal atoms and Cartesian Coordinates.

# B Training set libraries

In this chapter of the appendix we report the training libraries mentioned in this work. In table B.1 we report the list of sequences of the ABC training library composed of 39 18 base–pair long sequences designed by R. Lavery for the ABC consortium, see for example [37, 52]. In table B.2 the list of 13 18 base–pair long sequences in the miniABC library designed by M. Pasi and R. Lavery. In chapter 7 we discussed how to design a library which contains only palindromes with each possible tetramer sub–sequence appearing at least once. In table B.3 we report one palindromic library which satisfies the aforementioned conditions. Each palindromic sequence has $5'$ and $3'$ GC ends and is 24 base–pair long. Finally in table B.4 we report also the so called *ends sequence library* which contains all the 15 independent end dimers different then GC. Each sequence is designed to have a stable end, GC, and another dimer. The ends sequence library has been designed by D. Petckeviciute.

ABC library

| Number | Sequence |
|---|---|
| 1 | GCTATATATATATATAGC |
| 2 | GCATTAATTAATTAATGC |
| 3 | GCGCATGCATGCATGCGC |
| 4 | GCCTAGCTAGCTAGCTGC |
| 5 | GCCGCGCGCGCGCGCGGC |
| 6 | GCGCCGGCCGGCCGGCGC |
| 7 | GCTACGTACGTACGTAGC |
| 8 | GCGATCGATCGATCGAGC |
| 9 | GCAAAAAAAAAAAAAAGC |
| 10 | GCCGAGCGAGCGAGCGGC |
| 11 | GCGAAGGAAGGAAGGAGC |
| 12 | GCGTAGGTAGGTAGGTGC |
| 13 | GCTGAGTGAGTGAGTGGC |
| 14 | GCAGCAAGCAAGCAAGGC |
| 15 | GCAAGAAAGAAAGAAAGC |
| 16 | GCGAGGGAGGGAGGGAGC |
| 17 | GCGGGGGGGGGGGGGGGC |
| 18 | GCAGTAAGTAAGTAAGGC |
| 19 | GCGATGGATGGATGGAGC |
| 20 | GCTCTGTCTGTCTGTCGC |
| 21 | GCACAAACAAACAAACGC |
| 22 | GCAGAGAGAGAGAGAGGC |
| 23 | GCGCAGGCAGGCAGGCGC |
| 24 | GCTCAGTCAGTCAGTCGC |
| 25 | GCATCAATCAATCAATGC |
| 26 | GCGTCGGTCGGTCGGTGC |
| 27 | GCTGCGTGCGTGCGTGGC |
| 28 | GCACGAACGAACGAACGC |
| 29 | GCTAGATAGATAGATAGC |
| 30 | GCGCGGGCGGGCGGGCGC |
| 31 | GCGTGGGTGGGTGGGTGC |
| 32 | GCACTAACTAACTAACGC |
| 33 | GCGCTGGCTGGCTGGCGC |
| 34 | GCTATGTATGTATGTAGC |
| 35 | GCTGTGTGTGTGTGTGGC |
| 36 | GCGTTGGTTGGTTGGTGC |
| 37 | GCATAAATAAATAAATGC |
| 38 | GCATGAATGAATGAATGC |
| 39 | GCGACGGACGGACGGAGC |

Table B.1 – The ABC training library.

miniABC library

| Number | Sequence |
|---|---|
| 1 | GCAACGTGCTATGGAAGC |
| 2 | GCAATAAGTACCAGGAGC |
| 3 | GCAGAAACAGCTCTGCGC |
| 4 | GCAGGCGCAAGACTGAGC |
| 5 | GCATTGGGGACACTACGC |
| 6 | GCGAACTCAAAGGTTGGC |
| 7 | GCGACCGAATGTAATTGC |
| 8 | GCGGAGGGCCGGGTGGGC |
| 9 | GCGTTAGATTAAAATTGC |
| 10 | GCTACGCGGATCGAGAGC |
| 11 | GCTGATATACGATGCAGC |
| 12 | GCTGGCATGAAGCGACGC |
| 13 | GCTTGTGACGGCTAGGGC |

Table B.2 – The miniABC training library.

Palindromic library

| Number | Sequence |
|---|---|
| 1 | GCTTAGTTCAAATTTGAACTAAGC |
| 2 | GCTCTCTGTATTAATACAGAGAGC |
| 3 | GCCCTTGGCGATATCGCCAAGGGC |
| 4 | GCTAAAGCCTTATAAGGCTTTAGC |
| 5 | GCGGTAGAAAACGTTTTCTACCGC |
| 6 | GCCAAGACATTGCAATGTCTTGGC |
| 7 | GCAGATGGTCAGCTGACCATCTGC |
| 8 | GCCTCACCGCTCGAGCGGTGAGGC |
| 9 | GCAGTGGAATCATGATTCCACTGC |
| 10 | GCTTTACTTCGTACGAAGTAAAGC |
| 11 | GCTACCTATGCTAGCATAGGTAGC |
| 12 | GCGCACTGGGGATCCCCAGTGCGC |
| 13 | GCTGAGGAGTCCGGACTCCTCAGC |
| 14 | GCTGCCGTCGGGCCCGACGGCAGC |
| 15 | GCGCACAACACGCGTGTTGTGCGC |
| 16 | GCCTAACCCTGCGCAGGGTTAGGC |

Table B.3 – The Palindromic training library.

End sequence library

| Number | Sequence |
|---|---|
| 1 | AAGCAAACTAGC |
| 2 | TTGCGTTTAAGC |
| 3 | ACCACCCTAGGC |
| 4 | TGCTCGGAGCGC |
| 5 | AGCTCAGGTCGC |
| 6 | TCGTCTCCAGGC |
| 7 | ATGCCCGACGGC |
| 8 | TACGATTCTGGC |
| 9 | GATATGGCGTGC |
| 10 | CTACACCCTAGC |
| 11 | CGATGAAGTAGC |
| 12 | GGTATAAATAGC |
| 13 | CCTGCCCGCGGC |
| 14 | GTGTACAATCGC |
| 15 | CAGATGCTTGGC |

Table B.4 – The ends sequence library.

# C The ABC molecular dynamic protocol

In table C.1 we report the most important parameters of the molecular dynamics simulation, along with their definitions, which where used in all the simulations done for the ABC project. In this work we used the ABC protocol as base–line in all the MD simulations performed. The only changes to the original ABC protocol we have made concern only: simulation duration, force field, and ion type. In table 6.2 we report the values of the latter three parameters used for simulating different training libraries.

| Parameter | detail/value |
|---|---|
| Software | AMBER |
| Force field | param99 with parambsc0 modification |
| Ion type | potassium K |
| Concentration | 150 mM |
| Water model | SPC/E |
| Time step integration | 2 $f$s |
| Simulation duration | 50–100 $n$s |
| Writing rate | 1$p$s |

Table C.1 – Selection of the some values of the MD parameter used in the ABC project. More details can be founded in the articles [37, 52].

| Label | $St$ | $Lb$ | $Ff$ | $Io$ |
|---|---|---|---|---|
| $\mu$ABC | 1$\mu$s | ABC | bsc0 | K+ |
| MABC$_0^K$ | 1$\mu$s | miniABC | bsc0 | K+ |
| MABC$_1^K$ | 1$\mu$s | miniABC | bsc1 | K+ |
| MABC$_1^{KNa}$ | 1$\mu$s | miniABC | bsc1 | $\frac{1}{2}$ K+ $\frac{1}{2}$ Na+ |
| MABC$_1^{Na}$ | 1$\mu$s | miniABC | bsc1 | Na+ |

Table C.2 – Characteristics of the sets of simulations used in the comparisons. The main protocol is the ABC protocol [37, 52] where simulation duration, force field, and ions type have been changed one–at–a–time.

# D cgDNA+: complement to the simulation convergence discussion

In this chapter we report in tables D.1 and D.2 the values of the palindromic errors on the mean and on the covariance estimators for the six-teen palindromic sequences B.3 when considering the cgDNA+ degree of freedom. In table 7.7 we list the percentage of accepted snapshots, after hydrogen bond filtering, for different simulation duration.

| # $\mathcal{S}$ | $100ns$ | $1\mu$s | $2\mu$s | $3\mu$s |
|---|---|---|---|---|
| 1 | 2.674 | 1.7078 | 0.75098 | 0.72846 |
| 2 | 5.0809 | 2.954 | 1.68 | 1.4631 |
| 3 | 4.1417 | 1.72 | 0.89638 | 0.72616 |
| 4 | 3.0172 | 3.1105 | 1.6639 | 1.6975 |
| 5 | 11.667 | 3.6407 | 2.1639 | 1.6331 |
| 6 | 11.355 | 2.3982 | 1.259 | 1.02 |
| 7 | 2.8166 | 3.7586 | 2.4523 | 2.0041 |
| 8 | 3.4284 | 1.3654 | 1.0436 | 0.86127 |
| 9 | 7.8926 | 1.2537 | 1.4452 | 1.048 |
| 10 | 4.7561 | 1.7066 | 1.2032 | 1.7445 |
| 11 | 10.036 | 4.523 | 2.4275 | 2.2984 |
| 12 | 6.199 | 2.7985 | 2.0227 | 1.8517 |
| 13 | 3.3691 | 1.1925 | 1.1706 | 0.83372 |
| 14 | 2.4161 | 1.8326 | 1.0583 | 0.80795 |
| 15 | 6.2364 | 3.9084 | 1.3505 | 1.3305 |
| 16 | 4.2067 | 1.8224 | 2.438 | 1.5513 |

Table D.1 – Palindromic error in the estimator of the mean as function of simulation length for cgDNA+ degree of freedom. In table 7.7 one can find the actual percentage of accepted trajectories for each simulation length and each sequence.

| # $\mathcal{S}$ | 100$ns$ | 1$\mu$s | 2$\mu$s | 3$\mu$s |
|---|---|---|---|---|
| 1 | 10.308 | 9.1031 | 4.217 | 3.7444 |
| 2 | 24.951 | 13.675 | 8.4661 | 7.7868 |
| 3 | 15.972 | 6.9944 | 4.8431 | 3.8754 |
| 4 | 13.984 | 14.61 | 8.097 | 7.9871 |
| 5 | 21.751 | 16.708 | 10.828 | 8.4623 |
| 6 | 15.533 | 12.717 | 7.0401 | 5.5072 |
| 7 | 13.344 | 13.173 | 10.478 | 9.2184 |
| 8 | 14.835 | 6.2428 | 5.4558 | 3.9188 |
| 9 | 25.878 | 7.0195 | 7.3681 | 6.022 |
| 10 | 17.471 | 8.3987 | 5.6304 | 7.8423 |
| 11 | 25.225 | 17.808 | 11.116 | 10.82 |
| 12 | 22.966 | 13.361 | 10.751 | 8.7409 |
| 13 | 14.504 | 5.6737 | 6.0145 | 4.7198 |
| 14 | 12.103 | 9.5461 | 5.5763 | 4.5541 |
| 15 | 17.945 | 15.639 | 7.2471 | 6.564 |
| 16 | 19.339 | 9.2026 | 12.241 | 8.074 |

Table D.2 – Palindromic error in the estimator of the covariance as function of simulation length for cgDNA+ degree of freedom. In table 7.7 one can find the actual percentage of accepted trajectories for each simulation length and each sequence.

| # $\mathcal{S}$ | 100$ns$ | 500$n$s | 1$\mu$s | 1.5$\mu$s | 2$\mu$s | 2.5$\mu$s | 3$\mu$s |
|---|---|---|---|---|---|---|---|
| 1 | 0.8439 | 0.9030 | 0.7895 | 0.8196 | 0.8461 | 0.8581 | 0.8654 |
| 2 | 0.9065 | 0.9217 | 0.9166 | 0.9179 | 0.8959 | 0.9008 | 0.8960 |
| 3 | 0.8098 | 0.9117 | 0.8428 | 0.8623 | 0.8812 | 0.8927 | 0.9000 |
| 4 | 0.9127 4 | 0.9241 | 0.9239 | 0.9236 | 0.9247 | 0.9167 | 0.9164 |
| 5 | 0.6198 | 0.8775 | 0.9064 | 0.9176 | 0.9211 | 0.9236 | 0.9264 |
| 6 | 0.6253 | 0.8629 | 0.9041 | 0.9189 | 0.9044 | 0.7884 | 0.7930 |
| 7 | 0.9450 | 0.9442 | 0.9291 | 0.9270 | 0.8971 | 0.9071 | 0.9111 |
| 8 | 0.9372 | 0.9364 | 0.9395 | 0.9373 | 0.9345 | 0.9362 | 0.9340 |
| 9 | 0.9518 | 0.9324 | 0.9387 | 0.8668 | 0.8819 | 0.8946 | 0.8997 |
| 10 | 0.9425 | 0.9038 | 0.9201 | 0.9241 | 0.9268 | 0.9231 | 0.9254 |
| 11 | 0.8865 | 0.8912 | 0.8906 | 0.8953 | 0.9008 | 0.9039 | 0.9068 |
| 12 | 0.8490 | 0.9248 | 0.9177 | 0.9238 | 0.9285 | 0.9306 | 0.9324 |
| 13 | 0.8793 | 0.9166 | 0.9160 | 0.9181 | 0.7846 | 0.8020 | 0.8236 |
| 14 | 0.9441 | 0.9423 | 0.9409 | 0.9319 | 0.9344 | 0.9358 | 0.9367 |
| 15 | 0.9518 | 0.8921 | 0.9049 | 0.9132 | 0.9223 | 0.9209 | 0.9259 |
| 16 | 0.9233 | 0.9288 | 0.9287 | 0.9289 | 0.9276 | 0.9282 | 0.9275 |

Table D.3 – Percentage of accepted trajectories after HB filetring per simulation lengths considered in the error computations D.1, D.2.

| $\#\mathcal{S}$ | $1\mu$s | $2\mu$s | $3\mu$s | $4\mu$s | $5\mu$s |
|---|---|---|---|---|---|
| 1 | 0.9023 | 0.9134 | 0.8958 | 0.8985 | 0.9035 |
| 5 | 0.9424 | 0.8912 | 0.8837 | 0.8980 | 0.9064 |
| 11 | 0.7441 | 0.8295 | 0.8419 | 0.8528 | 0.8248 |
| $\#\mathcal{S}$ | $6\mu$s | $7\mu$s | $8\mu$s | $9\mu$s | $10\mu$s |
| 1 | 0.9057 | 0.8986 | 0.8998 | 0.9012 | 0.8963 |
| 5 | 0.8874 | 0.8861 | 0.8890 | 0.8933 | 0.8977 |
| 11 | 0.8391 | 0.8494 | 0.8525 | 0.8560 | 0.8620 |

Table D.4 – Percentage of accepted trajectories after HB filetring per simulation lengths considered in the error computations reported in tables: 9.3 9.4 9.5.

# E Derivatives of KLd, Gradient, and Hessian matrix

We start by introducing some useful notation. Let $F : \mathbb{R}^{n\times m} \to \mathbb{R}$ be a real–valued function with $n, m \geq 1$ . The derivative of $F$ in the direction $\mathbf{d} \in \mathbb{R}^{n\times m}$ is

$$\nabla_{\mathbf{d}} F(A) = \frac{\partial F(A)}{\partial A} : \mathbf{d} = \partial_A F(A) : \mathbf{d} = \text{trace}\left(\mathbf{d}^T \partial_A F(A)\right),$$

and denotes the partial entry–by–entry derivative of the function $F$ with respect to the Frobenius inner product (2.1). The special cases we will be using are when $n = m$, $n > 1$ and when $n > 1$ and $m = 1$. For the second case the Frobenius inner produce coincide with the Euclidean inner product and is denoted by $\cdot$. The second directional derivative of $F$ is defined by

$$\nabla_{\mathbf{d'}} \nabla_{\mathbf{d}} F(A) := \nabla_{\mathbf{d}'} \text{trace}\left(\mathbf{d}^T \partial_A F(A)\right),$$

for the two directions $\mathbf{d}, \mathbf{d}' \in \mathbb{R}^{n\times m}$. Finally, the Hessian of the real–valued function $F$ is the directional derivative of the real–valued function $G(A, \mathbf{d}) := \text{tr}\left(\mathbf{d}^T \partial_A F(A)\right)$.
In this section we will consider two multivariate normal distributions $\rho(x; \mu, K)$ and $\rho(x; \mu_d, K_d)$. For sake of compactness we use the short notations $\rho_m := \rho(x; \mu, K)$ and $\rho_d = \rho(x; \mu_d, K_d)$. The goal of this chapter is to compute the gradient vector and the Hessian matrix of the Kullback-Leibler divergence (KLd) with respect to the parameters of the Gaussian in first place of the argument. Let us recall the algebraic expression of KLd for $\rho_m$ and $\rho_d$

$$\begin{aligned} D_{KL}(\rho_m, \rho_d) &= \frac{1}{2}\left(K^{-1} : K_d + \ln\frac{|K|}{|K_d|} - N + (\mu_d - \mu)K_d(\mu_d - \mu)\right) \\ &= D^{\dagger}(\rho_m, \rho_d) + \mathcal{M}(\rho_m, \rho_d), \end{aligned}$$

where

$$D^{\dagger}(\rho_m, \rho_d) = \frac{1}{2}\left(K^{-1} : K_d + \ln\frac{|K|}{|K_d|} - N\right),$$
$$\mathcal{M}(\rho_m, \rho_d) = \frac{1}{2}(\mu_d - \mu)K_d(\mu_d - \mu).$$

Before presenting all the computations we recall that, in the context of the cgDNA+ parameter set extraction, we consider the vector $\sigma := K\mu$ called the weighted shape vector, instead of the vector $\mu$. Thus, for the computation of gradient and Hessian matrix we replace all the mean vector $\mu$ with $\sigma$ and we will take the partial derivative with respect to $\sigma$ and $K$.

## E.1 First derivative of the Kullback–Leibler divergence

The first derivative of the KLd with respect to $\sigma$ and $K$ for the direction $\mathbf{d} := (\lambda, \Lambda) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$ is defined by

$$\nabla_{\mathbf{d}} D_{KL}(\rho_m, \rho_d) = \partial_K D_{KL}(\rho_m, \rho_d) : \Lambda + \partial_\sigma D_{KL}(\rho_m, \rho_d) \cdot \lambda. \tag{E.1}$$

For sake of compactness we will drop the arguments $(\rho_m, \rho_d)$ in all the following computations.
For the stiffness part, we can split the derivation into two quantities, the one coming from the pure stiffness part, $D^{\dagger}$ and the one coming from the Mahalanobis part, $\mathcal{M}$, thus we can compute

$$\partial_K D_{KL} : \Lambda = \left(\partial_K D^{\dagger} + \partial_K \mathcal{M}\right) : \Lambda,$$
$$\partial_K D^{\dagger} : \Lambda = \frac{1}{2}(K^{-1} - K^{-1}K_dK^{-1}) : \Lambda$$
$$\partial_K \mathcal{M} : \Lambda = -K^{-1}K_d(\mu - \mu_d)\mu^T : \Lambda$$

We hence obtain that

$$\partial_K D_{KL} : \Lambda = \frac{1}{2}\left(K^{-1} - K^{-1}K_dK^{-1}\right) - K^{-1}K_d(\mu - \mu_d))\mu^T : \Lambda. \tag{E.2}$$

For the $\sigma$ part we obtain simply

$$\begin{aligned}\partial_\sigma D_{KL} \cdot \lambda &= \partial_\sigma \mathcal{M} \cdot \lambda \\ &= (K^{-1}K_d(\mu - \mu_d)) \cdot \lambda\end{aligned} \tag{E.3}$$

For the above computations we have used, in particular, the following useful expressions [54]: Let $A, B, X \in \mathbb{R}^{n\times n}$

$$\begin{aligned}\partial_X \text{trace}\left(AX^{-1}B\right) &= -\left(X^{-1}BAX^{-1}\right)^T.\\ \partial_X|X| &= |X|X^{-1}\end{aligned}$$

## E.2 Gradient of the Kullback–Leibler divergence

Before defining the gradient we use the relation between the Frobenius inner product and the Euclidean inner product (2.4) to rewrite the first directional derivative of KLd

$$\nabla_{\mathbf{d}} D_{KL} = \begin{bmatrix}\partial_\sigma D_{KL}\\ \mathbf{vec}\left(\partial_K D_{KL}\right)\end{bmatrix}\cdot\begin{bmatrix}\lambda\\ \mathbf{vec}(\Lambda)\end{bmatrix} = \text{param}\left(\partial_\sigma D_{KL}, \partial_K D_{KL}\right)\cdot\text{param}(\mathbf{d}), \quad \text{(E.4)}$$

where $\mathbf{vec}(\cdot)$ is defined in (2.3) and $\text{param}(\cdot,\cdot)$ is defined in (2.9). Now, without loss of generality we can consider a direction $\mathbf{d} \in \mathbb{R}^{n+n^2}$ and thus we drop the notation $\text{param}(\cdot)$ in the second term of the right–hand side expression in (E.4). The Gradient of the KLd is then defined, element–wise by evaluating (E.4) in the canonical base of $\mathbb{R}^{n+n^2}$, which simply leads to

$$\text{Grad} D_{KL} = \text{param}\left(\partial_\sigma D_{KL}, \partial_K D_{KL}\right). \quad \text{(E.5)}$$

## E.3 Second derivative of the Kullback–Leibler divergence

The second order directional derivative of KLd, evaluated in the direction $\mathbf{d} = (\lambda, \Lambda) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$ and $\mathbf{d}' = (\lambda', \Lambda') \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$ is defined by

$$\begin{aligned}\nabla_{\mathbf{d}'}\nabla_{\mathbf{d}} D_{KL} &= \partial_K \text{trace}\left(\Lambda^T \partial_K D_{KL}\right) : \Lambda'\\ &+ \partial_\sigma \text{trace}\left(\Lambda^T \partial_K D_{KL}\right)\cdot\lambda'\\ &+ \partial_K \text{trace}\left(\lambda^T \partial_\sigma D_{KL}\right) : \Lambda'\\ &+ \partial_\sigma \text{trace}\left(\lambda^T \partial_\sigma D_{KL}\right)\cdot\lambda'. \quad \text{(E.6)}\end{aligned}$$

Hereafter we give the explicit term of each second directional derivatives in E.6:

$$\partial_\sigma \text{trace}\left(\Lambda^T \partial_K D_{KL}\right)\cdot\lambda' = \left(-\mathcal{K}\Lambda\mu - K^{-1}\Lambda^T K^{-1} K_d(\mu-\mu_d)\right)\cdot\lambda'. \quad \text{(E.7)}$$

$$\partial_\sigma \text{trace}\left(\lambda^T \partial_K D_{KL}\right)\cdot\lambda' = \mathcal{K}\lambda\cdot\lambda', \quad \text{(E.8)}$$

where $\mathcal{K} = K^{-1}K_dK^{-1}$. For the second directional derivative with respect to the stiffness we have

$$\partial_K \text{trace}\left(\Lambda^T \partial_K D_{KL}\right) : \Lambda' = \tag{E.9}$$

$$\left(\frac{1}{2}\left(K^{-1}\Lambda\mathcal{K} + \mathcal{K}\Lambda K^{-1} - K^{-1}\Lambda K^{-1}\right) + \mathcal{K}\Lambda\mu\mu^T\right) : \Lambda' \tag{E.10}$$

$$+ \left(K^{-1}\Lambda\mu(\mu-\mu_d)^T K_d K^{-1} + \mu(\mu-\mu_d)^T K_d K^{-1}\Lambda K^{-1}\right) : \Lambda' \tag{E.11}$$

and

$$\partial_K \text{trace}\left(\lambda^T \partial_\sigma D_{KL}\right) : \Lambda' = -\left(K^{-1}\lambda(\mu-\mu_d)^T K_d K^{-1} + \mathcal{K}\lambda\mu^T\right) : \Lambda'. \tag{E.12}$$

For the above computations we have used the following matrix calculus identities [54]: Let $a, b \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times n}$,

$$\partial_X a^T X^{-1} b = -\left(X^{-1} b a^T X^{-1}\right)^T,$$

## E.4 Hessian matrix of the Kullback–Leibler divergence

As we have already done for the Gradient in the (E.5) we vectorise the expression of second derivatives of the Kullback–Leibler divergence. The Hessian matrix will then be defined column–wise by evaluating the terms of (E.6) on an element of the canonical base of $\mathbb{R}^{n+n^2}$. More precisely, the $j$–th column of the Hessian will be related to the base element $\mathbf{e}_j \in \mathbb{R}^{n+n^2}$. Before giving the explicit expression of the Hessian we point out that the term (E.11) of $\partial_K \text{trace}\left(\Lambda\partial_K D_{KL}\right) : \Lambda'$ is not symmetric under the change of order between $\Lambda$ and $\Lambda'$, meaning that a choice should be made between $\mathbf{d} = (\lambda, \Lambda)$ and $\mathbf{d}' = (\lambda', \Lambda')$ in order to define the column element of the Hessian. We have made the decision that the directions $\mathbf{d}'$ define the columns. Hereafter we explicitly give each term of the second derivative after the swap between directions:

$$\begin{aligned}\mathcal{H}_K(\rho_m, \rho_d; \mathbf{d}') := \frac{1}{2}&\left(K^{-1}\Lambda'\mathcal{K} + \mathcal{K}\Lambda' K^{-1} - K^{-1}\Lambda' K^{-1}\right) + \mathcal{K}\Lambda'\mu\mu^T \\ &+ K^{-1}\Lambda' K^{-1} K_d(\mu-\mu_d)\mu^T + K^{-1}K_d(\mu-\mu_d)\mu^T\Lambda' K^{-1} \\ &- \mathcal{K}\lambda'\mu^T - K^{-1}K_d(\mu-\mu_d)(\lambda')^T K^{-1},\end{aligned}$$

and

$$\mathcal{H}_\sigma(\rho_m, \rho_d; \mathbf{d}') := -K^{-1}\Lambda' K^{-1} K_d(\mu-\mu_d) - \mathcal{K}\Lambda'\mu + \mathcal{K}\lambda'. \tag{E.13}$$

The Hessian is then defined, column–wise, by

$$[\text{Hess}D_{KL}]_{(\cdot,j)} = \text{param}\left(\mathcal{H}_\sigma(\rho_m,\rho_d;\mathbf{e}_j),\mathcal{H}_K(\rho_m,\rho_d;\mathbf{e}_j)\right) \tag{E.14}$$

## E.5 Derivation of KLd with respect to the parameter set

Let $\mathcal{P} = \{\sigma^{\alpha\beta},\sigma^{5'\alpha\beta},K^{\alpha\beta},K^{5'\alpha\beta}\}_{\alpha\beta\in D}$ be a cgDNA+ parameter set and let $\rho_m := \rho(x,\mathcal{P},\mathcal{S})$ be a Gaussian predicted by cgDNA+ and $\rho_d = \rho(x,\mathcal{S})$ be a Gaussian observed from MD simulations of sequence $\mathcal{S}$, and consider again the KLd between $\rho_m$ and $\rho_d$

We can now compute the first derivative of the Kullback–Leibler divergence with respect to the parameter set $\mathcal{P}\in\mathbb{P}$. The variation with respect to the parameter set $\mathcal{P}$ corresponds to the variation of each single element in $\mathcal{P}$ with respect to the considered sequence $\mathcal{S}$ in $\rho_m$. More precisely, assume that the dimer $\alpha\beta$ is not present in the sequence $\mathcal{S}$, we then expect the gradient with respect to both cgDNA+ parameter set elements $\sigma^{\alpha\beta}$ and $K^{\alpha\beta}$ to be zero. These remark help in understand that the extra chain rule that we need to introduce in the computation of the variation of the KLd with respect to the parameter set $\mathcal{P}$ is linear and is basically a dimer dependent extraction of vector block and matrix block from the variation of the KLd with respect to $\sigma_m$ and $K_m$. More precisely, let us introduce the first variation of KLd with respect to $\mathcal{P}$ via the following notation

$$\nabla_{\mathcal{P}}D_{KL} = \left\{\nabla_{\mathcal{P}}\sigma^{\alpha\beta},\nabla_{\mathcal{P}}\sigma^{5'\alpha\beta},\nabla_{\mathcal{P}}K^{\alpha\beta},\nabla_{\mathcal{P}}K^{5'\alpha\beta}\right\}_{\alpha\beta\in D,5'\alpha\beta\in D'} \tag{E.15}$$

where $D$ and $D'$ are respectively the set of 10 independent dimer step and the set of all the 16 end dimers. For sake of compactness we will present just the following definition:

$$\nabla_{\mathcal{P}}\sigma^{XY} = \sum_{id\in\mathcal{S}(XY)}\mathbb{E}^T(id)\partial_\sigma D_{KL} \tag{E.16}$$

where $\partial_\sigma D_{KL}$ is defined in (E.3), $\mathcal{S}(XY)$ is the set of all the positions, in number of junctions, where there is the dimer step $XY$ in the sequence $\mathcal{S}$. The matrix $\mathbb{E}(id)\in\mathbb{R}^{24N-18\times 42}$ is called the *dimer extraction* matrix and has the properties that if $\mathcal{S}(XY)$ is empty then $\mathbb{E}(\emptyset)$ is a trivial matrix otherwise for $id\in\mathcal{S}(XY)$, the matrix $\mathbb{E}(id)$ is

defined by

$$\mathbb{E}(id) = \begin{bmatrix} 0 \\ \vdots \\ I_{42} \\ \vdots \\ 0 \end{bmatrix} \tag{E.17}$$

where the identity matrix is of dimension $42 \times 42$. Finally we get that the gradient with respect to the parameter set in vectorial format $P$, see for instance (9.44), is simply

$$\text{Grad}_P D_{KL} := \text{vec}(\nabla_{\mathcal{P}} D_{KL}) \tag{E.18}$$

where the operator vec$(\cdot)$ has been definite in (9.44).
For the second derivative of the KLd with respect to the parameter set in vectorial format $P$ we can combine the computation already done for E.14 and the latter methodology used to compute (E.15) and consequently (E.18). In fact, column wise we will define the second derivative of the KLd with respect to $P$ in the following way

$$[\text{Hess}_P D_{KL}]_{(\cdot,j)} := \text{vec}(\mathcal{H}_{\mathcal{P}}(\rho_m, \rho_d; \mathbf{p}_j)), \tag{E.19}$$

where $\mathbf{p}_j) = (\sigma(\mathcal{P}_j, \mathcal{S}), K(\mathcal{P}_j, \mathcal{S}))$ and $\mathcal{P}_j$ is in bijection with $P_j$ which satisfies

$$[P_i]_j = 1 \text{ if and only if } i = j,\ [P_i]_j = 0 \text{ otherwise.}$$

Then the definition of the right–hand side of (E.19) should again be interpreted as of the set of cgDNA+ parameter set, denoted by $\mathbb{P}$, which account of the entries of the Hessian matrix with respect to $P$ (vectorial form). Formally

$$\mathcal{H}_{\mathcal{P}}(\rho_m, \rho_d; \mathbf{p}_j)) = \left\{ \mathcal{H}_{\mathcal{P}} \sigma^{\alpha\beta}, \mathcal{H}_{\mathcal{P}} \sigma^{5'\alpha\beta}, \mathcal{H}_{\mathcal{P}} K^{\alpha\beta}, \mathcal{H}_{\mathcal{P}} K^{5'\alpha\beta} \right\}_{\alpha\beta \in D, 5'\alpha\beta \in D'} \tag{E.20}$$

where, for examples

$$\mathcal{H}_{\mathcal{P}} \sigma^{XY} = \sum_{id \in \mathcal{S}(XY)} \mathbb{E}^T(id) \mathcal{H}_{\sigma}(\theta_m, \theta_d; \delta\theta_j), \tag{E.21}$$

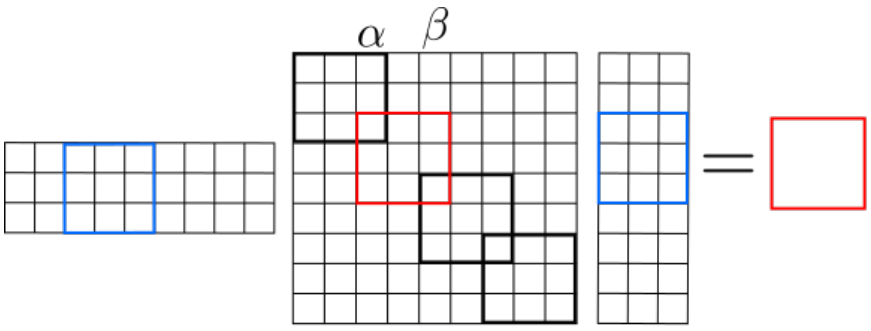
with $\mathcal{H}_{\sigma}(\rho_m, \rho_d; \cdot)$ that as been defined in (E.13). We can now define the Hessian matrix of the KLd with respect to the parameter set $P$ by

$$[\text{Hess}_P D_{KL}]_{(\cdot,j)} = \text{vec}(\mathcal{H}_{\mathcal{P}}(\rho_m, \rho_d; \mathbf{p}_j)), \tag{E.22}$$

with $\mathbf{p}_j) = (\sigma(\mathcal{P}_j, \mathcal{S}), K(\mathcal{P}_j, \mathcal{S}))$ and $\mathcal{P}_j$ is in bijection with $P_j$ which satisfies

$$[P_i]_j = 1 \text{ if and only if } i = j,\ [P_i]_j = 0 \text{ otherwise.}$$

For sake of completeness, in figure (E.1), we report a schematic representation of the dimer extraction matrix $\mathbb{E}(\alpha\beta,\mathcal{S}) \in \mathbb{R}^{N\times 42}$, where $N = 24n - 18$ for a $n$ long sequence $\mathcal{S}$.




$$\mathbb{E}(\alpha\beta,\mathcal{S})^T K(\mathcal{S})\,\mathbb{E}(\alpha\beta,\mathcal{S}) = [K(\mathcal{S})]_{\alpha\beta}$$

Figure E.1 – Schematic representation of the product $\mathbb{E}(\alpha\beta,\mathcal{S})^T \partial_K D_{KL} \mathbb{E}(\alpha\beta,\mathcal{S})$, $\mathbb{E}(\alpha\beta,\mathcal{S})$ is zero but in the blue square which is in fact the identity matrix and $\partial_K D_{KL}$ is in general dense.

# F External load acting on a single base

We first start by presenting some useful computations. Let $\mathbf{g} \in \mathrm{SE}(3)$, we want to find the relation between the infinitesimal perturbation of $\mathbf{g}$ and infinitesimal perturbation of $\mathbf{g}^{-1}$. In the case of left infinitesimal perturbation, let define $\delta\mathbf{g} = \mathbf{g}\mathcal{T}\boldsymbol{\theta}$ and $\delta\mathbf{g}^{-1} = \mathbf{g}^{-1}\mathcal{T}\boldsymbol{\phi}$ and compute

$$\begin{aligned} \delta(\mathbf{g}^{-1}\mathbf{g}) &= \delta\mathbf{g}^{-1}\mathbf{g} + \mathbf{g}^{-1}\delta\mathbf{g} = \mathbf{g}^{-1}\mathcal{T}\boldsymbol{\phi}\mathbf{g} + \mathbf{g}^{-1}\mathbf{g}\mathcal{T}\boldsymbol{\theta} && \text{(F.1)} \\ &= \mathcal{T}\mathrm{Ad}_g^{-1}\boldsymbol{\phi} + \mathcal{T}\boldsymbol{\theta} && \text{(F.2)} \end{aligned}$$

as $\delta(\mathbf{g}^{-1}\mathbf{g}) = I$ we have that the above perturbation should vanish, thus we obtain that $\boldsymbol{\phi} = -\mathrm{Ad}_{\mathbf{g}}\boldsymbol{\theta}$. Now, we can compute

$$\begin{aligned} \delta a_n &= -\mathbf{g}_n^{-1}\mathcal{T}\left(\mathrm{Ad}_{\mathbf{g}_n}\boldsymbol{\theta}_n\right)\mathbf{g}_{n+1} + \mathbf{g}_n^{-1}\mathcal{T}\left(\mathrm{Ad}_{\mathbf{g}_{n+1}}\boldsymbol{\theta}_{n+1}\right)\mathbf{g}_{n+1} \\ &= \mathbf{g}_n^{-1}\mathcal{T}\left(\mathrm{Ad}_{g_{n+1}}\boldsymbol{\theta}_{n+1} - \mathrm{Ad}_{g_n}\boldsymbol{\theta}_n\right)\mathbf{g}_{n+1} \\ &= a_n\mathcal{T}\left(\boldsymbol{\theta}_{n+1} - \mathrm{Ad}_{a_n}^{-1}\boldsymbol{\theta}_n\right). \end{aligned}$$

In the above computation we have used the relation between left and right infinitesimal perturbations (2.29).

## F.1 The reading strand case

Let us fix the $m$–th base on the reading strand. The key point for deriving the total force and couple acting on a single base, is to use the chain rule on the bichain energy (2.72) with the additional conditions:

1. the base–pair frames $\mathbf{g}_n$ does not have to move $\forall n \neq m$,
2. the base frames $\mathbf{g}_n^+$ does not have to move $\forall n \neq m$,

3. the base frames $\mathbf{g}_n^-$ does not have to move $\forall n$.

The latter conditions implies that the perturbation of these rigid–bodies must be equal to zeros. More over conditions 2) and 3) implies that the perturbation for the intra–base–pair rigid–body displacement should also be zeros for $n \neq m$. For $n = m$ the perturbation of the intra rigid–body motion $\mathcal{P}_m$ will be written simply as

$$\delta\mathcal{P}_m = \mathcal{P}_m\mathcal{T}(\phi_m^+ - \mathrm{Ad}_{\mathcal{P}_m}^{-1}\phi_m^-) = \mathcal{P}_m\mathcal{T}\phi_m^+ \Rightarrow \phi_m^{\mathcal{P}} = \phi_m^+. \tag{F.3}$$

Under the conditions 1)-3) the first order perturbation of the bichain energy (2.72) becomes

$$D_rE(\mathbf{g},\mathcal{P})\Phi^+ = (-\zeta_{m+1} + \mathrm{Ad}_{a_{m-1}}^T\zeta_m)\cdot\phi_m + \zeta_m^{\mathcal{P}}\cdot\phi_m^+, \tag{F.4}$$

where $\Phi_m^+ = (\phi_m, \phi_m^+)$. The chain rule is not finished yet as we need to find the linear change of variable that maps the perturbation $\phi_m$ to the perturbation $\phi_m^+$. This is possible using the following equation

$$\mathbf{g}_m^+ = \mathbf{g}_m\mathcal{P}_m^H, \tag{F.5}$$

where

$$\mathcal{P}_m^H \equiv \mathcal{P}^H(\eta_m, \mathrm{w}_m) = \begin{bmatrix} \mathrm{P}^{\frac{1}{2}}(\eta_m) & \frac{1}{2}\mathrm{w}_m \\ 0 & 1 \end{bmatrix}, \tag{F.6}$$

is the "half" rigid body motion defining the base–pair frame. Using equation (F.5) we want to relate the perturbation of $\mathbf{g}_m^+$ to the perturbation of $\mathbf{g}_m$ but we need to find the relation between perturbation of the rigid body motion $\mathcal{P}_m^H$ and the perturbation of the $\mathcal{P}_m$. For sake of simplicity let us first drop the subscripts $m$ in all the following computations.

### F.1.1 SE(3) perturbation of the matrix $\mathcal{P}^H$

We first use entry–entry matrix derivation to compute the derivative of $\mathcal{P}^H$ with respect to the coordinates $y = (\eta, \mathrm{w})$:

$$\partial_y\mathcal{P}^H(\delta\eta,\delta\mathrm{w}) = \begin{bmatrix} \partial_y\mathrm{P}^{\frac{1}{2}}(\eta_m) & \partial_y\frac{1}{2}\mathrm{w} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} [M\mathbb{P}_1\delta\eta\times]\mathrm{P}^{\frac{1}{2}} & \frac{1}{2}\delta\mathrm{w} \\ 0 & 0 \end{bmatrix}. \tag{F.7}$$

We can now write the matrix $\partial_y\mathcal{P}^H(\delta\eta,\delta\mathrm{w})$ as follow

$$\partial_y\mathcal{P}^H(\delta\eta,\delta\mathrm{w}) = \begin{bmatrix} [M\mathbb{P}_1\delta\eta\times] & X \\ 0 & 0 \end{bmatrix}\mathcal{P}^H, \tag{F.8}$$

where $X = \frac{1}{2}\left(\delta \mathrm{w} + [\mathrm{w}\times] M\mathbb{P}_1 \delta\eta\right)$. We recognize now that the matrix (F.8) can be written using the linear transformation $\mathcal{T}$ in the following way

$$\begin{bmatrix} [M\mathbb{P}_1\delta\eta\times] & X \\ 0 & 0 \end{bmatrix} = \mathcal{T}\left(\begin{array}{c} M\mathbb{P}_1\delta\eta \\ \frac{1}{2}\left(\delta \mathrm{w} + [\mathrm{w}\times] M\mathbb{P}_1\delta\eta\right) \end{array}\right) = \mathcal{T}\left(\begin{bmatrix} M\mathbb{P}_1 & 0 \\ \frac{1}{2}[\mathrm{w}\times] M\mathbb{P}_1 & \frac{1}{2}\mathrm{I} \end{bmatrix}\begin{bmatrix} \delta\eta \\ \delta\mathrm{w} \end{bmatrix}\right). \quad \text{(F.9)}$$

We recall that $y = (\eta, \mathrm{w})$ are the coordinates for the intra rigid body displacement $\mathcal{P} \equiv \mathcal{P}(y)$, and we also know that if $\delta\mathcal{P} = \mathcal{T}\theta^{\mathcal{P}}\mathcal{P}$, the relation between the perturbation of the coordinates and the perturbation in the group are relate through the linear mapping, i.e, $\theta^{\mathcal{P}} = \mathbb{L}_y \delta y$, see for instance (2.87). We expand then the very right part of equation (F.9) to obtain the perturbation of the matrix $\mathcal{P}^H$ in term of the perturbation of $\mathcal{P}$.

$$\mathcal{T}\left(\begin{bmatrix} M\mathbb{P}_1 & 0 \\ \frac{1}{2}[\mathrm{w}\times] M\mathbb{P}_1 & \frac{1}{2}\mathrm{I} \end{bmatrix}\begin{bmatrix} \delta\eta \\ \delta\mathrm{w} \end{bmatrix}\right)\mathcal{P}^H = \mathcal{T}(\mathbb{T}_y^+ \mathbb{L}_y \delta y)\mathcal{P}^H = \mathcal{T}(\mathbb{T}_y^+ \theta^{\mathcal{P}})\mathcal{P}^H, \quad \text{(F.10)}$$

where

$$\mathbb{T}_y^+ = \begin{bmatrix} M & 0 \\ 0 & \frac{1}{2}\mathrm{P}^{\frac{1}{2}} \end{bmatrix} \quad \text{(F.11)}$$

We thus obtain that

$$\theta^H = \mathbb{T}_y^+ \theta^{\mathcal{P}} \Longleftrightarrow \phi^H = \mathrm{Ad}_{\mathcal{P}^H}^{-1} \mathbb{T}_y^+ \mathrm{Ad}_{\mathcal{P}} \phi^{\mathcal{P}}. \quad \text{(F.12)}$$

### F.1.2 Closed form expression of force and torque on a base $\mathbf{g}^+$

Using the relation (F.12) for the perturbation *on the right* of the matrix $\mathcal{P}^H$ and the fact that under the conditions 1)-3) the perturbation of the intra displacement is totally defined by the perturbation of the base $\mathbf{g}^+$, we obtain

$$\phi_m = \left(\mathrm{Ad}_{\mathcal{P}_m^H} - \mathbb{T}_{y_m}^+ \mathrm{Ad}_{\mathcal{P}_m}\right)\phi_m^+ = \mathbb{W}_m \phi_m^+. \quad \text{(F.13)}$$

We obtain finally the expression for the total force and torque acting on a single base by replacing (F.13) into the perturbation (F.4),

$$D_r E(\mathbf{g}, \mathcal{P})\phi_m^+ = \left\{\mathbb{W}_m^T(-\zeta_{m+1} + \mathrm{Ad}_{a_{m-1}}^T \zeta_m) + \zeta_m^{\mathcal{P}}\right\} \cdot \phi_m^+, \quad \text{(F.14)}$$

where

$$\begin{aligned} \zeta_{m+1} &= \mathbb{L}_{x_m}^{-T} \partial_{x_m} w_n^x(x_n), \\ \zeta_m^{\mathcal{P}} &= \mathrm{Ad}_{\mathcal{P}_m}^T \mathbb{L}_{y_m}^{-T} \partial_{y_m}\left(w_n^y(y_n) + w_{n-1}^y(y_{n-1})\right),\ w_0 = 0 \text{ and } w_N = 0. \\ \mathbb{W}_m &= \mathrm{Ad}_{\mathcal{P}_m^H} - \mathbb{T}_{y_m}^+ \mathrm{Ad}_{\mathcal{P}_m}. \end{aligned}$$

## F.2 The complementary strand case

To get the force and torque on a base on the complementary we will first observe state the new conditions on the bichain :

1c) the basepair frames $\mathbf{g}_n$ does not have to move $\forall n \neq m$,

2c) the base frames $\mathbf{g}_n^-$ does not have to move $\forall n \neq m$,

3c) the base frames $\mathbf{g}_n^+$ does not have to move $\forall n$.

The first and simple implication is that the perturbation of the intra displacement will be written as

$$\phi_m^{\mathcal{P}} = -\mathrm{Ad}_{\mathcal{P}_m}^{-1} \phi_m^-. \tag{F.15}$$

Under the conditions 1c)-3c) and the previous remark the first order perturbation of the bichain energy become

$$D_r E(\mathbf{g}, \mathcal{P})\Phi^- = (-\zeta_{m+1} + \mathrm{Ad}_{a_{m-1}}^T \zeta_m) \cdot \phi_m - \mu_m^{\mathcal{P}} \cdot \phi_m^-, \tag{F.16}$$

where $\Phi_m^- = (\phi_m, \phi_m^-)$. We will now relate the perturbation of $\phi_m$ to the perturbation $\phi_m^-$ using the relation

$$\mathbf{g}_m^- = \mathbf{g}_m \mathbf{p}_m^{-H}, \tag{F.17}$$

where

$$\mathcal{P}_m^{-H} \equiv \mathcal{P}^{-H}(\eta_m, \mathrm{w}_m) = \begin{bmatrix} \mathrm{P}^{\frac{T}{2}}(\eta_m) & -\frac{1}{2}\mathrm{w}_m \\ 0 & 1 \end{bmatrix} \tag{F.18}$$

By using the same procedure used to obtain equation (F.12) we find that

$$\mathbb{T}_y^- = \begin{bmatrix} -\mathrm{P}^{\frac{T}{2}} M & 0 \\ \frac{1}{2}[\mathrm{w}\times] M^{-T} M & -\frac{1}{2}\mathrm{P}^{\frac{T}{2}} \end{bmatrix}, \tag{F.19}$$

and thus that

$$\theta^{-H} = \mathbb{T}_y^- \theta^{\mathcal{P}} \Longleftrightarrow \phi^{-H} = \mathrm{Ad}_{\mathcal{P}^{-H}}^{-1} \mathbb{T}_y^- \mathrm{Ad}_{\mathcal{P}} \phi^{\mathcal{P}} = -\mathrm{Ad}_{\mathcal{P}^{-H}}^{-1} \mathbb{T}_y^- \phi^- \tag{F.20}$$

### F.2.1 Closed form expression of force and torque on a base $\mathbf{g}^-$

Using the relation (F.20) for the perturbation on the right of the matrix $\mathcal{P}_m^{-H}$ and the fact that under the conditions 1c)-3c) the perturbation of the intra displacement is

totally defined by the perturbation of the base $\mathbf{g}_m^-$, we obtain

$$\phi_m = (\mathrm{Ad}_{\mathcal{P}_m^{-H}} + \mathbb{T}_{y_m}^-)\phi_m^- = \mathbb{C}_m\phi_m^-. \tag{F.21}$$

We obtain finally the expression for the total force and torque at a single base by replacing (F.21) into the perturbation (F.16),

$$D_r E(\mathbf{g}, \mathcal{P})\phi_m^- = \left\{\mathbb{C}_m^T(-\zeta_{m+1} + \mathrm{Ad}_{a_{m-1}}^T\zeta_m) - \mu_m^{\mathcal{P}}\right\} \cdot \phi_m^-, \tag{F.22}$$

where

$$\begin{aligned}
\zeta_{m+1} &= \mathbb{L}_{x_m}^{-T}\partial_{x_m} w_m^x(x_m), \\
\mu_m^{\mathcal{P}} &= \mathbb{L}_{y_m}^{-T}\partial_{y_m}\left(w_m^y(y_m) + w_{m-1}^y(y_m)\right),\ w_0 = 0 \text{ and } w_N = 0. \\
\mathbb{C}_m &= \mathrm{Ad}_{\mathcal{P}_m^{-H}} + \mathbb{T}_{y_m}^-.
\end{aligned}$$

# Index

# Bibliography


[1] H. Akaike. “A new look at the statistical model identification”. In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. ISSN: 0018-9286. DOI: 10.1109/TAC.1974.1100705.

[2] H. Akaike. “Information theory and an extension of the maximum likelihood principle”. In: *2nd International Symposium on Information Theory*. Budapest, Hungary: Akadmiai Kiad, 1971.

[3] C. Altona and M. Sundaralingam. “Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation”. In: *Journal of the American Chemical Society* 94.23 (1972), pp. 8205–8212. DOI: 10.1021/ja00778a043.

[4] A. Balaceanu et al. “The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA.” In: *The journal of physical chemistry letters* 8 1 (2017), pp. 21–28.

[5] F. Battistini et al. “Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA”. In: *Nucleic Acids Research* 42.18 (Sept. 2014), pp. 11304–11320. ISSN: 0305-1048. DOI: 10.1093/nar/gku809. eprint: http://oup.prod.sis.lan/nar/article-pdf/42/18/11304/17423804/gku809.pdf. URL: https://dx.doi.org/10.1093/nar/gku809.

[6] N. B. Becker and R. Everaers. “DNA nanomechanics: How proteins deform the double helix”. In: *J. Chem. Phys.* 130.13 (2009). ISSN: 00219606. DOI: 10.1063/1.3082157. arXiv: arXiv:0809.3938v2.

[7] N. B. Becker and R. Everaers. “DNA Nanomechanics in the Nucleosome”. In: *Structure* 17.4 (2009), pp. 579–589. ISSN: 09692126. DOI: 10.1016/j.str.2009.01.013. URL: http://linkinghub.elsevier.com/retrieve/pii/S0969212609000860.

[8] D.L. Beveridge et al. “Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonuleotides. I. Research Design and Results on d(CpG) Steps”. In: *Biophysical Journal* 87 (2004), pp. 3799 –3813.

[9] C. A. Bewley, A. M. Gronenborn, and G. M. Clore. "Minor Grooce-binding architectural proteins: Structure, Function, and DNA Recognition". In: *Annual Review of Biophysics and Biomolecular Structure* 27.1 (1998), pp. 105–131. DOI: 10.1146/annurev.biophys.27.1.105.

[10] B. R. Brooks et al. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". In: *Journal of Computational Chemistry* 4.2 (1983), pp. 187–217. DOI: 10.1002/jcc.540040211. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540040211. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540040211.

[11] D. A. Case et al. "The Amber biomolecular simulation programs". In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1668–1688. DOI: 10.1002/jcc.20290. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20290. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20290.

[12] G. S. Chirikjian. *Stochastic Models, Information Theory, and Lie Groups, Volume 2*. Birkhuser Basel, 2012.

[13] L. De Bruin and J. H. Maddocks. "cgDNAweb: a web interface to the cgDNA sequence-dependent coarse-grain model of double-stranded DNA". In: *Nucleic Acids Research* 46 (2018). http://cgdnaweb.epfl.ch, W5 –W10. DOI: 10.1093/nar/gky351.

[14] R. De Maesschalck, D Jouan-Rimbaud, and D. L. Massart. "The Mahalanobis distance". In: *Chemometrics and Intelligent Laboratory Systems* 50.1 (2000), pp. 1 –18. ISSN: 0169-7439. DOI: https://doi.org/10.1016/S0169-7439(99)00047-7. URL: http://www.sciencedirect.com/science/article/pii/S0169743999000477.

[15] S. B. Dixit et al. "Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonuleotides. II. Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps". In: *Biophysical Journal* 87 (2005), pp. 3721 –3740.

[16] L. Dostl et al. "Partial B-to-A DNA Transition upon Minor Groove Binding of Protein Sac7d Monitored by Raman Spectroscopy". In: *Biochemistry* 43.30 (2004), pp. 9600–9609. DOI: 10.1021/bi049192r.

[17] P. J. Flory. "Moments of end-to-end vector of a chain molecule, its persistence and distribution". In: *National Academy of Sciences* 70 (6 1973), pp. 1818–1823.

[18] J. Glowacki. "Computation and Visualization for Multiscale Modelling of DNA Mechanics". # 6977. PhD thesis. EPFL, 2016.

[19] O. Gonzalez, D. Petkevičiute, and J. H. Maddocks. "A sequence-dependent rigid-base model of DNA". In: *J. Chem. Phys.* 138 (2013). ISSN: 00219606. DOI: 10.1063/1.4789411.

[20] O. Gonzalez et al. “Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA”. In: *Multiscale Modeling and Simulation* 15.3 (2017), pp. 1073–1107. DOI: 10.1137/16M1086091.

[21] A. E. Grandchamp. “On the statistical physics of chains and rods, with application to multi-scale sequence-dependent DNA modelling”. # 7062. PhD thesis. EPFL, 2016.

[22] T. E. Haran and U. Mohanty. “The unique structure of A-tracts and intrinsic DNA bending”. In: *Quarterly Reviews of Biophysics* 42.1 (2009), 41–81. DOI: 10.1017/S0033583509004752.

[23] B. Hartmann, D. Piazzola, and R. Lavery. “BI-BII transitions in B-DNA”. In: *Nucleic acids research* 21.3 (1993), pp. 561–568. DOI: 09798797.

[24] I. Ivani et al. “Parmbsc1: a refined force field for DNA simulations”. In: *Nat. Methods* 13.1 (2015). ISSN: 1548-7091. DOI: 10.1038/nmeth.3658. URL: http://www.nature.com/doifinder/10.1038/nmeth.3658.

[25] E. T. Jaynes. *Information Theory and Statistical Mechanics*. Vol. 106. American Physical Society, 1957, pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL: http://link.aps.org/doi/10.1103/PhysRev.106.620.

[26] E. T. Jaynes. *Information Theory and Statistical Mechanics. II*. Vol. 108. American Physical Society, 1957. DOI: 10.1103/PhysRev.108.171.

[27] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Apr. 2003. ISBN: 0521592712.

[28] H. Jeffreys. “An invariant form for the prior probability in estimation problems”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186.1007 (1946). DOI: 10.1098/rspa.1946.0056.

[29] F. Jianqing and Y. Qiwei. *Nonlinear Time Series*. Springer-Verlag New York, 2003.

[30] A. Jones et al. “Protein-DNA interactions: a structural analysis”. In: *Journal of Molecular Biology* 287.5 (1999), pp. 877 –896. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1999.2659. URL: http://www.sciencedirect.com/science/article/pii/S0022283699926591.

[31] D. Jost et al. “6 - A Polymer Physics View on Universal and Sequence-Specific Aspects of Chromosome Folding”. In: *Nuclear Architecture and Dynamics*. Ed. by C. Lavelle and J-M. Victor. Vol. 2. Translational Epigenetics. Boston: Academic Press, 2018, pp. 149 –169. DOI: https://doi.org/10.1016/B978-0-12-803480-4.00006-5. URL: http://www.sciencedirect.com/science/article/pii/B9780128034804000065.

[32] O. Kratky and G. Porod. “Rntgenuntersuchung gelster Fadenmolekle”. In: *Recueil des Travaux Chimiques des Pays-Bas* 68 (1949), pp. 1106–1122.

[33] S. Kullback. *Information Theory and Statistics*. Wiley, 1959.

[34] S. Kullback and R. A. Leibler. “On information and sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[35] F. Lankas et al. “DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations”. In: *Biophysical Journal* 85.5 (2003), pp. 2872–2883.

[36] F. Lankas et al. “On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations.” In: *Phys. Chem. Chem. Phys.* 11.45 (2009), pp. 10565–88. ISSN: 1463-9084. DOI: 10.1039/b919565n. URL: http://www.ncbi.nlm.nih.gov/pubmed/20145802.

[37] R. Lavery et al. “A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA”. In: *Nucleic Acids Res.* 38.1 (2009), pp. 299–313. ISSN: 03051048. DOI: 10.1093/nar/gkp834.

[38] R. Lavery et al. “Conformational analysis of nucleic acids revisited: Curves+”. In: *Nucleic Acids Res.* 37 (2009), pp. 5917–5929. ISSN: 03051048. DOI: 10.1093/nar/gkp608.

[39] A. R. Leach. *Molecular modelling : principles and applications*. 2nd ed. Pearson/Prentice Hall, Apr. 2009. ISBN: 0582382106. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0582382106.

[40] L. Lercher et al. “Structural insights into how 5-hydroxymethylation influences transcription factor binding”. In: *Chem. Commun.* 50 (15 2014), pp. 1794–1796.

[41] X-J. Lu and W. K. Olson. “3DNA: a software package for the analysis, rebuilding and visualization of threedimensional nucleic acid structures”. In: *Nucleic Acids Research* 31.17 (Sept. 2003), pp. 5108–5121. ISSN: 0305-1048. DOI: 10.1093/nar/gkg680. eprint: http://oup.prod.sis.lan/nar/article-pdf/31/17/5108/4023165/gkg680.pdf. URL: https://dx.doi.org/10.1093/nar/gkg680.

[42] A. J. Majda and X. Wang. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge: Cambridge Univ. Press, 2006. URL: https://cds.cern.ch/record/992213.

[43] J. C. Marini et al. “Bent helical structure in kinetoplast DNA”. In: *Proceedings of the National Academy of Sciences* 79.24 (1982), pp. 7664–7668. ISSN: 0027-8424. DOI: 10.1073/pnas.79.24.7664. eprint: https://www.pnas.org/content/79/24/7664.full.pdf. URL: https://www.pnas.org/content/79/24/7664.

[44] R. C. Maroun and W. K. Olson. “Sequence-Dependent Persistence Lengths of DNA”. In: *Biopolymers* 13.27 (1988), p. 585603.

[45] J. S. Mitchell et al. “Sequence-Dependent Persistence Lengths of DNA”. In: *Journal of Chemical Theory and Computation* 13 (2017), pp. 1539–1555. DOI: 10.1021/acs.jctc.6b00904.

[46] Z. Morvek, S. Neidle, and B. Schneider. “Protein and drug interactions in the minor groove of DNA”. In: *Nucleic Acid Research* 30.5 (2002), pp. 1182–1191.

[47] L. Nekludova and C. O. Pabo. “Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes”. In: *Proceedings of National Academy of Sciences of the United States of America* 91.15 (1994), pp. 6948–6952.

[48] T. T. Ngo et al. “Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability”. In: *Nature Communication* 7 (Feb. 2016), p. 10813.

[49] C. Oguey, N. Foloppe, and B. Hartmann. “Understanding the Sequence-Dependence of DNA Groove Dimensions: Implications for DNA Interactions”. In: *PLoS One* 5.12 (2010), e15931. DOI: 10.1371/journal.pone.0015931.

[50] W. K. Olson et al. “A standard reference frame for the description of nucleic acid base-pair geometry11Edited by P. E. Wright22This is a document of the Nomenclature Committee of IUBMB (NC-IUBMB)/IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN), whose members are R. Cammack (chairman), A. Bairoch, H.M. Berman, S. Boyce, C.R. Cantor, K. Elliott, D. Horton, M. Kanehisa, A. Kotyk, G.P. Moss, N. Sharon and K.F. Tipton.” In: *Journal of Molecular Biology* 313.1 (2001), pp. 229 –237. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.2001.4987. URL: http://www.sciencedirect.com/science/article/pii/S0022283601949873.

[51] W. K. Olson et al. “DNA sequence-dependent deformability deduced from protein-DNA crystal complexes”. In: *Proceedings of the National Academy of Sciences* 95.19 (1998), pp. 11163–11168. ISSN: 0027-8424. DOI: 10.1073/pnas.95.19.11163. eprint: https://www.pnas.org/content/95/19/11163.full.pdf. URL: https://www.pnas.org/content/95/19/11163.

[52] M. Pasi et al. “ ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA”. In: *Nucleic Acids Res.* 42.19 (2014), pp. 12272–12283. ISSN: 0305-1048. DOI: 10.1093/nar/gku855. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku855.

[53] D. A. Pearlman et al. “AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules”. In: *Computer Physics Communications* 91.1 (1995), pp. 1 –41. ISSN: 0010-4655. DOI: https://doi.org/10.1016/0010-4655(95)00041-D. URL: http://www.sciencedirect.com/science/article/pii/001046559500041D.

[54] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Version 20121115. Nov. 2012. URL: http://www2.imm.dtu.dk/pubdb/p.php?3274.

[55] D Petkeviciute. “A DNA Coarse-Grain Rigid Base Model and Parameter Estimation from Molecular Dynamics Simulations”. # 5520. PhD thesis. EPFL, 2012.

[56] G. Portella, F. Battistini, and M. Orozco. “Understanding the Connection between Epigenetic DNA Methylation and Nucleosome Positioning from Computer Simulations”. In: *PLOS Computational Biology* 9.11 (Nov. 2013), pp. 1–7. DOI: 10.1371/journal.pcbi.1003354. URL: https://doi.org/10.1371/journal.pcbi.1003354.

[57] P. Prabakaran et al. “Classification of Protein-DNA Complexes Based on Structural Descriptors”. In: *Structure* 14.9 (2006), pp. 1355 –1367. ISSN: 0969-2126. DOI: https://doi.org/10.1016/j.str.2006.06.018. URL: http://www.sciencedirect.com/science/article/pii/S0969212606003029.

[58] A. Prez, F. J. Luque, and M. Orozco. “Dynamics of B-DNA on the Microsecond Time Scale”. In: *Journal of the American Chemical Society* 129.47 (2007), pp. 14739–14745. DOI: 10.1021/ja0753546.

[59] A. Prez et al. “Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of $\alpha/\gamma$ Conformers”. In: *Biophysical Journal* 92.11 (2007), pp. 3817 –3829. ISSN: 0006-3495. DOI: https://doi.org/10.1529/biophysj.106.097782. URL: http://www.sciencedirect.com/science/article/pii/S0006349507711827.

[60] C. Rivetti, C. Walker, and C.J Bustamante. “Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility”. In: *Journal of Molecular Biology* 280 (1998), p. 4159.

[61] O. Rodrigues. “Des lois gomtriques qui rgissent les dplacements d’un systme solide dans l’space, et de la variation des coordonnes provenant de ces dplacement considrs indpendamment des causes qui peuvent les produire”. In: *Journal de mathmatique pures et appliques* 5 (1840), pp. 380–400.

[62] R. Rohs et al. “Origins of Specificity in Protein-DNA Recognition”. In: *Annual Review of Biochemistry* 79.1 (2010). PMID: 20334529, pp. 233–269. DOI: 10.1146/annurev-biochem-060408-091030.

[63] R. Rohs et al. “The role of DNA shape in proteinDNA recognition”. In: *Nature* 461 (2009), pp. 1248–1253.

[64] J. A. Schellman. “Flexibility of DNA”. In: *Biopolymers* 13 (1 1974), pp. 217–226.

[65] J. A. Schellman and S. C. Harvey. “Static contributions to the persistence length of DNA and dynamic contributions to DNA curbature”. In: *Biophysical Chemistry* 55.1-2 (1995), pp. 95–114.

[66] P. M. D. Severin et al. “Cytosine methylation alters DNA mechanical properties”. In: *Nucleic Acids Research* 39.20 (July 2011), pp. 8740–8751. ISSN: 0305-1048. DOI: 10.1093/nar/gkr578. eprint: http://oup.prod.sis.lan/nar/article-pdf/39/20/8740/16778983/gkr578.pdf. URL: https://dx.doi.org/10.1093/nar/gkr578.

[67] O. Sorkine-Hornung and M. Rabinovich. *Least-Squares Rigid Motion Using SVD*. Tech. rep. Department of Computer Science, ETH Zurich, 2017.

[68] "Structure of a B-DNA dodecamer: conformation and dynamics". In: *Proceedings of the National Academy of Sciences of the United States of America* 78.4 (1981), pp. 2179–2183.

[69] M. Suzuki and N. Yagi. "An In-the-Groove View of DNA Structures in Complexes with Proteins". In: *Journal of Molecular Biology* 255.5 (1996), pp. 677 –687. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1996.0055. URL: http://www.sciencedirect.com/science/article/pii/S0022283696900558.

[70] "The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids". In: *Biophysical journal* 63.3 (1992), pp. 751–759.

[71] B. Theveny et al. "Local variations of curvature and flexibility along DNA molecules analyzed from electron micrographs". In: *Structure and Expression (Vol. 3): DNA Bending and Curvature*. New York: Adenine Press, 1988.

[72] M. Y Tolstorukov, R. L. Jernigan, and V. B. Zhurkin. "ProteinDNA Hydrophobic Recognition in the Minor Groove is Facilitated by Sugar Switching". In: *Journal of Molecular Biology* 337.1 (2004), pp. 65 –76. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2004.01.011. URL: http://www.sciencedirect.com/science/article/pii/S0022283604000531.

[73] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997. ISBN: 0898713617.

[74] E. N. Trifonov, R. K. Tan, and S. C. Harvey. "Static persistence length of DNA". In: *Structure and Expression (Vol. 3): DNA Bending and Curvature*. New York: Adenine Press, 1988.

[75] Cédric Vaillant et al. "DNA physical properties determine nucleosome occupancy from yeast to fly". In: *Nucleic Acids Research* 36.11 (May 2008), pp. 3746–3756. ISSN: 0305-1048. DOI: 10.1093/nar/gkn262. eprint: http://oup.prod.sis.lan/nar/article-pdf/36/11/3746/16749095/gkn262.pdf. URL: https://dx.doi.org/10.1093/nar/gkn262.

[76] M. Vologodskaia and A. Vologodskii. "Contribution of the intrinsic curvature to measured DNA persistence length11Edited by I. Tinoco". In: *Journal of Molecular Biology* 317.2 (2002), pp. 205 –213. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.2001.5366. URL: http://www.sciencedirect.com/science/article/pii/S0022283601953665.

[77] E. Westhof and M. Sundaralingam. "A method for the analysis of puckering disorder in five-membered rings: the relative mobilities of furanose and proline rings and their effects on polynucleotide and polypeptide backbone flexibility". In: *Journal of the American Chemical Society* 105.4 (1983), pp. 970–976. DOI: 10.1021/ja00342a054.

[78] L. Wolff, N. B. Becker, and R. Everaers. “Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials”. In: *Nucleic Acids Research* 34.19 (Oct. 2006), pp. 5638–5649. ISSN: 0305-1048. DOI: 10.1093/nar/gkl683. eprint: http://oup.prod.sis.lan/nar/article-pdf/34/19/5638/16761115/gkl683.pdf. URL: https://dx.doi.org/10.1093/nar/gkl683.

# ALESSANDRO PATELLI

Rue de la Paix 19 ⋄ Renens, 1020, Vaud
alessandro.patelli@epfl.ch

## EDUCATION

**PhD Thesis in mathematics, EPFL** February 2015–April 2019
*A sequence—dependent coarse–grain model of B–DNA with explicit description of bases and phosphate groups parametrised from large scale Molecular Dynamics simulations*
Under the supervisions of: Prof. J. H. Maddocks.

**Master degree in applied Mathematics, EPFL** September 2012–July 2014
Master Thesis: *Isogeometric Analysis of Electrophysiological Models on Surfaces.*
Under the supervisions of: Prof. A. Quarteroni, Dr. L. Dede', Dr. T. Lassila.

**Bachelor degree in Mathematics, EPFL** September 2008–2012

## EXPERIENCE

**EPFL** February 2014-present
*Teaching Assistant* *Lausanne, VD*

- Mathematical modelling of DNA, Master, (principal assistant).
- Analyse 3, Bachelor (principal assistant).
- Analyse Numerique, Bachelor, (principal assistant).

**CHUV, Department of Clinical Neuroscience** September 2014–December 2014
*Scientific collaborator* *Lausanne, VD*

- Acquisition of clinical data with patients.
- Design of a methodology for EEG data analysis using signal processing and graph theory.

**St. Jude Medical, (previously Endosense)** August 2013–December 2013
*Internship* *Meyrin, GE*

- Statistical analysis of lesion formation in ablation technique.

## TECHNICAL STRENGTHS

| | |
|---|---|
| **Computer Languages** | Matlab, phyton, C/C++ , bash, SLURM |
| **Software** | latex, Windows Office, Open office |
| **Operation System** | Linux, macOS, Window |
| **Tools** | Vim, Emacs |

## LANGUES

| | |
|---|---|
| **Italian** | mother tongue |
| **French** | Fluent |
| **English** | Fluent |

## INTEREST

Home brewing, photography, culinary travel.

**Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.**