# A DNA Coarse-Grain Rigid Base Model and Parameter Estimation from Molecular Dynamics Simulations

THÈSE N$^{O}$ 5520 (2012)

PRÉSENTÉE LE 19 OCTOBRE 2012
À LA FACULTÉ DES SCIENCES DE BASE
CHAIRE D'ANALYSE APPLIQUÉE
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Daiva PETKEVIČIŪTĖ

acceptée sur proposition du jury:

Prof. F. Eisenbrand, président du jury
Prof. J. Maddocks, directeur de thèse
Prof. C. Benham, rapporteur
Prof. G. Chirikjian, rapporteur
Prof. D. Kressner, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2012

# Acknowledgements

First of all, I would like to thank Prof. John Maddocks for giving me a great opportunity to work with him. I am grateful for his guidance and support all through this project and for so many things I could learn from him during these years.

The main part of this thesis is joint work also with Prof. Oscar Gonzalez of the University of Texas, and I would like to thank him for this collaboration. Dr. Richard Lavery and Dr. Krystyna Zakrzewska of the Institute for the Biology and Chemistry of Proteins at CNRS were advising me multiple times about running molecular dynamics simulations and data processing, and I really appreciate their help. I would also like to thank the jury members of my thesis, Professors Craig Benham, Gregory Chirikjian and Daniel Kressner, for their interesting comments and valuable suggestions.

It was a pleasure for me to work in the LCVMM group and to share the ideas, excitements and problems with the group members. Carine Tschanz, the secretary of the group, deserves gratitude for making our life at EPFL (and for some of us in Switzerland) easier and nicer. Philippe Caussignac was always very helpful solving my computer related problems and answering my questions. I learned a lot from Filip Lankaš and Jérémy Curuksu about molecular dynamics and other computations that had been done in the group, and I was lucky to be able to profit from their experience. In fact, part of the parameter estimation described in Chapter 4 was done in collaboration with Jérémy Curuksu. I would like to thank Jonathan Mitchell for reading and correcting parts of my thesis, Alexandre Grandchamp for his help with my French, Jarek Głowacki for his figures that I used in Chapter 2 of the thesis and, together with Marco Pasi, Ludovica Cotta-Ramusino, Mathias Carlen and Jessika Walter for various advise and help during this time. Also thanks to Julien Delafonfaine, who did his master project in our group, for the permission to use his figures. Some of the schematic figures in Chapter 2 were based on his presentation. While still naming members of the LCVMM group, I would like to thank very much to Henryk Gerlach for many kinds of technical, mathematical and moral support while writing and polishing this thesis.

I am thankful to all my friends, who made my stay at EPFL and in Lausanne so nice. Especially to Audrius Alkauskas, for all the help at my arrival to Lausanne, and to Maya Shevlyakova, for our everyday coffee breaks and intresting discussions.

Finally, I would like to thank my family for all the love and support that I always felt from them.

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

# Abstract

Sequence dependent mechanics of DNA is believed to play a central role in the functioning of the cell through the expression of genetic information. Nucleosome positioning, gene regulation, DNA looping and packaging within the cell are only some of the processes that are believed to be at least partially governed by mechanical laws. Therefore it is important to understand how the sequence of DNA affects its mechanical properties.

For exploring the mechanical properties of DNA, various discrete and continuum models have been, and continue to be, developed. A large family of these models, including the model considered in this work, assume that bases or base pairs of DNA are rigid bodies. The most standard are rigid base pair models, with parameters either obtained directly from experimental data or from Molecular Dynamics (MD) simulations. The drawback of current experimental data, such as crystal structures, is that only small ensembles of configurations are available for a small number of sequences. In contrast, MD simulations allow a much more detailed view of a larger number of DNA sequences. However, the drawback is that the results of these simulations depend on the choice of the simulation protocol and force field parameters. MD simulations also have sequence length limitations and are currently too intensive for (linear) molecules longer than a few tens of base pairs. The only way to simulate longer sequences is to construct a coarse-grain model.

The goal of this work is to construct a small parameter set that can model a sequence-dependent equilibrium probability distribution for rigid base configurations of a DNA oligomer with any given sequence of any length. The model parameter sets previously available were for rigid base pair models ignoring all the couplings beyond nearest neighbour interactions. However it was shown in previous work, that this standard model of rigid *base pair* nearest neighbour interactions is inconsistent with a (then) large scale MD simulation of a single oligomer [36]. In contrast we here show that a rigid *base* nearest neighbour, dimer sequence dependent model is a quite good fit to many MD simulations of different duration and sequence.

In fact a hierarchy of rigid base models with different interaction range and length of sequence-dependence is discussed, and it is concluded that the nearest neighbour, dimer based model is a good compromise between accuracy and complexity of the model. A full parameter set for this model is estimated. An interesting feature is that despite the dimer dependence of the parameter set, due to the phenomenon of frustration, our model predicts non local changes in the oligomer shape as a function of local changes in the sequence, down to the level of a point mutation.

**Keywords**: coarse-grain DNA, rigid base model, Jacobian, parameter extraction, molecular dynamics simulations, frustration energy.

# Résumé

La comportement mécanique de l'ADN basée sur la dépendence de sa séquence est pressentie de jouer un rôle central dans le fonctionnement de la cellule au travers de l'expression de l'information génétique. Le positionnement des nucléosomes, la régulation des gènes, la formation de boucles d'ADN et son empaquetage dans la cellule ne sont que quelques-uns des processus qui sont soupçonnés d'être au moins partiellement régis par des lois mécaniques. Par conséquent, il est important de comprendre comment la séquence de l'ADN affecte ses propriétés mécaniques.

Pour explorer les propriétés mécaniques de l'ADN, différents modèles discrets et continus ont été, et continuent d'être développés. Une grande famille de ces modèles, notamment le modèle considéré dans ce travail, suppose que les bases ou paires de bases de l'ADN sont des corps rigides. Les plus standards sont les modèles ou les paires de bases sont rigides, avec des paramètres obtenus soit directement à partir des données expérimentales, soit des simulations de dynamique moléculaire (MD). L'inconvénient de données expérimentales actuelles, telles que les structures cristallines, est que seuls les petits ensembles de configurations sont disponibles pour un petit nombre de séquences. En revanche, les simulations MD permettent une vision beaucoup plus détaillée d'un plus grand nombre de séquences d'ADN. Cependant, l'inconvénient est que les résultats de ces simulations dépendent du choix du protocole de simulation et des paramètres du potentiel. Les simulations MD ont aussi des limites de longueur de séquence et sont actuellement trop intensives pour des molécules linéaires plus longs que les quelques dizaines de paires de bases. La seule façon de prédire les valeurs des paramètres nécessaires pour des longs séquences est de construire un modèle gros grains.

L'objectif de ce travail est de construire un modèle qui génère les paramètres, dépendents de la séquence, pour une distribution stationnaire de probabilités des configurations l'ADN, et cela pour toute séquence d'ADN donnée de n'importe quelle longueur. L'ADN est modélisé comme un système de bases rigides. L'ensembles de paramètres disponibles précédemment étaient que pour des modèles qui ignorent tous les couplages au-delà de paires des bases rigides les plus proches. Toutefois, dans un travail précédent il a été montré, que ce modèle standard des interactions locales des *paires de bases* rigides est incompatible avec une simulation MD d'échelle importante (a l'époque) d'un seul oligomère [36]. Ici nous montrons que le modèle des *bases* rigides interagissants avec les bases voisines les plus proches, et avec la dépendence locale d'une séquence (dimère), donne une assez bonne approximation de distribution de probabilités pour les nombreuses simulations de dynamique moléculaire de durée différente et des séquences différentes.

En fait, une hiérarchie de modèles avec une gamme d'interactions différentes et des dépendances différentes en séquence est discutée. Il est conclu que le modèle ne regardant que les interactions des voisins les plus proches et les paramètres dépendents en séquence de dimère est un bon compromis entre la précision et la complexité du modèle. Un ensemble complet des paramètres pour ce modèle est estimé. Une caractéristique intèressante est que, malgré la dépendance en dimère de l'ensemble des paramètres, en raison du phénomène de frustration, notre modèle prévoit des modifications non locales de la forme d'un oligomère en fonction des variations locales de la séquence, jusqu'au niveau d'une mutation ponctuelle.

**Mots-clés**: modèle d'ADN gros grains, modèle des bases rigides, jacobien, extraction des paramètres, simulations de dynamique moléculaire, énergie de frustration.

# Santrauka

Manoma, kad DNR molekulės mechaninės savybės, priklausančios nuo jos sekos, veikia genų raišką ir todėl yra reikšmingos ląstelės funkcionavimui. Nukleosomų išsidėstymas, genų reguliavimas, DNR kilpų susidarymas ir DNR pakavimas lastelėje - tai tik keletas procesų, kurie, kaip manoma, yra bent iš dalies nulemti mechaninių dėsnių. Taigi svarbu suprasti kaip DNR seka veikia jos mechanines savybes.

Mechaninių DNR savybių tyrimui yra sukurta ir vis dar kuriama daug įvairių diskrečių ir tolydžių modelių. Didelė modelių grupė, kuriai priklauso ir šiame darbe aprašytas modelis, remiasi prielaida, kad DNR bazės arba bazių poros yra kieti kūnai. Dažniausiai naudojami kietų bazių porų modeliai, kurių parametrai gali buti gaunami tiesiogiai iš eksperimentinių duomenų arba iš molekulinės dinamikos (MD) simuliacijų. Šiuolaikinių eksperimentų, pavyzdžiui, kristalinių strukturų tyrimo, trūkumas yra tai, kad iš jų duomenų galima gauti tik nedidelę DNR konfiguracijų aibę, ir tik nedideliam skaičiui DNR sekų. Tuo tarpu MD simuliacijos leidžia susidaryti daug detalesnį daugelio DNR sekų vaizdą. Tačiau šio metodo rezultatai priklauso nuo pasirinkto simuliacijų protokolo ir nuo potencialo parametrų. Be to, MD simuliacijas galima atlikti tik riboto ilgio sekoms, kadangi šiuo metu skaičiavimų apimtis tiesinėms molekulėms, ilgesnėms nei keletas desimčių bazių porų, yra per didelė. Taigi norint tirti ilgesnes sekas, vienintelė išeitis - sukurti stambių komponentų modelį, kuriame pasirinkta DNR atomų grupė modeliuojama kaip vienas kietas kūnas.

Šio darbo tikslas - sukurti stambių komponentų modelį ir įvertinti jo parametrus, kurie leistų gauti stacionarų DNR konfiguracijų tikimybių skirstinį bet kokiai norimo ilgio DNR sekai. DNR molekulė yra modeliuojama kaip kietų bazių sistema. Iki šiol buvo įvertinti tik parametrai kietų bazių porų modeliams, ignoruojantiems visas sąveikas, tolimesnes nei tarp kaimyninių bazių porų. Ankstesniame darbe [36] buvo parodyta, kad šis standartinis kietų *bazių porų* modelis, vertinantis tik lokalias sąveikas, neatitiko didelės apimties MD simuliacijos vienam DNR oligomerui rezultatų. Šiame darbe parodoma, kad artimiausių kietų *bazių* sąveikų modelis, kurio parametrai priklauso nuo dimero sekos, gerai atitinka skirtingos trukmės ir įvairių DNR sekų MD simulacijų duomenis.

Darbe pateikiama hierachinė DNR kietų bazių modelių su įvairaus nuotolio sąveikomis ir parametrų priklausomybe nuo skirtingo ilgio DNR sekos šeima. Daroma išvada, kad artimiausių bazių sąveikų modeliui būdingas geras tikslumo ir sudėtingumo santykis. Darbe įvertinami visi šio modelio parametrai. Įdomi modelio savybė, kad nepaisant parametrų priklausomybės tik nuo dimero sekos, dėl frustracijos fenomeno jis gali numatyti platesnius oligomero formos pasikeitimus, sukeltus lokalių DNR sekos pokyčių, pavyzdžiui, taškinės mutacijos.

**Raktiniai žodžiai**: DNR stambių komponentų modelis, DNR kietų bazių modelis, jakobianas, parametrų vertinimas, molekulinės dinamikos simuliacijos, frustracijos energija.

# Contents

1 Introduction 1

2 Kinematics of rigid base and rigid base pair models of DNA 5
2.1 DNA sequence and configuration of rigid bases 5
2.2 Parametrisation of rotations 6
2.3 Internal coordinates of rigid bases and base pairs 8
2.4 Non-dimensionalisation and scaling 12
2.5 Change of reference strand 13
2.6 Changing the rule of embedding base frames 15

3 An oligomer dependent rigid base model 17
3.1 Distribution of DNA configurations 17
3.2 Material assumptions and symmetries 19
3.3 Moment-parameter relations 20
3.4 Rigid base pair marginal distribution 21

4 Molecular dynamics and data 23
4.1 Molecular dynamics simulations 23
4.2 ABC data set and simulation protocol 26
4.3 Oligomers with different ends, breaking hydrogen bonds 27
4.4 Estimates of the oligomer based model parameters and their convergence 29

5 The Jacobian and a Gaussian approximation 37
5.1 Bimodality in parameter distributions 37
5.2 Can the Jacobian explain bimodality? A one dimensional example 38
5.3 Can the Jacobian explain bimodality? The multidimensional case 39
5.4 A Gaussian approximation 41
5.5 Comparison of probability densities 49

6 A nearest-neighbour rigid base model 51
6.1 Mechanical assumptions 51
6.2 Oligomer-based, nearest-neighbour model 54
6.3 Fitting model parameters for given oligomer data 55
6.4 Model fitting results, approximation errors 56

**7 Dimer dependent nearest-neighbour rigid base model** **59**
7.1 Material assumptions and model parameters . . . . . . . . . . . . . . . . . . . . 59
7.2 Structure and properties of oligomer parameters . . . . . . . . . . . . . . . . . 61
7.3 Compatibility between model assumptions and the MD data . . . . . . . . . 63
7.4 Least-squares system without end data . . . . . . . . . . . . . . . . . . . . . . . . 66
7.5 Weighted least-squares system with end data . . . . . . . . . . . . . . . . . . . . 68
7.6 Dimer-based fitting . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71

**8 Parameters and tests for the dimer dependent nearest-neighbour rigid base model** **75**
8.1 The $\mathcal{P}^\dagger$ parameter set . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75
8.2 Reconstructions of example oligomers from the training set . . . . . . . . . . 82
8.3 Example reconstruction of an oligomer not from the training set . . . . . . . 94

**9 Summary, conclusions and discussion** **99**

**Appendicies:** **105**

**A A parameter set for the nearest-neighbour model of DNA** **105**
A.1 Monomer stiffness blocks $\mathsf{K}_1^\gamma$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . 105
A.2 Eigenvalues of monomer stiffness blocks $\mathsf{K}_1^\gamma$ . . . . . . . . . . . . . . . . . . . 106
A.3 Monomer weighted shape vectors $\sigma_1^\gamma$ . . . . . . . . . . . . . . . . . . . . . . . . 106
A.4 Dimer stiffness blocks $\mathsf{K}_2^{\alpha\beta}$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107
A.5 Eigenvalues of dimer stiffness blocks $\mathsf{K}_2^{\alpha\beta}$ . . . . . . . . . . . . . . . . . . . . . 111
A.6 Dimer weighted shape vectors $\sigma_2^{\alpha\beta}$ . . . . . . . . . . . . . . . . . . . . . . . . 111

**B Plots of sequence averaged nearest-neighbour stiffness blocks** **113**
B.1 Dimer averages of the nearest-neighbour stiffness matrix blocks with an assumed dimer dependence . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114
B.2 Trimer averages of the diagonal intra base pair stiffness matrix blocks . . . . 115

**Bibliography** **117**

# Chapter 1

# Introduction

Deoxyribonucleic acid (DNA) is one of the key molecules for life on our planet. As discovered in 1953 [76], it carries genetic information encoded in its sequence. In addition, the sequence influences mechanical properties of the molecule, which are believed to be important in the expression of the genetic information. Various biological processes where the sequence dependent intrinsic curvature and flexibility of DNA are believed to play a role include nucleosome positioning [68], and gene regulation [50], as well as DNA looping [67] and packaging in the cell. For this reason, understanding sequence dependent DNA mechanics is an important question and is an active research topic.

For example, in eukaryotic cells DNA is packaged by wrapping around proteins. The elementary unit of this compact structure is the nucleosome, consisting of a histone protein core with a 147 base pair DNA segment tightly wound around it a bit less than two times. Between every two nucleosomes there is a 10-50 base pair long linker DNA segment, which means that there exists some freedom for where a nucleosome is centred. It has been shown that the exact location of a nucleosome along the DNA depends on the sequence induced DNA flexibility [68, 51]. While wrapped on a nucleosome, DNA is not easily accessible to copying enzymes and other proteins. Hence the placement of nucleosomes influences gene regulation, transcription, replication and recombination [50].

In prokaryotes and viruses, instead of forming nucleosomes, DNA is folded and wrapped about itself, forming toroidal or interwound structures [26]. The formation of these supercoiled DNA structures is generally also governed by the mechanical properties of the molecule, which again influence the way proteins can approach and process the genetic material.

In general, many of the DNA-protein interactions occurring during various important biological tasks in both eukaryotic and prokaryotic cells involve a combination of chemistry and geometry, and so depend upon the mechanical ability of DNA to adopt necessary configurations [11].

There exist various methods for experimentally quantifying the sequence dependent flexibility of DNA. One popular approach is cyclisation experiments, where the rate of formation of closed loops from initially linear molecules with cohesive ends is observed [71], [70]. Other experimental studies of DNA mechanics include observing distributions of DNA configurations in crystal structures or using cryo-electron microscopy, also separating circular DNA molecules with different linking numbers, using the electrophoretic band-shift method [5], [63], [11]. At larger length scales, where sequence dependence is believed to be less important, DNA mechanical properties are investigated by atomic force microscopy [34] or by force spectroscopies, i.e. by pulling or twisting a single DNA molecule attached to a bead using magnetic or optical “tweezers" or a micropipette [63]. To be able to explain the results of such experiments, and to further use these results for predicting processes that are not experimentally observable, it is good to have an accurate mechanical model.

DNA is a very thin and long molecule: its width is approximately two nanometres ($10^{-9}$ m), while the total length of DNA in each human cell is about 2 m [11]. Modelling its mechanical properties and behaviour at different length scales accordingly requires different approaches [73]. For the length scales from several hundred to thousands of base pairs a uniform elastic rod (or worm like chain) model is considered to be a good approximation [5]. However, for shorter polymers sequence-dependence cannot be ignored. In this case sequence dependent rod models or discrete rigid base or rigid base pair models are more appropriate. Finally, for some purposes atomistic representations of DNA are preferred, for example to explore DNA interactions with ions or water. Various connexions between different models can be made [2] to relate the experimental and computational results for various length scales of DNA.

One large family of coarse-grain DNA models is based on the idea of representing several atoms by one bead or rigid body. Different such models may have a different level of detail, include different types of interactions, and may or may not account for sequence-dependent behaviour [35]. This approach enables the exploration, usually by computer simulations, of a range of DNA behaviours that would be too numerically intensive to explore using a fully atomistic model. Such behaviour includes duplex formation and fraying, hairpin formation etc. [56]

A standard coarse-grain way to model DNA is to assume that its bases or base pairs are rigid bodies. The configuration of a molecule in such a model is described by relative variables that have generally accepted, standard names, tilt, roll, twist, etc, although it should be noted that they have some variation in their precise definitions [35]. Often a probability distribution function of these internal configuration variables is considered, with a common assumption being that this distribution function admits a quadratic potential (or free) energy with sequence dependent shape and stiffness parameters, so that the associated probability distribution is Gaussian. The most widely used description is a rigid base pair model with local nearest-neighbour parameters, i.e. a model where the only interactions considered are those between neighbouring base pairs, which are assumed to be rigid. For one particular MD simulation this model was recently shown to poorly represent the data [36]. In contrast, a local rigid base model was shown to be a quite good approximation for that simulation of DNA [op. cit.].

Another family of models involve elastic rods or birods, which can actually be regarded as continuum limits of respectively rigid base pair and rigid base DNA models [3], [53]. An elastic rod model is a powerful tool to study long DNA polymers, with a possibility to account for external loads and moments, and a great deal of work has been done in this direction [3], [72], [27], [15]. For large length scales (thousands of base pairs) it can be appropriate to approximate DNA by a uniform rod and ignore the sequence dependence. However for shorter length scales (tens of base pairs), the sequence dependence can play an important role [16], and thus has to be included in the model. Accordingly, the rigid base pair and rigid base models are also useful for obtaining sequence dependent parameters for rods and birods.

Parameter extraction for coarse-grain DNA models represents a different area of research. Experimental data from crystal structure experiments with naked DNA and DNA in complexes with proteins can be used to parametrise rigid base and rigid base pair DNA models [55]. However, the existing experimental data only allow the computation of average shape parameters and on-diagonal blocks of covariance matrices assuming local dependence. In addition, the data is available only for a relatively small number of oligomers. An alternative to experimental observations for parameter fitting is provided by data taken from molecular dynamics (MD) simulations, which have been used for modelling DNA since 1983 [44]. In fact MD simulation is already quite widely used for coarse-grain DNA parameter estimation [38], [39], [40], [8].

In this work we present a new dimer-dependent nearest-neighbour rigid base model of the equilibrium probability distribution for configurations of a DNA oligomer with arbitrary sequence at length scales up to several hundreds of base pairs. The model that we develop assumes a local nearest-neighbour stiffness dependence, but predicts non local dependence of average shapes, due to the fact that individual DNA bases have to compromise in their interactions with their neighbours to find their minimal energy configuration. This non local shape dependence is not possible in a local rigid base pair model.

We then develop a method for extracting a sequence-dependent parameter set for the rigid base model, and find, in contrast to the nearest-neighbour rigid base pair description, that a rather good approximation of MD data can be obtained. For the parameter extraction we start from a large Molecular Dynamics data set, the main part of which was produced by the Ascona B-DNA Consortium [42]. As a test for our parameter set, we reconstruct a Gaussian probability distribution function for each of our training set oligomers, and then compare these reconstructions with the distributions observed from MD. We conclude that the two distributions are close for every oligomer, i.e. our reconstructions are quite good. We are also able to predict non local changes of the intrinsic shape of an oligomer, caused by a local (point) modification of its sequence. Consequently, we conclude that we have a model and associated parameter set, which is dependent only on local dimer and monomer steps, and which allows us to construct the average shape and stiffness of DNA sequences that are too long to be explored directly using MD simulations. Our model parameters can then be used, for example, to compute sequence dependent persistence lengths and looping probabilities of DNA.

This thesis has the following structure. Chapters 2 and 3 contain background material, which was mainly developed prior to our work, but which is necessary for the presentation of our model. In Chapter 2 we give a definition of the rigid base and rigid base pair DNA configuration parameters, and discuss their symmetry properties. The material of this chapter, except Section 2.6, appears in the publications [41] (which is a description of the program *Curves+*) and [36]. Then in Chapter 3 we introduce an oligomer dependent probability density function and the internal (free) energy of the rigid base DNA configurations, and explain the relations for estimating the shape and stiffness parameters, appearing in this probability density function. The material of this chapter comes from [23] and [36].

In Chapter 4 we describe the Molecular Dynamics simulations, which are the source of our training data set. The estimation of coarse-grain oligomer-dependent parameters from atomistic data, and convergence issues are also discussed.

Our assumed probability density is not Gaussian because of a Jacobian factor, which inevitably arises due to the appearance of rotation group $SO(3)$ in a rigid body description of DNA [74]. In Chapter 5 we raise and address the question: can the presence of this Jacobian factor explain the bimodal distribution of some coarse-grain DNA configuration parameters, as observed in the MD data? We also discuss some errors associated with suppressing the Jacobian in our subsequent developments. To quantify distances between probability distributions the Kullback-Leibler divergence was chosen; its properties are reviewed in Section 5.5.

The material presented in Chapters 6, 7 and 8 (except for sections 7.4, 7.5 and part of Section 7.3) is part of the joint publication [24]. In Chapters 6 and 7 a rigid base DNA mechanical model with nearest-neighbour interactions is introduced. We show that the nearest-neighbour assumptions imply a special structure in the manner in which the dimer-dependent stiffness and shape parameters are manifested in the oligomer-dependent training set data. Accordingly, we describe the procedure for, and the results of, deconvolving the best fit parameters for this model. In Chapter 8 we discuss a parameter set for our dimer-dependent nearest-neighbour model, and illustrate its properties with example data from our training set, as well as from additional simulations.

A summary of the thesis, conclusions and discussion of further work are contained in Chapter 9. In addition, there are two appendices with A) a full tabulation of our model parameter set, and B) statistics of sequence variation in the training set data. The full parameter set, together with corresponding Matlab® scripts that can be used to construct stiffness matrices and intrinsic shapes of arbitrary length oligomers, are available from the webpage http://lcvmwww.epfl.ch/cgDNA.

# Chapter 2

# Kinematics of rigid base and rigid base pair models of DNA

The notations and configuration parameters for a rigid base and rigid base pair DNA model, as presented in this chapter, have been described in [36] and [41].

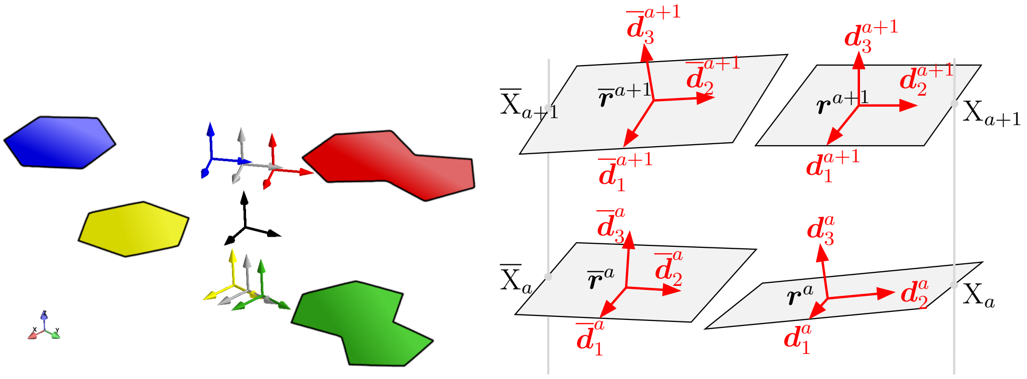
## 2.1 DNA sequence and configuration of rigid bases

In this work we consider a right-handed, double-stranded, linear B form DNA in which bases `T`, `A`, `C` and `G` are attached to two, oriented, anti-parallel backbone strands and form only the standard Watson-Crick pairs $(\mathtt{A},\mathtt{T})$ and $(\mathtt{C},\mathtt{G})$. Choosing one backbone strand as a reference, a DNA oligomer consisting of $n$ base pairs is identified with a sequence of bases $\mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$, listed in the $5'$ to $3'$ direction along the strand, where $\mathtt{X}_a \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$. The base pairs associated with this sequence are denoted by $(\mathtt{X},\overline{\mathtt{X}})_1,(\mathtt{X},\overline{\mathtt{X}})_2,\ldots,(\mathtt{X},\overline{\mathtt{X}})_n$, where $\overline{\mathtt{X}}$ is defined as the Watson-Crick complement of $\mathtt{X}$ in the sense that $\overline{\mathtt{A}}=\mathtt{T}$, $\overline{\mathtt{T}}=\mathtt{A}$, $\overline{\mathtt{C}}=\mathtt{G}$ and $\overline{\mathtt{G}}=\mathtt{C}$. The notation $(\mathtt{X},\overline{\mathtt{X}})_a$ for a base pair indicates that base $\mathtt{X}$ is attached to the reference strand while $\overline{\mathtt{X}}$ is attached to the complementary strand.

In our coarse-grain model of DNA the bases are assumed to be rigid and the configuration of an oligomer is described by position (location and orientation) of all its bases. The orientation of a rigid body can be described by fixing an orthonormal frame in it, and the location can be specified by choosing a reference point of that body. This means that the configuration of a rigid base $n$-mer $\mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$ can be given by a set of $2n$ base reference points with $2n$ attached base frames.

We use the rules of the Tsukuba convention [54] to assign a reference point $\boldsymbol{r}^a$ and a right-handed, orthonormal frame $\{\boldsymbol{d}_i^a\}$ $(i=1,2,3)$ to each base $\mathtt{X}_a \in \{\mathtt{T},\mathtt{A},\mathtt{C},\mathtt{G}\}$. The vector $\boldsymbol{d}_1^a$ points in the direction of the major groove, $\boldsymbol{d}_2^a$ points in the direction of the reference strand and $\boldsymbol{d}_3^a$ is perpendicular to the plane of the base $\mathtt{X}_a$ and is pointing towards the base $\mathtt{X}_{a+1}$, as shown in Figure 2.1. Once the reference points and frame of each base are specified, the positions of all the non-hydrogen atoms of the base are known. Likewise, we assign a point $\overline{\boldsymbol{r}}^a$ and frame $\{\overline{\boldsymbol{d}}_i^a\}$ to each complementary base $\overline{\mathtt{X}}_a$. However, because the two strands are anti-parallel, we flip the frame $\{\overline{\boldsymbol{d}}_i^a\}$ (change the signs of the vectors $\overline{\boldsymbol{d}}_2$ and $\overline{\boldsymbol{d}}_3$ to opposite), so that the two frames overlap when the base pair is in its ideal shape.

We will also consider matrices $\boldsymbol{D}^a$ and $\overline{\boldsymbol{D}}^a$ with vectors $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$ as columns, i.e. $\boldsymbol{D}^a \ = \{\boldsymbol{d}_i^a\} = [\boldsymbol{d}_1^a\ \boldsymbol{d}_2^a\ \boldsymbol{d}_3^a]$ and $\overline{\boldsymbol{D}}^a \ = \{\overline{\boldsymbol{d}}_i^a\} = [\overline{\boldsymbol{d}}_1^a\ \overline{\boldsymbol{d}}_2^a\ \overline{\boldsymbol{d}}_3^a]$.


$\overline{\boldsymbol{d}}_3^{a+1}$
$\overline{\boldsymbol{d}}_2^{a+1}$
$\overline{\boldsymbol{r}}^{a+1}$
$\overline{\mathrm{X}}_{a+1}$
$\overline{\boldsymbol{d}}_1^{a+1}$
$\boldsymbol{d}_3^{a+1}$
$\boldsymbol{d}_2^{a+1}$
$\boldsymbol{r}^{a+1}$
$\mathrm{X}_{a+1}$
$\boldsymbol{d}_1^{a+1}$
$\overline{\boldsymbol{d}}_3^{a}$
$\overline{\boldsymbol{d}}_2^{a}$
$\overline{\boldsymbol{r}}^{a}$
$\overline{\mathrm{X}}_{a}$
$\overline{\boldsymbol{d}}_1^{a}$
$\boldsymbol{d}_3^{a}$
$\boldsymbol{r}^{a}$
$\boldsymbol{d}_2^{a}$
$\mathrm{X}_{a}$
$\boldsymbol{d}_1^{a}$

(a) (b)

Figure 2.1: (a) Rigid bases of DNA with reference points and frames, corresponding to the Tsukuba convention [54]. Base pair and junction frames, as introduced in Section 2.3, are drawn in grey and black respectively. (Thanks to J. Głowacki for this figure.) (b) A schematic view of rigid bases with their reference point and frames with the indicated directions of vectors $\{\boldsymbol{d}_i^a\}$ and $\{\overline{\boldsymbol{d}}_i^a\}$, $i = 1, 2, 3$. Note that, despite the two strands of DNA being antiparallel, the frames $\boldsymbol{D}^a$ and $\overline{\boldsymbol{D}}^a$ are oriented in the same way.

For modelling purposes, it is convenient to describe the configuration of DNA in coordinates that do not depend on the absolute translations and rotations of the molecule. In particular, we would like to have a set of parameters and rules, that would still allow us to reconstruct the position and orientation of each base, once the position and orientation of one chosen base is given. To do this, we first need to parametrise rotations between orthonormal frames.

## 2.2 Parametrisation of rotations

To describe rotations we will use Cayley (or Euler - Rodrigues) parameters. One other popular choice is Euler angles, used, for example, in the 3DNA software package [46].

Having a unit axis $\boldsymbol{k}$ and an angle of rotation $\varphi \in [0, \pi)$, the corresponding rotation matrix $\boldsymbol{Q} \in SO(3)$ can be obtained using the Euler-Rodrigues formula

$$\boldsymbol{Q} = \cos\varphi\,\boldsymbol{I} + (1 - \cos\varphi)\boldsymbol{k} \otimes \boldsymbol{k} + \sin\varphi\,\boldsymbol{k}^{\times}, \tag{2.1}$$

where $\boldsymbol{I} \in \mathbb{R}^{3\times 3}$ is an identity matrix, $\boldsymbol{k} \otimes \boldsymbol{k} = \boldsymbol{k}\boldsymbol{k}^T$ is the rank-one outer product matrix and $\boldsymbol{k}^{\times}$ is the skew matrix

$$\boldsymbol{k}^{\times} = \begin{pmatrix} 0 & -k_3 & k_2 \\ k_3 & 0 & -k_1 \\ -k_2 & k_1 & 0 \end{pmatrix}, \tag{2.2}$$

satisfying $(\boldsymbol{k}^{\times})\boldsymbol{v} = \boldsymbol{k} \times \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbb{R}^3$. Then $\boldsymbol{Q}\boldsymbol{v}$ gives the absolute coordinates of the vector

$\boldsymbol{v}$ rotated around the axis $\boldsymbol{k}$ by the angle $\varphi$ for all $\boldsymbol{v} \in \mathbb{R}^3$. Moreover, from (2.1) we get

$$\boldsymbol{Q} - \boldsymbol{Q}^T = 2\sin\varphi\, \boldsymbol{k}^\times \quad \text{and} \quad \operatorname{tr}\boldsymbol{Q} = 1 + 2\cos\varphi, \tag{2.3}$$

which allows us to compute $\varphi$ and $\boldsymbol{k}$ as

$$\begin{aligned} \varphi &= \arccos\left(\frac{\operatorname{tr}\boldsymbol{Q} - 1}{2}\right) \in [0, \pi) \quad \text{and} \\ \boldsymbol{k} &= \frac{1}{2\sin\varphi}\begin{pmatrix} Q_{32} - Q_{23} \\ Q_{13} - Q_{31} \\ Q_{21} - Q_{12} \end{pmatrix} = \frac{1}{2\sin\varphi}\operatorname{vec}[\boldsymbol{Q} - \boldsymbol{Q}^T], \end{aligned} \tag{2.4}$$

when the rotation matrix $\boldsymbol{Q} = \{Q_{ij}\}$ is known. Here $\operatorname{vec}\boldsymbol{A}$ for a skew matrix $\boldsymbol{A} \in \mathbb{R}^{3\times3}$ is defined as

$$\operatorname{vec}\boldsymbol{A} = \begin{pmatrix} A_{32} \\ A_{13} \\ A_{21} \end{pmatrix}. \tag{2.5}$$

On the other hand, for any two orthonormal frames $\boldsymbol{D} = [\boldsymbol{d}_1\ \boldsymbol{d}_2\ \boldsymbol{d}_3]$ and $\widetilde{\boldsymbol{D}} = [\widetilde{\boldsymbol{d}}_1\ \widetilde{\boldsymbol{d}}_2\ \widetilde{\boldsymbol{d}}_3]$, $\boldsymbol{D}, \widetilde{\boldsymbol{D}} \in \mathbb{R}^{3\times3}$, a rotation from one frame to another can be written as

$$\widetilde{\boldsymbol{D}} = \boldsymbol{Q}\boldsymbol{D} = [\boldsymbol{Q}\boldsymbol{d}_1\ \boldsymbol{Q}\boldsymbol{d}_2\ \boldsymbol{Q}\boldsymbol{d}_3], \tag{2.6}$$

where

$$\boldsymbol{Q} = \widetilde{\boldsymbol{D}}\boldsymbol{D}^T. \tag{2.7}$$

Note that because $\boldsymbol{Q}$ is a proper right-handed rotation matrix (an orthonormal matrix with a determinant 1), it has eigenvalues $(1, e^{-i\varphi}, e^{i\varphi})$, where $\varphi$ is the angle of rotation, and the real eigenvector of $\boldsymbol{Q}$ is its rotation axis $\boldsymbol{k}$.

When $\boldsymbol{Q}$ is symmetric, which corresponds to the cases $\varphi = 0$ and $\varphi = \pi$, then $\boldsymbol{Q} - \boldsymbol{Q}^T$ is a zero matrix, and we can not use (2.4) for computing $\boldsymbol{k}$. In the case $\varphi = 0$, $\boldsymbol{Q} = \boldsymbol{I}$ and any unit vector can be chosen to be the rotation axis. In the case $\varphi = \pi$, the rotation axis can be computed as an eigenvector of the matrix $\boldsymbol{Q} + \boldsymbol{I}$, which has eigenvalues $(0, 0, 2)$. However, in B-DNA, which we want to model, rotations through an angle $\pi$ are very unlikely for relative rotations between adjacent bases.

We can choose a scaling $f(\varphi)$ for the rotation axis (where $f$ is an invertible function), to have a three parameter vector $f(\varphi)\boldsymbol{k}$ fully describing the rotation. Then, as $\boldsymbol{k}$ is a unit vector, we can recover the angle $\varphi$ from the relation $f(\varphi) = ||f(\varphi)\boldsymbol{k}||$. For example, the scaling $f(\varphi)$ used in *Curves+* [41] is the value of $\varphi$ in degrees, in [36] $f(\varphi) = 2\tan\frac{\varphi}{2}$, and we will use $f(\varphi) = 10\tan\frac{\varphi}{2}$ as in [24]. We explain our choice of parameter scaling in Section 2.4.

If we introduce the Cayley vector $\eta$, defined by

$$\eta = 2\tan\frac{\varphi}{2}\boldsymbol{k}, \tag{2.8}$$

which is well defined for $\varphi \in [0, \pi)$, then (2.1) is equivalent to

$$\boldsymbol{Q} = (\boldsymbol{I} + \frac{1}{2}\eta^{\times})(\boldsymbol{I} - \frac{1}{2}\eta^{\times})^{-1} =: \mathrm{cay}(\eta). \tag{2.9}$$

For defining DNA configuration parameters, described in the next section, we will also introduce a middle frame between any two frames $\boldsymbol{D}$ and $\widetilde{\boldsymbol{D}}$. We define the middle frame $\boldsymbol{M}$ as

$$\boldsymbol{M} = \sqrt{\boldsymbol{Q}}\boldsymbol{D}, \tag{2.10}$$

where $\sqrt{\boldsymbol{Q}}$ is a matrix of half rotation from $\boldsymbol{D}$ to $\widetilde{\boldsymbol{D}}$, i.e. rotation about the same axis $\boldsymbol{k}$ by the angle $\frac{\varphi}{2}$.

The same rotation as in (2.6) can be expressed multiplying $\boldsymbol{D}$ by a rotation matrix on the right:

$$\widetilde{\boldsymbol{D}} = \boldsymbol{D}\boldsymbol{R} = \left[ \sum_{i=1}^{3} \boldsymbol{R}_{i1}\boldsymbol{d}_i \ \ \sum_{i=1}^{3} \boldsymbol{R}_{i2}\boldsymbol{d}_i \ \ \sum_{i=1}^{3} \boldsymbol{R}_{i3}\boldsymbol{d}_i \right], \tag{2.11}$$

where

$$\boldsymbol{R} = \boldsymbol{D}^T\widetilde{\boldsymbol{D}} = \boldsymbol{D}^T\boldsymbol{Q}\boldsymbol{D} = \widetilde{\boldsymbol{D}}^T\boldsymbol{Q}\widetilde{\boldsymbol{D}}. \tag{2.12}$$

The normalised axial vector $\boldsymbol{k}_R$ of the skew matrix $\boldsymbol{R} - \boldsymbol{R}^T$ is

$$\boldsymbol{k}_R = \boldsymbol{D}\boldsymbol{k} = \widetilde{\boldsymbol{D}}\boldsymbol{k}, \tag{2.13}$$

i.e. $\boldsymbol{k}_R$ is equal to $\boldsymbol{k}$ written in the coordinates of the $\boldsymbol{D}$ frame (which are the same as the coordinates in the $\widetilde{\boldsymbol{D}}$ frame, because $\boldsymbol{k}$ is the axis of rotation). Also, the eigenvalues of $\boldsymbol{Q}$ and of $\boldsymbol{R}$ are the same, which means that the two matrices indeed correspond to the same rotation, but expressed in different coordinates. We call $\boldsymbol{R}$ relative rotation matrix and $\boldsymbol{Q}$ the absolute rotation matrix. It will be convenient to use relative rotations in our work.

## 2.3 Internal coordinates of rigid bases and base pairs

Having chosen a parametrisation of rotations, one can replace the absolute coordinates of bases by vectors of relative translations (difference vectors) between the reference points and vectors of relative rotations (Cayley vectors or scaled Cayley vectors) between the reference frames. As we want these coordinates to be independent of global translations and rotations of the whole molecule, they have to be expressed with respect to the frames attached to the molecule, rather than to a fixed lab coordinate frame. We could choose an order of tracing all the bases of an oligomer, for example $\mathtt{X}_1, \overline{\mathtt{X}}_1, \mathtt{X}_2, \overline{\mathtt{X}}_2 \ldots, \mathtt{X}_n, \overline{\mathtt{X}}_n$, and compute the vectors of translation and rotation between each consecutive pair of them, expressing both vectors in the reference frame of one of the two bases. However, the values of the configuration parameters of this kind depend on the choice of the reference strand. It would be convenient if the relations between two parameter vectors, corresponding to the two different choices of reference strand, were simple. In order to define relative coordinates with this property, it is helpful to introduce base-pair reference points together with base pair and junction frames.

We define the base pair frame $\boldsymbol{G}^a = \{\boldsymbol{g}_i^a\}$ as the average orientation of the two base frames,

$$\boldsymbol{G}^a = \overline{\boldsymbol{D}}^a \sqrt{\Lambda^a}, \tag{2.14}$$

where

$$\Lambda^a = (\overline{\boldsymbol{D}}^a)^T \boldsymbol{D}^a \in \mathbb{R}^{3\times 3}, \tag{2.15}$$

is the relative rotation matrix that describes the orientation of the frame $\{\boldsymbol{d}_i^a\}$ with respect to $\{\overline{\boldsymbol{d}}_i^a\}$.

The intra base pair rotational coordinate vector, describing the relative rotation between $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$, is defined as the Cayley vector of the matrix $\Lambda^a$:

$$\vartheta^a = \frac{2}{\mathrm{tr}[\Lambda^a] + 1} \mathrm{vec}[\Lambda^a - (\Lambda^a)^T]. \tag{2.16}$$

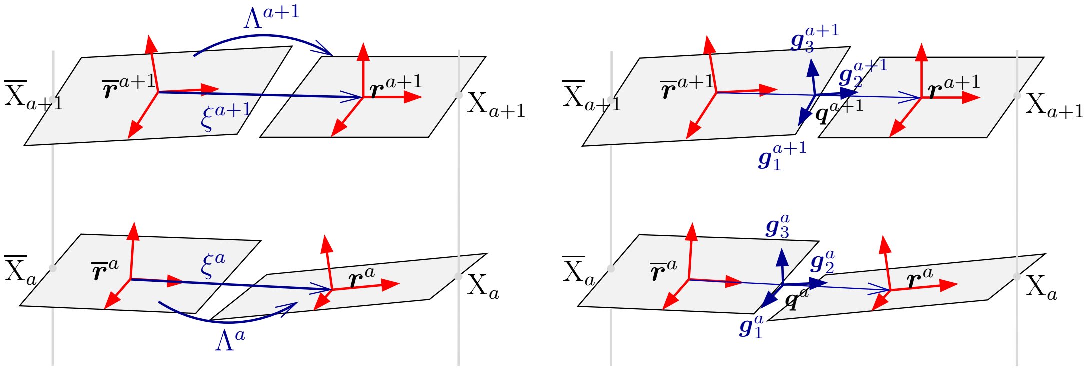
Then

$$||\vartheta^a|| = \frac{2\,\sin\varphi^a}{1 + \cos\varphi^a} = 2\tan\frac{\varphi^a}{2}, \tag{2.17}$$

where $\varphi^a$ is the angle of rotation corresponding to $\Lambda^a$. Note that the coordinates of this vector are the same with respect to all of the three frames $\overline{\boldsymbol{D}}^a$, $\boldsymbol{D}^a$ and $\boldsymbol{G}^a$. We used a relative rotation matrix in order to directly get the Cayley vector expressed in these frames. We then use $\boldsymbol{G}^a$ as a reference frame for defining the intra base pair translational coordinate vector between $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$:

$$\xi^a = (\boldsymbol{G}^a)^T(\boldsymbol{r}^a - \overline{\boldsymbol{r}}^a). \tag{2.18}$$

An illustration of intra base pair rotational and translational coordinates is shown in Figure 2.2.



(a) (b)

Figure 2.2: (a) A schematic illustration of the intra base pair translational and rotational coordinates defined in (2.16) and (2.18). The base frames are shown in red, as in Figure 2.1(b). Note that the intra coordinates are expressed in the base pair frames, shown in (b) and defined in (2.14).

To describe the relative rotation and displacement between neighbouring base pairs along the strands we consider the base pair frame $\boldsymbol{G}^a$ and a reference point $\boldsymbol{q}^a$, defined as the average position of the two base reference points:

$$\boldsymbol{q}^a = \frac{1}{2}(\boldsymbol{r}^a + \overline{\boldsymbol{r}}^a). \tag{2.19}$$

Similarly, we define a frame $\boldsymbol{G}^{a+1} = \{\boldsymbol{g}_i^{a+1}\}$ and a point $\boldsymbol{q}^{a+1}$ associated with the base pair $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$.

Finally, the inter base pair, or junction, rotational coordinate vector, describing the relative rotation between base pairs $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$ is

$$\theta^a = \frac{2}{\mathrm{tr}[\boldsymbol{L}^a] + 1}\mathrm{vec}[\boldsymbol{L}^a - (\boldsymbol{L}^a)^T], \tag{2.20}$$

where

$$\boldsymbol{L}^a = (\boldsymbol{G}^a)^T \boldsymbol{G}^{a+1} \in \mathbb{R}^{3\times 3}, \tag{2.21}$$

is the relative rotation matrix that describes the orientation of frame $\{\boldsymbol{g}_{i+1}^a\}$ with respect to $\{\boldsymbol{g}_i^a\}$. Note that

$$||\theta^a|| = \frac{2\,\sin\phi^a}{1 + \cos\phi^a} = 2\tan\frac{\phi^a}{2}, \tag{2.22}$$

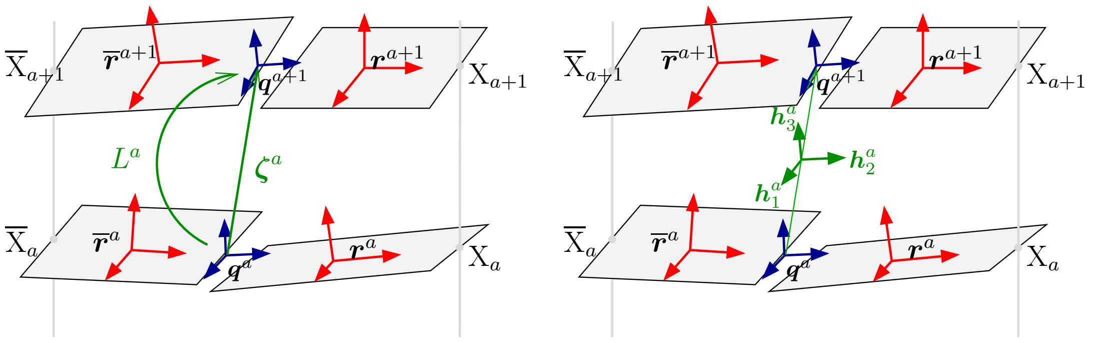
where $\phi^a$ is the angle of rotation corresponding to $\boldsymbol{L}^a$. The inter base pair translational vector, describing the relative rotation between base pairs $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$ is

$$\zeta^a = (\boldsymbol{H}^a)^T(\boldsymbol{q}^{a+1} - \boldsymbol{q}^a), \tag{2.23}$$

where the junction frame $\boldsymbol{H}^a = \{\boldsymbol{h}_i^a\}$ is defined as

$$\boldsymbol{H}^a = \boldsymbol{G}^a\sqrt{\boldsymbol{L}^a}. \tag{2.24}$$

An illustration of inter base pair rotational and translational coordinates is shown in Figure 2.3.



(a) (b)

Figure 2.3: (a) A schematic illustration of the inter base pair (junction) translational and rotational coordinates defined in (2.20) and (2.23). The base frames are shown in red, as in Figure 2.1(b). and the base pair frames in blue, as in Figure 2.2(b). Note that the inter coordinates are expressed in the junction frames, shown in (b) and defined in (2.24).

Thus the relative rotation and displacement between bases $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$ across the strands is described by the intra-base pair coordinates $\boldsymbol{y}^a = (\vartheta, \xi)^a$, whereas the relative rotation and displacement between the pairs $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}, \overline{\mathtt{X}})_{a+1}$ along the strands is described by the inter base pair coordinates $\boldsymbol{z}^a = (\theta, \zeta)^a$. We define the vector of relative coordinates as

$$\mathsf{w} = (\boldsymbol{y}^1, \boldsymbol{z}^1, \boldsymbol{y}^2, \boldsymbol{z}^2, ..., \boldsymbol{y}^{n-1}, \boldsymbol{z}^{n-1}, \boldsymbol{y}^n) \in \mathbb{R}^{12n-6}. \tag{2.25}$$

Note that $\mathsf{w}$ does not depend on the global translations and rotations of the DNA molecule.

The definitions of these coordinates can be shown to satisfy all the qualitative guidelines set forth in the Cambridge convention [19] for nucleic acid structures, including the symmetry conditions associated with a change of reference strand. Accordingly, we refer to the components of the intra rotational parameters $\vartheta^a$ as Buckle-Propeller-Opening, the intra translational parameters $\xi^a$ as Shear-Stretch-Stagger, the inter rotational parameters $\theta^a$ as Tilt-Roll-Twist, and the inter translational parameters $\zeta^a$ as Shift-Slide-Rise (see Figure 2.4). We remark that $\vartheta^a$ and $\theta^a$ are components of a Cayley rotation vector and so are not conventional angular coordinates about various axes as employed by many authors; however, they can be put into correspondence with conventional angular coordinates, and are nearly identical in the case of small rotations when the angular ones are measured in radians (or fifths of a radian for the scaling described in Section 2.4).

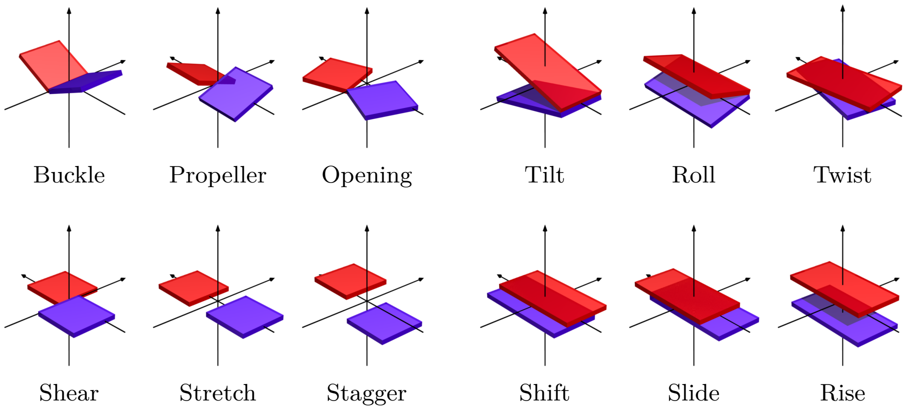


Figure 2.4: Intra base pair (left) and inter base pair (right) parameters of a coarse-grain DNA model; rotational parameters are on the top, translation parameters on the bottom; bases/base pairs are represented as cuboids. (Thanks to J. Głowacki for this figure.)

The complete configuration of a DNA oligomer is specified by introducing six additional coordinates $\boldsymbol{z}^0 = (\theta, \zeta)^0$ for the first base pair frame and reference point with respect to an external, lab-fixed frame. Ignoring these six degrees of freedom exactly corresponds to eliminating the overall symmetry of rigid body motion that exists when there is no external potential field. Given $\boldsymbol{z}^0 = (\theta, \zeta)^0 \in \mathbb{R}^6$, the external base pair frame $\{\boldsymbol{g}_i^0\}$ (e.g. $\{\boldsymbol{g}_i^0\} = \{\boldsymbol{e}_i\}$), the reference point $\boldsymbol{q}^0$ (e.g. $\boldsymbol{q}^0 = (0, 0, 0)$), and the vector of relative coordinates $\mathsf{w} \in \mathbb{R}^{12n-6}$, we can recover all the frames and reference points for the bases $\mathtt{X}_a$ and $\overline{\mathtt{X}}_a$, $a \in 1, \dots, n$, using the formulas

$$\overline{\boldsymbol{d}}^a_j = \sum_{i=1}^{3} (\sqrt{\varLambda^a})_{ji} \boldsymbol{g}^a_i, \qquad \overline{\boldsymbol{r}}^a = \boldsymbol{q}^a - \frac{1}{2} \sum_{i=1}^{3} \xi^a_i \boldsymbol{g}^a_i, \tag{2.26}$$

$$\boldsymbol{d}^a_j = \sum_{i=1}^{3} \varLambda^a_{ij} \overline{\boldsymbol{d}}^a_i, \qquad \boldsymbol{r}^a = \overline{\boldsymbol{r}}^a + \sum_{i=1}^{3} \xi^a_i \boldsymbol{g}^a_i, \tag{2.27}$$

$$\boldsymbol{g}^{a+1}_j = \sum_{i=1}^{3} L^a_{ij} \boldsymbol{g}^a_i, \qquad \boldsymbol{q}^{a+1} = \boldsymbol{q}^a + \sum_{i=1}^{3} \zeta^a_i \boldsymbol{h}^a_i, \tag{2.28}$$

where $\varLambda^a = \text{cay}[\vartheta^a]$ is the rotation matrix corresponding to $\vartheta^a$, $\boldsymbol{L}^a = \text{cay}[\theta^a]$ and $\{\boldsymbol{h}^a_i\}$ is defined by (2.24).

One could go further in coarse-graining DNA and make an assumption that the base pairs of DNA are rigid bodies. In this case, it is enough to have base pair frames $\{\boldsymbol{g}^a_i\}$ and base pair reference points $\boldsymbol{q}^a$ for describing the configuration of a molecule. Thus the vector of relative coordinates for the rigid base pair DNA model is

$$\mathsf{w}^{\text{bp}} = (\boldsymbol{z}^1, \boldsymbol{z}^2, ..., \boldsymbol{z}^{n-1}) \in \mathbb{R}^{6n-6}. \tag{2.29}$$

## 2.4 Non-dimensionalisation and scaling

For the purposes of numerics and analysis, it will be convenient to introduce scales and transform the rigid-base model into a dimensionless form. To define dimensionless quantities associated with the kinematics of the model, we introduce a characteristic scale $\ell$ for the translational coordinates, a characteristic scale $g$ for the rotational coordinates, and define dimensionless variables by

$$\underline{y}^a = \mathsf{G}^{-1} y^a, \quad \underline{z}^a = \mathsf{G}^{-1} z^a, \quad \underline{\mathsf{w}} = \mathsf{G}_n^{-1} \mathsf{w}, \tag{2.30}$$

where $\mathsf{G} = \text{diag}(g, g, g, \ell, \ell, \ell) \in \mathbb{R}^{6\times 6}$ is a constant, diagonal matrix and $\mathsf{G}_n \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is the diagonal matrix formed by $2n-1$ copies of the matrix $\mathsf{G}$.

In any given application, the scales $\ell$ and $g$ would normally be set by the phenomena of interest. The scale $\ell$ is ideally chosen to describe the magnitude of the variation in the intra- and inter-base pair translational variables, whereas the scale $g$ is ideally chosen to describe the magnitude of the variation in the intra- and inter-base pair rotational variables. In the analysis of molecular dynamics data of DNA, the phenomena of interest are fluctuations of atomic positions on the order of 1Å. Hence a reasonable scale for the translational variables is $\ell = 1$Å because variations in these variables are in direct correspondence to variations in atomic positions. Moreover, a reasonable scale for the rotational variables is $g = 1/5$ (radians) because variations of this size in these rotation variables, about a zero reference value, correspond to a variation of about 1Å in atomic positions of the atoms making up a base. This follows from the fact that the characteristic size of both a base and the junction between base pairs is approximately 5Å, so that rotational variations of $1/5$ gives rise to variations in atomic positions of approximately 1Å, as shown in Figure 2.5.

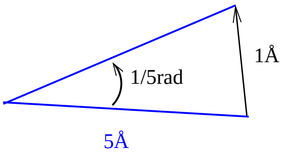


Figure 2.5: Scaling of rotational parameters: a rotation of a 5Å length vector by $\frac{1}{5}$ of a radian corresponds to translating its endpoint by approximately 1Å.

In particular, with this choice of scaling we have

$$\begin{aligned} \underline{\vartheta}^a &= 5\,\vartheta^a, \quad \underline{\xi}^a = \xi^a, \quad a = 1 \dots n, \\ \underline{\theta}^a &= 5\,\theta^a, \quad \underline{\zeta}^a = \zeta^a, \quad a = 1 \dots n-1 \end{aligned} \tag{2.31}$$

so that

$$||\underline{\vartheta}^a|| = 10 \tan \frac{\varphi^a}{2} \quad \text{and} \quad ||\underline{\theta}^a|| = 10 \tan \frac{\phi^a}{2}, \tag{2.32}$$

where $\underline{\xi}^a$ and $\underline{\zeta}^a$ are the scaled intra and inter base pair translational parameters, $\underline{\vartheta}^a$ and $\underline{\theta}^a$ are the scaled intra and inter base pair rotational parameters, and $\varphi^a$, $\phi^a$ are the corresponding angles of rotation (in radians). Further in this work we will use the nondimensionalised configuration parameters with these scales and drop the underline notation for convenience.

## 2.5 Change of reference strand

The sequence identifying a DNA molecule depends on the choice of reference strand. There are two strands in B-DNA, therefore there are two sequences, corresponding to the same molecule. If the sequence along one strand is $\mathtt{X}_1 \cdots \mathtt{X}_n$, along the other, anti parallel, strand it is $\mathtt{X}_1^* \cdots \mathtt{X}_n^*$, where $\mathtt{X}_1^* = \overline{\mathtt{X}}_n$, $\mathtt{X}_2^* = \overline{\mathtt{X}}_{n-1}$, …, $\mathtt{X}_n^* = \overline{\mathtt{X}}_1$. This means that $(\mathtt{X}, \overline{\mathtt{X}})_a$ and $(\mathtt{X}^*, \overline{\mathtt{X}}^*)_{a_*}$ denote the same base pair when $a_* = n - a + 1$.

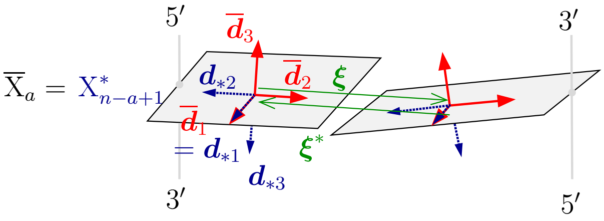

5′
3′
$\overline{\mathrm{X}}_a = \mathrm{X}^*_{n-a+1}$
$\overline{\boldsymbol{d}}_3$
$\boldsymbol{d}_{*2}$
$\overline{\boldsymbol{d}}_2$
$\boldsymbol{\xi}$
$\overline{\boldsymbol{d}}_1$
$= \boldsymbol{d}_{*1}$
$\boldsymbol{\xi}^*$
$\boldsymbol{d}_{*3}$
3′
5′

Figure 2.6: A schematic illustration of how the coordinates change when choosing a different reference strand. Here the base frames, corresponding to one initial choice of strand are shown in red, and the inverted base frames, corresponding to the other choice of reference strand are blue. Note that the internal coordinates are expressed in the middle frames, that are also inverted.

The base reference points and base frames for the two different choices of reference strand are related in a simple way, as shown in Figure 2.6:

$$\begin{aligned}
\boldsymbol{r}^a_* &= \overline{\boldsymbol{r}}^{n-a+1}, & \overline{\boldsymbol{r}}^a_* &= \boldsymbol{r}^{n-a+1},\\
\boldsymbol{d}^a_{*1} &= \overline{\boldsymbol{d}}_1^{n-a+1}, & \overline{\boldsymbol{d}}^a_{*1} &= \boldsymbol{d}_1^{n-a+1},\\
\boldsymbol{d}^a_{*2} &= -\overline{\boldsymbol{d}}_2^{n-a+1}, & \overline{\boldsymbol{d}}^a_{*2} &= -\boldsymbol{d}_2^{n-a+1},\\
\boldsymbol{d}^a_{*3} &= -\overline{\boldsymbol{d}}_3^{n-a+1}, & \overline{\boldsymbol{d}}^a_{*3} &= -\boldsymbol{d}_3^{n-a+1},
\end{aligned} \tag{2.33}$$

i.e. the reference points stay the same in both cases (but get assigned a different index) and the base frames are flipped when passing from one case to another, because, as described in Section 2.1, we want the two base frames to be aligned and to be pointing along the reference strand for the perfect shape of B-DNA, despite its anti parallel nature. From (2.33) we get the relation between the orthogonal frame matrices

$$\boldsymbol{D}^a_* = \boldsymbol{D}^{n-a+1}\,\mathsf{E}^*, \tag{2.34}$$

where $\mathsf{E}^*$ is a constant, diagonal rotation matrix $\mathsf{E}^* = \mathrm{diag}(1,-1,-1) \in \mathbb{R}^{3\times 3}$, and $\mathsf{E}^* = (\mathsf{E}^*)^T = (\mathsf{E}^*)^{-1}$. Similarly, we get the formulas for the base pair points $\boldsymbol{q}^a$, base pair frames $\boldsymbol{G}^a$ and junction frames $\boldsymbol{H}^a$:

$$\boldsymbol{q}^a_* = \boldsymbol{q}^{n-a+1}, \quad \boldsymbol{G}^a_* = \boldsymbol{G}^{n-a+1}\,\mathsf{E}^* \quad \text{and} \quad \boldsymbol{H}^a_* = \boldsymbol{H}^{n-a+1}\,\mathsf{E}^*. \tag{2.35}$$

And finally, from the definitions of the relative rotation and displacement coordinates, we obtain the expressions

$$\begin{aligned}
\vartheta^a_* &= -\mathsf{E}^*\,\vartheta^{n-a+1}, & \xi^a_* &= -\mathsf{E}^*\,\xi^{n-a+1}, & a &= 1\dots n,\\
\theta^a_* &= -\mathsf{E}^*\,\theta^{n-a+1}, & \zeta^a_* &= -\mathsf{E}^*\,\zeta^{n-a+1}, & a &= 1\dots n-1,
\end{aligned} \tag{2.36}$$

that can be written as a relation between the two shape vectors:

$$\mathsf{w}_* = \mathsf{E}_n\,\mathsf{w}. \tag{2.37}$$

Here $\mathsf{E}_n \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a block, trailing-diagonal matrix formed by $2n-1$ copies of the constant, diagonal matrix $\mathsf{E} = \mathrm{diag}(-\mathsf{E}^*, -\mathsf{E}^*) = \mathrm{diag}(-1,1,1,-1,1,1) \in \mathbb{R}^{6\times 6}$, with the property that $\mathsf{E}_n = \mathsf{E}_n^T = \mathsf{E}_n^{-1}$. Specifically, we have

$$\mathsf{E}_n = \begin{pmatrix} & & & \mathsf{E}\\ & & \mathsf{E} & \\ & \iddots & & \\ \mathsf{E} & & & \end{pmatrix}. \tag{2.38}$$

These relations can be described as follows. Characterise the twelve types of internal coordinates as being odd or even, where the odd coordinates are Buckle, Shear, Tilt and Shift, (one each of intra and inter and translation and rotation) and the remaining eight are all even. Then under a change of reference strand the odd coordinates at any location along the sequence change sign, whereas the even coordinates remain unaltered.

## 2.6 Changing the rule of embedding base frames

In this section we will discuss the impact of the choice of the reference points $\boldsymbol{r}^a$ and base frames $\{\boldsymbol{d}_i^a\}$ on the values of the internal coordinate vector $\mathsf{w}$.

First let us consider introducing a new rule of embedding frames and reference points for every one of the four possible bases $\mathtt{T}, \mathtt{A}, \mathtt{C}$ and $\mathtt{G}$. We refer to the new frames as $\underline{\boldsymbol{D}}^a = \{\underline{\boldsymbol{d}}_i^a\}$ (and $\underline{\overline{\boldsymbol{D}}}^a = \{\underline{\overline{\boldsymbol{d}}}_i^a\}$ on the complementary strand) and to the new reference points as $\underline{\boldsymbol{r}}^a$ (and $\underline{\overline{\boldsymbol{r}}}^a$), where

$$\begin{aligned} \underline{\boldsymbol{r}}^a = \boldsymbol{r}^a + \boldsymbol{s}^a, &\quad \underline{\overline{\boldsymbol{r}}}^a = \overline{\boldsymbol{r}}^a + \overline{\boldsymbol{s}}^a, \\ \underline{\boldsymbol{D}}^a = \boldsymbol{D}^a \boldsymbol{R}^a, &\quad \text{and} \quad \underline{\overline{\boldsymbol{D}}}^a = \overline{\boldsymbol{D}}^a \overline{\boldsymbol{R}}^a. \end{aligned} \tag{2.39}$$

Here $\boldsymbol{R}^a$ and $\overline{\boldsymbol{R}}^a$ correspond to rotations smaller than $\pi$, and $\boldsymbol{s}^a$, $\overline{\boldsymbol{s}}^a$ are small enough for the reference points $\boldsymbol{r}^a$ and $\overline{\boldsymbol{r}}^a$ to stay close to their corresponding bases, as shown in Figure 2.7.

Note that we assume that the rotation $\overline{\boldsymbol{R}}^a$ is applied to the flipped frame $\overline{\boldsymbol{D}}^a$, i.e. when the two strands are parallel.

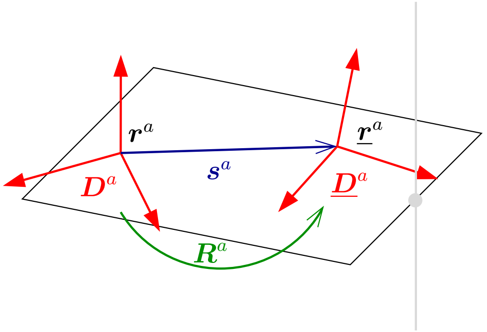


Figure 2.7: A schematic illustration of changing the rule of embedding base frames and reference points. Here $\overline{\boldsymbol{D}}^a$ is the new reference frame and $\overline{\boldsymbol{r}}^a$ is the new reference point on the reference strand, as defined in (2.39).

Then the new base pair frame is

$$\begin{aligned} \underline{\boldsymbol{G}}^a &= \underline{\overline{\boldsymbol{D}}}^a \sqrt{\underline{\Lambda}^a} = \underline{\overline{\boldsymbol{D}}}^a \sqrt{(\underline{\overline{\boldsymbol{D}}}^a)^T \underline{\boldsymbol{D}}^a} \\ &= \overline{\boldsymbol{D}}^a \overline{\boldsymbol{R}}^a \sqrt{(\overline{\boldsymbol{R}}^a)^T (\overline{\boldsymbol{D}}^a)^T \boldsymbol{D}^a \boldsymbol{R}^a} \\ &= \overline{\boldsymbol{D}}^a \overline{\boldsymbol{R}}^a \sqrt{(\overline{\boldsymbol{R}}^a)^T \Lambda^a \boldsymbol{R}^a}, \end{aligned} \tag{2.40}$$

where $\Lambda^a$ is defined in (2.15). In the special case, when $\overline{\boldsymbol{R}}^a = \boldsymbol{R}^a$, we have the identity $\sqrt{(\boldsymbol{R}^a)^T \Lambda^a \boldsymbol{R}^a} = (\boldsymbol{R}^a)^T \sqrt{\Lambda^a} \boldsymbol{R}^a$, which follows from the fact that if $\Lambda^a$ has a rotation axis $\boldsymbol{v}$ then both $(\boldsymbol{R}^a)^T \sqrt{\Lambda^a} \boldsymbol{R}^a$ and $(\boldsymbol{R}^a)^T \Lambda^a \boldsymbol{R}^a$ have the same rotation axis $\boldsymbol{R}^a \boldsymbol{v}$. Therefore

$$\begin{aligned} \underline{\boldsymbol{G}}^a &= \overline{\boldsymbol{D}}^a \boldsymbol{R}^a (\boldsymbol{R}^a)^T \sqrt{\Lambda^a} \boldsymbol{R}^a = \overline{\boldsymbol{D}}^a \sqrt{\Lambda^a} \boldsymbol{R}^a \\ &= \boldsymbol{G}^a \boldsymbol{R}^a, \end{aligned} \tag{2.41}$$

$$\begin{aligned}
\underline{\vartheta}^a &= \frac{2}{\mathrm{tr}[\underline{\mathit{\Lambda}}^a]+1}\mathrm{vec}[\underline{\mathit{\Lambda}}^a - (\underline{\mathit{\Lambda}}^a)^T] \\
&= \frac{2}{\mathrm{tr}[\mathit{\Lambda}^a]+1}\,(\boldsymbol{R}^a)^T\,\mathrm{vec}[\mathit{\Lambda}^a - (\mathit{\Lambda}^a)^T] \qquad (2.42)\\
&= (\boldsymbol{R}^a)^T\vartheta^a
\end{aligned}$$

and

$$\begin{aligned}
\underline{\xi}^a &= (\underline{\boldsymbol{G}}^a)^T(\underline{\boldsymbol{r}}^a - \underline{\overline{\boldsymbol{r}}}^a) \\
&= (\boldsymbol{R}^a)^T(\boldsymbol{G}^a)^T(\boldsymbol{r}^a - \overline{\boldsymbol{r}}^a + \boldsymbol{s}^a - \overline{\boldsymbol{s}}^a) \qquad (2.43)\\
&= (\boldsymbol{R}^a)^T\xi^a + (\boldsymbol{R}^a)^T(\boldsymbol{G}^a)^T(\boldsymbol{s}^a - \overline{\boldsymbol{s}}^a).
\end{aligned}$$

The new base pair reference point is

$$\underline{\boldsymbol{q}}^a = \boldsymbol{q}^a + \frac{1}{2}(\boldsymbol{s}^a + \overline{\boldsymbol{s}}^a) \qquad (2.44)$$

and the new junction frame, still with $\overline{\boldsymbol{R}}^a = \boldsymbol{R}^a$, becomes

$$\underline{\boldsymbol{H}}^a = \underline{\boldsymbol{G}}^a\sqrt{\underline{\boldsymbol{L}}^a} = \underline{\boldsymbol{G}}^a\sqrt{(\underline{\boldsymbol{G}}^a)^T\underline{\boldsymbol{G}}^{a+1}} = \boldsymbol{G}^a\boldsymbol{R}^a\sqrt{(\boldsymbol{R}^a)^T\boldsymbol{L}^a\boldsymbol{R}^{a+1}}, \qquad (2.45)$$

where $\boldsymbol{L}^a$ is defined in (2.21). In the further simple special case when $\forall a : \boldsymbol{R}^a = \boldsymbol{R}$,

$$\underline{\boldsymbol{H}}^a = \boldsymbol{H}^a\boldsymbol{R}, \qquad (2.46)$$

$$\underline{\theta}^a = \boldsymbol{R}^T\theta^a \qquad (2.47)$$

and

$$\underline{\zeta}^a = \boldsymbol{R}^T\zeta^a + \frac{1}{2}\boldsymbol{R}^T(\boldsymbol{H}^a)^T(\boldsymbol{s}^{a+1} + \overline{\boldsymbol{s}}^{a+1} - \boldsymbol{s}^a - \overline{\boldsymbol{s}}^a). \qquad (2.48)$$

With the further simplification when $\forall a : \boldsymbol{s}^a = \overline{\boldsymbol{s}}^a = 0$, we get

$$\underline{\mathsf{w}} = [\mathrm{diag}(\boldsymbol{R}, \boldsymbol{R}, \ldots, \boldsymbol{R})]^T\mathsf{w} = \boldsymbol{R}_n^T\mathsf{w}, \qquad (2.49)$$

where $\boldsymbol{R}_n \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a block diagonal matrix consisting of $4n-2$ blocks $\boldsymbol{R} \in \mathbb{R}^{3\times3}$, and $n$ is the number of base pairs of the oligomer.

If we allow moving reference points, i.e. $\boldsymbol{s}^a \neq \overline{\boldsymbol{s}}^a \neq 0$, or apply different rotations to each base frame, i.e. $\overline{\boldsymbol{R}}^a \neq \boldsymbol{R}^a \neq \boldsymbol{R}$, the relation between the elements of $\underline{\mathsf{w}}$ and $\mathsf{w}$ becomes nonlinear, so there is no simple expression relating these two vectors.

# Chapter 3

# An oligomer dependent rigid base model

In this chapter we describe a mechanical model of a rigid base DNA, which was presented in [36] and [24]. This model is the starting point for developing a family of models, discussed in Chapters 6 and 7.

## 3.1 Distribution of DNA configurations

Let $\mathsf{w} \in \mathbb{R}^{12n-6}$ be a vector of internal coordinates of a rigid-base model of DNA for an oligomer of $n$ base pairs, with rotations scaled by 1/5 of a radian and translations scaled by 1Å, as described in the previous chapter. We assume that the equilibrium distribution of $\mathsf{w}$ for a DNA molecule in contact with a heat bath at absolute temperature $T$ is described by the density function [36, 74]

$$\rho(\mathsf{w}) = \frac{1}{Z_J} e^{-\mathsf{U}(\mathsf{w})/k_B T}\, J(\mathsf{w})\, d\mathsf{w}, \quad Z_J = \int e^{-\mathsf{U}(\mathsf{w})/k_B T}\, J(\mathsf{w})\, d\mathsf{w}. \tag{3.1}$$

Here $k_B$ is the Boltzmann constant, $Z_J$ is a normalising constant, $\mathsf{U}(\mathsf{w}) \in \mathbb{R}$ is called the internal elastic energy function (or the potential energy, or the free energy of the molecule) and $J$ is a Jacobian factor which arises due to the non-Cartesian nature of the rotational coordinates, namely [36, 74]

$$J(\mathsf{w}) = \left[\prod_{a=1}^{n-1} \frac{1}{(1 + \frac{1}{100}|\theta^a|^2)^2}\right] \left[\prod_{a=1}^{n} \frac{1}{(1 + \frac{1}{100}|\vartheta^a|^2)^2}\right]. \tag{3.2}$$

The $\frac{1}{100}$ multiplier in the Jacobian factor, in contrast to $\frac{1}{4}$ used in [36], arises due to the choice of parameter scaling, as shown later in (3.6).

Further in this work we will only consider the case when the internal energy function $\mathsf{U}$ is of the general quadratic form

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}(\mathsf{w} - \widehat{\mathsf{w}}) \cdot \mathsf{K}(\mathsf{w} - \widehat{\mathsf{w}}) + \widehat{\mathsf{U}}, \tag{3.3}$$

where $\mathsf{K} \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a symmetric, positive-definite matrix of stiffness parameters, $\widehat{\mathsf{w}} \in \mathbb{R}^{12n-6}$ is a vector of shape parameters that defines the ground or minimum energy state. Notice that because the coordinate vector $\widehat{\mathsf{w}}$ is invariant under global rigid body transformations of the whole DNA molecule, the energy (3.3) is also invariant under these transformations, i.e. if the whole molecule is rigidly moved, its energy $\mathsf{U}$ remains unchanged.

The constant $\widehat{\mathsf{U}} \geq 0$ represents the energy of the ground state $\widehat{\mathsf{w}}$ compared to an unstressed state. Thus $\widehat{\mathsf{U}} = 0$ implies that the ground state is unstressed, whereas $\widehat{\mathsf{U}} > 0$ implies that it is stressed. In the latter case, the oligomer is referred to as being pre-stressed or frustrated. We are unaware of a prior discussion of the concept of of pre-stress in the context of a coarse-grain model of DNA. Indeed it cannot arise in a local rigid base pair description with their topologically linear chain of interactions. However the possibility of pre-stress or frustration is a natural consequence of the double chain topology of interactions associated with a rigid base description.

Notice that the internal configuration density $\rho$, and hence the statistical mechanical average of any internal state function $\phi$, is invariant under shifts of the oligomer internal energy $\mathsf{U}$. Hence the actual value of the oligomer frustration energy $\widehat{\mathsf{U}}$ appearing in (3.3) does not explicitly affect the statistical properties of the system. This is consistent with the fact that the internal energy of the system is itself only defined up to an arbitrary constant. In (3.3), the arbitrary constant is chosen so that zero energy corresponds to an unstressed state, which for a given oligomer may or may not correspond to an accessible state.

To define dimensionless quantities associated with the energies, we use the scale $k_BT$ and define a dimensionless internal energy by $\underline{\mathsf{U}} = \mathsf{U}/k_BT$. Substituting this relation into (3.3), we obtain

$$\underline{\mathsf{U}}(\mathsf{w}) = \frac{1}{2}(\mathsf{w} - \widehat{\mathsf{w}}) \cdot \underline{\mathsf{K}}(\mathsf{w} - \widehat{\mathsf{w}}) + \underline{\widehat{\mathsf{U}}}, \tag{3.4}$$

where

$$\underline{\mathsf{K}} = \mathsf{K}/(k_BT), \quad \underline{\widehat{\mathsf{U}}} = \widehat{\mathsf{U}}/(k_BT). \tag{3.5}$$

If we had decided to pick a new scaling for configuration parameters, so that $\underline{\mathsf{w}} = \mathsf{G}_n^{-1}\mathsf{w}$ and $\underline{\widehat{\mathsf{w}}} = \mathsf{G}_n^{-1}\widehat{\mathsf{w}}$, with $\mathsf{G} = \mathrm{diag}(g, g, g, \ell, \ell, \ell) \in \mathbb{R}^{6\times6}$, then the scaled and nondimensionalised stiffness matrix would become $\underline{\mathsf{K}} = \mathsf{G}_n\mathsf{K}\mathsf{G}_n/(k_BT)$ and the transformed Jacobian factor would be

$$\underline{J}(\underline{\mathsf{w}}) = \left[\prod_{a=1}^{n-1}\left(1 + \frac{1}{100}g^2|\underline{\theta}^a|^2\right)^{-2}\right]\left[\prod_{a=1}^{n}\left(1 + \frac{1}{100}g^2|\underline{\vartheta}^a|^2\right)^{-2}\right]. \tag{3.6}$$

The dimensionless form of the internal configuration density is then obtained as

$$\underline{\rho}(\mathsf{w}) = \frac{1}{\underline{Z}_J}e^{-\underline{\mathsf{U}}(\underline{\mathsf{w}})}\,\underline{J}(\underline{\mathsf{w}}), \quad \underline{Z}_J = \int e^{-\underline{\mathsf{U}}(\underline{\mathsf{w}})}\,\underline{J}(\underline{\mathsf{w}})\,d\underline{\mathsf{w}}. \tag{3.7}$$

In our further developments, we will only use the dimensionless formulation (3.7) with the scale, described earlier, and we will drop the underline notation for convenience.

Now let us consider the impact of rotating every base frame by the same matrix $\boldsymbol{R}$, as described in the Chapter 2, Section 2.6, on the configuration density function. We know that the new configuration vector, after performing these rotations, is

$$\underline{\mathsf{w}} = \boldsymbol{R}_n^T\mathsf{w}. \tag{3.8}$$

Therefore the new average shape vector is

$$\underline{\widehat{\mathsf{w}}} = \boldsymbol{R}_n^T \widehat{\mathsf{w}}. \tag{3.9}$$

(And the underline notation is independent of the scaling notation in Section 2.4.) Moreover, the Jacobian factor stays the same, i.e. $\underline{J} = J$, because rotation of a vector does not change its norm. As the rotated frames still present orientations of the same rigid bases, we want the internal energy of the molecule to stay the same, as before the rotations, so that $\underline{\mathsf{U}} = \mathsf{U}$. Then the transformed stiffness matrix is

$$\underline{\mathsf{K}} = \boldsymbol{R}_n^T \mathsf{K} \boldsymbol{R}_n. \tag{3.10}$$

Note that multiplying by a rotation matrix does not change the norm of a stiffness block, therefore this transformation cannot change the general sparsity of a stiffness matrix. This means that applying the same rotation to all reference frames we could diagonalise some of the $3 \times 3$ off-diagonal stiffness blocks, but not set them to zero.

## 3.2 Material assumptions and symmetries

We will assume that the material parameters $\widehat{\mathsf{U}}$, $\widehat{\mathsf{w}}$ and $\mathsf{K}$ are completely determined by the oligomer length $n$ and sequence $\mathtt{X}_1 \cdots \mathtt{X}_n$ along the reference strand. This means that there exist functions $\mathbb{U}$, $\mathbb{W}$ and $\mathbb{K}$ such that

$$\begin{gathered} \widehat{\mathsf{U}} = \mathbb{U}(n, \mathtt{X}_1, \dots, \mathtt{X}_n), \\ \widehat{\mathsf{w}} = \mathbb{W}(n, \mathtt{X}_1, \dots, \mathtt{X}_n), \quad \mathsf{K} = \mathbb{K}(n, \mathtt{X}_1, \dots, \mathtt{X}_n). \end{gathered} \tag{3.11}$$

Let $\mathsf{U}(\mathsf{w})$ and $\mathsf{U}_*(\mathsf{w}_*)$ denote the internal energies of an oligomer computed using the two different choices of reference strand: the sequence along one is $\mathtt{X}_1 \cdots \mathtt{X}_n$, and along the other it is $\mathtt{X}_1^* \cdots \mathtt{X}_n^*$. Thus $\mathsf{U}(\mathsf{w})$ is given by the expression in (3.3) with the parameters in (3.11), and $\mathsf{U}_*(\mathsf{w}_*)$ is given by an exactly analogous expression with the parameters

$$\begin{gathered} \widehat{\mathsf{U}}_* = \mathbb{U}(n, \mathtt{X}_1^*, \dots, \mathtt{X}_n^*), \\ \widehat{\mathsf{w}}_* = \mathbb{W}(n, \mathtt{X}_1^*, \dots, \mathtt{X}_n^*), \quad \mathsf{K}_* = \mathbb{K}(n, \mathtt{X}_1^*, \dots, \mathtt{X}_n^*). \end{gathered} \tag{3.12}$$

We would like the value of the internal energy function to be independent of the choice of a reference strand, so then $\mathsf{U}(\mathsf{w})$ must equal $\mathsf{U}_*(\mathsf{w}_*)$ for all possible configurations. Using the change of reference strand relations, described in Chapter 2, we deduce that the functions $\mathbb{U}$, $\mathbb{W}$ and $\mathbb{K}$ must satisfy

$$\begin{aligned} \mathbb{U}(n, \mathtt{X}_1, \dots, \mathtt{X}_n) &= \mathbb{U}(n, \overline{\mathtt{X}}_n, \dots, \overline{\mathtt{X}}_1), \\ \mathbb{W}(n, \mathtt{X}_1, \dots, \mathtt{X}_n) &= \mathsf{E}_n \mathbb{W}(n, \overline{\mathtt{X}}_n, \dots, \overline{\mathtt{X}}_1), \\ \mathbb{K}(n, \mathtt{X}_1, \dots, \mathtt{X}_n) &= \mathsf{E}_n \mathbb{K}(n, \overline{\mathtt{X}}_n, \dots, \overline{\mathtt{X}}_1) \mathsf{E}_n, \end{aligned} \tag{3.13}$$

with the matrix $\mathsf{E}_n \in \mathbb{R}^{(12n-6)\times(12n-6)}$, defined in Chapter 2, Section 2.5. We conclude that the stiffness parameters delivered by $\mathbb{K}$ must behave in concordance to the shape parameters $\mathbb{W}$; namely, parameters for odd-even couplings change sign under a change of reference strand, whereas parameters for odd-odd and even-even couplings remain unaltered.

In the special case of a palindromic molecule, when the sequence is the same for both choices of reference strand, i.e. when $\mathtt{X}_1 \cdots \mathtt{X}_n$ is the same as $\overline{\mathtt{X}}_1 \cdots \overline{\mathtt{X}}_n$, the relations (3.13) imply special symmetry conditions on the functions $\mathbb{W}$ and $\mathbb{K}$. Specifically, the components of $\widehat{\mathsf{w}} = \mathbb{W}(n, \mathtt{X}_1, \ldots, \mathtt{X}_n) = (\hat{\boldsymbol{y}}^1, \hat{\boldsymbol{z}}^1, \hat{\boldsymbol{y}}^2, \hat{\boldsymbol{z}}^2, ..., \hat{\boldsymbol{y}}^n, \hat{\boldsymbol{z}}^n, \hat{\boldsymbol{y}}^{n+1}) \in \mathbb{R}^{12n-6}$ have to be such that $\hat{\boldsymbol{y}}^a = \mathsf{E}\, \hat{\boldsymbol{y}}^{n-a+1}$ and $\hat{\boldsymbol{z}}^a = \mathsf{E}\, \hat{\boldsymbol{z}}^{n-a}$, $a = 1 \ldots n$. This means that the equilibrium values of the odd coordinates (Buckle, Shear, Tilt and Shift) have to be anti-symmetric about the middle of the molecule, whereas the equilibrium values of the even coordinates must be symmetric. We remark that for a palindromic oligomer the number of base pairs $n$ has to be even. Then for $a = n/2$ we get $\hat{\boldsymbol{z}}^a = \mathsf{E}\, \hat{\boldsymbol{z}}^a$, so the equilibrium values of Tilt and Shift in the middle junction must be zero.

Similarly, we deduce that the stiffness parameters, corresponding to odd-odd and even-even couplings, must be symmetric, while the ones corresponding to odd-even couplings must be antisymmetric. Also the stiffness parameters, associated with the odd-even couplings of the middle junction have to vanish for a palindromic molecule.

## 3.3 Moment-parameter relations

Suppose we have a big enough data sample of configurations from the distribution (3.1). Then the values of the parameters $\widehat{\mathsf{w}}$ and $\mathsf{K}$ can be found by computing the statistical mechanical averages of the appropriately chosen functions of the configuration vector $\mathsf{w}$. The method for extracting DNA average shape and stiffness parameters for a distribution with a Jacobian was described in [23].

The statistical mechanical average of any state function $\phi = \phi(\mathsf{w})$ with respect to the density $\rho(\mathsf{w})$ is defined as

$$\langle \phi \rangle = \int_{\mathbb{R}^{12n-6}} \phi(\mathsf{w}) \rho(\mathsf{w})\, d\mathsf{w}. \tag{3.14}$$

It is not always possible to get an explicit expression of the integral (3.14). However, using the standard results of the Gaussian integrals [29], it can be easily verified that

$$\left\langle \frac{\mathsf{w}_i}{J} \right\rangle = \frac{1}{Z_J} \int_{\mathbb{R}^{12n-6}} \mathsf{w}_i\, e^{-\frac{1}{2}(\mathsf{w}-\widehat{\mathsf{w}})\cdot\mathsf{K}(\mathsf{w}-\widehat{\mathsf{w}})}\, d\mathsf{w} = \frac{Z_N}{Z_J}\, \widehat{\mathsf{w}}_i, \tag{3.15}$$

where $1 \leq i \leq 12n-6$, $\mathsf{K}$ is a symmetric, positive-definite matrix and $Z_N$ is a normalising constant of the Gaussian distribution with parameters $(\widehat{\mathsf{w}}, \mathsf{K}^{-1})$, $Z_N = (2\pi)^{-6n+3} \sqrt{\det[K]}$, and $Z_J$ is defined in (3.1). Similarly, it can be shown that

$$\begin{aligned} \left\langle \frac{(\mathsf{w}_i - \widehat{\mathsf{w}}_i)(\mathsf{w}_j - \widehat{\mathsf{w}}_j)}{J} \right\rangle &= \frac{1}{Z_J} \int_{\mathbb{R}^{12n-6}} (\mathsf{w}_i - \widehat{\mathsf{w}}_i)(\mathsf{w}_j - \widehat{\mathsf{w}}_j)\, e^{-\frac{1}{2}(\mathsf{w}-\widehat{\mathsf{w}})\cdot\mathsf{K}(\mathsf{w}-\widehat{\mathsf{w}})}\, d\mathsf{w} \\ &= \frac{Z_N}{Z_J}\, [\mathsf{K}^{-1}]_{ij}, \end{aligned} \tag{3.16}$$

for $1 \leq i, j \leq n$. And finally, one can get the missing constant:

$$\frac{Z_N}{Z_J} = \frac{1}{Z_J} \int_{\mathbb{R}^{12n-6}} e^{-\frac{1}{2}(\mathsf{w}-\widehat{\mathsf{w}})\cdot\mathsf{K}(\mathsf{w}-\widehat{\mathsf{w}})}\, d\mathsf{w} = \left\langle \frac{1}{J} \right\rangle. \tag{3.17}$$

These results can be written in more compact vector and matrix forms, namely

$$\frac{\langle \mathsf{w}/J\rangle}{\langle 1/J\rangle} = \widehat{\mathsf{w}} \tag{3.18}$$

and

$$\frac{\langle \Delta\mathsf{w} \otimes \Delta\mathsf{w}/J\rangle}{\langle 1/J\rangle} = \mathsf{K}^{-1}, \tag{3.19}$$

where $\Delta\mathsf{w} = \mathsf{w} - \widehat{\mathsf{w}}$.

We can discretise the relations (3.18) and (3.19) to get formulas for estimating parameters $\widehat{\mathsf{w}}$ and $\mathsf{K}$ from a data sample. Let $\{\mathsf{w}^{(i)}\}$, $i = 1\dots N$ be a set of coordinate vectors, corresponding to a sufficiently large number $N$ of observed configurations of the oligomer at a fixed temperature $T$. Then an estimate for equilibrium shape parameters $\widehat{\mathsf{w}}$, denoted $\widehat{\mathsf{w}}_E$, can be obtained as

$$\widehat{\mathsf{w}}_E = \frac{\sum\limits_{i=1}^{N} \mathsf{w}^{(i)}/J^{(i)}}{\sum\limits_{i=1}^{N} 1/J^{(i)}} \tag{3.20}$$

where the Jacobian factor $J^{(i)}$ is given by (3.2). Thus, $\widehat{\mathsf{w}}_E$ is computed using a discrete analogy of (3.18). Similarly, the computed estimate for the stiffness matrix $\mathsf{K}$, denoted $\mathsf{K}_E$, can be obtained from a discrete version of (3.19):

$$\frac{\sum\limits_{i=1}^{N} \left(\Delta\mathsf{w}^{(i)} \otimes \Delta\mathsf{w}^{(i)}/J^{(i)}\right)}{\sum\limits_{i=1}^{N} 1/J^{(i)}} = [\mathsf{K}_E]^{-1}. \tag{3.21}$$

where $\Delta\mathsf{w}^{(i)} = \mathsf{w}^{(i)} - \widehat{\mathsf{w}}_E$.

## 3.4 Rigid base pair marginal distribution

One can decide to ignore the intra base pair configuration parameters, and consider a marginal distribution, described by the density function

$$\widetilde{\rho}(\mathsf{w}^{\mathrm{bp}}) = \frac{1}{Z_J}\int_{\mathbb{R}^{6n}} e^{-\mathsf{U}(\mathsf{w})/k_B T}\, J(\mathsf{w})\, d\mathsf{w}^{\mathrm{b}}, \tag{3.22}$$

where $\mathsf{w}^{\mathrm{bp}} = (\boldsymbol{z}^1, \boldsymbol{z}^2, ..., \boldsymbol{z}^{n-1}) \in \mathbb{R}^{6n-6}$ is a vector containing all inter base pair configuration parameters and $\mathsf{w}^{\mathrm{b}} = (\boldsymbol{y}^1, \boldsymbol{y}^2, ..., \boldsymbol{y}^n) \in \mathbb{R}^{6n}$ is a vector of all intra base pair configuration parameters. Unfortunately, in the presence of the Jacobian factor, we do not know how to integrate (3.22) to find an explicit expression for the function $\widetilde{\rho}(\mathsf{w}^{\mathrm{bp}})$.

If we assume that the base pairs are always close to their ideal shape, defined by the Tsukuba convention in [54], then the values of $\boldsymbol{y}^a$ and $\hat{\boldsymbol{y}}^a$ are close to zero for all $a$. The intra base pair term $(1 + |\vartheta^a|^2/100)^{-2}$ in the Jacobian (3.2) can then be approximated by one, so the Jacobian becomes

$$J^{\mathrm{bp}}(\mathsf{w}^{\mathrm{bp}}) = \prod_{a=1}^{n-1} \frac{1}{(1+\frac{1}{100}|\theta^a|^2)^2}, \tag{3.23}$$

and the distribution of $\mathsf{w}^{\mathrm{b}}$ is Gaussian.

Another standard result of Gaussian integrals is that a marginal of a Gaussian distribution is also a Gaussian. Namely, if $\boldsymbol{x} \sim N(\hat{\boldsymbol{x}}, \boldsymbol{\Sigma})$, then $\boldsymbol{x}_1 \sim N(\hat{\boldsymbol{x}}_1, \boldsymbol{\Sigma}_{11})$, i.e.

$$\frac{\left(\frac{\beta}{\pi}\right)^{\frac{n}{2}}}{\sqrt{\det[K^{-1}]}} \int_{\mathbb{R}^m} e^{-\beta(\boldsymbol{x}-\hat{\boldsymbol{x}})\cdot K(\boldsymbol{x}-\hat{\boldsymbol{x}})}\, d\boldsymbol{x}_2 = \frac{\left(\frac{\beta}{\pi}\right)^{\frac{k}{2}}}{\sqrt{\det \boldsymbol{\Sigma}_{11}}}\, e^{-\beta(\boldsymbol{x}_1-\hat{\boldsymbol{x}}_1)\cdot\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\hat{\boldsymbol{x}}_1)}. \tag{3.24}$$

Here $\beta > 0$, $n \geq 0$, $K = \boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{n\times n}$ is a symmetric, positive - definite matrix,

$$\begin{aligned} \boldsymbol{x} &= \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}, \quad \hat{\boldsymbol{x}} = \begin{bmatrix} \hat{\boldsymbol{x}}_1 \\ \hat{\boldsymbol{x}}_2 \end{bmatrix}, \quad \boldsymbol{x}_1, \hat{\boldsymbol{x}}_1 \in \mathbb{R}^k, \quad \boldsymbol{x}_2, \hat{\boldsymbol{x}}_2 \in \mathbb{R}^m, \\ \boldsymbol{\Sigma} &= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{11}^T \in \mathbb{R}^{k\times k}, \quad \boldsymbol{\Sigma}_{12} \in \mathbb{R}^{k\times m}, \\ & \boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_{22}^T \in \mathbb{R}^{m\times m}, \quad \text{and} \quad k+m=n. \end{aligned}$$

Therefore, using (3.24) we deduce that, when base pairs are close to rigid, the marginal probability density function of the inter base pair variables is

$$\widetilde{\rho}(\mathsf{w}^{\mathrm{bp}}) = \frac{1}{Z_J^{\mathrm{bp}}} e^{-\mathsf{U}(\mathsf{w}^{\mathrm{bp}})}\, J^{\mathrm{bp}}(\mathsf{w}^{\mathrm{bp}}), \quad Z_J^{\mathrm{bp}} = e^{-\mathsf{U}(\mathsf{w}^{\mathrm{bp}})}\, J^{\mathrm{bp}}(\mathsf{w}^{\mathrm{bp}})\, d\mathsf{w}. \tag{3.25}$$

Here

$$\mathsf{U}^{\mathrm{bp}}(\mathsf{w}^{\mathrm{bp}}) = \frac{1}{2}(\mathsf{w}^{\mathrm{bp}} - \widehat{\mathsf{w}}^{\mathrm{bp}}) \cdot \mathsf{K}^{\mathrm{bp}}(\mathsf{w}^{\mathrm{bp}} - \widehat{\mathsf{w}}^{\mathrm{bp}}) + \widehat{\mathsf{U}}^{\mathrm{bp}}, \tag{3.26}$$

and a symmetric, positive definite matrix $\mathsf{K}^{\mathrm{bp}} \in \mathbb{R}^{(6n-6)\times(6n-6)}$, which is, as in (3.19), the inverse of a “weighted" covariance matrix of $\mathsf{w}^{\mathrm{bp}}$.

# Chapter 4

# Molecular dynamics and data

In this chapter we describe the origin and properties of the molecular dynamics data set, used for fitting the parameters of our model.

## 4.1 Molecular dynamics simulations

To be able to estimate the shape and stiffness parameters in the probability density function (3.1) of a given oligomer, one needs a sufficient number of observations of its configurations. There exists data from crystal structure and NMR experiments of short DNA oligomers [55, 1]. However, the existing methods only allow the computation of average shape parameters and diagonal blocks of covariance matrices, assuming their local sequence dependence. In addition, the data is available only for a relatively small number of oligomers.

An alternative to direct experimental observations is molecular dynamics (MD) simulations, which have been used for modelling DNA since 1983 [44]. MD simulations consist of numerically modelling the interactions of particles over a period of time, by creating trajectories via integration of Newton's equations. There exist different MD software packages (AMBER, CHARMM, GROMACS, GROMOS), the most widely used ones for atomistic simulations of DNA being AMBER [57, 12] and CHARMM [10, 47, 9]. We used the AMBER suite of programs in this work, mainly because it was used by the ABC consortium [7, 20, 42] that provided a large part of our data set. However, it has been shown that AMBER and CHARMM produce similar results for B-DNA [59, 61].

The idea of molecular dynamics simulations is based on Newton's second law of motion

$$\boldsymbol{F}_i = m_i \boldsymbol{a}_i = m_i \frac{\mathrm{d}^2 \boldsymbol{r}_i}{\mathrm{d}t^2}, \tag{4.1}$$

applied to every particle $i$ of the system, $i = 1 \dots N$. Here $m_i \in \mathbb{R}$ is the mass of the particle, $\boldsymbol{r}_i \in \mathbb{R}^3$ is its position in Cartesian coordinates and $\boldsymbol{a}_i$ is acceleration. We performed atomistic MD simulations, where the particles are atoms. The force $\boldsymbol{F}_i \in \mathbb{R}^3$, acting on the atom $i$, is given as a derivative of the potential energy of the system $U \in \mathbb{R}$ with respect to its position $\boldsymbol{r}_i$:

$$\boldsymbol{F}_i = -\frac{\partial U(\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_N)}{\partial \boldsymbol{r}_i}. \tag{4.2}$$

Knowing the force, it is possible to integrate numerically the equations of motion (4.1) for every $i$, to get the trajectories $\boldsymbol{r}_i(t)$ of all the atoms in the system.

The potential energy function (which is also called the force field) is usually the sum of bonded and non-bonded interaction potentials,

$$U = U_{\text{bonded}} + U_{\text{non-bonded}}, \tag{4.3}$$

where $U_{\text{bonded}}$ is the interaction energy of particles connected by covalent bonds and $U_{\text{non-bonded}}$ is the interaction energy of particles that are not covalently bonded (electrostatic and van der Waals energy). The AMBER potential, which we used for our simulations, is given by

$$\begin{aligned} U(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) &= \sum_{\text{bonds}} k_b (r_b - \hat{r}_b)^2 + \sum_{\text{angles}} k_a (\theta_a - \hat{\theta}_a)^2 \\ &+ \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\,\phi_d - \delta_d)] \\ &+ \sum_{i<j} \left[ \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right] + \sum_{i<j} \left[ \frac{q_i\, q_j}{\epsilon\, r_{ij}} \right], \end{aligned} \tag{4.4}$$

where the first three terms describe the bonded interactions and the last two the non-bonded.

What do the terms in this equation mean? The first term corresponds to the energy of stretching a covalent bond, modelled by a harmonic potential. Here $k_b$ is a stiffness constant, $\hat{r}_b$ is an equilibrium distance and $r_b$ is the observed distance between the two bonded atoms for each bond $b$. The second term is the harmonic energy of the angle between two bonds, that share a common atom (valence angle). Here, again, for each such angle $a$, $k_a$ is a stiffness constant, $\hat{\theta}_a$ is the equilibrium value and $\theta_a$ is the observed value of the angle. The third term depends on four successively bonded atoms and corresponds to a dihedral (torsion) angle potential. A torsion angle for four such atoms $i, j, k, l$, associated with the bond between the middle atoms $j$ and $k$, is the angle between the two planes, going through the atoms $i, j, k$ and $j, k, l$. Then $\delta_d$ is the value corresponding to the minimal value of the energy term, $n$ is the number of energy minima as the angle $d$ changes from 0 to $2\pi$, $V_n$ is the energy barrier height and $\phi_d$ is the observed value of the angle $d$.

The fourth and fifth terms of equation (4.4) represent the van der Waals and electrostatic interactions between atoms. The van der Waals force is the sum of attractive and repulsive forces between a pair of atoms. One simple model, that approximates this interaction, is the the Lennard-Jones potential (also called the 12-6 potential). It is the fourth term of equation (4.4). In this model, the $1/r_{ij}^{12}$ term corresponds to a short-range repulsive force and the $1/r_{ij}^{6}$ term to a long-range attractive force. Here $A_{ij} = \varepsilon_{ij}\, \hat{r}_{ij}^{12}$ and $B_{ij} = 2\,\varepsilon_{ij}\, \hat{r}_{ij}^{6}$, where $\varepsilon_{ij}$ is the potential well depth, $\hat{r}_{ij}$ is the distance between the atoms $i, j$ at which the potential is zero, and $r_{ij}$ is the observed distance. The electrostatic interaction between two atoms $i, j$ is represented by Coulomb potential, the last term of equation (4.4). Here $\epsilon$ is the Coulomb constant, $r_{ij}$ is the distance between the two atoms and $q_i, q_k$ are their charges.

The non-bonded energy terms are defined for every pair of atoms in the system, which makes evaluating these terms for big systems very expensive. One solution to this problem would be to ignore interactions between atoms separated by a distance greater than a chosen cutoff. However, non-bonded, especially electrostatic, interactions sometimes can be important even over quite long distances. A number of models have been developed for efficiently including an approximation of long-range interactions to the potential. AMBER uses the Particle Mesh Ewald procedure [21] for estimating the long-range electrostatic interactions, and a continuous cutoff for the long-range van der Waals interactions. Even still, evaluating the function $U$ is the most computationally intensive task during an MD simulation.

Setting the values of the parameters $k_b, \hat{r}_b, k_a, \hat{\theta}_a, V_n$, etc. of the force field $U$ is a whole area of research. Usually these values are optimised to fit experimental data and results from quantum mechanical simulations of a small number of particles. Subsequently, they are tested on MD simulations for different systems. The parameter sets used for our MD simulations are further detailed in Section 4.2.

To avoid modelling forces acting on the surface particles and effects of the container walls, periodic boundary conditions, are imposed. This corresponds to surrounding the simulation box by an infinite number of its copies. Then the particles close to the border of the box interact to the images of the particles that are close to the opposite border.

Finally, having estimated a potential and chosen the boundary conditions, we can numerically solve the system of ordinary differential equations (4.1). The $N$ second order differential equations in (4.1) can be transformed to $2N$ first order equations:

$$\begin{aligned} \frac{\mathrm{d}\,\boldsymbol{v}_i}{\mathrm{d}t} &= \boldsymbol{a}_i \left( = \frac{\boldsymbol{F}_i}{m_i} \right), \\ \frac{\mathrm{d}\boldsymbol{r}_i}{\mathrm{d}t} &= \boldsymbol{v}_i, \end{aligned} \tag{4.5}$$

where $\boldsymbol{v}_i$ is the velocity of a particle $i$. Then these equations can be discretised by choosing a small enough integration time step $\Delta t$. For solving them, AMBER uses the Leapfrog integration algorithm, with a recursive scheme

$$\begin{aligned} \boldsymbol{r}_i^{k+1} &= \boldsymbol{r}_i^k + \boldsymbol{v}_i^k\,\Delta t + \tfrac{1}{2}\,\boldsymbol{a}_i^k\,\Delta t^{\,2}, \\ \boldsymbol{v}_i^{k+1} &= \boldsymbol{v}_i + \tfrac{1}{2}\,(\boldsymbol{a}_i^k + \boldsymbol{a}_i^{k+1})\,\Delta t. \end{aligned} \tag{4.6}$$

The time step must be chosen small enough to avoid discretisation errors, it should be smaller than the fastest internal vibrations of the system, that are in the order of a femtosecond. Placing constraints on the bonds which oscillate with the highest frequency allows to the time step to be shortened and this reduces the computational cost.

The integration method is deterministic: given the positions and velocities of each atom at a chosen time, they can be predicted for any future time. However, in practice, because of the accumulating round-off errors these predictions become inaccurate. Having computed the trajectories $\{\boldsymbol{r}_i^k\}$ of all the atoms $i = 1 \ldots N$, the macroscopic properties of the system, such as temperature and pressure, can be obtained using statistical mechanics. More details about the method of MD simulations can be found in the books [43] and [47].

To start a molecular dynamics simulation, the initial positions of all the atoms in the system have to be assigned. It is important to choose the initial configuration not far from the minimum of the potential. Otherwise, if we start from a configuration that is unrealistically stressed, the structure of our molecule can get unrealistically distorted during the simulation.

There exist databases with experimentally determined X-ray crystal or NMR structures for some biomolecules. Otherwise, one can use various programs to build a starting structure using experimentally obtained configurations of structural parts (e.g. nucleotides) of the molecule that one wishes to model. After building a structure, an energy minimisation is usually carried out to remove the largest strains and possible overlaps of atoms.

In simulations of DNA the solvent (usually water) plays an important role as it screens electrostatic interactions and enforces the stability of the helical structure. There exist implicit solvent models which estimate the approximate effect of the solvent without modelling its every molecule. However in this work we used explicit solvent, which means that each DNA oligomer was put in a sufficiently large box of water molecules. DNA is negatively charged, and it is necessary to add neutralising counterions to balance this charge, again for reasons of stability. Often some neutral pairs of salt ions are added to the system as well to reproduce a physiologically realistic salt concentration. The final system for simulating a solvated DNA 18-mer consists of about 36 000 atoms (this means 72 000 equations in the system (4.5)) of which more than 90% is water (see Figure 4.1).

When the initial positions are set for all the atoms of DNA, water and ions, small velocities are assigned to each atom, and the simulation is started. Then the velocities are periodically increased to heat the system, until the chosen temperature is reached. Before starting the final, or production, stage of the simulation, some equilibration, involving energy minimisation, is performed.

Usually, in part to model the conditions of laboratory experiments, some chosen macroscopic values of the system, like temperature, volume, pressure or energy, are kept constant during the simulation. This can be again done by rescaling the velocities of the atoms. A set of all possible systems which have the same macroscopic or thermodynamic state, but different microscopic states, is called a statistical ensemble: in the microcanonical ensemble (NVE) the number of atoms N, volume V and energy E are fixed, in the canonical ensemble (NVT) the properties held constant are the number of atoms N, the volume V and the temperature T, and the isobaric-isothermal ensemble (NPT) is characterised by a constant number of atoms N, a constant pressure P, and a constant temperature T.

The length of an MD simulation is chosen to reach an acceptable convergence of the parameters of interest. However, as the number of particles in the simulated systems is big, the simulations are computationally expensive. To accelerate the simulations, the computations are usually run in parallel on distributed processors. Moreover, a special supercomputer has been designed for molecular dynamics simulations [69], and various collaborations have been established to carry out consistent sets of MD simulations.

## 4.2 ABC data set and simulation protocol

The main part of the data set, used to build our coarse-grain model of DNA, was a shared pool of DNA MD trajectories produced by the ABC collaboration within a consortium of groups [7, 20, 42]. The ABC collaboration was initiated precisely because of interest in coarse-grain parameter extraction, although the data set is also being exploited in a variety of other ways by other members of the consortium. The ABC sequence set are the 39 18-mers $\mu = 1, \ldots, 36$ and $\mu = 54, 55, 56$ listed in Table 4.1, that were designed to include multiple instances of all

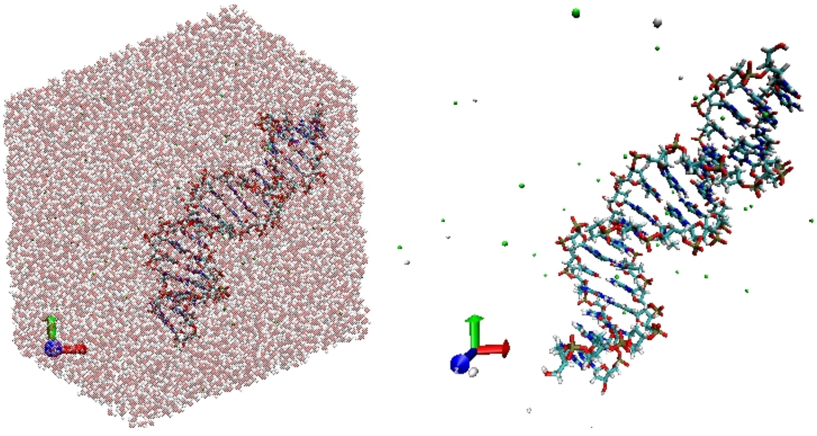

(a) (b)

Figure 4.1: Two snapshots from an MD simulation of the sequence $\mathtt{S}_3$, given in Table 4.1 with atoms marked as beads and covalent bonds as tubes. The system consists of 35 882 atoms of which only 1 138 are DNA, 106 are neutralising and salt ions, and the rest is water. In (a) a cubic simulation box containing all the atoms is shown, with water atoms being transparent. In (b) only DNA and ions are left. Here green beads mark K+ ions and white beads are Cl- ions. The figure was made with VMD [30].

136 possible tetramer sub-sequences away from the ends.

Each DNA sequence $\mathtt{S}_\mu$ was simulated using atomic resolution, explicit solvent, molecular dynamics. The AMBER suite of programs together with the *parmbsc0* [62] modifications to the *parm99* force field [13, 12] was used. Simulations were run in a solution of 150 mM KCl with the parameters from Dang [17], water was modelled using the SPC/E parameters [28] and periodic boundary conditions were imposed. In order to have stable simulations with a 2 fs time step, all chemical bonds involving hydrogen atoms were restrained using the SHAKE [65] algorithm. For each sequence, the DNA duplex was built, neutralised, hydrated and equilibrated as described in [42]. The simulations comprised approximately 36 000 atoms. Each simulation was in the *NPT* ensemble with temperature 300 K and pressure 1 atm, controlled by the Berendsen algorithm [6]. A trajectory of between 50-100 ns was generated (with the length depending largely on the group running the simulation), and a snapshot was saved every 1 ps. Further details of the simulation protocol can be found in [42].

The ABC database consists in total of almost 3 million snapshots of DNA configurations, or 30 gigabytes of data in compressed format, sampled from approximately 2.75 microseconds of (deterministic) simulation. In particular, the configuration of the solvent is discarded, and correspondingly the DNA configurations are assumed to be sampling a stochastic dynamics.

## 4.3 Oligomers with different ends, breaking hydrogen bonds

Every ABC oligomer was chosen to have $5'\mathtt{G}p\mathtt{C}$ and $\mathtt{G}p\mathtt{C}3'$ ends. That choice was made to minimise possible convergence issues in the simulations because the $\mathtt{G}p\mathtt{C}$ end dimers were known to be amongst the most stable against end fraying. In the development of the theory presented here, and as explained more below, we realised that while we did not need all tetramer sub-sequences to be present in our training set, we did however need a greater variety of sequences at the ends. We therefore enhanced the original ABC data set with the oligomers labelled $\mu = 37, \ldots, 53$ in Table 4.1, which, when considering both the reference and complementary strands, contain all 16 possible $5'$-dimer-step ends, and all 16 possible dimer-step-$3'$ ends. As the additional sequences were focused on enhancing the range of end sequences, it was significantly faster to simulate 12-mers rather than 18-mers (the simulations of $n_\mu = 12$ base pair sequences comprised approximately 17 000 atoms, which is more than twice less than approximately 36 000 atoms for $n_\mu = 18$). We did also include three additional 18-mers, in part to verify consistency between the original simulations and those of the extended oligomer set. The simulation protocol of these simulations was exactly the same as ABC and their duration was between 100 ns to 200 ns.

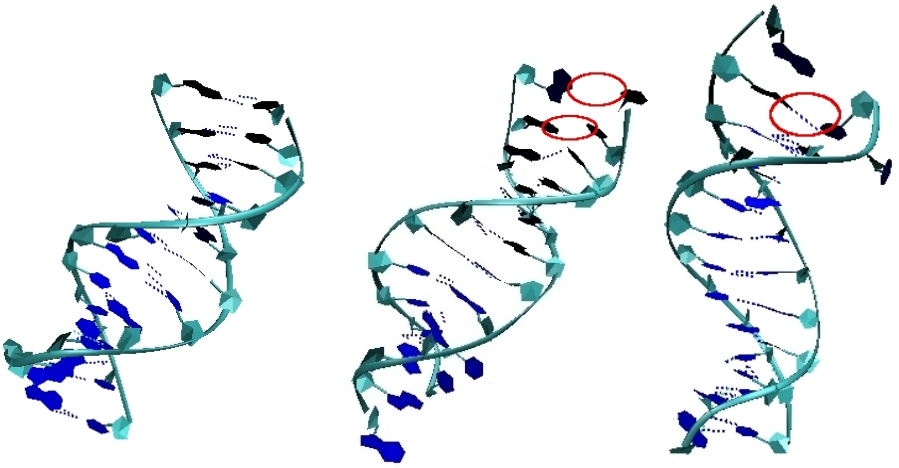

(a) (b) (c)

Figure 4.2: Three snapshots from an MD simulation of the sequence $\mathtt{S}_7$, where backbones are drawn as ribbons, bases are drawn as blue polygons and hydrogen bonds are marked as dotted lines. In (a) all the bases are connected by hydrogen bonds, in (b) all the hydrogen bonds are broken for two base pairs at each end (the two separated base pairs at the top end are marked with a red circle) and in (c) two non Watson-Crick hydrogen bonds are formed between bases `A` and `T` across the junction at the top end (inside the red circle ). The figure was made with VMD [30].

One observation from these simulations is that indeed the $5'\mathtt{G}p\mathtt{C}$ ends tend to break apart much less than the other ones. The least stable are the oligomers, ending with an `AT` (or `TA`) base pair. This base pair has only two hydrogen bonds. For some simulations, at least one hydrogen bond at the end was broken for more than 190 000 snapshots out of 200 000 (200 ns). An interesting case are the sequences $\mu = 37, 38, 39, 45, 46$, that have two `AT` base pairs at one end and three at the other one. The hydrogen bonds at the ends of these oligomers were often breaking and then re-forming. However, sometimes a hydrogen bond was forming between bases across the junction (from neighbouring base pairs), as shown in Figure 4.2. We excluded these configurations from our analysis, as we were only interested in the equilibrium distribution of B-DNA.

| $\mu$ | $\mathsf{S}_\mu$ | $\mathsf{S}_\mu$ | $\mu$ |
|---|---|---|---|
| 1 | GCTATATATATATATAGC | GCTAGATAGATAGATAGC | 29 |
| 2 | GCATTAATTAATTAATGC | GCGCGGGCGGGCGGGCGC | 30 |
| 3 | GCGCATGCATGCATGCGC | GCGTGGGTGGGTGGGTGC | 31 |
| 4 | GCCTAGCTAGCTAGCTGC | GCACTAACTAACTAACGC | 32 |
| 5 | GCCGCGCGCGCGCGCGGC | GCGCTGGCTGGCTGGCGC | 33 |
| 6 | GCGCCGGCCGGCCGGCGC | GCTATGTATGTATGTAGC | 34 |
| 7 | GCTACGTACGTACGTAGC | GCTGTGTGTGTGTGTGGC | 35 |
| 8 | GCGATCGATCGATCGAGC | GCGTTGGTTGGTTGGTGC | 36 |
| 9 | GCAAAAAAAAAAAAAAGC | AAACAATAAGAA | 37 |
| 10 | GCCGAGCGAGCGAGCGGC | AAAGAACAATAA | 38 |
| 11 | GCGAAGGAAGGAAGGAGC | AAATAACAAGAA | 39 |
| 12 | GCGTAGGTAGGTAGGTGC | GGGAGGTGGCGG | 40 |
| 13 | GCTGAGTGAGTGAGTGGC | GGGCGGAGGTGG | 41 |
| 14 | GCAGCAAGCAAGCAAGGC | GGGCGGTGGAGG | 42 |
| 15 | GCAAGAAAGAAAGAAAGC | GGGTGGAGGCGG | 43 |
| 16 | GCGAGGGAGGGAGGGAGC | GGGTGGCGGAGG | 44 |
| 17 | GCGGGGGGGGGGGGGGGC | AAATAAAAATAAGAACAA | 45 |
| 18 | GCAGTAAGTAAGTAAGGC | AAATAACAATAAGAACAA | 46 |
| 19 | GCGATGGATGGATGGAGC | GGGAGGGGGAGGCGGTGG | 47 |
| 20 | GCTCTGTCTGTCTGTCGC | GACATGGTACAG | 48 |
| 21 | GCACAAACAAACAAACGC | ACGATCCTAGCA | 49 |
| 22 | GCAGAGAGAGAGAGAGGC | ATGCTAATCGTA | 50 |
| 23 | GCGCAGGCAGGCAGGCGC | AGCTGAAGTCGA | 51 |
| 24 | GCTCAGTCAGTCAGTCGC | CGAACTTCAAGC | 52 |
| 25 | GCATCAATCAATCAATGC | GTCTACCATCTG | 53 |
| 26 | GCGTCGGTCGGTCGGTGC | GCATAAATAAATAAATGC | 54 |
| 27 | GCTGCGTGCGTGCGTGGC | GCATGAATGAATGAATGC | 55 |
| 28 | GCACGAACGAACGAACGC | GCGACGGACGGACGGAGC | 56 |

Table 4.1: Sequences $\mathsf{S}_\mu$ contained in the MD data set.

## 4.4 Estimates of the oligomer based model parameters and their convergence

We will now discuss the process and some results of oligomer based model parameter estimation from the MD data. The first step of this process was to analyse each snapshot coming from the MD simulations with the program *Curves+* [41], to get a set of reference points and base frames, as described in Section 2.1. From these points and frames we computed a vector of our non dimensional coarse-grain configuration parameters $\mathsf{w}$, as defined in Sections 2.3 and 2.4. To obtain the desired oligomer parameters, we assume ergodicity of these time series, i.e. we assume that the time series that stays in, but explores all of the quadratic potential well where the assumed form of $\mathsf{U}$ is valid. Then the statistical mechanical averages over configuration space appearing in the moment relations (3.18) and (3.19) can be replaced with averages over the time series, so that the average shape vector $\widehat{\mathsf{w}}$ and the stiffness matrix $\mathsf{K}$ that can be estimated using the formulas (3.20) and (3.21). To obtain parameters consistent with the B-form DNA structural family, we follow the treatment in [36] excluding structures with broken H-bonds. Broken bonds provide a signal that a structure is defective, for example, it may have frayed ends. Following previous work [36, 46], we consider a hydrogen bond to be broken if the distance between donor and acceptor is greater than 4 Å.

Various scripts for computing and plotting oligomer parameters were developed from those used in [36]. The oligomer based parameter estimation from the ABC data set was done in collaboration with J. Curuksu.

As mentioned previously, the values of the configuration parameters $\mathsf{w}$ and therefore also the values of the average vector $\widehat{\mathsf{w}}$ and the stiffness matrix $\mathsf{K}$ depend on the choice of a reference strand, the two sets of parameters being related by simple formulas (3.13). For example, the 12-mer described by the sequence $\mathsf{S}_{51}$ = **AGCTGAAGTCGA** from Table 4.1, can alternatively be described by the sequence **TCGACTTCAGCT**, choosing a different reference strand. Figure 4.3 shows the intra rotational variables of the vector $\widehat{\mathsf{w}}_E$ (Buckle, Propeller and Opening), computed from the same 200 ns simulation, but for two different choices of reference strand. The base sequence along the reference strand is indicated on the abscissa, with parameters evaluated at each base-pair and interpolated with piecewise-linear curves for visualisation. While passing from one choice to another, the values of even coordinates (Propeller and Opening) are just reordered, while the values of Buckle also change sign. In our further examples, we use the sequences listed in Table 4.1 to describe the simulated oligomers. However, we will be considering both choices of reference strand for computing various statistics of parameters, dependent on dimer or trimer sequence, as described in later chapters.

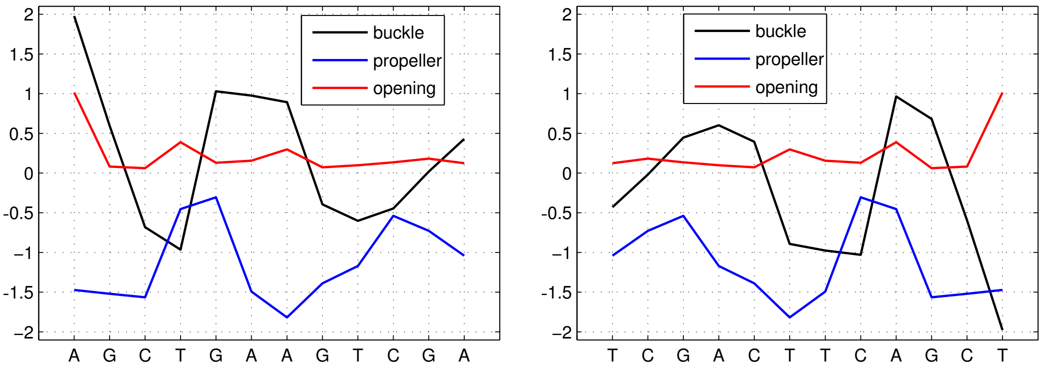


Figure 4.3: Intra rotational average shape parameters of the same oligomer, computed from the same 200 ns simulation, but for two different choices of reference strand, the sequence on the left being $\mathsf{S}_{51}$. Note that Buckle changes sign while passing from one case to the other.

Figure 4.4 presents a standard set of parameter plots, that will be used throughout this thesis. There are eight panels in the figure. The top four of them contain plots of the shape vector $\widehat{\mathsf{w}}_{4,E}$ for the sequence $\mathsf{S}_4$ versus sequence position, with values interpolated by solid lines. The intra variables are evaluated at each base-pair and inter variables at each junction. At the bottom of the page, the remaining four panels show the diagonal entries of the stiffness matrix $\mathsf{K}_{4,E}$. Even though the stiffness matrix has many non zero entries, as shown in Figures 4.6 and 4.7, it is interesting to observe the sequence dependence of the diagonal entries. The diagonal stiffness parameters are named according to the couplings that they represent, i.e. Buckle-Buckle, Propeller-Propeller, etc. Their presentation in the plots is analogous to the shape parameters above.

The goal of the current stage of the ABC project is to have a data set, which consists of at least $1\mu$s simulated trajectories for each of the 39 sequences $\mathsf{S}_\mu$, $\mu = 1, \ldots, 36$ and $\mu = 54, 55, 56$. Even though the complete microsecond data set was not available yet by the time of writing this thesis, we already had the simulations accomplished for some of the sequences, so the parameter values estimated from the original data set could be compared to the ones estimated from the prolonged (ten or twenty times) simulations. The dashed lines in Figure 4.4 show shape and diagonal stiffness parameters for the same sequence $\mathsf{S}_4$, computed from a $1\,\mu$s simulation, where the first 100 ns are the same as before. The biggest differences between the two estimates are for the Twist-Twist diagonal stiffness parameters close to the ends of the oligomer. Otherwise the two lines are almost overlapping, indicating that the 100 ns simulation was converged enough to estimate our parameters of interest. In later examples of this thesis we will not use the $1\mu$s simulation data for consistency, unless stated differently.

Another convergence check was done together with J. Curuksu. We ran ten independent 50 ns simulations of a palindromic sequence $\mathsf{S}_3$, with different random initial velocities assigned in each of them. Then we could compare the average shapes and diagonal stiffness parameters computed from these simulations with the values obtained from the 100 ns ABC simulation. The plots of these values are shown in Figure 4.5, the ABC simulation parameters being interpolated by piecewise-linear curves and the other ten estimates marked with crosses. To quantify the spread of these ten estimated values, we computed the standard deviations of each parameter at each base pair or each junction. As can be noticed from the plots, the deviation of the values is bigger close to the ends of the oligomer, which implies that the estimates close to ends should be trusted less than the ones in the middle. The value of standard deviation, averaged over all the shape parameters, is 0.04, and the analogous value for the diagonal stiffness parameters is 0.02, which are quite small compared to the average values of the estimates. The biggest average standard deviation, both for shapes and diagonal stiffnesses, is in intra rotational parameters (0.05 for the shapes and 0.03 for the stiffnesses). The smallest average standard deviation, again both for shapes and diagonal stiffnesses, is in intra translational parameters (0.01 for the shapes and 0.00 for the stiffnesses). The conclusion from this experiment is that simulations of 50 ns and 100 ns (the lengths of ABC simulations) are acceptable for estimating our model parameters, however the estimates are obtained with an error.

Figures 4.6 and 4.7 contain plots of estimated stiffness matrices for the sequences $\mathsf{S}_7$ and $\mathsf{S}_{51}$. The plots on the left are the matrices $\mathsf{K}_E$ (the plot on the bottom of Figure 4.6 being just a zoomed portion of the plot above), and on the right are the plots of the matrices $\mathsf{K}_E^{\mathrm{bp}}$, each computed as an inverse of the “weighted" covariance of the variables $\mathsf{w}^{\mathrm{bp}}$, as discussed in Section 3.4. The ordering of the parameters in the plots is from top to bottom and from left to right, the black lines separate $12 \times 12$ blocks of the matrices $\mathsf{K}_E$ and $6 \times 6$ blocks of the matrices $\mathsf{K}_E^{\mathrm{bp}}$. The observation that can be made from these plots is that the rigid base model stiffness matrix is much more banded than the stiffness matrix for the rigid base pair model, i.e. the interactions in the rigid base model are much more local than in the rigid base pair model, which is in agreement with the conclusions of [36], made from different MD simulation data of a different sequence and using a different force field. Moreover, the same conclusion holds for all the 56 sequences in our training set. This motivates us to consider rigid bases, rather than base pairs, for the development of our local nearest neighbour model.

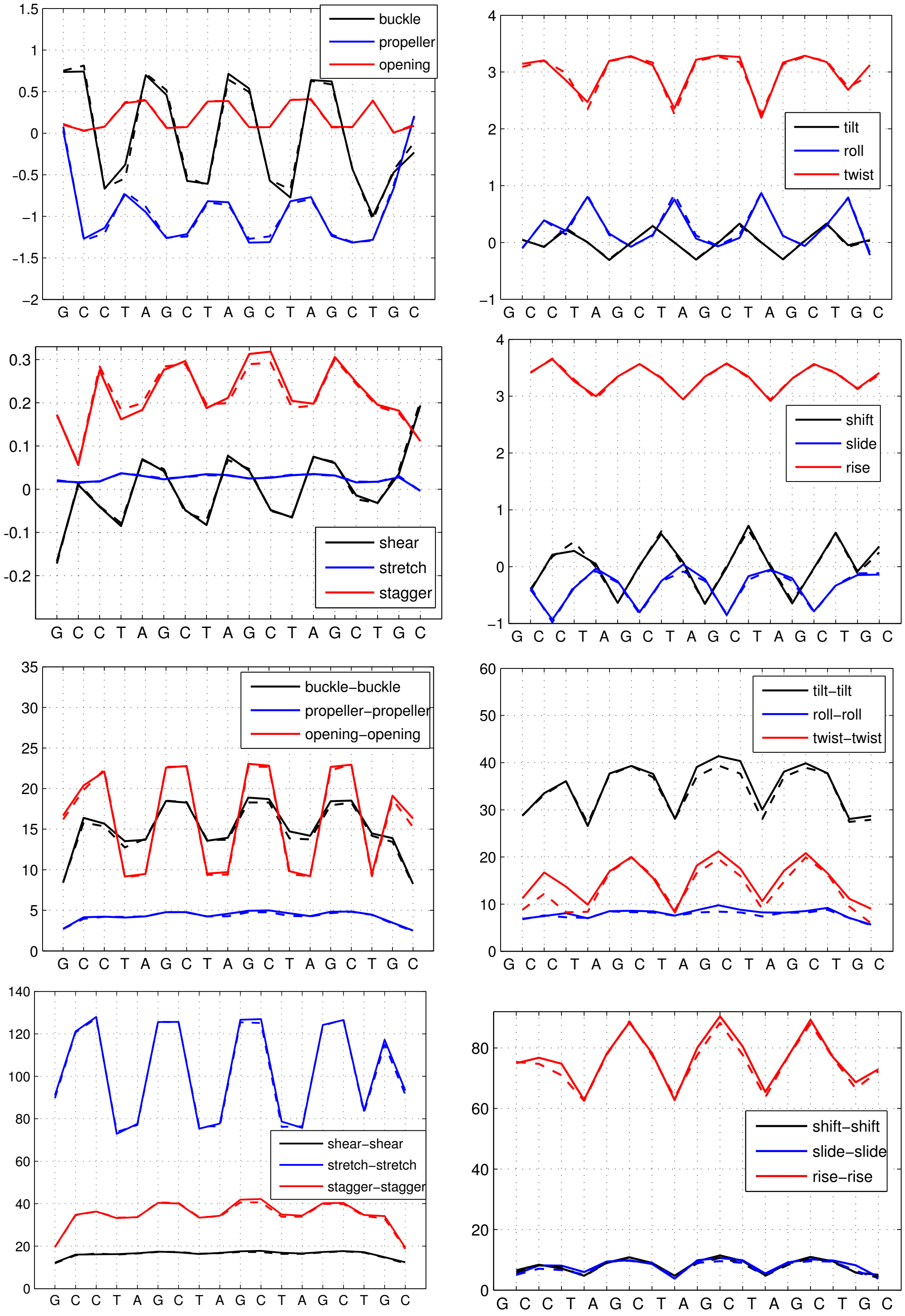


Figure 4.4: The shape parameter vector $\widehat{\mathsf{w}}_{4,E}$ (top four plots) and the diagonal entries of the stiffness parameter matrix $\mathsf{K}_{4,E}$ for the 18-base-pair sequence $\mathsf{S}_4$, where the parameter values are interpolated by piecewise-linear curves. Solid lines correspond to the parameters, computed from a 100 ns trajectory, dashed lines to the parameters, computed from a 1$\mu$s trajectory. Note different coordinate scales in different panels.

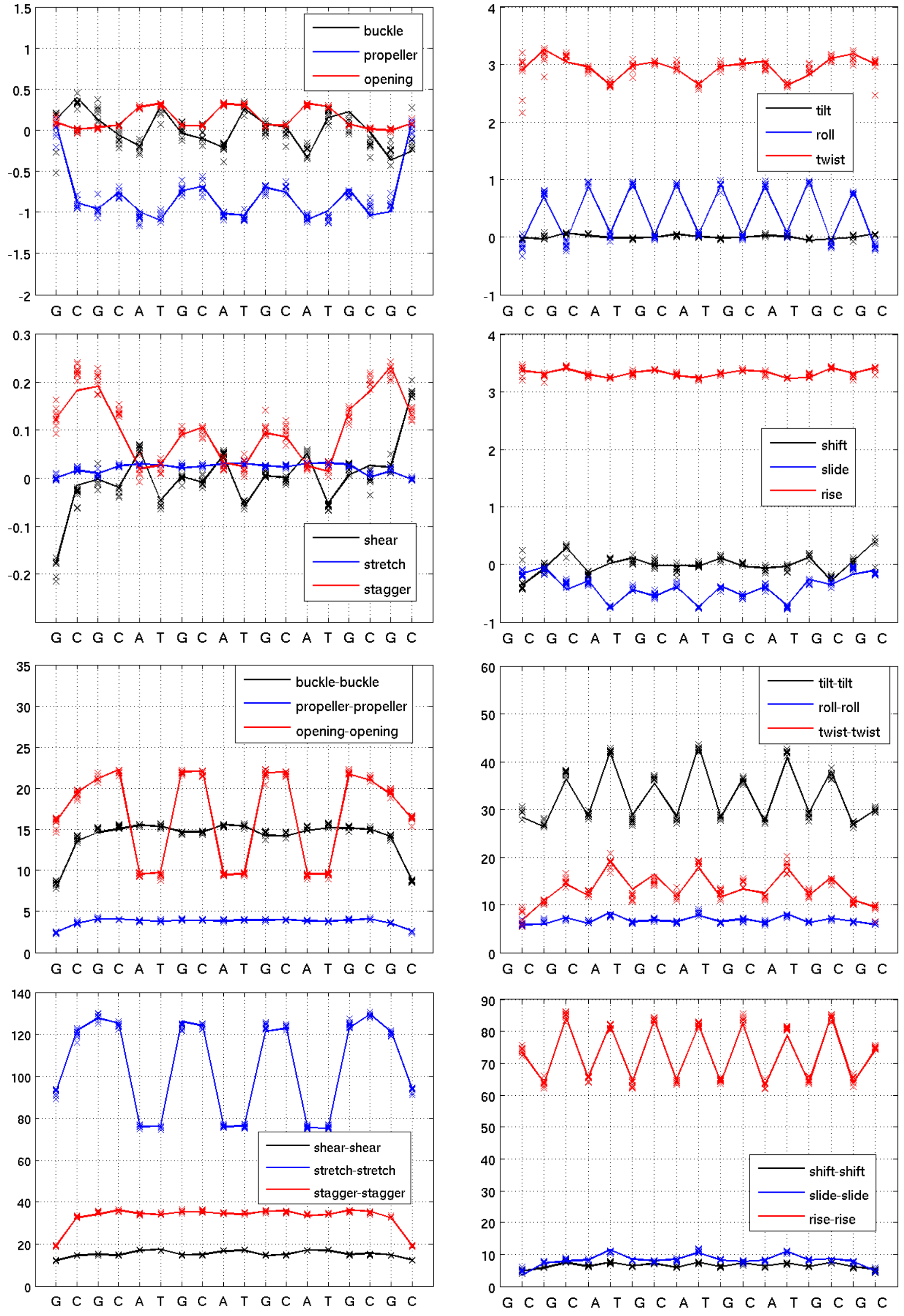


Figure 4.5: Shape parameters $\widehat{w}_{3,E}$ (top four plots) and diagonal entries of the stiffness matrices $\mathsf{K}_{3,E}$ (bottom plots), estimated from ten independent 50 ns simulations of the sequence $\mathsf{S}_3$ and one 100 ns ABC simulation of the same sequence. The parameter values, computed from the ABC simulation, are interpolated by piecewise-linear curves, the other ten are marked with crosses. We remark that the values are wider spread close to the ends of the oligomer.

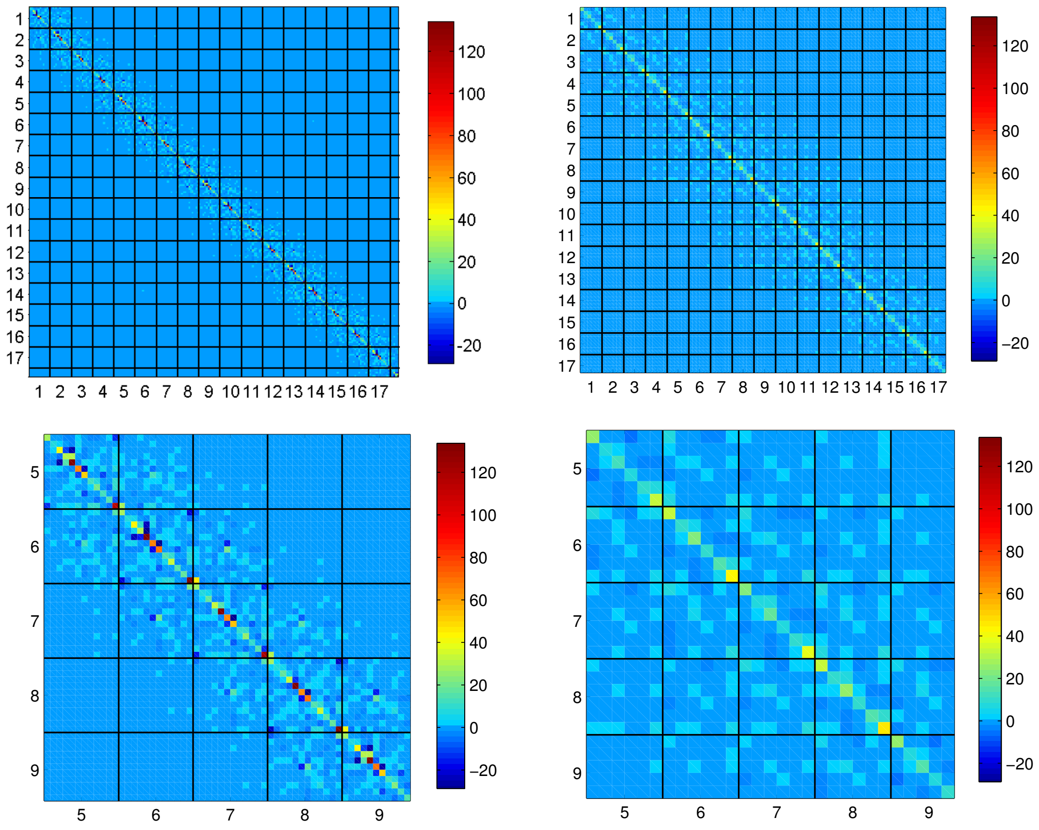


Figure 4.6: Rigid base (left) and rigid base pair (right) stiffness matrices for the sequence $\mathsf{S}_7$, computed from a 100 ns simulation. The two plots on the bottom are zoomed portions of the corresponding figures on the top. The black lines in the rigid base figures limit 12 × 12 blocks (intra and inter parameters), while in the rigid base pair figures 6 × 6 inter blocks are limited by black lines. The rigid base model stiffness matrix is much more banded than the stiffness matrix for the rigid base pair model, which is in agreement with the results of [36].

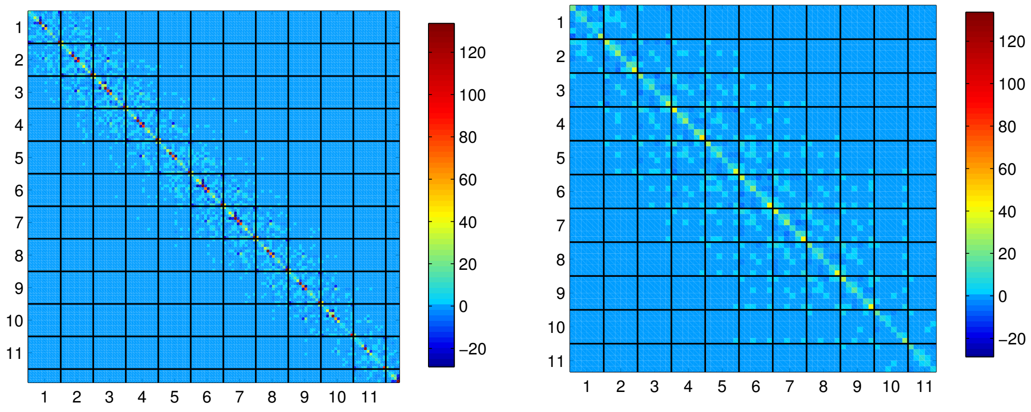


Figure 4.7: Rigid base (left) and rigid base pair (right) stiffness matrices for the sequence $\mathsf{S}_{51}$, computed from a 200 ns simulation. The black lines in the rigid base figures limit 12 × 12 blocks (intra and inter parameters), while in the rigid base pair figures 6 × 6 blocks are limited by black lines. The rigid base model stiffness matrix again is more banded than the stiffness matrix for the rigid base pair model. This conclusion holds for all the 56 sequences in our training set.

# Chapter 5

# The Jacobian and a Gaussian approximation

Here we explore the properties of the probability density function, introduced in Chapter 3, and explain our choice to approximate this density with a multivariate Gaussian.

## 5.1 Bimodality in parameter distributions

We have assumed that our rigid base DNA configuration parameter vector $\mathsf{w} \in \mathbb{R}^{12n-6}$ follows a distribution $\rho(\mathsf{w})$, defined in (3.7). The marginal distribution of a single scalar parameter $\mathsf{w}_k$, which is an element of $\mathsf{w}$, is then given by

$$\rho^{(k)}(\mathsf{w}_k) = \frac{1}{Z_k} \int_{\mathbb{R}^{12n-6}} \rho(\mathsf{w}) \;\; d\mathsf{w}_1 \dots d\mathsf{w}_{k-1} \, d\mathsf{w}_{k+1} \; \dots \; d\mathsf{w}_{12n-6}, \tag{5.1}$$

where $Z_k$ is the normalising constant. Due to the presence of the Jacobian in $\rho$, we do not know how to compute this integral explicitly. However one can look at the distributions of parameters $\mathsf{w}_k$, computed from a Molecular Dynamics trajectory, and then try to answer the question: can this distribution be described by the function (5.1)?

Examples of histograms of all the parameters $\mathsf{w}_k$, $k = 1 \dots 12n-6$, along the oligomer $\mathsf{S}_5$ of our data set are shown in Figures 5.2, 5.6 and 5.7 and their wider description is in Section 5.4. Also some additional histograms for other sequences can be found in Chapter 8. It can be noticed, that most of these distributions are symmetric bell-shaped curves, looking a lot like Gaussian curves. However, some parameters have an asymmetric or even a bimodal distribution, particularly for Twist and Slide, and particularly for the steps `C`$p$`G`, as discussed at length in [42] and [60].

There exist various definitions of bimodality. Here we call a probability density function bimodal if it has multiple maxima. We now pose a more specific question: can the marginal probability distribution (5.1) be bimodal, as are the observed distributions of some parameters $\mathsf{w}_k$? We can deduce some properties of $\rho^{(k)}(\mathsf{w}_k)$, while exploring the behaviour of $\rho$.

## 5.2 Can the Jacobian explain bimodality? A one dimensional example

Let us first consider a simple case: a one dimensional function

$$f(x) = \frac{1}{Z}\underbrace{e^{-\frac{1}{2}k\,(x-h)^2}}_{E(x)}\underbrace{\frac{1}{\left(1+\frac{1}{100}\,x^2\right)^2}}_{J(x)}, \qquad k>0, \quad x,h\in\mathbb{R}, \tag{5.2}$$

with a normalising constant $Z$. $f(x)$ can be seen as a product of $\frac{1}{Z}$ and two functions: a Gaussian function $E(x)$, and $J(x)$, which is of the same form as the Jacobian $J(\mathsf{w})$ in (3.2). Can this product have multiple extrema? If $f$ has multiple extremes, its derivative

$$f'(x) = -\frac{1}{Z}e^{-\frac{1}{2}k\,(x-h)^2}\frac{1}{\left(1+\frac{1}{100}\,x^2\right)^2}\left(k(x-h)+\frac{x}{25\left(1+\frac{1}{100}x^2\right)}\right) \tag{5.3}$$

must vanish for multiple values of $x$. The equation $f'(x)=0$ is equivalent to a cubic equation

$$x^3 - hx^2 + \left(100+\frac{4}{k}\right)x - 100\,h = 0, \tag{5.4}$$

which can have one, two or three distinct real roots. We are interested in the case when there

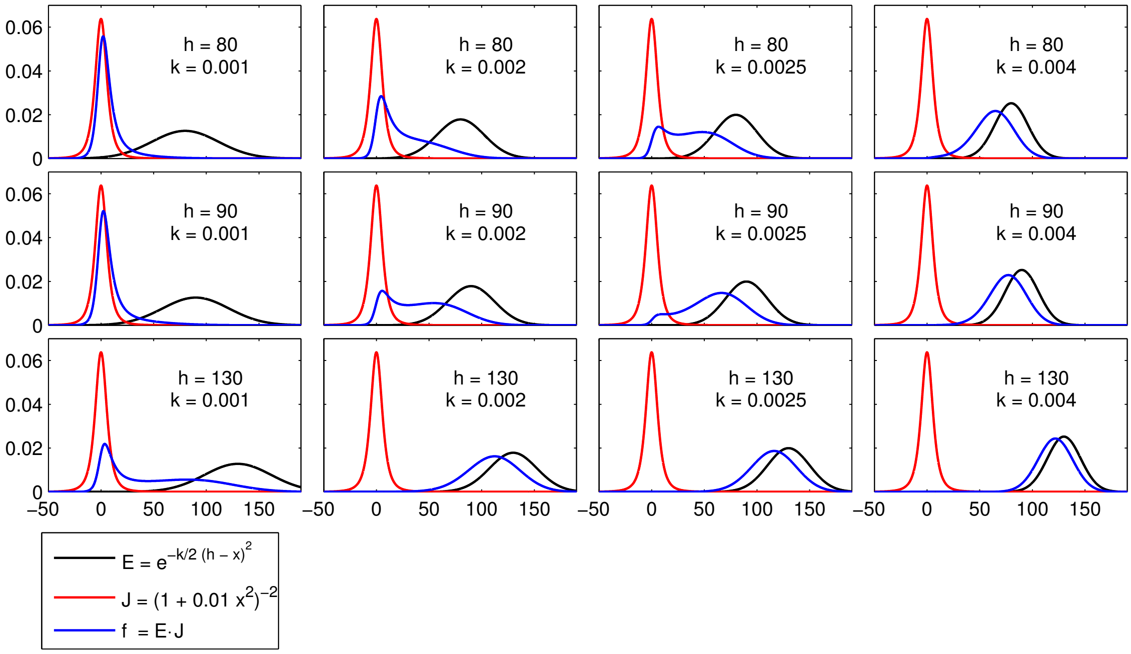


Figure 5.1: Normalised plots of the functions $E(x)$, $J(x)$ and $f(x)$, where $f(x)$ is the product of the first two functions. The values of the parameter $k$ increase from left to right and the values of $h$ increase from top to bottom, when passing from one plot to another. For some combinations of the parameter values, $f(x)$ has two distinguished peaks.

are three distinct real roots. In this case the discriminant $\Delta$ is bigger than zero, where

$$\Delta = -\frac{16}{k^3}\left(25\,k^3h^4 + \left(5000\,k^3 - 500\,k^2 - k\right)h^2 + 25\cdot 10^4\,k^3 + 3\cdot 10^4\,k^2 + 1200\,k + 16\right). \tag{5.5}$$

For $k > 0$, $\Delta$ is positive when $h_1 < h < h_2$, with

$$h_1, h_2 = \left(\frac{1 + 500k - 5000k^2 \mp \sqrt{1 - 600\,k + 12\cdot 10^4\,k^2 - 8\cdot 10^6\,k^3}}{50k^2}\right)^{\frac{1}{2}}, \quad k \in (0, 0.005). \tag{5.6}$$

This implies that the function $f$ can have two peaks only for $|h| > 51.96$ and $k < 0.005$.

The behaviour of $f(x)$ as a product of two single-peaked functions $E(x)$ and $J(x)$ is illustrated in Figure 5.1. For the values of the parameters $h$ and $k$, satisfying $\Delta > 0$, $f(x)$ can have two peaks of equal height, e.g. for $(h; k) = (80; 0.0025)$, or two peaks of which one is dominating, e.g. for $(h; k) = (90; 0.0025)$. When $\Delta < 0$, the curve $f(x)$ can be largely asymmetric with a shoulder, e.g. when $(h; k) = (80; 0.004)$, or it can get close to either to $E(x)$ or $J(x)$, depending on the values of the parameters. Note that all three functions were normalised before plotting, so that the area under each plots equals one. The normalising function for $E$ is $Z_1 = \sqrt{\frac{2\pi}{k}}$, for $J$ it is $Z_2 = 5\pi$ and for $f(x)$ it was evaluated numerically.

From this simple example one could hope that the Jacobian factor could indeed explain the bimodal distribution of some configuration parameters of rigid base DNA. Even though the values of $|h|$, that allow the function $f$ to be two-peaked, are far from the possible rigid base DNA average shape values, it is still worth exploring the multidimensional case.

## 5.3 Can the Jacobian explain bimodality? The multidimensional case

Consider a multidimensional probability density function of rigid base DNA configurations

$$\rho(\mathsf{w}) = \frac{1}{Z_J}\, e^{-\frac{1}{2}(\mathsf{w}-\widehat{\mathsf{w}})\cdot\mathsf{K}(\mathsf{w}-\widehat{\mathsf{w}})} J(\mathsf{w}), \tag{5.7}$$

with $\mathsf{w} \in \mathbb{R}^{12n-6}$, $n$ the number of base pairs, and the Jacobian $J(\mathsf{w})$ given in (3.2). To simplify further expressions, we reorder the variables in $\mathsf{w}$ to get

$$\underline{\mathsf{w}} = \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix}, \tag{5.8}$$

where $\boldsymbol{u} = (\vartheta^1, \theta^1, \dots, \vartheta^{n-1}, \theta^{n-1}, \vartheta^n) = (u^1, u^2, \dots, u^{2n-1})$ with $u^i \in \mathbb{R}^3$, $i = 1\dots 2n-1$, is the vector of all the rotational parameters and $\boldsymbol{v} = (\xi^1, \zeta^1, \dots, \xi^{n-1}, \zeta^{n-1}, \xi^n)$ is the vector of all translational parameters in $\mathsf{w}$. The matrix of the transformation, that reorders the variables in this way, can be written as

$$\underline{\mathsf{w}} = R\,\mathsf{w}, \tag{5.9}$$

where

$$R = \left(\begin{array}{ccccccccc}
I_3 & 0 & 0 & & & & & & \\
0 & 0 & I_3 & 0 & & & & & \\
0 & 0 & 0 & 0 & I_3 & & & & \\
 & & & & & \ddots & & 0 & 0 \\
 & & & & & & 0 & I_3 & 0 \\
\hline
0 & I_3 & 0 & 0 & & & & & \\
0 & 0 & 0 & I_3 & 0 & & & & \\
 & & & & & \ddots & & 0 & 0 \\
 & & & & & & 0 & 0 & I_3
\end{array}\right), \tag{5.10}$$

with the property that $R^{-1} = R^T$. Then the new stiffness matrix can be expressed as

$$\underline{\mathsf{K}} = R\,\mathsf{K}\,R^T, \tag{5.11}$$

and the probability density function $\rho$ can be written as

$$\begin{aligned}
\rho(\underline{\mathsf{w}}) &= \frac{1}{Z_J} e^{-\frac{1}{2}(\underline{\mathsf{w}}-\widehat{\underline{\mathsf{w}}})\cdot\underline{\mathsf{K}}(\underline{\mathsf{w}}-\widehat{\underline{\mathsf{w}}})} \prod_{i=1}^{2n-1}\left(1+\frac{1}{100}|u^i|^2\right)^{-2} \\
&= \frac{1}{Z_J} e^{-\frac{1}{2}(\underline{\mathsf{w}}-\widehat{\underline{\mathsf{w}}})\cdot\underline{\mathsf{K}}(\underline{\mathsf{w}}-\widehat{\underline{\mathsf{w}}}) - 2\sum_{i=1}^{2n-1}\ln\left(1+\frac{1}{100}|u^i|^2\right)}.
\end{aligned} \tag{5.12}$$

A necessary condition for the function $\rho$ to have multiple peaks, is that the equation $\nabla\rho = 0$ has multiple solutions, or, equivalently, that the equation $\nabla \ln\rho = 0$ has multiple solutions. The gradient of $\ln\rho$ can be explicitly expressed as

$$\nabla \ln\rho(\underline{\mathsf{w}}) = -\underline{\mathsf{K}}(\underline{\mathsf{w}} - \widehat{\underline{\mathsf{w}}}) - \frac{1}{25}\widetilde{D}(\boldsymbol{u})\underline{\mathsf{w}}, \tag{5.13}$$

where

$$\widetilde{D}(\boldsymbol{u}) = \begin{bmatrix} D_n(\boldsymbol{u}) & 0_n \\ 0_n & 0_n \end{bmatrix} \quad \text{with} \quad D_n(\boldsymbol{u}) = \begin{bmatrix} D(u^1) & 0 & \cdots & 0 \\ 0 & D(u^2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & D(u^{6n-3}) \end{bmatrix} \in \mathbb{R}^{(6n-3)\times(6n-3)},$$

$$D(u^i) = \left(\frac{1}{1+\frac{1}{100}|u^i|^2}\, I_3\right) \in \mathbb{R}^{3\times 3}, I_3 \text{ is an identity and } 0_n \in \mathbb{R}^{(6n-3)\times(6n-3)} \text{ a zero matrix.} \tag{5.14}$$

Then $\nabla \ln\rho = 0$ implies

$$\underline{\mathsf{K}}(\underline{\mathsf{w}} - \widehat{\underline{\mathsf{w}}}) + \frac{1}{25}\widetilde{D}(\boldsymbol{u})\underline{\mathsf{w}} = 0, \tag{5.15}$$

which is a system of $(12n - 6)$ equations. Actually, one can show that the function $\ln\rho$ is concave in our domain of interest, so there can be only one solution to (5.15), which means that the distribution $\rho$ can only have one peak. To show this, we compute the Hessian of $\ln\rho$:

$$\nabla^2 \ln\rho = \underline{\mathsf{K}} + \frac{1}{25}\left[\widetilde{D}(\boldsymbol{u}) + \widetilde{G}(\boldsymbol{u})\right] \tag{5.16}$$

where

$$\widetilde{G}(\boldsymbol{u}) = \begin{bmatrix} G_n(\boldsymbol{u}) & 0_n \\ 0_n & 0_n \end{bmatrix} \quad \text{with} \quad G_n(\boldsymbol{u}) = \begin{bmatrix} G(u^1) & 0 & \cdots & 0 \\ 0 & G(u^2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & G(u^{6n-3}) \end{bmatrix}, \tag{5.17}$$

$$G(u^i) = \tfrac{1}{50}\frac{-u^i {u^i}^T}{\left(1+\frac{1}{100}|u^i|^2\right)^2} \in \mathbb{R}^{3\times 3}.$$

As $\underline{\mathsf{K}} > 0$, a sufficient condition for the Hessian (5.16) to be positive definite is that the block diagonal matrix $[D_n(\boldsymbol{u}) + G_n(\boldsymbol{u})]$ is positive definite. The eigenvalues of this matrix are equal to the eigenvalues of its blocks, which are

$$\begin{aligned} \operatorname{eig}[D(u^i) + G(u^i)] &= \operatorname{eig}\left[\frac{1}{1+\frac{1}{100}|u^i|^2}\left(I_3 - \frac{\frac{1}{50}u^i {u^i}^T}{1+\frac{1}{100}|u^i|^2}\right)\right] \\ &= \left[\frac{1}{1+\frac{1}{100}|u^i|^2}, \frac{1}{1+\frac{1}{100}|u^i|^2}, \frac{1-\frac{1}{50}|u^i|^2}{(1+\frac{1}{100}|u^i|^2)^2}\right], \end{aligned} \tag{5.18}$$

and are all positive for $|u^i| < 5\sqrt{2} \approx 7$. The value $|u^i| = 7$ corresponds to a rotation angle of 1.22 radians or about 70 degrees, which does not correspond to a likely configuration (a maximum of a probability density) of B-DNA, for $u^i$ being a triple of either intra or inter rotation parameters. And indeed, all the peaks in the observed marginal distributions of our MD data, are in the domain $|u^i| < 7$, $i = 1 \dots 6n-3$, where the assumed density $\rho$ is concave.

There is still one remaining question: can the function $\ln \rho$ be concave, and the marginals of $\rho$ be bimodal? It is shown in [64] that if $\rho$ is a logarithmic concave multivariate probability density, i.e. if $\ln \rho$ is a concave function, then all the marginal densities of $\rho$ are also logarithmic concave. Therefore we conclude that our assumed probability density function with a Jacobian cannot explain the bimodal distribution of some DNA configuration parameters.

The bimodality of DNA equilibrium configuration is observed not only in the Molecular Dynamics, but also in the crystallography experiments [32, 14, 48]. These experiments indicate, that bimodality is a physical property of DNA, and not an outcome of the parametrisation of its configuration or the MD force field.

## 5.4 A Gaussian approximation

Having shown that $\rho$ is a concave function for $|u^i| < 7$, and having noticed that the Jacobian $J$ is close to one in this domain, it is reasonable to expect that variations in $J$ can be neglected and the distribution $\rho$ can be approximated by a Gaussian. By the Gaussian approximation of the internal configuration density $\rho$ we mean the density obtained by assuming $J$ to be constant. In this approximation, we have

$$\rho(\mathsf{w}) = \frac{1}{Z} e^{-\frac{1}{2}(\mathsf{w}-\widehat{\mathsf{w}})\cdot\mathsf{K}(\mathsf{w}-\widehat{\mathsf{w}})}, \quad Z = \sqrt{\frac{(2\pi)^{12n-6}}{\mathsf{K}}}, \quad \mathsf{w} \in \mathbb{R}^{12n-6}. \tag{5.19}$$

The moment-parameter relations (3.18) and (3.19) then become

$$\langle \mathsf{w} \rangle = \widehat{\mathsf{w}}, \quad \langle \Delta \mathsf{w} \otimes \Delta \mathsf{w} \rangle = \mathsf{K}^{-1}, \tag{5.20}$$

and their discrete versions, that can be compared to (3.20) and (3.21), are

$$\widehat{\mathsf{w}}_{\mu,\mathrm{o}} = \frac{1}{N} \sum_{l=1}^{N} \mathsf{w}_{\mu}^{(l)}, \quad \mathsf{K}_{\mu,\mathrm{o}}^{-1} = \frac{1}{N} \sum_{l=1}^{N} \Delta \mathsf{w}_{\mu}^{(l)} \otimes \Delta \mathsf{w}_{\mu}^{(l)}, \tag{5.21}$$

where $\Delta \mathsf{w}_{\mu} = \mathsf{w}_{\mu} - \widehat{\mathsf{w}}_{\mu,\mathrm{o}}$, $\mu$ is the number of an oligomer, $l$ is the index of the observation (the snapshot of the MD time series) and $N$ is the total number of the observations used for the oligomer $\mathsf{S}_{\mu}$. The notations $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$ and $\mathsf{K}_{\mu,\mathrm{o}}$ are chosen to be different from $\widehat{\mathsf{w}}_{\mu,E}$ and $\mathsf{K}_{\mu,E}$, denoting the parameters computed using the formulas (3.20) and (3.21) that include the Jacobian factor.

The marginal densities can be explicitly computed for the Gaussian measure (5.19) and they are also Gaussian. The expression for the one dimensional marginal density functions (5.1) is

$$\rho^{(k)}(\mathsf{w}_k) = \frac{1}{Z_k} e^{-\frac{1}{2}\left(\frac{\mathsf{w}_k - \widehat{\mathsf{w}}_k}{\sigma_k}\right)^2}, \quad Z_k = \sigma_k \sqrt{2\pi}, \quad \sigma_k = \sqrt{\left(\mathsf{K}^{-1}\right)_{kk}}, \quad \mathsf{w}_k, \sigma_k \in \mathbb{R}. \tag{5.22}$$

One can plot the graph of $\rho^{(k)}(\mathsf{w}_k)$ for each parameter $\mathsf{w}_k$ and compare it to the histogram of this variable, as it is shown in Figures 5.6, 5.7, and the Figures 8.9, 8.10, 8.11, 8.12 and 8.13 in Chapter 8. The three examples in Figure 5.2 summarise the different cases that occurred in our MD data. If the distribution of $\mathsf{w}_k$ is symmetric, the Gaussian fits it very well. The most likely value of the normal curve gets shifted a bit with respect to the observed one, if the observed distribution is asymmetric. And, of course, the Gaussian fit is the worst for the two peaked parameter distribution. Fortunately, most of the parameters have a symmetric distribution, so $\rho^{(k)}$ in general seems to be a good fit.

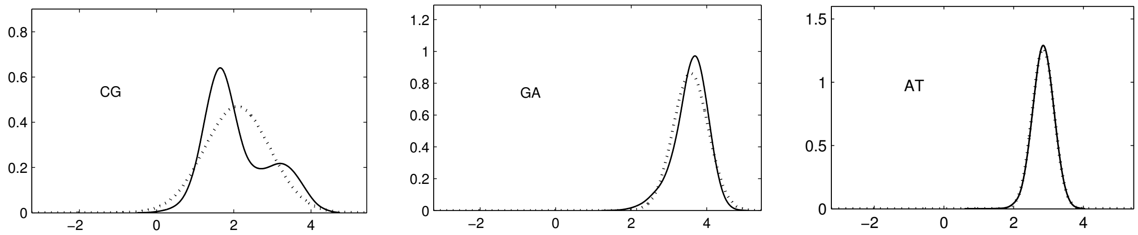


Figure 5.2: Examples of Gaussian fits to MD data. Plots of a marginal distribution of Twist for three selected dimer steps ($\mathtt{X}_6\mathtt{X}_7 = \mathtt{CG}$, $\mathtt{X}_7\mathtt{X}_8 = \mathtt{GA}$ and $\mathtt{X}_8\mathtt{X}_9 = \mathtt{AT}$) of the oligomer $\mathsf{S}_8 = \mathtt{GCGATCGATCGATCGAGC}$. Solid line: marginal, observed from MD. Dotted line: Gaussian marginal determined from the estimated density $\rho_{8,\mathrm{o}}$.

How different are the estimated shape and stiffness parameters for the distributions with and without Jacobian? From the equations (3.18) and (3.19), one can remark that the vector $\widehat{\mathsf{w}}_E$ is the arithmetic average of all the observed values $\frac{\mathsf{w}_k}{J\langle 1/J \rangle}$ and the stiffness matrix $\mathsf{K}_E$ is obtained as the inverse of the matrix $\left\langle \frac{\mathsf{w}}{\sqrt{J\langle 1/J \rangle}} \otimes \frac{\mathsf{w}}{\sqrt{J\langle 1/J \rangle}} \right\rangle - \widehat{\mathsf{w}} \otimes \widehat{\mathsf{w}}$. We therefore first decided to look to the histograms of the functions $\frac{\mathsf{w}_k}{J\langle 1/J\rangle}$ and $\mathsf{w}_k/\sqrt{J\langle 1/J\rangle}$ for the oligomers of our training set. As shown in Figure 5.3 for an example oligomer $\mathsf{S}_5$, the distribution of these functions are unimodal for all the parameters $\mathsf{w}_k$, both unimodal and bimodal ones. One can notice, that the centres (not the peaks) of all three histograms are close, and the width of the distributions of $\mathsf{w}_k/\sqrt{J\langle 1/J\rangle}$ is close to the ones of $\mathsf{w}_k$. Therefore, one could expect the estimates $(\widehat{\mathsf{w}}_{\mu,\mathrm{o}}, \mathsf{K}_{\mu,\mathrm{o}})$ to be not very different from $(\widehat{\mathsf{w}}_{\mu,E}, \mathsf{K}_{\mu,E})$. In addition we observe that the distribution of the scalar function $\langle 1/J\rangle J$ is quite concentrated around one.

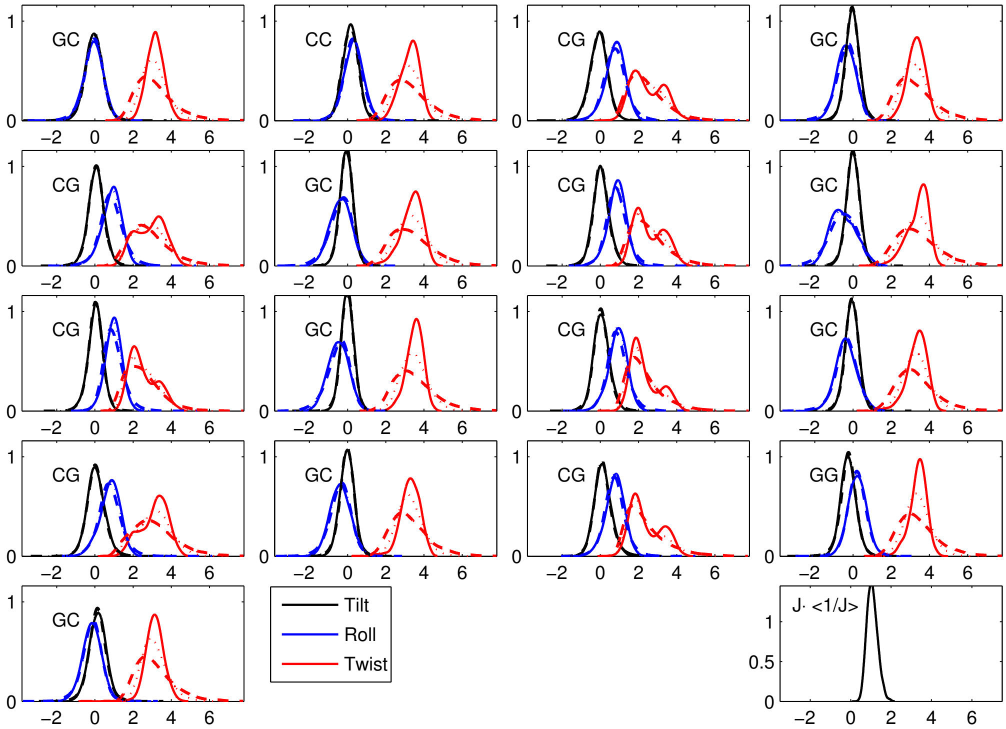


Figure 5.3: Histograms of inter rotational parameters $\mathsf{w}_k$ of the oligomer $\mathsf{S}_5$ (solid lines) compared to histograms of functions $\frac{\mathsf{w}_k}{J\langle 1/J\rangle}$ (dashed lines), that are used to estimate the shape vector $\widehat{\mathsf{w}}_{5,E}$. The stiffness matrix $\mathsf{K}_{5,\mathrm{o}}$ is estimated from a different function, $\mathsf{w}_k/\sqrt{J\langle 1/J\rangle}$, the histograms of which are plotted in dotted lines. Somewhat surprisingly, even though the distribution of Twist is bimodal for the $\mathsf{C}p\mathsf{G}$ steps, the distribution of these functions are unimodal. The plot in the right bottom corner is the histogram of the scalar function $J\langle 1/J\rangle$.

The potential of the Gaussian distribution (5.19),

$$\mathsf{U} = \frac{1}{2}(\mathsf{w} - \widehat{\mathsf{w}}) \cdot \mathsf{K}(\mathsf{w} - \widehat{\mathsf{w}}), \tag{5.23}$$

can be considered as a quadratic approximation to

$$\bar{\mathsf{U}} = \frac{1}{2}(\mathsf{w} - \bar{\mathsf{w}}) \cdot \bar{\mathsf{K}}(\mathsf{w} - \bar{\mathsf{w}}) - \ln J(\mathsf{w}). \tag{5.24}$$

at the point $\widehat{\mathsf{w}}$. We know that (5.24) has only one minimum in our domain of interest. We want this minimum to coincide with the minimum of (5.23), so

$$\widehat{\mathsf{w}} = \underset{\mathsf{w}}{\operatorname{argmin}} \left( \frac{1}{2}(\mathsf{w} - \bar{\mathsf{w}}) \cdot \bar{\mathsf{K}}(\mathsf{w} - \bar{\mathsf{w}}) - \ln J(\mathsf{w}) \right). \tag{5.25}$$

This minimum is achieved when $\widehat{\mathsf{w}}$ satisfies

$$\bar{\mathsf{K}}(\widehat{\mathsf{w}} - \bar{\mathsf{w}}) - \nabla \ln J(\widehat{\mathsf{w}}) = 0$$

which gives

$$\bar{\mathsf{w}} = \widehat{\mathsf{w}} + \bar{\mathsf{K}}^{-1} \frac{1}{J(\widehat{\mathsf{w}})} \nabla J(\widehat{\mathsf{w}}). \tag{5.26}$$

Equation (5.23) is the quadratic approximation to (5.24) at the point $\widehat{\mathsf{w}}$, if

$$\nabla^2 \left( \frac{1}{2}(\mathsf{w} - \widehat{\mathsf{w}}) \cdot \mathsf{K}(\mathsf{w} - \widehat{\mathsf{w}}) \right) = \nabla^2 \left( \frac{1}{2}(\mathsf{w} - \bar{\mathsf{w}}) \cdot \bar{\mathsf{K}}(\mathsf{w} - \bar{\mathsf{w}}) - \ln J(\mathsf{w}) \right). \tag{5.27}$$

This implies

$$\bar{\mathsf{K}} = \mathsf{K} + \nabla^2 \ln J(\widehat{\mathsf{w}}). \tag{5.28}$$

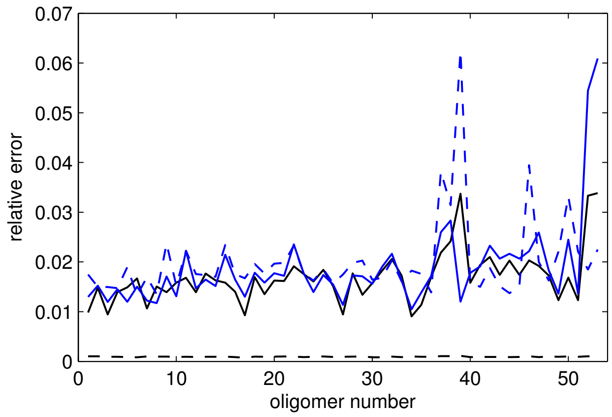


Figure 5.4: Plots of relative distances of parameters, estimated in the following three different ways: ($\widehat{\mathsf{w}}_{\mu,E}$, $\mathsf{K}_{\mu,E}$) are oligomer based parameters estimated with the Jacobian, ($\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$, $\mathsf{K}_{\mu,\mathrm{o}}$) are oligomer based Gaussian parameters, estimated assuming that the Jacobian is equal to a constant, and ($\bar{\mathsf{w}}_{\mu,\mathrm{o}}$, $\bar{\mathsf{K}}_{\mu,\mathrm{o}}$) are parameters of the quadratic approximation to the distribution with the Jacobian. Solid black line is $||\widehat{\mathsf{w}}_{\mu,E} - \widehat{\mathsf{w}}_{\mu,\mathrm{o}}||/||\widehat{\mathsf{w}}_{\mu,\mathrm{o}}||$, dashed black line is $||\mathsf{K}_{\mu,E} - \mathsf{K}_{\mu,\mathrm{o}}||/||\mathsf{K}_{\mu,\mathrm{o}}||$, solid blue line is $||\bar{\mathsf{w}}_{\mu,\mathrm{o}} - \widehat{\mathsf{w}}_{\mu,\mathrm{o}}||/||\widehat{\mathsf{w}}_{\mu,\mathrm{o}}||$ and dashed blue line is $||\bar{\mathsf{K}}_{\mu,\mathrm{o}} - \mathsf{K}_{\mu,\mathrm{o}}||/||\mathsf{K}_{\mu,\mathrm{o}}||$. The number of an oligomer is indicated on the horizontal axis.

Using the results of Section 5.3, we obtain

$$\begin{aligned} \bar{\mathsf{w}} &= \left( I - \tfrac{1}{25} \bar{\mathsf{K}}^{-1} R^T \widetilde{D}(\hat{\boldsymbol{u}}) R \right) \widehat{\mathsf{w}}, \\ \bar{\mathsf{K}} &= \mathsf{K} + R^T \widetilde{G}(\hat{\boldsymbol{u}}) R. \end{aligned} \tag{5.29}$$

where $I \in \mathbb{R}^{(12^n-6)\times(12n-6)}$ is an identity matrix, $G_n(\boldsymbol{u})$ is defined in (5.17) and $\widetilde{D}(\boldsymbol{u})$ is defined in (5.14). Note that in general, evaluating $(5.29)_2$ can lead to a matrix $\bar{\mathsf{K}}$ that is not positive definite, which would mean that (5.23) cannot be a quadratic approximation to a function of the form (5.24). However, for all the oligomers in our data set $\bar{\mathsf{K}}$ was positive definite.

Comparing the values of the parameters $(\widehat{\mathsf{w}}_{\mu,\mathrm{o}}, \mathsf{K}_{\mu,\mathrm{o}})$ to $(\bar{\mathsf{w}}_{\mu,\mathrm{o}}, \bar{\mathsf{K}}_{\mu,\mathrm{o}})$ and to $(\widehat{\mathsf{w}}_{\mu,E}, \mathsf{K}_{\mu,E})$ is one way to measure the error, introduced by suppressing the Jacobian. Figure 5.4 shows relative Euclidean distances from vector $\widehat{\mathsf{w}}_{\mu,o}$ to the other two shape vector estimates and from the matrix $\mathsf{K}_{\mu,\mathrm{o}}$ to the other two stiffness matrix estimates for every oligomer $\mu$. For the oligomers these relative distances were smaller than 3%, which suggests that the Gaussian approximation is appropriate for our assumed distribution. Figure 5.5 contains plots of the shape parameters and diagonal entries of the stiffness matrices for the sequence $\mathsf{S}_5$ computed in these three ways. The plots are organised in the standard way, introduced in Section 4.4, with values interpolated with piecewise linear curves. The three curves for three different ways of estimation are almost indistinguishable, indicating that the difference between the three probability densities is very small. Therefore we conclude that approximating our observed parameter distributions by Gaussians and ignoring the presence of the Jacobian is a reasonable choice.

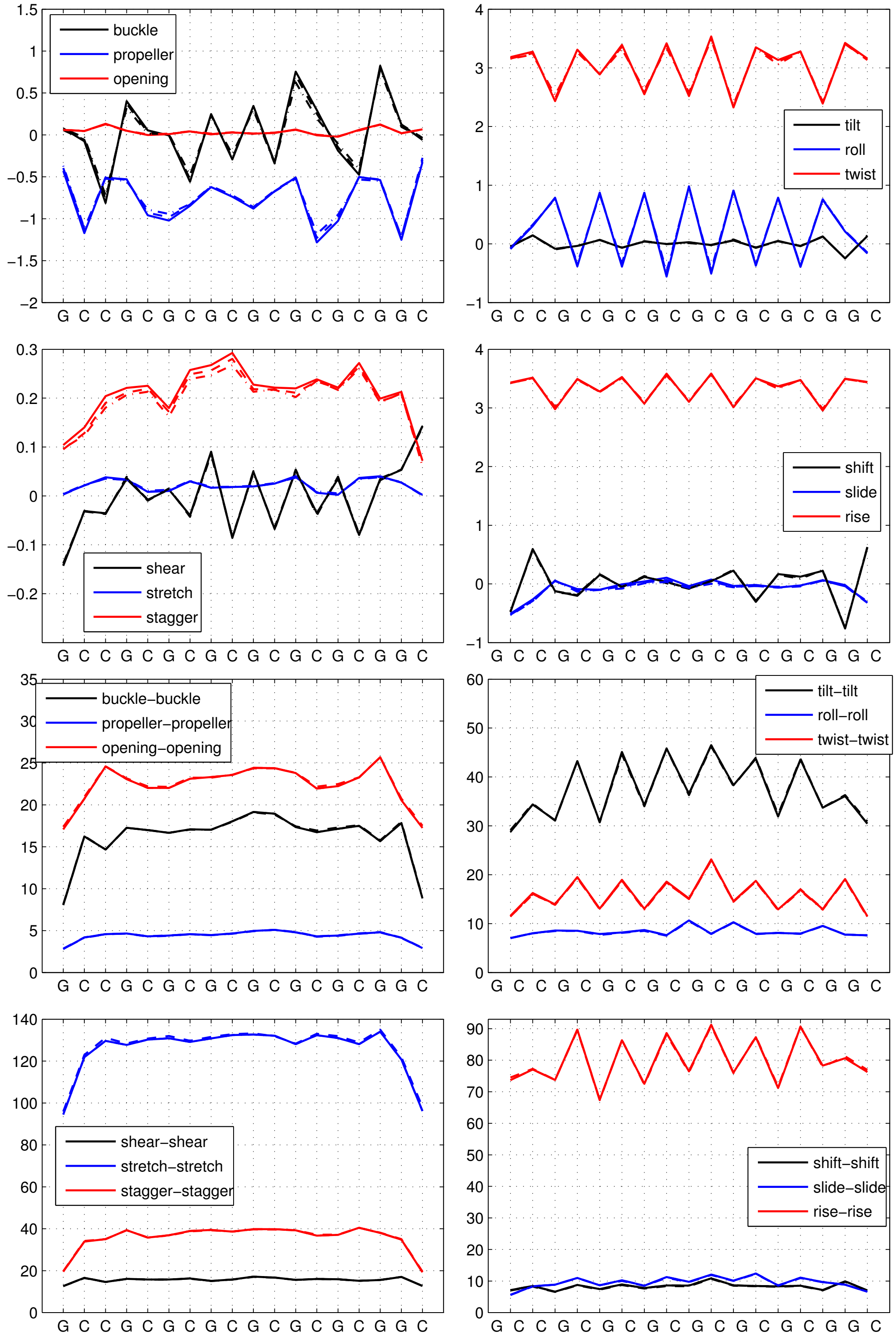


Figure 5.5: Shape parameters (top four plots) and diagonal entries of the stiffness matrices (bottom plots) for the sequence $\mathsf{S}_5$. Solid line: elements of $\widehat{\mathsf{w}}_{5,E}$ and $\mathrm{diag}(\mathsf{K}_{5,E})$, estimated assuming the distribution with a Jacobian, dashed line: $\widehat{\mathsf{w}}_{5,\mathrm{o}}$ and $\mathrm{diag}(\mathsf{K}_{5,\mathrm{o}})$, estimated assuming the Gaussian distribution, dashed-dotted line: $\bar{\mathsf{w}}_{5,\mathrm{o}}$ and $\mathrm{diag}(\bar{\mathsf{K}}_{5,\mathrm{o}})$, computed from $\widehat{\mathsf{w}}_{5,\mathrm{o}}$ and $\mathsf{K}_{5,\mathrm{o}}$ as (5.29).

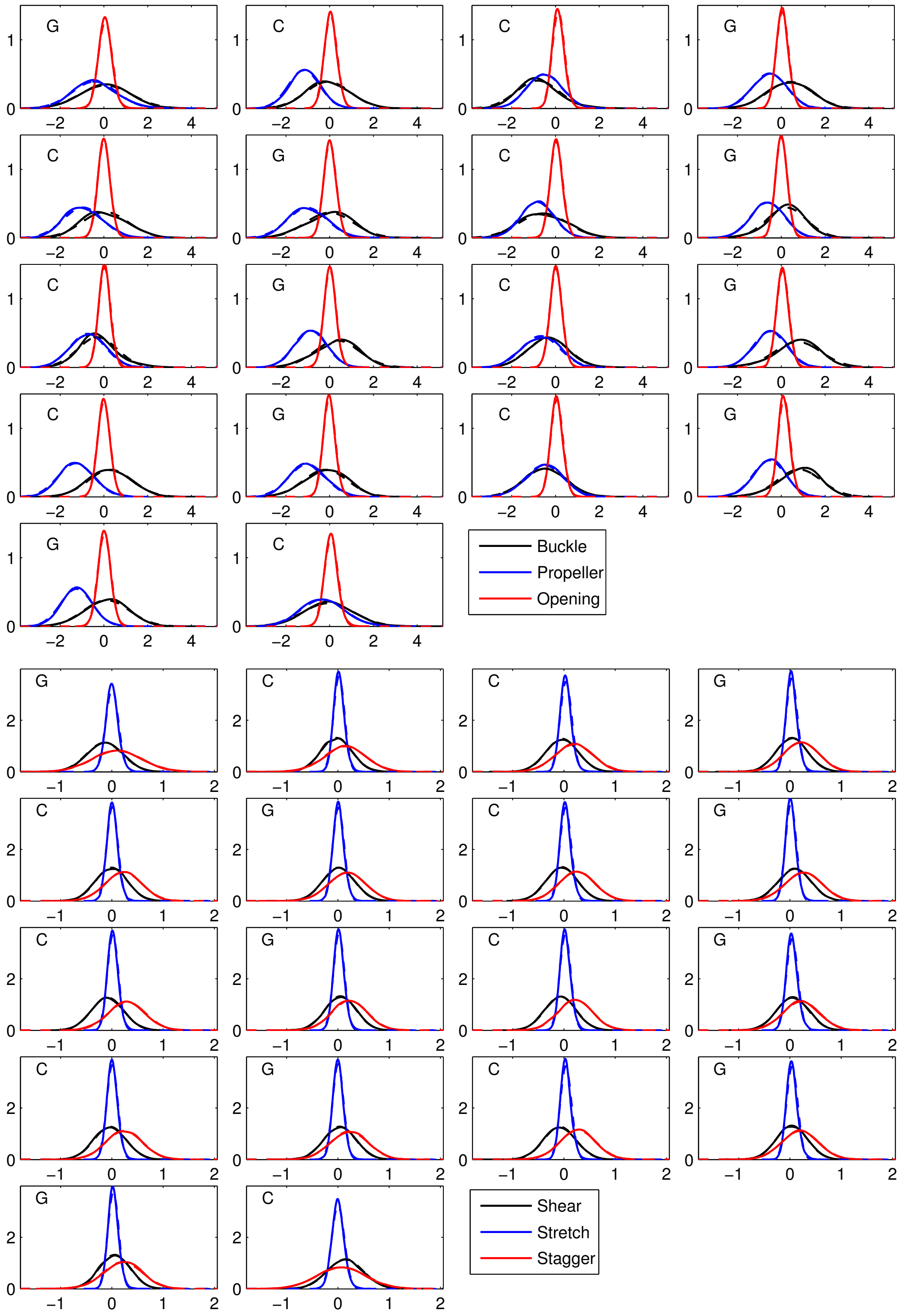


Figure 5.6: Examples of Gaussian fits to MD data. Plots of marginal distributions of intra base pair parameters of the oligomer $\mathsf{S}_5$. Solid line: marginal densities, observed from MD. Dotted line: Gaussian marginals, determined from the estimated density $\rho_{5,\mathrm{o}}$.

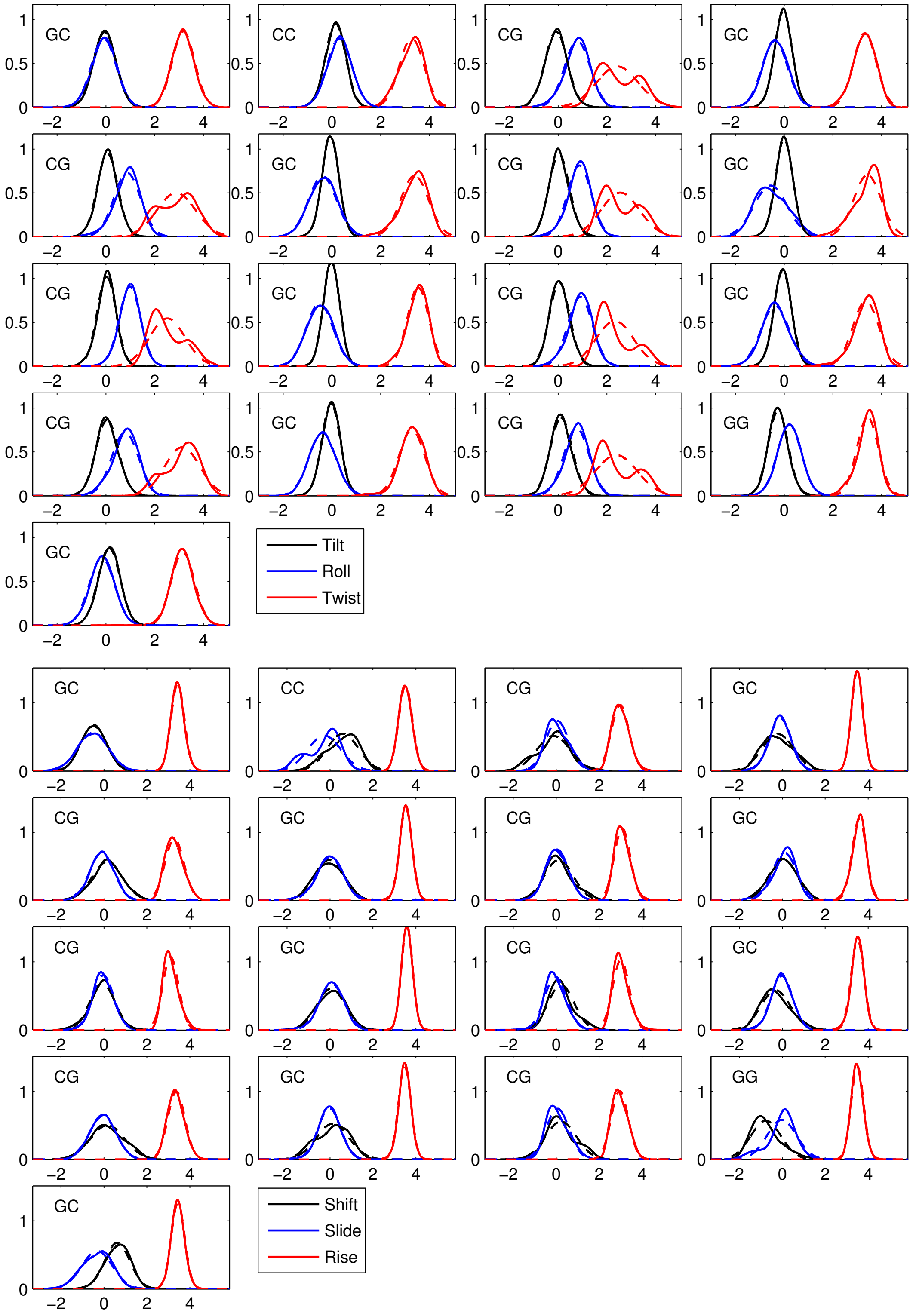


Figure 5.7: Examples of Gaussian fits to MD data. Plots of marginal distributions of inter base pair parameters of the oligomer $\mathsf{S}_5$. Solid line: marginal densities, observed from MD. Dotted line: Gaussian marginals, determined from the estimated density $\rho_{5,\mathrm{o}}$. Histograms of the inter rotational variables $\frac{\mathsf{w}_k}{J\langle 1/J\rangle}$ and $\frac{\mathsf{w}}{\sqrt{J\langle 1/J\rangle}}$ for the same sequence are shown in Figure 5.3.

## 5.5 Comparison of probability densities

The equilibrium statistical properties of the internal configuration of an oligomer are described by the density $\rho(\mathsf{w})$, which is completely defined by the internal energy parameters $\mathsf{K}$ and $\widehat{\mathsf{w}}$. Throughout our developments it will be necessary to quantify the difference in two densities $\rho_{\rm m}(\mathsf{w})$ and $\rho_{\rm o}(\mathsf{w})$ defined by two different sets of parameters $\{\mathsf{K}_{\rm m}, \widehat{\mathsf{w}}_{\rm m}\}$ and $\{\mathsf{K}_{\rm o}, \widehat{\mathsf{w}}_{\rm o}\}$. To this end, we appeal to results from general probability theory and employ the Kullback-Leibler divergence [33]

$$D(\rho_{\rm m}, \rho_{\rm o}) := \int \rho_{\rm m}(\mathsf{w}) \ln\left[\frac{\rho_{\rm m}(\mathsf{w})}{\rho_{\rm o}(\mathsf{w})}\right] \, d\mathsf{w}. \tag{5.30}$$

This quantity is a non-symmetric measure of the difference between two normalised probability densities: it satisfies $D(\rho_{\rm m}, \rho_{\rm o}) \geq 0$ for any $\rho_{\rm m}$ and $\rho_{\rm o}$, and $D(\rho_{\rm m}, \rho_{\rm o}) = 0$ if and only if $\rho_{\rm m} = \rho_{\rm o}$. The divergence $D(\rho_{\rm m}, \rho_{\rm o})$ defines a pre-metric on the space of probability densities, but is not a metric since it is non-symmetric and does not satisfy the triangle inequality.

It is not always possible to integrate (5.30) to find an explicit expression for $D(\rho_{\rm m})$, for example when $\rho_{\rm m}$ or $\rho_{\rm o}$ are distributions of the form (5.7) with a non-constant Jacobian factor. In the special case when $\rho_{\rm m}$ and $\rho_{\rm o}$ are both Gaussian, the integral (5.30) can be explicitly evaluated [33] to obtain

$$D(\rho_{\rm m}, \rho_{\rm o}) = \frac{1}{2}\left[\mathsf{K}_{\rm m}^{-1} : \mathsf{K}_{\rm o} - \ln(\det \mathsf{K}_{\rm o}/\det \mathsf{K}_{\rm m}) - I : I\right] + \frac{1}{2}(\widehat{\mathsf{w}}_{\rm m} - \widehat{\mathsf{w}}_{\rm o}) \cdot \mathsf{K}_{\rm o}(\widehat{\mathsf{w}}_{\rm m} - \widehat{\mathsf{w}}_{\rm o}), \tag{5.31}$$

where a colon denotes the standard Euclidean inner product for square matrices and $I$ denotes the identity matrix of the same dimension $n$ as $\mathsf{K}_{\rm m}$ and $\mathsf{K}_{\rm o}$. The term in parentheses involves only the two stiffness matrices, and can be rewritten in the form

$$D^{\dagger}(\mathsf{K}_{\rm m}, \mathsf{K}_{\rm o}) := \Sigma_{i=1}^{n}(\nu_i - \ln \nu_i - 1), \tag{5.32}$$

where the $\nu_i$ are the $n$ positive eigenvalues of the symmetric generalised eigenvalue problem

$$\mathsf{K}_{\rm o}\xi_i = \nu_i \mathsf{K}_{\rm m}\xi_i. \tag{5.33}$$

It is evident that $D^{\dagger}$ defined in (5.32) is non-negative and vanishes only when $\nu_1 = \cdots = \nu_n = 1$, i.e. when $\mathsf{K}_{\rm o} = \mathsf{K}_{\rm m}$, so that it is an appropriate measure of the difference between two symmetric positive definite matrices [52]. Similarly it is evident that the eigenvalues defined in (5.33) are non dimensional and so independent of the choice of length, rotation and energy scales, and that

$$D^{\dagger}(\mathsf{K}_{\rm m}, \mathsf{K}_{\rm o}) = D^{\dagger}(\mathsf{K}_{\rm o}^{-1}, \mathsf{K}_{\rm m}^{-1}). \tag{5.34}$$

Analogously the second term in (5.31) is non-negative and vanishes only when the means $\widehat{\mathsf{w}}_{\rm m}$ and $\widehat{\mathsf{w}}_{\rm o}$ coincide. It is also independent of length and rotation scales, but depends linearly on the energy scale. Thus the divergence (5.31) is a linear combination of differences in stiffnesses and averages of two Gaussians, with the relative weighting dependent on the energy scale, or equivalently temperature.

The divergence $D(\rho_{\rm m}, \rho_{\rm o})$ is employed in various standard parameter estimation methods in statistics. When $\rho_{\rm o}$ is interpreted as an observed density and $\rho_{\rm m}$ is interpreted as a model density, then the minimisation of $D(\rho_{\rm m}, \rho_{\rm o})$ over a space of admissible $\rho_{\rm m}$ yields a best-fit model density $\rho_{\mathrm{m}}^{*}$. Alternatively, due to the lack of symmetry, minimisation of $D(\rho_{\mathrm{o}}, \rho_{\mathrm{m}})$ over $\rho_{\mathrm{m}}$ yields a generally different best-fit model density $\rho_{\mathrm{m}}^{**}$. In [49] the first approach is described as model fitting via the maximum relative entropy principle, whereas the second approach, in the Gaussian case, can be shown to correspond to model fitting via the maximum likelihood principle. In the developments that follow we adopt the maximum relative entropy principle when fitting models to data. However, we checked that the two ways of fitting the data (i.e. the two choices of an objective function for a fitting procedure) gave very similar results.

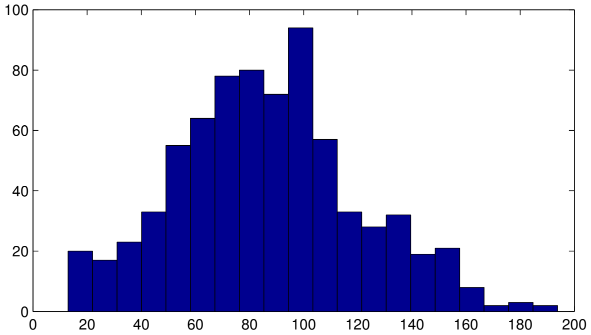


Figure 5.8: A histogram of the symmetrised pairwise Kullback-Leibler divergences between the 39 18-mers in the adopted training set. The average of this distribution $D_{\mathrm{o}} \approx 85$ sets a scale for variation in probability densities due to different sequences, and is used to estimate closeness of model reconstructions.

To quantify various modelling and approximation errors it is desirable to set a scale for the Kullback-Leibler divergence (5.30). To do so we use the training set (4.1) in the following way. Figure 5.8 presents a histogram of the symmetrised divergences $(D(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}}) + D(\rho_{\mu_2,\mathrm{o}}, \rho_{\mu_1,\mathrm{o}}))/2$ for all distinct pairs amongst the 39 18-mers retained in the training set. Then we introduce the scale $D_{\mathrm{o}}$ to be the average over this distribution, namely

$$D_{\mathrm{o}} = \operatorname*{avg}_{\substack{n_{\mu_1}=n_{\mu_2}=18 \\ \mu_1 \neq \mu_2}} D(\rho_{\mu_1,\mathrm{o}}, \rho_{\mu_2,\mathrm{o}}), \tag{5.35}$$

where $n_{\mu_i}$ is the number of base pairs in the sequence $\mathsf{S}_{\mu_i}$. Notice that $D_{\mathrm{o}} \approx 85$ provides a characteristic scale for the variation in probability density due to variation in sequence for 18-mers. Thus, if for any given oligomer we find that the divergence between a training set distribution $\rho_{\mu,\mathrm{o}}$ and a model distribution $\rho_{\mu,\mathrm{m}}$ is small on this scale, then the modelling error is small compared to variations due to sequence effects.

# Chapter 6

# A nearest-neighbour rigid base model

In this chapter we describe a model in which the assumed local interactions of the DNA rigid bases imply a special structure of the stiffness and configuration parameters. Together with the following two chapters, it forms part of the joint publication [24].

## 6.1 Mechanical assumptions

For the purposes of building an internal energy model, we consider different partitions of an oligomer into different types of structural units. Here we restrict attention to a nearest-neighbour model built on two types of units: 1-mers (or monomers) and 2-mers (or dimers). However we intentionally establish a notation that generalises naturally to trimers, tetramers, and so on to facilitate extensions in later work. Specifically, an oligomer of $n$ base pairs and arbitrary sequence $\mathtt{X}_1 \cdots \mathtt{X}_n$ can be partitioned into 1-mer units $\mathtt{X}_a$, $a = 1, \ldots, n$, and 2-mer units $\mathtt{X}_a\mathtt{X}_{a+1}$, $a = 1, \ldots, n-1$. Just as the internal configuration of the oligomer is specified by the coordinate vector $\mathsf{w} = (y^1, z^1, y^2, z^2, ..., y^{n-1}, z^{n-1}, y^n) \in \mathbb{R}^{12n-6}$, the internal configuration of each 1-mer $\mathtt{X}_a$ is specified by the coordinate vector $\mathsf{w}_1^a = y^a \in \mathbb{R}^6$, and the internal configuration of each 2-mer $\mathtt{X}_a\mathtt{X}_{a+1}$ is specified by the vector $\mathsf{w}_2^a = (y^a, z^a, y^{a+1}) \in \mathbb{R}^{18}$. For convenience, the collection of all 1-mer coordinates will be denoted by $\mathsf{w}_1 = (\mathsf{w}_1^1, \ldots, \mathsf{w}_1^n) \in \mathbb{R}^{6n}$, and the collection of all 2-mer coordinates will be denoted by $\mathsf{w}_2 = (\mathsf{w}_2^1, \ldots, \mathsf{w}_2^{n-1}) \in \mathbb{R}^{18(n-1)}$. We stress that there is considerable redundancy in this notation in that the intra variables $y^a$ appear in both $\mathsf{w}_2^a$ and $\mathsf{w}_2^{(a-1)}$. This redundancy is both notationally convenient and physically pertinent; it gives rise to the phenomenon of frustration.

In our developments, it will be necessary to consider various linear maps between the vectors $\mathsf{w}$, $\mathsf{w}_1$ and $\mathsf{w}_2$. The matrix which copies elements of $\mathsf{w}$ into the 1-mer vector $\mathsf{w}_1$ is denoted by $\mathsf{P}_1 \in \mathbb{R}^{6n} \times \mathbb{R}^{12n-6}$, so that $\mathsf{w}_1 = \mathsf{P}_1\mathsf{w}$, and the matrix which copies elements of $\mathsf{w}$ into the 2-mer vector $\mathsf{w}_2$ is denoted by $\mathsf{P}_2 \in \mathbb{R}^{18(n-1)} \times \mathbb{R}^{12n-6}$, so that $\mathsf{w}_2 = \mathsf{P}_2\mathsf{w}$.

$$\mathsf{P}_1 = \begin{pmatrix} I & 0 & 0 & & & \\ 0 & 0 & I & 0 & & \\ 0 & 0 & 0 & 0 & I & \\ & & & & \ddots & \end{pmatrix} \quad \mathsf{P}_2 = \begin{pmatrix} I & 0 & 0 & 0 & & \\ 0 & I & 0 & 0 & & \\ 0 & 0 & I & 0 & & \\ 0 & 0 & I & 0 & & \\ 0 & 0 & 0 & I & & \\ & & & & I & \\ & & & & \ddots & \end{pmatrix}, \tag{6.1}$$

where $I \in \mathbb{R}^{6\times 6}$ is the identity matrix and $0 \in \mathbb{R}^{6\times 6}$ is the zero matrix. It will also be necessary to consider the transpose matrix $\mathsf{P}_1^T$, which maps an arbitrary 1-mer vector $\mathsf{u}_1$ into a particular oligomer vector $\mathsf{u} = \mathsf{P}_1^T\mathsf{u}_1$. Here the entries in the 1-mer vector $\mathsf{u}_1$ (all of which are intra-base pair coordinates) are mapped to their corresponding location in the oligomer vector $\mathsf{u}$, and the remaining entries of $\mathsf{u}$ (all of which correspond to inter-base pair coordinates) are zero. The transpose matrix $\mathsf{P}_2^T$ maps an arbitrary 2-mer vector $\mathsf{u}_2$ into the oligomer vector $\mathsf{u} = \mathsf{P}_2^T\mathsf{u}_2$ with the entries in the 2-mer vector $\mathsf{u}_2$ mapped to their corresponding location in the oligomer vector $\mathsf{u}$ with overlapping contributions summed.

We consider an internal energy model based on local energies that describe physically distinct interactions within the 1-mer and 2-mer units. Specifically, to any 1-mer $\mathtt{X}_a$ we associate an energy of the form

$$\mathsf{U}_1^a(\mathsf{w}_1^a) = \frac{1}{2}(\mathsf{w}_1^a - \widehat{\mathsf{w}}_1^a)\cdot \mathsf{K}_1^a(\mathsf{w}_1^a - \widehat{\mathsf{w}}_1^a), \tag{6.2}$$

where $\widehat{\mathsf{w}}_1^a \in \mathbb{R}^6$ is a vector of shape parameters that defines the minimum energy or ground state of the interaction, and $\mathsf{K}_1^a \in \mathbb{R}^{6\times 6}$ is a symmetric matrix of stiffness parameters that describes the elastic stiffness associated with each internal coordinate and couplings between them. The energy $\mathsf{U}_1^a$ is to be interpreted as a model for the intra-base pair interactions between the two bases of the base pair $(\mathtt{X},\overline{\mathtt{X}})_a$. The description of these interactions involves only the intra-base pair coordinates $\mathsf{w}_1^a = y^a$, and the stiffness matrix $\mathsf{K}_1^a$ may in general be dense.

Similarly, to any 2-mer $\mathtt{X}_a\mathtt{X}_{a+1}$ we associate an energy

$$\mathsf{U}_2^a(\mathsf{w}_2^a) = \frac{1}{2}(\mathsf{w}_2^a - \widehat{\mathsf{w}}_2^a)\cdot \mathsf{K}_2^a(\mathsf{w}_2^a - \widehat{\mathsf{w}}_2^a), \tag{6.3}$$

where $\widehat{\mathsf{w}}_2^a \in \mathbb{R}^{18}$ is a vector of shape parameters and $\mathsf{K}_2^a \in \mathbb{R}^{18\times 18}$ is a symmetric matrix of stiffness parameters analogous to before. The energy $\mathsf{U}_2^a$ is to be interpreted as a model for all the inter-base pair interactions involving a base of the base pair $(\mathtt{X},\overline{\mathtt{X}})_a$ and a base of the base pair $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$, in other words any nearest-neighbour base-base interaction across the junction between the base pairs $(\mathtt{X},\overline{\mathtt{X}})_a$ and $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$. The description of all of these interactions naturally involves the internal coordinates $\mathsf{w}_2^a = (y^a, z^a, y^{a+1})$. The stiffness matrix $\mathsf{K}_2^a$ may in general be dense and has the natural block form

$$\mathsf{K}_2^a = \begin{pmatrix} \mathsf{K}_{2,11}^a & \mathsf{K}_{2,12}^a & \mathsf{K}_{2,13}^a \\ \mathsf{K}_{2,21}^a & \mathsf{K}_{2,22}^a & \mathsf{K}_{2,23}^a \\ \mathsf{K}_{2,31}^a & \mathsf{K}_{2,32}^a & \mathsf{K}_{2,33}^a \end{pmatrix}, \tag{6.4}$$

where each entry is an element of $\mathbb{R}^{6\times 6}$. The assumption that the overall matrix is symmetric implies that the diagonal blocks $\mathsf{K}^a_{2,11}$, $\mathsf{K}^a_{2,22}$ and $\mathsf{K}^a_{2,33}$ are each symmetric, and implies that the off-diagonal blocks satisfy $[\mathsf{K}^a_{2,12}]^T = \mathsf{K}^a_{2,21}$, $[\mathsf{K}^a_{2,13}]^T = \mathsf{K}^a_{2,31}$ and $[\mathsf{K}^a_{2,23}]^T = \mathsf{K}^a_{2,32}$.

The local 1-mer and 2-mer energies can be summed in a natural way to obtain an overall oligomer energy. Specifically, we define the oligomer energy as

$$\mathsf{U}(\mathsf{w}) = \sum_{j=1}^{2} \sum_{a=1}^{n-j+1} \mathsf{U}^a_j(\mathsf{w}^a_j) = \frac{1}{2} \sum_{j=1}^{2} \sum_{a=1}^{n-j+1} (\mathsf{w}^a_j - \widehat{\mathsf{w}}^a_j) \cdot \mathsf{K}^a_j(\mathsf{w}^a_j - \widehat{\mathsf{w}}^a_j). \tag{6.5}$$

Here the index $j = 1, 2$ is a label for the 1-mer and 2-mer interactions, and for each $j$, the index $a = 1, \dots, n - j + 1$ runs over all the $j$-mers along the oligomer. The oligomer energy can be written in a more convenient matrix form as the sum of two shifted quadratic forms (of different dimensions)

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2} \sum_{j=1}^{2} (\mathsf{P}_j\mathsf{w} - \widehat{\mathsf{w}}_j) \cdot \mathsf{K}_j(\mathsf{P}_j\mathsf{w} - \widehat{\mathsf{w}}_j), \tag{6.6}$$

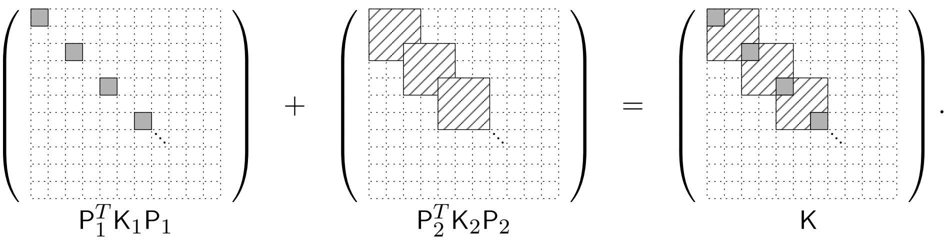
where $\widehat{\mathsf{w}}_j = (\widehat{\mathsf{w}}^1_j, \dots, \widehat{\mathsf{w}}^{n-j+1}_j)$ is a vector of all the $j$-mer shape parameters (for $j = 1, 2$), $\mathsf{K}_j = \mathrm{diag}(\mathsf{K}^1_j, \dots, \mathsf{K}^{n-j+1}_j)$ is a block-diagonal matrix of all the $j$-mer stiffness parameters, and $\mathsf{P}_j$ is the matrix which copies oligomer coordinates into $j$-mer coordinates. By completing squares this energy can be expressed in the standard form introduced in Chapter 3:

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}(\mathsf{w} - \widehat{\mathsf{w}}) \cdot \mathsf{K}(\mathsf{w} - \widehat{\mathsf{w}}) + \widehat{\mathsf{U}}, \tag{6.7}$$

where

$$\begin{gathered} \mathsf{K} = \sum_{j=1}^{2} \mathsf{P}^T_j \mathsf{K}_j \mathsf{P}_j, \qquad \widehat{\mathsf{w}} = \mathsf{K}^{-1}\Big(\sum_{j=1}^{2} \mathsf{P}^T_j \mathsf{K}_j \widehat{\mathsf{w}}_j\Big), \\ \widehat{\mathsf{U}} = \frac{1}{2} \sum_{j=1}^{2} (\mathsf{P}_j\widehat{\mathsf{w}} - \widehat{\mathsf{w}}_j) \cdot \mathsf{K}_j(\mathsf{P}_j\widehat{\mathsf{w}} - \widehat{\mathsf{w}}_j). \end{gathered} \tag{6.8}$$

The above expressions provide the relations between the oligomer based energy parameters $\mathsf{K}$, $\widehat{\mathsf{w}}$ and $\widehat{\mathsf{U}}$ and the $j$-mer based energy parameters contained in $\mathsf{K}_j$ and $\widehat{\mathsf{w}}_j$. Indeed, the parameters of an oligomer are determined by those of its constituent $j$-mers in the following way. From $(6.8)_1$ and the definitions of $\mathsf{P}_j$ and $\mathsf{K}_j$, we deduce that $\mathsf{K}$ is a banded, block matrix whose entries depend locally on the entries of $\mathsf{K}_j$ $(j = 1, 2)$ as illustrated below:


+
=
.
$\mathsf{P}^T_1\mathsf{K}_1\mathsf{P}_1$
$\mathsf{P}^T_2\mathsf{K}_2\mathsf{P}_2$
$\mathsf{K}$

(6.9)

In the illustration, the shaded entries denote the blocks $\mathsf{K}^1_1, \dots, \mathsf{K}^n_1 \in \mathbb{R}^{6\times 6}$ of $\mathsf{K}_1$, the ruled entries denote the blocks $\mathsf{K}_2^1, \dots, \mathsf{K}_2^{n-1} \in \mathbb{R}^{18\times18}$ of $\mathsf{K}_2$, grid lines denote elements of $\mathbb{R}^{6\times6}$, and entries in the triple overlaps are summed in the obvious way.

In contrast to the stiffness, the entries of the oligomer shape vector $\widehat{\mathsf{w}}$ do not depend locally on the entries of $\widehat{\mathsf{w}}_j$ $(j = 1, 2)$. Indeed, from $(6.8)_2$ we see that the vector $\widehat{\mathsf{w}}$ is related to the vectors $\widehat{\mathsf{w}}_j$ through the inverse matrix $\mathsf{K}^{-1}$. Specifically, the entries in the product $\mathsf{K}_j\widehat{\mathsf{w}}_j$ depend locally on the entries of $\widehat{\mathsf{w}}_j$; the product is a $j$-mer vector of weighted $j$-mer shape parameters. Moreover, by definition of $\mathsf{P}_j^T$, the entries of the product $\mathsf{P}_j^T\mathsf{K}_j\widehat{\mathsf{w}}_j$ also depend locally on those of $\widehat{\mathsf{w}}_j$; the matrix $\mathsf{P}_j^T$ maps contributions from each $j$-mer into its corresponding location in the oligomer, with overlapping contributions being summed. However, the matrix $\mathsf{K}^{-1}$ will in general be dense; its diagonal blocks will be dominant and its off-diagonal blocks will decay at an exponential rate in accordance with the bandwidth of $\mathsf{K}$ [18]. From this we deduce that the oligomer shape parameters in $\widehat{\mathsf{w}}$ will be a convolution of the $j$-mer shape parameters in $\widehat{\mathsf{w}}_j$. The convolution window will be peaked along the diagonal of $\mathsf{K}^{-1}$ and will decay at a rate as described above.

The oligomer energy introduced here provides a natural model for frustration. Indeed, the term $\widehat{\mathsf{U}}$ in $(6.8)_3$ will in general be non-zero. This reflects the fact that, in the minimum energy or ground state $\widehat{\mathsf{w}}$ of the oligomer, each base cannot simultaneously minimise all of its $j$-mer interactions. Explicitly each base cannot simultaneously minimise its intra base pair interaction energy, and its two inter base pair cross-junction interactions. Instead, each base must find a compromise, which provides the physical explanation for the non-local nature of $\widehat{\mathsf{w}}$. We shall refer to the forces between bases required to maintain this state of compromise as the frustration forces; their collective energy content over the entire oligomer is precisely the frustration energy $\widehat{\mathsf{U}}$. Notice that a local measure $\widehat{\mathsf{U}}^a$ of frustration for each base pair $(\mathtt{X}, \overline{\mathtt{X}})_a$ can be obtained by summing over all the $j$-mer interactions involving that base pair.

## 6.2 Oligomer-based, nearest-neighbour model

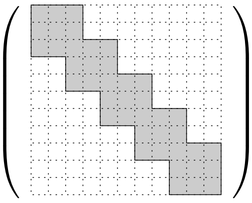
A nearest-neighbour internal energy for an oligomer of length $n$ is completely determined by the shape and stiffness parameters $\{\widehat{\mathsf{w}}_1^a, \mathsf{K}_1^a\}$ $(a = 1, \dots, n)$ and $\{\widehat{\mathsf{w}}_2^a, \mathsf{K}_2^a\}$ $(a = 1, \dots, n-1)$ introduced above. In general, these parameters may depend on the oligomer length $n$, the entire oligomer sequence $\mathtt{X}_1 \cdots \mathtt{X}_n$, and the location $a$ within the sequence. Of course we seek the simplest possible sequence-dependence of the parameter set compatible with a desired accuracy. In Chapter 7 we introduce a model, that assumes a local parameter dependence. However, we will make use of the general model as well, as an intermediate step for passing to the simpler one.

By an oligomer-based model we mean one in which the parameters $\{\widehat{\mathsf{w}}_1^a, \mathsf{K}_1^a\}$ and $\{\widehat{\mathsf{w}}_2^a, \mathsf{K}_2^a\}$ depend on the oligomer length, sequence and location in the most general way. Specifically, we assume there exists functions $\mathbb{W}_1$, $\mathbb{K}_1$, $\mathbb{W}_2$ and $\mathbb{K}_2$ such that

$$\begin{aligned} \widehat{\mathsf{w}}_1^a &= \mathbb{W}_1(n, \mathtt{X}_1, \dots, \mathtt{X}_n, a) \in \mathbb{R}^6, & \mathsf{K}_1^a &= \mathbb{K}_1(n, \mathtt{X}_1, \dots, \mathtt{X}_n, a) \in \mathbb{R}^{6\times6}, \\ \widehat{\mathsf{w}}_2^a &= \mathbb{W}_2(n, \mathtt{X}_1, \dots, \mathtt{X}_n, a) \in \mathbb{R}^{18}, & \mathsf{K}_2^a &= \mathbb{K}_2(n, \mathtt{X}_1, \dots, \mathtt{X}_n, a) \in \mathbb{R}^{18\times18}. \end{aligned} \tag{6.10}$$

In view of the relations in (6.8), this assumption implies that the oligomer stiffness and shape parameters $\mathsf{K}$ and $\widehat{\mathsf{w}}$ would be arbitrary functions of the oligomer length and sequence, and moreover that $\mathsf{K}$ would have the nearest-neighbour sparsity structure as depicted below, where grid lines denote $6 \times 6$ blocks as before:

$\mathsf{K} =$



. (6.11)

In this model, every oligomer is described by a unique pair of parameters $\mathsf{K}$ and $\widehat{\mathsf{w}}$. While the parameters for an oligomer with sequence $\mathtt{X}_1 \cdots \mathtt{X}_n$ are necessarily related by objectivity to those for an oligomer with sequence $\overline{\mathtt{X}}_n \cdots \overline{\mathtt{X}}_1$, there is in general no other relation and each independent oligomer is described by an independent pair of parameters $\mathsf{K}$ and $\widehat{\mathsf{w}}$. Consequently, there is no finite set of parameters that describes all possible oligomers of all possible lengths.

## 6.3 Fitting model parameters for given oligomer data

In an oligomer-based model, each sequence $\mathsf{S}_\mu$ is modelled by a Gaussian probability density

$$\rho_{\mu,\mathrm{M}}(\mathsf{w}) = \frac{1}{Z_{\mu,\mathrm{M}}} e^{-(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{M}})\cdot\mathsf{K}_{\mu,\mathrm{M}}(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{M}})/2}, \tag{6.12}$$

with shape vector $\widehat{\mathsf{w}}_{\mu,\mathrm{M}}$ and positive definite, symmetric stiffness matrix $\mathsf{K}_{\mu,\mathrm{M}}$, where now the only difference from the training set distributions is that the stiffness matrices $\mathsf{K}_{\mu,\mathrm{M}}$ must satisfy a specified sparsity pattern corresponding to the assumed type and range of interactions. We will be primarily concerned with the nearest-neighbour rigid base sparsity pattern illustrated in (6.11), but we will briefly discuss both a smaller and a larger stencil corresponding respectively to a nearest-neighbour rigid base pair model, and to a next to nearest-neighbour rigid base model.

For a specified sparsity pattern a best-fit oligomer model for each sequence $\mathsf{S}_\mu$ is defined by parameters $\mathsf{K}^*_{\mu,\mathrm{M}}$ and $\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$ with corresponding density $\rho^*_{\mu,\mathrm{M}}$ satisfying

$$\rho^*_{\mu,\mathrm{M}} = \underset{\rho_{\mu,\mathrm{M}}}{\operatorname{argmin}}\, D(\rho_{\mu,\mathrm{M}}, \rho_{\mu,\mathrm{o}}). \tag{6.13}$$

That is, among all model densities of the form (6.12) with specified sparsity pattern, the best-fit density has a minimum divergence to the training set density. Using the explicit expression in (5.31) for the divergence between Gaussian densities, we find that the best-fit parameters defined by (6.13) must satisfy the necessary conditions

$$\begin{aligned} \widehat{\mathsf{w}}^*_{\mu,\mathrm{M}} &= \widehat{\mathsf{w}}_{\mu,\mathrm{o}}, \\ \mathsf{K}^*_{\mu,\mathrm{M}} &= \underset{\mathsf{K}_{\mu,\mathrm{M}}}{\operatorname{argmin}}\ \frac{1}{2}\left[\mathsf{K}^{-1}_{\mu,\mathrm{M}} : \mathsf{K}_{\mu,\mathrm{o}} - \ln(\det \mathsf{K}_{\mu,\mathrm{o}} / \det \mathsf{K}_{\mu,\mathrm{M}}) - I : I\right], \end{aligned} \tag{6.14}$$

where the minimum is taken over the set of symmetric matrices with the specified sparsity pattern. This minimum can be found by numerically searching in the direction of the gradient

$$\nabla_{\mathsf{K}_{\mu,\mathrm{M}}} D(\rho_{\mu,\mathrm{M}}, \rho_{\mu,\mathrm{o}}) = \frac{1}{2}(\mathsf{K}_{\mu,\mathrm{M}}^{-1} - \mathsf{K}_{\mu,\mathrm{M}}^{-1}\,\mathsf{K}_{\mu,\mathrm{o}}\,\mathsf{K}_{\mu,\mathrm{M}}^{-1}). \tag{6.15}$$

Because of the identity (5.34) this optimisation problem can also be regarded as a best fit of the training set covariance by the inverse of a stiffness matrix of specified sparsity. However the optimisation procedure is simpler in terms of the stiffness matrix because the inverse of a sparse matrix is in general not itself sparse. Since the functional in the minimisation problem (6.14) is continuous and bounded below among positive-definite matrices, and becomes unbounded above as a definite matrix approaches a semi-definite one, we expect that a minimum exists within the set of symmetric, positive-definite matrices of prescribed sparsity pattern. Indeed, using two different numerical procedures, a numerical gradient flow, custom written by O. Gonzalez, and an implementation using the optimisation code Hanso [31, 45], we have been able to find a minimum for each sequence in the training set for each of three prescribed sparsity patterns. While such minima may only be local, our computations suggest that any possible multiple minima are rather isolated: for each sequence and sparsity pattern, small changes to the initial condition did not change the location of the minima, and the outputs of the two different codes were extremely close in all cases.

## 6.4 Model fitting results, approximation errors

Before presenting the results of our computations we explain in more detail the three sparsity patterns that we consider. Figure 6.2 presents sub matrices of two of the training set stiffness matrices $\mathsf{K}_{\mu,\mathrm{o}}$ for $\mu = 1, 3$. The qualitative features are similar for all sub matrices of all of the training set. All of the largest entries lie within a $6 \times 6$ block diagonal stencil (not explicitly shown). Such a stencil can be interpreted as a rigid base pair model with an internal energy that is an entirely local function of the inter coordinates at each junction. Each set of junction variables are then decoupled from all other configuration variables, either inter or intra. While such a model is rather standard in much of the literature of coarse-graining DNA, such a localised rigid base pair internal energy is a rather poor fit with MD simulations at the scale of tens of base pairs, as discussed in Chapter 4. Moreover we are now in a position to quantify this observation by comparing with other stencils, and using the Kullback-Leibler divergence. It is evident that considerably more of the signal in the training set stiffnesses lie within the overlapping $18 \times 18$ nearest-neighbour rigid base stencil marked in red in Figure 6.2, and yet more within the overlapping $30 \times 30$ next to nearest-neighbour rigid base stencil marked in blue. We remark that the next to nearest-neighbour stencil still excludes entries of the same magnitude to those that it captures beyond the nearest-neighbour stencil. This observation suggests that it would be interesting to consider an even longer range interaction model. But here we content ourselves with discussing only three stencils.

For each of the 53 training set oligomers, and for each of the three sparsity patterns, Figure 6.1 provides the values of the Kullback-Leibler divergence between the best oligomer based fit of that prescribed sparsity pattern (computed as described immediately above) and the training set distribution, scaled by $D_{\mathrm{o}}$ as defined in (5.35) for 18-mers and, to account for

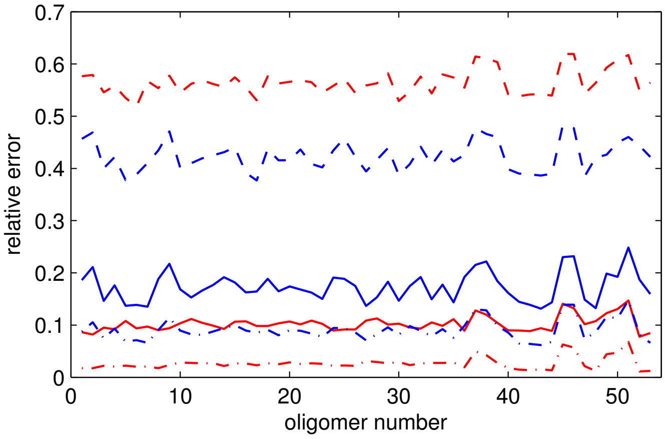


Figure 6.1: Relative errors in oligomer based models. As a function of sequence index $\mu$ in the training set the red curves indicate the relative divergences in probability densities $D(\rho^*_{\mu,\mathrm{M}}, \rho_{\mu,\mathrm{o}})/D_\mathrm{o}$ (with $D_\mathrm{o} \approx 85$ for 18-mers and $D_\mathrm{o} \approx 57$ for 12-mers) while the blue curves represent the relative Frobenius error in the stiffness matrices $||\mathsf{K}^*_{\mu,\mathrm{M}} - \mathsf{K}_{\mu,\mathrm{o}}||/||\mathsf{K}_{\mu,\mathrm{o}}||$. In each of three cases the line style indicates the enforced sparsity pattern: dashed, nearest-neighbour rigid base pair; solid, nearest-neighbour rigid base; dash-dot, next to nearest-neighbour rigid base.

the smaller number of degrees of freedom, by $2D_\mathrm{o}/3$ for 12-mers. Because it is only differences in stiffness matrices that contribute to the divergences, we also plot the relative errors in the Frobenius matrix norm of the differences in stiffness matrices. For the nearest-neighbour rigid base pair sparsity pattern the relative Kullback-Leibler divergence is 55% or more, with the relative Frobenius error somewhat smaller. For the nearest-neighbour rigid base stencil the relative Kullback-Leibler divergence is approximately 10% with the Frobenius error now larger, around 20%. A visualisation of the entries of the difference of stiffness matrices for the sequence $\mu = 3$ is given in Figure 6.2. And for the next to nearest-neighbour rigid base stencil the Kullback-Leibler divergence is less than 5% for nearly all sequences, with the Frobenius error approximately 10%.

In this work we make a compromise between the complexity of the model and the quality of the oligomer based enforced sparsity fit, and hereafter consider only the stencil with overlapping $18 \times 18$ blocks corresponding to a rigid base model with nearest-neighbour interactions.

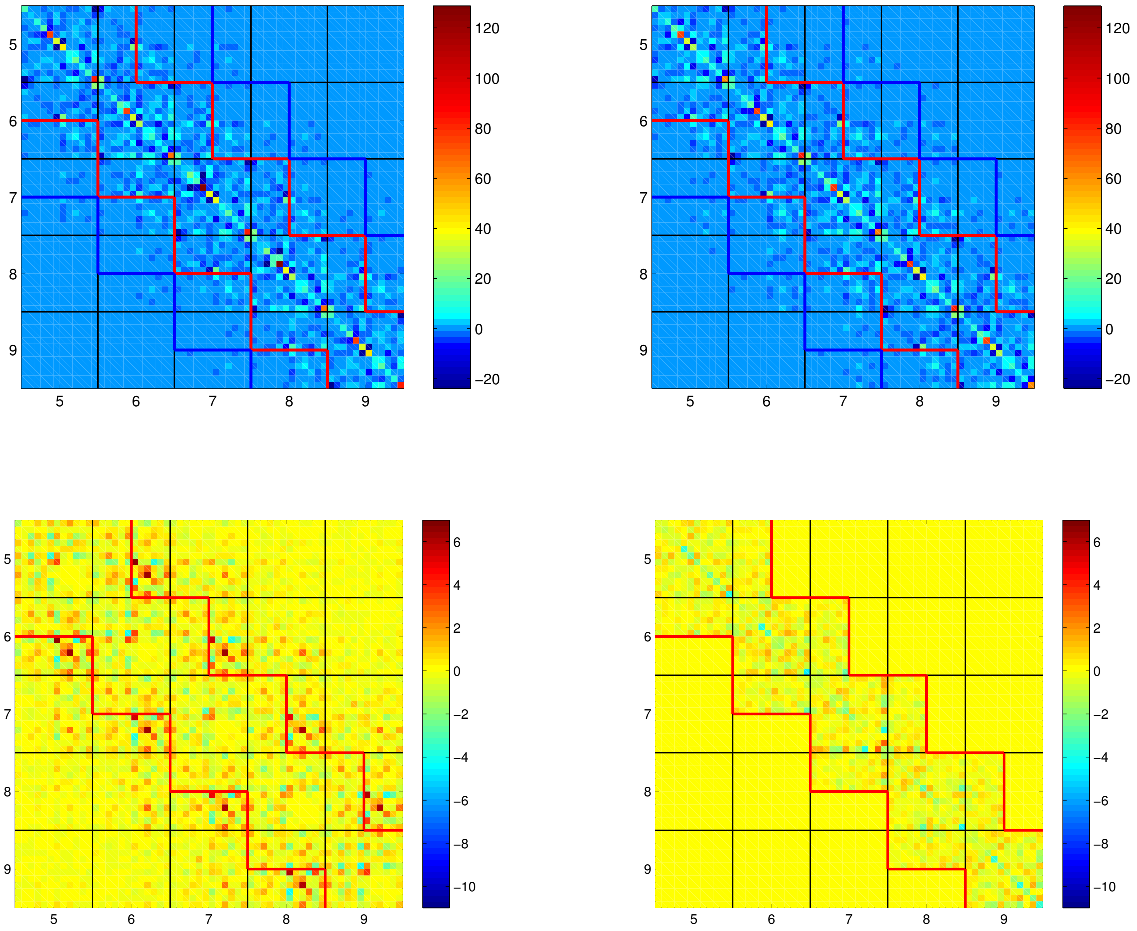


Figure 6.2: Top: A portion of two training set stiffness matrices ($\mu = 1, 3$ on right and left respectively) with stencils for nearest-neighbour (red) and next to nearest-neighbour (blue) versions of our rigid base internal energy model. $12 \times 12$ blocks marked by black lines contain entries corresponding to intra and inter base pair coordinates $(y^a, z^a)$ at positions $a = 5, \ldots, 9$. Bottom: On left the difference between the training set stiffness matrix $\mathsf{K}_{3,\mathrm{o}}$ and the oligomer based, best fit, nearest-neighbour stiffness $\mathsf{K}_{3,\mathrm{M}}$, on right the difference between $\mathsf{K}_{3,\mathrm{M}}$ and the stiffness $\mathsf{K}_{3,\mathrm{m}}$ reconstructed from the dimer dependent model described in Chapter 7. All entries outside the red stencil on the right vanish by definition, while on the left they are identical with the plot immediately above, but now the colour scale is much different.

# Chapter 7

# Dimer dependent nearest-neighbour rigid base model

In this chapter we present a special case of the model, presented in Chapter 6, and the procedure of getting model parameters from the data.

## 7.1 Material assumptions and model parameters

By a dimer-based model we mean one in which the parameters $\{\widehat{\mathsf{w}}_1^a, \mathsf{K}_1^a\}$ in (6.2) and $\{\widehat{\mathsf{w}}_2^a, \mathsf{K}_2^a\}$ in (6.3) depend only on the local dimer sequence $\mathtt{X}_a\mathtt{X}_{a+1}$, but not explicitly on either the oligomer length $n$ or the location $a$. Specifically, we assume there exist functions $\mathbb{W}_1$, $\mathbb{K}_1$, $\mathbb{W}_2$ and $\mathbb{K}_2$ such that

$$\begin{aligned} \widehat{\mathsf{w}}_1^a = \mathbb{W}_1(\mathtt{X}_a) \in \mathbb{R}^6, \quad \mathsf{K}_1^a = \mathbb{K}_1(\mathtt{X}_a) \in \mathbb{R}^{6\times 6}, \\ \widehat{\mathsf{w}}_2^a = \mathbb{W}_2(\mathtt{X}_a, \mathtt{X}_{a+1}) \in \mathbb{R}^{18}, \quad \mathsf{K}_2^a = \mathbb{K}_2(\mathtt{X}_a, \mathtt{X}_{a+1}) \in \mathbb{R}^{18\times 18}, \end{aligned} \tag{7.1}$$

Notice that the above relations are assumed to hold in the interior as well as the ends of an arbitrary oligomer; additional assumptions or parameters could be introduced to capture any exceptional end effects, but we do not explore that line here. With the assumption of dimer sequence-dependence there is a finite set of parameters that describes the energy of all possible oligomers of all possible lengths. Specifically, each of the functions $\mathbb{W}_1(\mathtt{X}_a)$ and $\mathbb{K}_1(\mathtt{X}_a)$ can assume only 4 possible values corresponding to the 4 possible choices of $\mathtt{X}_a \in \{\mathtt{T}, \mathtt{A}, \mathtt{C}, \mathtt{G}\}$. Similarly, each of the functions $\mathbb{W}_2(\mathtt{X}_a, \mathtt{X}_{a+1})$ and $\mathbb{K}_2(\mathtt{X}_a, \mathtt{X}_{a+1})$ can assume only 16 possible values corresponding to the 16 possible choices of the pair $\mathtt{X}_a, \mathtt{X}_{a+1} \in \{\mathtt{T}, \mathtt{A}, \mathtt{C}, \mathtt{G}\}$.

To satisfy objectivity, i.e. the relations (3.13), it is sufficient to assume that the functions $\mathbb{W}_1$, $\mathbb{K}_1$, $\mathbb{W}_2$ and $\mathbb{K}_2$ are locally objective in the following sense, for any $\mathtt{X}, \mathtt{Y} \in \{\mathtt{T}, \mathtt{A}, \mathtt{C}, \mathtt{G}\}$,

$$\begin{aligned} \mathbb{W}_1(\mathtt{X}) = \mathsf{E}_1\mathbb{W}_1(\overline{\mathtt{X}}), \quad \mathbb{K}_1(\mathtt{X}) = \mathsf{E}_1\mathbb{K}_1(\overline{\mathtt{X}})\mathsf{E}_1, \\ \mathbb{W}_2(\mathtt{X}, \mathtt{Y}) = \mathsf{E}_2\mathbb{W}_2(\overline{\mathtt{Y}}, \overline{\mathtt{X}}), \quad \mathbb{K}_2(\mathtt{X}, \mathtt{Y}) = \mathsf{E}_2\mathbb{K}_2(\overline{\mathtt{Y}}, \overline{\mathtt{X}})\mathsf{E}_2, \end{aligned} \tag{7.2}$$

where the matrix $\mathsf{E}_n$ is defined in (2.38). These conditions imply that the values of the functions are not all independent. Specifically, if we arrange the 4 possible values for $\mathtt{X}$ in a table as shown

$$\begin{array}{cc|cc} \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \end{array}$$

then the value of $\mathbb{W}_1$ for the entries on the right-half of the table are completely determined by those of the left-half, namely $\mathbb{W}_1(\mathtt{T}) = \mathsf{E}_1\mathbb{W}_1(\mathtt{A})$ and $\mathbb{W}_1(\mathtt{C}) = \mathsf{E}_1\mathbb{W}_1(\mathtt{G})$, and similarly for $\mathbb{K}_1$. Thus there are only 2 independent values of the monomer parameter functions $\{\mathbb{W}_1, \mathbb{K}_1\}$. Similarly, if we arrange the 16 possible values of $\mathtt{XY}$ in a table as shown, where $\mathtt{X}$ is vertical and $\mathtt{Y}$ is horizontal,

| | T | C | A | G |
|---|---|---|---|---|
| A | AT | AC | AA | AG |
| G | GT | GC | GA | GG |
| T | TT | TC | TA | TG |
| C | CT | CC | CA | CG |

then the value of $\mathbb{W}_2$ for the entries above the table diagonal are completely determined by those below, namely $\mathbb{W}_2(\mathtt{A},\mathtt{C}) = \mathsf{E}_2\mathbb{W}_2(\mathtt{G},\mathtt{T})$ and so on, and similarly for $\mathbb{K}_2$. Moreover, the value of $\mathbb{W}_2$ for each diagonal entry must satisfy the objectivity condition $\mathbb{W}_2(\mathtt{A},\mathtt{T}) = \mathsf{E}_2\mathbb{W}_2(\mathtt{A},\mathtt{T})$ and so on, and similarly for $\mathbb{K}_2$. From this we deduce that there are only 10 independent values of the dimer parameter functions $\{\mathbb{W}_2, \mathbb{K}_2\}$ corresponding to a triangular portion of the table with the diagonal included, and that values on the diagonal entries must be invariant under the transformation $\mathsf{E}_2$. Note that the four $2 \times 2$ blocks of the table correspond to purine-pyrimidine and pyrimidine-purine dimer steps on the diagonal, with the off-diagonal blocks being purine-purine and pyrimidine-pyrimidine dimer steps.

Thus a dimer-based model is completely defined by the parameters

$$\begin{aligned} \widehat{\mathsf{w}}_1^{\alpha} &:= \mathbb{W}_1(\alpha) \in \mathbb{R}^6, \quad \mathsf{K}_1^{\alpha} := \mathbb{K}_1(\alpha) \in \mathbb{R}^{6\times 6}, \quad \alpha \in \mathtt{M}, \\ \widehat{\mathsf{w}}_2^{\alpha\beta} &:= \mathbb{W}_2(\alpha,\beta) \in \mathbb{R}^{18}, \quad \mathsf{K}_2^{\alpha\beta} := \mathbb{K}_2(\alpha,\beta) \in \mathbb{R}^{18\times 18}, \quad \alpha\beta \in \mathtt{D}, \end{aligned} \tag{7.3}$$

where $\mathtt{M}$ is a set of any 2 independent monomers and $\mathtt{D}$ is a set of any 10 independent dimers.

In our later developments, it will be convenient to work with weighted shape parameters instead of the unweighted parameters $\widehat{\mathsf{w}}_1^{\alpha}$ and $\widehat{\mathsf{w}}_2^{\alpha\beta}$ above. Specifically, we introduce the weighted shape parameters

$$\sigma_1^{\alpha} := \mathsf{K}_1^{\alpha}\widehat{\mathsf{w}}_1^{\alpha} \in \mathbb{R}^6, \quad \sigma_2^{\alpha\beta} := \mathsf{K}_2^{\alpha\beta}\widehat{\mathsf{w}}_2^{\alpha\beta} \in \mathbb{R}^{18}. \tag{7.4}$$

We remark that the variables $\sigma_1^{\alpha}$ and $\sigma_2^{\alpha\beta}$ are stress-like in character, each being the product of a strain variable by a stiffness matrix. Their physical interpretation is not immediately evident, but they arise very naturally in the algebra of the problem. In particular a complete dimer-based parameter set is defined by specifying the values of $\{\sigma_1^{\alpha}, \mathsf{K}_1^{\alpha}\}$ for $\alpha \in \mathtt{M}$ and the values of $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ for $\alpha\beta \in \mathtt{D}$. Parameters for all other monomers and dimers can be obtained from the objectivity relations (7.2). One such explicit parameter set is presented below.

## 7.2 Structure and properties of oligomer parameters

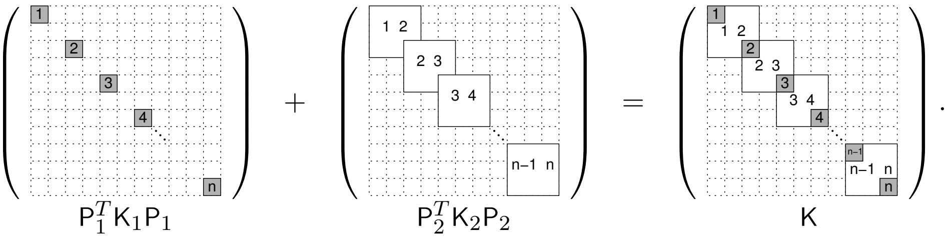
The shape, stiffness and frustration parameters $\widehat{\mathsf{w}}$, $\mathsf{K}$ and $\widehat{\mathsf{U}}$ for any given oligomer $\mathtt{X}_1 \cdots \mathtt{X}_n$ can be assembled from the monomer and dimer parameters $\{\sigma_1^\alpha, \mathsf{K}_1^\alpha\}$ and $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ using the relations in (6.8). Specifically, and omitting the expression for $\widehat{\mathsf{U}}$ for brevity, the expressions for the oligomer parameters $\widehat{\mathsf{w}}$ and $\mathsf{K}$ take the forms

$$\begin{array}{rcll} \mathsf{K} & = & \mathsf{P}_1^T \mathsf{K}_1 \mathsf{P}_1 + \mathsf{P}_2^T \mathsf{K}_2 \mathsf{P}_2 & \in \mathbb{R}^{(12n-6)\times(12n-6)}, \\ \sigma & = & \mathsf{P}_1^T \sigma_1 + \mathsf{P}_2^T \sigma_2 & \in \mathbb{R}^{12n-6}, \\ \widehat{\mathsf{w}} & = & \mathsf{K}^{-1}\sigma & \in \mathbb{R}^{12n-6}, \end{array} \tag{7.5}$$

where

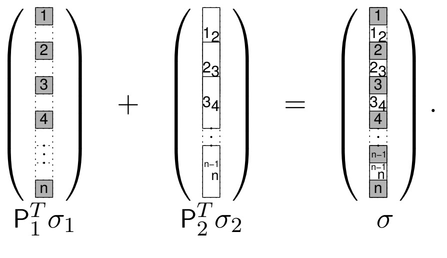
$$\begin{array}{rcllr} \mathsf{K}_1 & = & \mathrm{diag}(\mathsf{K}_1^1, \ldots, \mathsf{K}_1^n) & \in \mathbb{R}^{6n\times 6n}, & \\ \sigma_1 & = & (\sigma_1^1, \ldots, \sigma_1^n) & \in \mathbb{R}^{6n}, & \\ \mathsf{K}_2 & = & \mathrm{diag}(\mathsf{K}_2^1, \ldots, \mathsf{K}_2^{n-1}) & \in \mathbb{R}^{18(n-1)\times 18(n-1)}, & \\ \sigma_2 & = & (\sigma_2^1, \ldots, \sigma_2^{n-1}) & \in \mathbb{R}^{18(n-1)}, & \\ \mathsf{K}_1^a & = & \mathsf{K}_1^{\mathtt{X}_a} & \in \mathbb{R}^{6\times 6} & (a = 1, \ldots, n), \\ \sigma_1^a & = & \sigma_1^{\mathtt{X}_a} & \in \mathbb{R}^{6} & (a = 1, \ldots, n), \\ \mathsf{K}_2^a & = & \mathsf{K}_2^{\mathtt{X}_a\mathtt{X}_{a+1}} & \in \mathbb{R}^{18\times 18} & (a = 1, \ldots, n-1), \\ \sigma_2^a & = & \sigma_2^{\mathtt{X}_a\mathtt{X}_{a+1}} & \in \mathbb{R}^{18} & (a = 1, \ldots, n-1). \end{array} \tag{7.6}$$

The dependence of each block of the oligomer stiffness matrix $\mathsf{K}$ upon the oligomer sequence $\mathtt{X}_1 \cdots \mathtt{X}_n$ is illustrated below, where on the left hand side each single number in a block denotes a dependence on the monomer $\mathtt{X}_a$, while each pair of numbers in a block denotes a dependence on the dimer $\mathtt{X}_a\mathtt{X}_{a+1}$. On the right hand side the triple overlapping blocks denote sums as before, with a trimer sequence dependence given by the union of the two overlapping block dimer dependencies.


1
2
3
4
n
+
1 2
2 3
3 4
n–1 n
=
1
1 2
2
2 3
3
3 4
4
n–1
n–1 n
n
.
$\mathsf{P}_1^T \mathsf{K}_1 \mathsf{P}_1$
$\mathsf{P}_2^T \mathsf{K}_2 \mathsf{P}_2$
$\mathsf{K}$

(7.7)

Similarly, the dependence of each block of the weighted shape vector $\sigma$ upon the oligomer sequence $\mathtt{X}_1 \cdots \mathtt{X}_n$ is as follows:



(7.8)

In (7.5), notice that the oligomer stiffness matrix $\mathsf{K}$ and shape vector $\widehat{\mathsf{w}}$ depend directly on the local stiffness and weighted shape parameters $\{\sigma_1^\alpha, \mathsf{K}_1^\alpha\}$ and $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$, but only indirectly on the local unweighted shape parameters $\widehat{\mathsf{w}}_1^{\alpha}$ and $\widehat{\mathsf{w}}_2^{\alpha\beta}$ via the definition (7.4). For this reason, we henceforth restrict attention to the weighted parameters. The unweighted parameters are needed explicitly only when the oligomer frustration parameter $\widehat{\mathsf{U}}$ or the local energy functions $\mathsf{U}_1^{\alpha}$ and $\mathsf{U}_2^{\alpha}$ are needed explicitly.

The overlapping structure of the map from local to oligomer parameters defined in (7.5)–(7.6) and illustrated in (7.7)–(7.8) has some interesting implications. A first implication concerns the observability of the local parameters. In positions within the oligomer arrays where there are no overlaps, we find that the local parameters are directly observable: there is no coupling and the oligomer parameters are equal to the relevant local parameters. In particular the assumption of dimer sequence dependence of non-overlapping blocks in both the stiffness matrix and weighted shape variables can be tested directly from the statistics of the occurrences of dimer steps in the training set to be described below.

In contrast, in positions within the oligomer arrays where there are overlaps, we find that the local parameters are not directly observable: there is coupling and the oligomer parameters are sums of relevant local parameters. Specifically, for the stiffness parameters, the structure of the coupling at interior positions within an oligomer are all identical, of the form $\mathsf{K}_{2,33}^{\alpha\beta} + \mathsf{K}_1^{\beta} + \mathsf{K}_{2,11}^{\beta\gamma}$, whereas the structure of the coupling at each of the two ends is different, of the form $\mathsf{K}_1^{\beta} + \mathsf{K}_{2,11}^{\beta\gamma}$ at the leading end, and $\mathsf{K}_{2,33}^{\alpha\beta} + \mathsf{K}_1^{\beta}$ at the trailing end, for some $\alpha$, $\beta$ and $\gamma$. Exactly analogous couplings hold for the weighted shape parameters. In view of this, the inverse problem of determining local parameters from oligomer parameters requires data from the ends as well as the interior of an oligomer so that the couplings can be resolved.

A second implication of the structure of the map from local to oligomer parameters concerns the positivity of the local stiffness parameters. The set of local stiffness parameters $\{\mathsf{K}_1^{\alpha}, \mathsf{K}_2^{\alpha\beta}\}$ is called admissible if it yields a positive-definite oligomer stiffness matrix $\mathsf{K}$ for an arbitrary sequence $\mathtt{X}_1\mathtt{X}_2\cdots\mathtt{X}_n$ of arbitrary length $n \geq 2$. In view of the additive, overlapping structure of the map from local to oligomer parameters, a sufficient condition for admissibility is that each of the local stiffness parameter matrices be positive-definite. Alternatively, a weaker set of conditions is also sufficient, namely that each of the local stiffness parameter matrices be only semi-positive-definite, provided additionally that the reconstructed oligomer matrices for any ten independent sequences of length two, i.e. physical dimer oligomers, are positive-definite. This latter weaker set of sufficient conditions is mathematically more convenient than the former in dealing with certain optimisation problems associated with the estimation of local parameters. As it happens the parameter set we extract below has positive definite local stiffness parameter matrices, but some extremely small eigenvalues arise. While all the physical dimer stiffness matrices are positive definite with much larger, smallest eigenvalues. Thus the weakened sufficient conditions are of some importance to ensure robustness.

## 7.3 Compatibility between model assumptions and the MD data

For estimating a parameter set $\mathcal{P} = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$ from our training data set (4.1), we first considered the over-determined system of equations

$$\left.\begin{array}{rcl} \mathsf{K}_{\mu,\mathrm{m}} & = & \mathsf{K}^*_{\mu,\mathrm{M}} \\ \sigma_{\mu,\mathrm{m}} & = & \sigma^*_{\mu,\mathrm{M}} \end{array}\right\}, \quad \mu = 1, \dots, N. \tag{7.9}$$

The number of equations is determined by the number of training oligomers. The left hand sides are the explicit combinations of the parameter set given in (7.5). The matrices $\mathsf{K}^*_{\mu,\mathrm{M}}$ on right hand side of (7.9) are known data at this stage of the process, because they have been determined from the oligomer based fit. For the data on the right hand side of the vector equations in (7.9), we chose $\sigma^*_{\mu,\mathrm{M}} = \mathsf{K}_{\mu,\mathrm{m}}\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$, which means that we first solved the decoupled least squares system for $\mathsf{K}_{\mu,\mathrm{m}}$ in (7.9), and then used the resulting solution to construct the data on the right hand sides of the equations for $\sigma_{\mu,\mathrm{m}}$. Alternatively we could have instead used $\sigma^*_{\mu,\mathrm{M}} = \mathsf{K}^*_{\mu,\mathrm{M}}\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$, but we rejected this approach because we found that the resulting, intermediate, parameter set gave less good reconstructions of oligomer shapes. One explanation of this observation is that the choice $\sigma^*_{\mu,\mathrm{M}} = \mathsf{K}_{\mu,\mathrm{m}}\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$ eliminates a part of variation in the elements of $\sigma^*_{\mu,\mathrm{M}}$ that is due to possible errors in the matrix $\mathsf{K}^*_{\mu,\mathrm{M}}$.

There are two significant features of system (7.9): i) there is a block by block localised dependence of the parameter set on sequence, and ii) the equations decouple into independent least squares systems entry by entry in the matrix equations, with the vector equations decoupling in the same way, once the solution of the matrix least squares system has been found.

The localised dependence on sequence of the unknown parameter set is encoded in the overlapping structures illustrated in diagrams (7.7) and (7.8)—away from ends the parameters in non-overlap regions depend only on the local dimer sequence, while the parameters in overlap regions depend on the local trimer sequence. The extent to which these assumptions are compatible with the data can then be assessed.

For the sequences $\mu \leq 36$ (which had on average considerably more snapshots retained after filtering for broken hydrogen bonds), and staying two base pairs away from the ends, there were approximately 65 instances of each of the 10 independent dimers, and approximately 15 instances of each of the 32 independent trimers, reading along both backbones of each oligomer. To verify our hypotheses about dimer and trimer dependence of the entries of matrices $\mathsf{K}^*_{\mu,\mathrm{M}}$ and vectors $\sigma^*_{\mu,\mathrm{M}}$, we first computed and compared entry by entry for both stiffness matrices and weighted shape vectors, the averages and standard deviations for each independent instance of the dimer, respectively trimer.

Figure 7.1 shows plots of diagonal entries of inter base pair (non-overlap) stiffness blocks and weighted shape parameters, averaged over all instances of each independent dimer in the sequences $\mu \leq 36$, the error bars indicating standard deviations with a group of observations for each dimer. We observed that the average values of both stiffness and weighted shape parameters are quite similar for all purine-pyrimidine steps, all pyrimidine-purine steps and all purine-purine steps, but quite different between the groups. The standard deviations are relatively big for stiffness blocks, which could be due to convergence errors in the data, or could indicate the dependence of the corresponding entries of the wider base pair neighbourhood.

The weighted shape parameter error bars are relatively small, however, as mentioned before, they only account for variation in shape vectors $\widehat{\mathsf{w}}$, as we chose to use $\sigma^*_{\mu,\mathrm{M}} = \mathsf{K}_{\mu,\mathrm{m}}\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$. The same conclusions hold for the intra base pair (overlap) parameters, averaged over each independent trimer (see Figure 7.2). One can notice, that the parameter values are quite different for the XAZ steps than for the XGZ steps, $\mathtt{X}, \mathtt{Z} \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$, even though both A and G are purines. Inside these groups, the values are similar for all the trimers, where the pair $\mathtt{X}, \mathtt{Z}$ is purine-pyrimidine, where it is pyrimidine-purine and so on. The plots of dimer and trimer averages of complete stiffness blocks, centred by subtracting averages of each entry over all dimers/trimers can be found in the Appendix B. The observation that can be made from these plots is that the conclusions about the diagonal stiffness entries can be generalised for other entries as well. However, many of the off-diagonal stiffness entries are close to zero and have a small variation among all the instances in all the dimers/trimers.

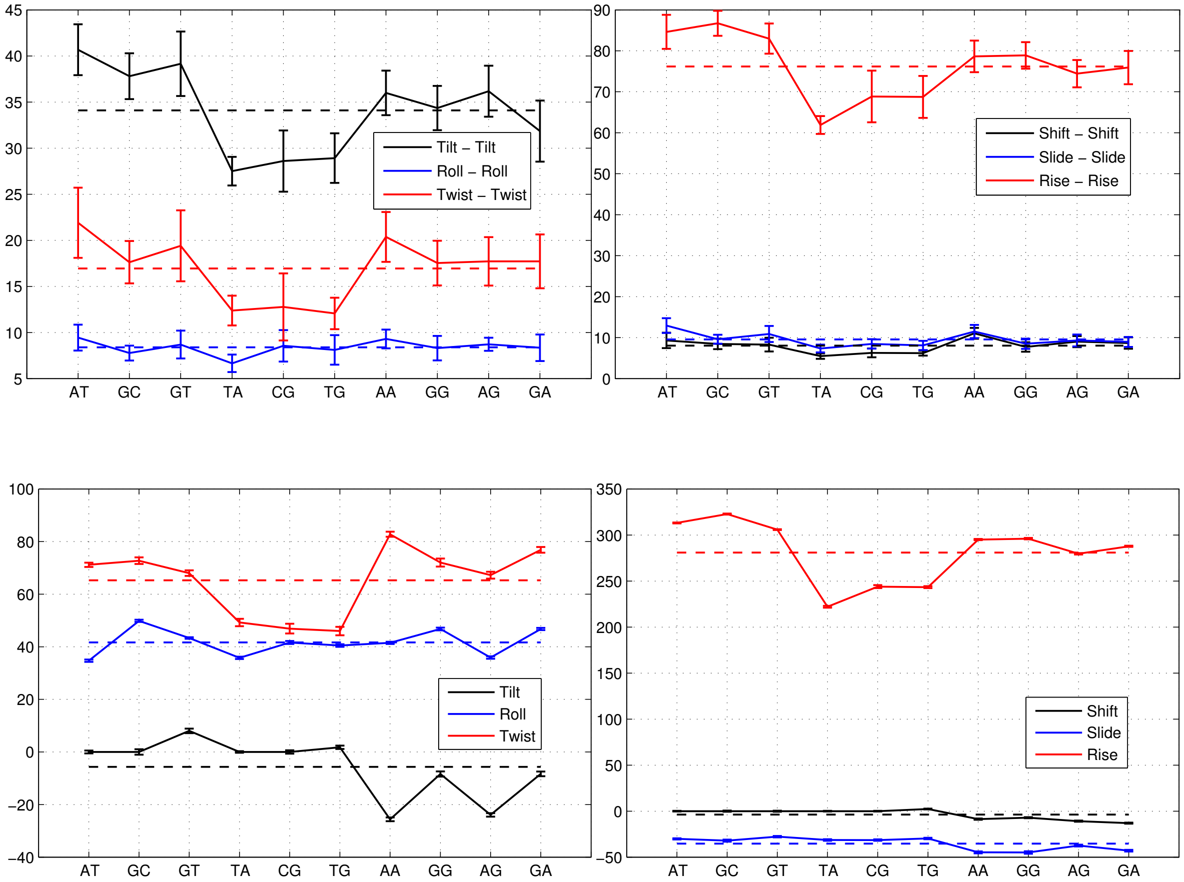


Figure 7.1: Diagonal elements of inter base pair stiffness blocks (top plots) and weighted shape parameters (bottom plots), averaged over all instances of each independent dimer step on both strands of the sequences $\mu \leq 36$. The dimers in each plot are ordered in three groups, the first group on the left being RY (three dimers), the second group YR (three dimers), the third RR (four dimers), where R denotes a purine and Y a pyrimidine. Error bars indicate standard deviations for each dimer, dashed lines mark averages over all dimers.

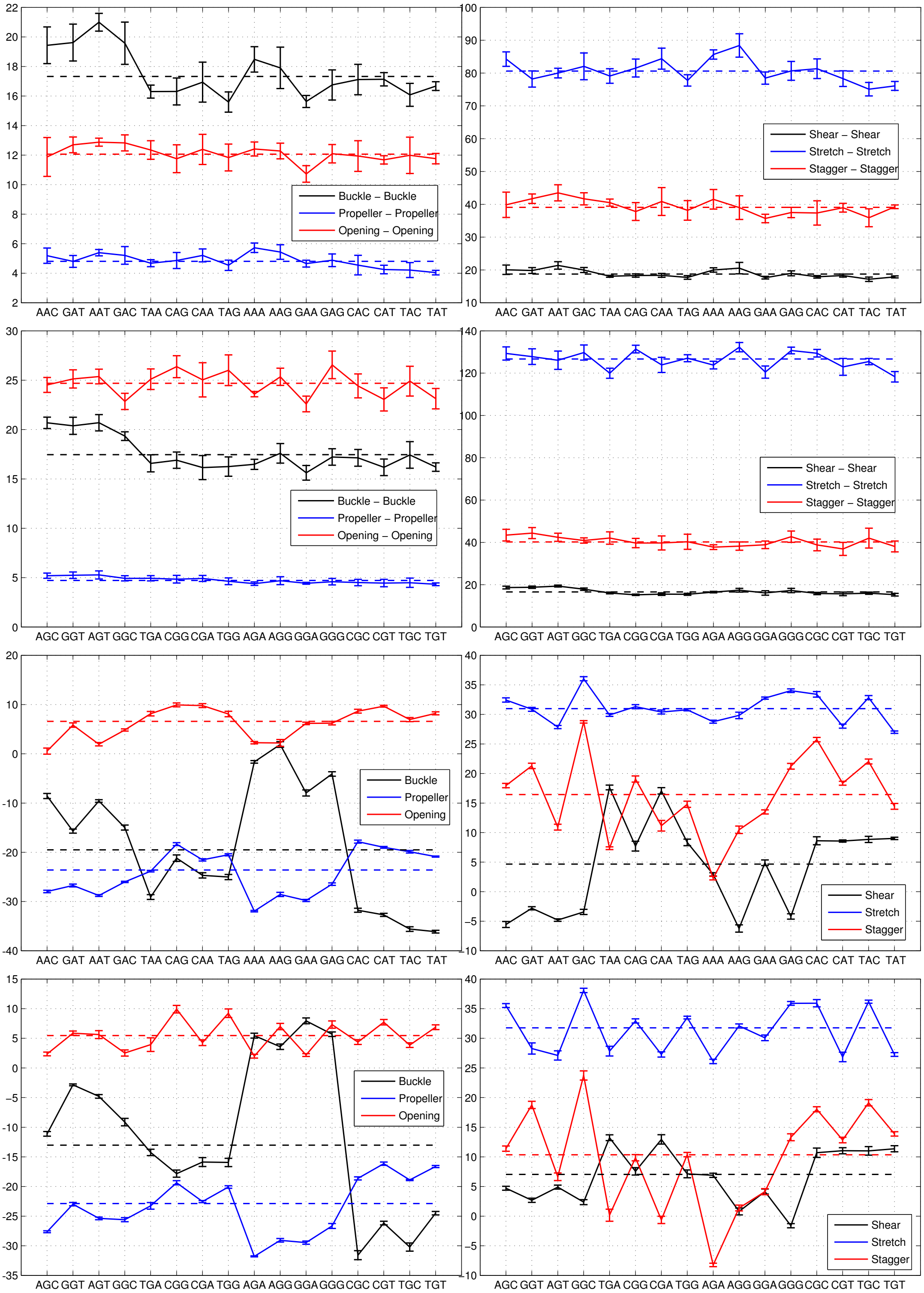


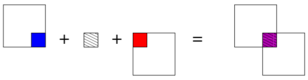
Figure 7.2: Diagonal elements of intra base pair stiffness blocks (top four plots) and weighted shape parameters (bottom four plots), averaged over all instances of each independent trimer on both strands of the sequences $\mu \leq 36$. The trimers in each plot are ordered in groups of four, the first group on the left being `RRY`, the second group `YRR`, the third `RRR` and the fourth `YRY`, where `R` denotes a purine and `Y` a pyrimidine. Error bars indicate standard deviations for each trimer, dashed lines mark averages over all trimers in a plot.

Finally, it can be concluded that the observed stiffness and weighted shape parameters are sequence dependent, as can already be seen by looking to individual oligomer parameters. Trying to approximate this dependence by dimer/trimer dependence of corresponding entries seems to be a reasonable choice, even though models with wider sequence dependence could also be considered.

## 7.4 Least-squares system without end data

In the over-determined system (7.9) we are free to consider only a subset of all blocks that appear, or to assign different weights to different blocks, reflecting differences in either importance of the fit or confidence in the data. For both reasons, it is quite natural to first consider only the system of equations that arise away from the ends. Then the least-squares solution of (7.9) for the stiffness and the weighted shape entries, corresponding to the non-overlapping regions of the oligomer parameters, are just the table averages of these entries over all oligomers in the training set, as we explain further now.

For the entries of the overlapping regions, the system (7.9) decouples into separate equations



$$\text{i.e.} \quad \mathsf{K}^{\alpha\beta}_{2,33} + \mathsf{K}^{\beta}_{1} + \mathsf{K}^{\beta\gamma}_{2,11} = \mathsf{K}^{\alpha\beta\gamma}, \tag{7.10}$$

for stiffness blocks and $\sigma^{\alpha\beta}_{2,33} + \sigma^{\beta}_{1} + \sigma^{\beta\gamma}_{2,11} = \sigma^{\alpha\beta\gamma}$ for weighted shape vectors. These equations further decouple into separate scalar equations for each entry of the stiffness and weighted shape parameters.

Consider first the equations for stiffnesses and let $k^{\alpha\beta}_{33} = [\mathsf{K}^{\alpha\beta}_{2,33}]_{ij}$, $k^{\beta\gamma}_{11} = [\mathsf{K}^{\beta\gamma}_{2,11}]_{ij}$, $k^{\beta} = [\mathsf{K}^{\beta}_{1}]_{ij}$ and $k^{\alpha\beta\gamma} = [\mathsf{K}^{\alpha\beta\gamma}]_{ij}$ for some fixed $i$ and $j$, $1 \leq i, j \leq 6$. Similarly, for weighted shape equations we would denote $k^{\alpha\beta}_{33} = [\sigma^{\alpha\beta}_{2,3}]_{i}$, $k^{\beta\gamma}_{11} = [\sigma^{\beta\gamma}_{2,1}]_{i}$, $k^{\beta} = [\sigma^{\beta}_{1}]_{i}$ and $k^{\alpha\beta\gamma} = [\sigma^{\alpha\beta\gamma}]_{i}$ for some fixed $i \leq 6$. Then (7.9) becomes a linear system $A\boldsymbol{x} = \boldsymbol{b}$, with $A \in \mathbb{R}^{16\times 9}$, $\boldsymbol{x} \in \mathbb{R}^{9}$ and $\boldsymbol{b} \in \mathbb{R}^{16}$, that can be explicitly written as

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0\\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0\\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0\\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0\\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0\\
0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0\\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0\\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0\\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0\\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1\\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1\\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1\\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
k^{A\beta}_{33}\\ k^{G\beta}_{33}\\ k^{C\beta}_{33}\\ k^{T\beta}_{33}\\ k^{\beta}\\ k^{\beta A}_{11}\\ k^{\beta G}_{11}\\ k^{\beta C}_{11}\\ k^{\beta T}_{11}
\end{pmatrix}
=
\begin{pmatrix}
k^{A\beta A}\\ k^{G\beta A}\\ \vdots\\ \vdots\\ \vdots\\ k^{T\beta T}
\end{pmatrix}. \tag{7.11}$$

The corresponding system of normal equations $A^T A\boldsymbol{x} = A^T\boldsymbol{b}$ is

$$\begin{pmatrix} 4 & 0 & 0 & 0 & 4 & 1 & 1 & 1 & 1 \\ 0 & 4 & 0 & 0 & 4 & 1 & 1 & 1 & 1 \\ 0 & 0 & 4 & 0 & 4 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 4 & 4 & 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & 4 & 16 & 4 & 4 & 4 & 4 \\ 1 & 1 & 1 & 1 & 4 & 4 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 4 & 0 & 4 & 0 & 0 \\ 1 & 1 & 1 & 1 & 4 & 0 & 0 & 4 & 0 \\ 1 & 1 & 1 & 1 & 4 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} k_{33}^{\mathtt{A}\beta} \\ k_{33}^{\mathtt{G}\beta} \\ k_{33}^{\mathtt{C}\beta} \\ k_{33}^{\mathtt{T}\beta} \\ k^{\beta} \\ k_{11}^{\beta\mathtt{A}} \\ k_{11}^{\beta\mathtt{G}} \\ k_{11}^{\beta\mathtt{C}} \\ k_{11}^{\beta\mathtt{T}} \end{pmatrix} = \begin{pmatrix} \sum_\gamma k^{\mathtt{A}\beta\gamma} \\ \sum_\gamma k^{\mathtt{G}\beta\gamma} \\ \sum_\gamma k^{\mathtt{C}\beta\gamma} \\ \sum_\gamma k^{\mathtt{T}\beta\gamma} \\ \sum_{\alpha,\gamma} k^{\alpha\beta\gamma} \\ \sum_\alpha k^{\alpha\beta\mathtt{A}} \\ \sum_\alpha k^{\alpha\beta\mathtt{G}} \\ \sum_\alpha k^{\alpha\beta\mathtt{C}} \\ \sum_\alpha k^{\alpha\beta\mathtt{T}} \end{pmatrix} \tag{7.12}$$

This system has a two dimensional null space, and its general solution can be expressed as

$$\boldsymbol{x}(c_1, c_2) = \begin{pmatrix} \frac{1}{4}\sum_\gamma k^{\mathtt{A}\beta\gamma} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} - c_2 - c_1 \\ \frac{1}{4}\sum_\gamma k^{\mathtt{G}\beta\gamma} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} - c_2 - c_1 \\ \frac{1}{4}\sum_\gamma k^{\mathtt{C}\beta\gamma} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} - c_2 - c_1 \\ \frac{1}{4}\sum_\gamma k^{\mathtt{T}\beta\gamma} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} - c_2 - c_1 \\ 2\,c_1 \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{A}} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} + c_2 - c_1 \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{G}} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} + c_2 - c_1 \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{C}} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} + c_2 - c_1 \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{T}} - \frac{1}{32}\sum_{\alpha\gamma} k^{\alpha\beta\gamma} + c_2 - c_1 \end{pmatrix}, \quad c_1, c_2 \in \mathbb{R}. \tag{7.13}$$

For each choice of $\beta$ we have 21 independent equations for entries of the symmetric $6 \times 6$ stiffness blocks, and 6 for weighted shape parameters. As there are two independent choices of $\beta$, e.g. $\beta = \mathtt{A}$ and $\beta = \mathtt{G}$, we have in total a rather high dimensional null space, of dimension 84 for the stiffness matrix parameters and of dimension 24 for the weighted shape parameters.

This means that if $\{\mathsf{K}_1^\alpha, \mathsf{K}_2^{\alpha\beta}, \sigma_1^\alpha, \sigma_2^{\alpha\beta}\}$, $\alpha, \beta \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$ is a parameter set, obtained as a least-squares solution to (7.9), ignoring the ends of oligomers, then a set $\{\widetilde{\mathsf{K}}_1^\alpha, \widetilde{\mathsf{K}}_2^{\alpha\beta}, \widetilde{\sigma}_1^\alpha, \widetilde{\sigma}_2^{\alpha\beta}\}$ is also a solution, where

$$\begin{aligned} \widetilde{\mathsf{K}}_1^\alpha &= \mathsf{K}_1^\alpha + 2\Delta^\alpha, & \widetilde{\mathsf{K}}_2^{\alpha\beta} &= \mathsf{K}_2^{\alpha\beta} + \Lambda^{\alpha,11} + \Lambda^{\beta,33}, \\ \widetilde{\sigma}_1^\alpha &= \sigma_1^\alpha + 2\delta^\alpha, & \widetilde{\sigma}_2^{\alpha\beta} &= \sigma_2^{\alpha\beta} + \lambda^{\alpha,1} + \lambda^{\beta,3}, \end{aligned} \tag{7.14}$$

with

$$\Lambda^{\alpha,11} = \begin{pmatrix} \Lambda^\alpha - \Delta^\alpha & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Lambda^{\alpha,33} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\Lambda^\alpha - \Delta^\alpha \end{pmatrix}, \quad \Lambda^\alpha = (\Lambda^\alpha)^T,\ \Delta^\alpha = (\Delta^\alpha)^T \in \mathbb{R}^{6\times 6},$$

$$\lambda^{\alpha,1} = \begin{pmatrix} \lambda^\alpha - \delta^\alpha \\ 0 \\ 0 \end{pmatrix}, \quad \lambda^{\alpha,3} = \begin{pmatrix} 0 \\ 0 \\ -\lambda^\alpha - \delta^\alpha \end{pmatrix}, \quad \lambda^\alpha, \delta^\alpha \in \mathbb{R}^6, \quad \text{for} \quad \alpha \in \{\mathtt{A}, \mathtt{G}\}, \quad \text{and}$$

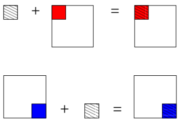
$$\Lambda^{\alpha,11} = \mathsf{E}_2\, \Lambda^{\bar\alpha,33}\, \mathsf{E}_2, \quad \Lambda^{\alpha,33} = \mathsf{E}_2\, \Lambda^{\bar\alpha,11}\, \mathsf{E}_2, \quad \Delta^\alpha = \mathsf{E}_1\, \Delta^{\bar\alpha}\, \mathsf{E}_1,$$
$$\lambda^{\alpha,1} = \mathsf{E}_2\, \lambda^{\bar\alpha,3}, \quad \lambda^{\alpha,3} = \mathsf{E}_2\, \lambda^{\bar\alpha,1}, \quad \delta^\alpha = \mathsf{E}_1\, \delta^{\bar\alpha}, \quad \text{for} \quad \alpha \in \{\mathtt{T}, \mathtt{C}\}.$$

Here $\bar\alpha$ denotes the complementary base to $\alpha$ and the matrix $\mathsf{E}_n$ was previously defined by (2.38) in Section 2.5.

Having obtained a general form of the least-squares solution, a search in the null space for choosing the best solution for our model parameter set can be done. Alternatively, one can consider a system with additional equations, leading to a unique solution.

## 7.5 Weighted least-squares system with end data

To eliminate the freedom in the null space of (7.12), we extended the original ABC set of oligomers $\mu = 1, \ldots, 36$ in the training set to include oligomer sequences with all possible dimer steps at the ends. There are then 72 instances of both types of GC ends, but only one instance for most of the other possible dimer ends in our training set. Then to our overdetermined linear system (7.11) we added the equations



$$\text{i.e} \quad \mathsf{K}_1^\beta + \mathsf{K}_{2,11}^{\alpha\beta} = \mathsf{K}^{5'\alpha\beta} \quad \text{and}$$
$$\text{i.e} \quad \mathsf{K}_{2,33}^{\alpha\beta} + \mathsf{K}_1^\beta = \mathsf{K}^{\alpha\beta 3'} \qquad (7.15)$$

for the stiffness blocks and, similarly, $\sigma_1^\beta + \sigma_{2,1}^{\beta\gamma} = \sigma^{5'\beta\gamma}$ and $\sigma_{2,3}^{\alpha\beta} + \sigma_1^\beta = \sigma^{\alpha\beta 3'}$ for the weighted shape vectors.

Again first consider the stiffness equations, and let $k^{5'\beta\gamma} = [\mathsf{K}^{5'\beta\gamma}]_{ij}$ and $k^{\alpha\beta 3'} = [\mathsf{K}^{\alpha\beta 3'}]_{ij}$ for some fixed $i$ and $j$, $1 \le i, j \le 6$. Similarly, in the weighted shape equations we would denote $k^{5'\beta\gamma} = [\sigma^{5'\beta\gamma}]_i$ and $k^{\alpha\beta 3'} = [\sigma^{\alpha\beta 3'}]_i$ for some fixed $i \le 6$.

Because we had much less data for the ends than for the interior of the oligomers, we assigned a unit weight to all interior $6 \times 6$ and $6 \times 1$ blocks, and a small, variable weight to the end-blocks corresponding to both leading and trailing ends of each sequence. The resulting weighted linear system $\bar A W \boldsymbol{y} = W \boldsymbol{b}$, with $\bar A \in \mathbb{R}^{24\times 9}$, $W = \mathrm{diag}(1, \ldots, 1, \epsilon, \ldots, \epsilon) \in \mathbb{R}^{24\times 24}$, and $\boldsymbol{y} \in \mathbb{R}^9$, can be explicitly written as

$$
\begin{pmatrix}
1&0&0&0&1&1&0&0&0\\
0&1&0&0&1&1&0&0&0\\
0&0&1&0&1&1&0&0&0\\
0&0&0&1&1&1&0&0&0\\
1&0&0&0&1&0&1&0&0\\
0&1&0&0&1&0&1&0&0\\
0&0&1&0&1&0&1&0&0\\
0&0&0&1&1&0&1&0&0\\
1&0&0&0&1&0&0&1&0\\
0&1&0&0&1&0&0&1&0\\
0&0&1&0&1&0&0&1&0\\
0&0&0&1&1&0&0&1&0\\
1&0&0&0&1&0&0&0&1\\
0&1&0&0&1&0&0&0&1\\
0&0&1&0&1&0&0&0&1\\
0&0&0&1&1&0&0&0&1\\
0&0&0&0&\epsilon&\epsilon&0&0&0\\
0&0&0&0&\epsilon&0&\epsilon&0&0\\
0&0&0&0&\epsilon&0&0&\epsilon&0\\
0&0&0&0&\epsilon&0&0&0&\epsilon\\
\epsilon&0&0&0&\epsilon&0&0&0&0\\
0&\epsilon&0&0&\epsilon&0&0&0&0\\
0&0&\epsilon&0&\epsilon&0&0&0&0\\
0&0&0&\epsilon&\epsilon&0&0&0&0
\end{pmatrix}
\begin{pmatrix}
k_{33}^{\mathtt{A}\beta}\\
k_{33}^{\mathtt{G}\beta}\\
k_{33}^{\mathtt{C}\beta}\\
k_{33}^{\mathtt{T}\beta}\\
k^{\beta}\\
k_{11}^{\beta\mathtt{A}}\\
k_{11}^{\beta\mathtt{G}}\\
k_{11}^{\beta\mathtt{C}}\\
k_{11}^{\beta\mathtt{T}}
\end{pmatrix}
=
\begin{pmatrix}
k^{\mathtt{A}\beta\mathtt{A}}\\
k^{\mathtt{G}\beta\mathtt{A}}\\
\vdots\\
\vdots\\
\vdots\\
k^{\mathtt{T}\beta\mathtt{T}}\\
\epsilon\, k^{5'\beta\mathtt{A}}\\
\vdots\\
\epsilon\, k^{5'\beta\mathtt{T}}\\
\epsilon\, k^{\mathtt{A}\beta 3'}\\
\vdots\\
\epsilon\, k^{\mathtt{T}\beta 3'}
\end{pmatrix}, \tag{7.16}
$$

and the corresponding weighted least-squares system $\bar{A}^T\, W\, \bar{A}\boldsymbol{y} = \bar{A}^T W \boldsymbol{b}$ is

$$
\begin{pmatrix}
s&0&0&0&s&1&1&1&1\\
0&s&0&0&s&1&1&1&1\\
0&0&s&0&s&1&1&1&1\\
0&0&0&s&s&1&1&1&1\\
s&s&s&s&8s-16&s&s&s&s\\
1&1&1&1&s&s&0&0&0\\
1&1&1&1&s&0&s&0&0\\
1&1&1&1&s&0&0&s&0\\
1&1&1&1&s&0&0&0&s
\end{pmatrix}
\begin{pmatrix}
k_{33}^{\mathtt{A}\beta}\\
k_{33}^{\mathtt{G}\beta}\\
k_{33}^{\mathtt{C}\beta}\\
k_{33}^{\mathtt{T}\beta}\\
k^{\beta}\\
k_{11}^{\beta\mathtt{A}}\\
k_{11}^{\beta\mathtt{G}}\\
k_{11}^{\beta\mathtt{C}}\\
k_{11}^{\beta\mathtt{T}}
\end{pmatrix}
=
\begin{pmatrix}
\epsilon\, k^{\mathtt{A}\beta 3'} + \sum_{\gamma} k^{\mathtt{A}\beta\gamma}\\
\epsilon\, k^{\mathtt{G}\beta 3'} + \sum_{\gamma} k^{\mathtt{G}\beta\gamma}\\
\epsilon\, k^{\mathtt{C}\beta 3'} + \sum_{\gamma} k^{\mathtt{C}\beta\gamma}\\
\epsilon\, k^{\mathtt{T}\beta 3'} + \sum_{\gamma} k^{\mathtt{T}\beta\gamma}\\
\epsilon \sum_{\alpha} k^{\alpha\beta 3'} + \epsilon \sum_{\gamma} k^{5'\beta\gamma} + \sum_{\alpha,\gamma} k^{\alpha\beta\gamma}\\
\epsilon\, k^{5'\beta\mathtt{A}} + \sum_{\alpha} k^{\alpha\beta\mathtt{A}}\\
\epsilon\, k^{5'\beta\mathtt{G}} + \sum_{\alpha} k^{\alpha\beta\mathtt{G}}\\
\epsilon\, k^{5'\beta\mathtt{C}} + \sum_{\alpha} k^{\alpha\beta\mathtt{C}}\\
\epsilon\, k^{5'\beta\mathtt{T}} + \sum_{\alpha} k^{\alpha\beta\mathtt{T}}
\end{pmatrix}, \tag{7.17}
$$

with $s = 4 + \epsilon$. The solution of this system is unique for any nonzero choice of end weighting $\epsilon$, and has the closed form

$$\boldsymbol{y}(\epsilon) = \frac{1}{4+\epsilon}\begin{pmatrix} \epsilon k^{\mathtt{A}\beta 3'} + \sum_\gamma k^{\mathtt{A}\beta\gamma} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{4+\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \\ \epsilon k^{\mathtt{G}\beta 3'} + \sum_\gamma k^{\mathtt{G}\beta\gamma} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{4+\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \\ \epsilon k^{\mathtt{C}\beta 3'} + \sum_\gamma k^{\mathtt{C}\beta\gamma} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{4+\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \\ \epsilon k^{\mathtt{T}\beta 3'} + \sum_\gamma k^{\mathtt{T}\beta\gamma} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{4+\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \\ (4+\epsilon)\left(\frac{1}{4}\sum_\alpha k^{\alpha\beta 3'} + \frac{1}{4}\sum_\gamma k^{5'\beta\gamma} - \frac{1}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma}\right) \\ \epsilon k^{5'\beta\mathtt{A}} + \sum_\alpha k^{\alpha\beta\mathtt{A}} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{4+\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \\ \epsilon k^{5'\beta\mathtt{G}} + \sum_\alpha k^{\alpha\beta\mathtt{A}} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{4+\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \\ \epsilon k^{5'\beta\mathtt{C}} + \sum_\alpha k^{\alpha\beta\mathtt{A}} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{4+\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \\ \epsilon k^{5'\beta\mathtt{T}} + \sum_\alpha k^{\alpha\beta\mathtt{A}} + \frac{\epsilon}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} - \frac{4+\epsilon}{4}\sum_\alpha k^{\alpha\beta 3'} - \frac{\epsilon}{4}\sum_\gamma k^{5'\beta\gamma} \end{pmatrix}. \tag{7.18}$$

Again, as we had comparatively little end data available, and the ultimate objective was a model of DNA away from ends, we decided to take as our least squares solution the limit of vanishing weight $\epsilon$ for the end data. This limiting vanishing weight solution has the rather simple expression:

$$\boldsymbol{y}(\epsilon)\big|_{\epsilon=0} = \begin{pmatrix} \frac{1}{4}\sum_\gamma k^{\mathtt{A}\beta\gamma} - \frac{1}{4}\sum_\gamma k^{5'\beta\gamma} \\ \frac{1}{4}\sum_\gamma k^{\mathtt{G}\beta\gamma} - \frac{1}{4}\sum_\gamma k^{5'\beta\gamma} \\ \frac{1}{4}\sum_\gamma k^{\mathtt{C}\beta\gamma} - \frac{1}{4}\sum_\gamma k^{5'\beta\gamma} \\ \frac{1}{4}\sum_\gamma k^{\mathtt{T}\beta\gamma} - \frac{1}{4}\sum_\gamma k^{5'\beta\gamma} \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta 3'} + \frac{1}{4}\sum_\gamma k^{5'\beta\gamma} - \frac{1}{16}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma} \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{A}} - \frac{1}{4}\sum_\alpha k^{\alpha\beta 3'} \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{G}} - \frac{1}{4}\sum_\alpha k^{\alpha\beta 3'} \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{C}} - \frac{1}{4}\sum_\alpha k^{\alpha\beta 3'} \\ \frac{1}{4}\sum_\alpha k^{\alpha\beta\mathtt{T}} - \frac{1}{4}\sum_\alpha k^{\alpha\beta 3'} \end{pmatrix}. \tag{7.19}$$

Moreover, (7.19) is also a particular solution of the least squares system (7.12) that does not use the end data, i.e. $\boldsymbol{y}(\epsilon)\big|_{\epsilon=0} = \boldsymbol{x}(c_1, c_2)$ with $c_1 = \frac{1}{8}\sum_\alpha k^{\alpha\beta 3'} + \frac{1}{8}\sum_\gamma k^{5'\beta\gamma} - \frac{1}{32}\sum_{\alpha,\gamma} k^{\alpha\beta\gamma}$ and $c_2 = \frac{1}{8}\sum_\alpha k^{\alpha\beta 3'} + \frac{1}{8}\sum_\gamma k^{5'\beta\gamma}$.

The least squares procedure described above leads to a parameter set that provides reasonably good reconstructions, at least away from the ends of the oligomers. However it takes no account of the constraint that the stiffness parameter matrices should be at least positive semi definite. Somewhat to our surprise all of the least squares solutions that we found had some indefinite stiffness parameter matrices with a small number of slightly negative eigenvalues. In particular even with the 84 dimensional freedom in the null space of the normal equations for the system using only interior data, we were unable to find a single parameter set that had all stiffness parameter matrices positive semi-definite. What we could do was to construct an ad hoc, low rank, perturbation scheme that numerically raised the negative eigenvalues to yield a set of positive definite stiffness parameter matrices $\mathsf{K}_1^\alpha$ and $\mathsf{K}_2^{\alpha\beta}$.

| | $\boldsymbol{x}(c_1, c_2)\vert_{c_1=c_2=0}$ | | | $\boldsymbol{y}(\epsilon)\vert_{\epsilon=0}$ | | | $\mathcal{P}^\dagger$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Block | $n_{\lambda<0}$ | $\min(\lambda)$ | $\max(\lambda)$ | $n_{\lambda<0}$ | $\min(\lambda)$ | $\max(\lambda)$ | $n_{\lambda<0}$ | $\min(\lambda)$ | $\max(\lambda)$ |
| AT | 1 | -1.22 | 103.67 | 0 | 0.92 | 106.00 | 0 | 0.00 | 105.17 |
| GC | 1 | -1.02 | 106.91 | 0 | 1.09 | 109.46 | 0 | 0.06 | 107.83 |
| TA | 1 | -0.59 | 74.66 | 0 | 0.51 | 77.78 | 0 | 0.00 | 77.33 |
| CG | 1 | -0.02 | 83.92 | 0 | 0.11 | 85.17 | 0 | 0.00 | 84.90 |
| GT | 1 | -1.13 | 101.95 | 0 | 1.03 | 104.47 | 0 | 0.02 | 103.91 |
| TG | 1 | -0.21 | 83.01 | 0 | 0.34 | 85.08 | 0 | 0.03 | 84.76 |
| AG | 1 | -1.10 | 93.05 | 0 | 0.42 | 95.96 | 0 | 0.00 | 95.59 |
| GA | 1 | -0.89 | 94.47 | 0 | 0.95 | 97.51 | 0 | 0.00 | 97.00 |
| AA | 1 | -0.90 | 97.92 | 0 | 0.98 | 101.08 | 0 | 0.07 | 100.82 |
| GG | 1 | -0.71 | 98.04 | 0 | 0.47 | 100.90 | 0 | 0.00 | 100.02 |
| A | 0 | 0 | 0 | 4 | -10.76 | 16.61 | 0 | 0.00 | 15.29 |
| G | 0 | 0 | 0 | 3 | -9.88 | 37.18 | 0 | 0.00 | 38.53 |

Table 7.1: Eigenvalues of the dimer based model parameter blocks, that are least squares solutions of the form $\boldsymbol{x}(c_1, c_2)\big|_{c_1=c_2=0}$, $\boldsymbol{y}(\epsilon)\big|_{\epsilon=0}$ and $\mathcal{P}^\dagger$. $n_{\lambda<0}$ denotes the number of negative eigenvalues in that block.

The smallest and largest eigenvalues of the ten independent dimer stiffness blocks $\mathsf{K}_2^{\alpha\beta}$ and two independent monomer blocks $\mathsf{K}_1^\gamma$ for three parameter sets are listed in Table 7.1. The dimers $\alpha\beta$ and the monomers $\gamma$ are given in the first column. The first two parameter sets in the table are obtained as least squares solutions $\boldsymbol{x}(c_1, c_2)\big|_{c_1=c_2=0}$ given in (7.13) and $\boldsymbol{y}(\epsilon)\big|_{\epsilon=0}$ 7.19 for each independent entry of the block. The first parameter set has 10 negative eigenvalues in total, and the second one has 7 negative eigenvalues. The third parameter set $\mathcal{P}^\dagger$, described in Section 7.6 and further in Chapter 8, is obtained by a numerical eigenvalue perturbation and optimisation procedure from the second parameter set in the table.

Finally, for the reasons that are described next, this least squares solution is only taken as an intermediate step before a numerical nonlinear fitting procedure, so that we do not believe the particular choice that is made to be crucial.

## 7.6 Dimer-based fitting

In the dimer-based nearest neighbour rigid base model, each sequence $\mathsf{S}_\mu$ is described by a model probability density

$$\rho_{\mu,\mathrm{m}}^{\mathcal{P}}(\mathsf{w}) = \frac{1}{Z_{\mu,\mathrm{m}}} e^{-(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{m}})\cdot\mathsf{K}_{\mu,\mathrm{m}}(\mathsf{w}-\widehat{\mathsf{w}}_{\mu,\mathrm{m}})/2}. \tag{7.20}$$

where, as described in Section 7.2, the stiffness matrix $\mathsf{K}_{\mu,\mathrm{m}}$ and shape vector $\widehat{\mathsf{w}}_{\mu,\mathrm{m}}$ can be reconstructed from a finite parameter set $\mathcal{P} = \{\sigma_1^\alpha, \mathsf{K}_1^\alpha, \sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$. By a best-fit parameter set for the collection of training set sequences $\mathsf{S}_\mu$ we mean a set $\mathcal{P}^*$ satisfying

$$\mathcal{P}^* = \underset{\mathcal{P}}{\operatorname{argmin}} \sum_{\mu=1}^{N} D(\rho_{\mu,\mathrm{m}}^{\mathcal{P}}, \rho_{\mu,\mathrm{M}}^*). \tag{7.21}$$

That is, among all dimer-dependent parameter sets, the best-fit parameters $\mathcal{P}^*$ should minimise the sum of the divergences between the Gaussian probability distributions with the dimer based reconstructed parameters and the best oligomer based nearest neighbour fit to the training set probability distributions. Presumably, if the collection of the training sequences is sufficiently rich, this best-fit parameter set $\mathcal{P}^*$ should not only provide good model reconstructions for the training set sequences, but also for oligomers of arbitrary length and sequence.

We used the modified least squares parameter set as an admissible initialisation for a numerical minimisation of the nonlinear Kullback-Leibler objective functional (7.21). We were thereby able to find a numerical approximation to a minimum using a numerical gradient flow algorithm. Our computations suggest that this optimisation problem is a rather delicate one that is worthy of further study. Nevertheless in Section 8.1 below we describe one numerical approximation $\mathcal{P}^\dagger$ to a best-fit parameter set, which seems typical of the optimal approximations found thus far. In the numerical scheme that we constructed the explicit constraint that the individual stiffness parameter matrices are all semi-definite was enforced by factoring each stiffness matrix in the form $K = A^2$ with the, possibly indefinite, symmetric matrices $A$ being adopted as the basic unknowns. The palindromic symmetries (7.2) of self-symmetric dimers were enforced explicitly.

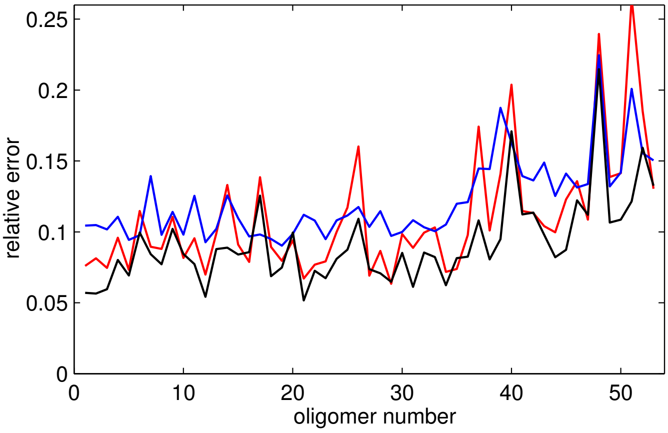


Figure 7.3: Relative errors in dimer-dependent reconstructions. Red: the divergence $D(\rho^*_{\mu,\mathrm{m}}, \rho^*_{\mu,\mathrm{M}})/D_{\mathrm{o}}$; blue, relative error of stiffness matrices in the Frobenius norm $||\mathsf{K}^*_{\mu,\mathrm{m}} - \mathsf{K}^*_{\mu,\mathrm{M}}||/||\mathsf{K}^*_{\mu,\mathrm{M}}||$; black, relative error of shape vectors in Euclidean norm $|\widehat{\mathsf{w}}^*_{\mu,\mathrm{m}} - \widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}|/|\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}|$, all versus training set index $\mu$. Here $\rho^*_{\mu,\mathrm{m}}$ denotes a reconstructed density and $\rho^*_{\mu,\mathrm{M}}$ denotes a nearest-neighbour best-fit oligomer density.

The modelling errors incurred in approximating an oligomer-based model density $\rho^*_{\mu,\mathrm{M}}$ by a dimer-based model density $\rho^*_{\mu,\mathrm{m}}$ constructed with our best-fit parameter set $\mathcal{P}^\dagger$ can now be quantified for each sequence $\mathsf{S}_\mu$, and are shown in Figure 7.3. Here the modelling error for each sequence is due both to the difference between the oligomer stiffness matrices $\mathsf{K}^*_{\mu,\mathrm{M}}$ and $\mathsf{K}^*_{\mu,\mathrm{m}}$, and the oligomer shape vectors $\widehat{\mathsf{w}}^*_{\mu,\mathrm{M}}$ and $\widehat{\mathsf{w}}^*_{\mu,\mathrm{m}}$. Both of these differences arise due to the finiteness of the parameter set $\mathcal{P}^\dagger$. In physical terms, this modelling error reflects the difference between a quadratic, nearest-neighbour internal energy model in which the parameters of the model are allowed to depend on the sequence composition of the entire oligomer, and one in which the parameters depend only on the composition of the local dimer. As can be seen, the relative errors between the stiffness matrices, shape vectors and probability density functions are all in the range $5-15\%$ on the scale of $D_{\mathrm{o}}$ for indices $\mu \leq 36$, with slightly higher errors for higher indices. One possible explanation for these higher errors is that for $\mu \geq 37$ the oligomers all have non `GC` dimer ends, which are much less represented in the training set data. Nevertheless our conclusion is that this particular parameter set $\mathcal{P}^{\dagger}$ of the dimer-based nearest neighbour rigid base model is able to well resolve sequence variations across the training set. Particular examples of reconstructions are discussed in more detail in Sections 8.2 and 8.3 below.

# Chapter 8

# Parameters and tests for the dimer dependent nearest-neighbour rigid base model

In this chapter we present a full parameter set for the dimer dependent nearest-neighbour model. We give examples of stiffness and shape parameter reconstructions for sequences from the training data set and also show that we are able to predict non local changes in shapes caused by local changes in sequence.

## 8.1 The $\mathcal{P}^{\dagger}$ parameter set

In this section we make a visual presentation of some properties of the first optimal fit parameter set $\mathcal{P}^{\dagger}$. Figure 8.1 provides colour plots of the 1-mer stiffness matrices and weighted shape parameters. To be consistent with the analogous 2-mer parameters we plot the averages and standard deviations over all four possible bases occurring on the reference strand, along with differences $\Delta$`A` and $\Delta$`G` to two independent cases. As was to be expected there are marked differences between the `A` and `G` parameters.

The 2-mer parameters have more interesting structure. In Figure 8.2 the sequence averaged weighted shape and stiffness matrices are presented, along with their standard deviations, over all sixteen possible steps. Then in the first column the difference between the values for three independent purine-pyrimidine steps and the average are given, and analogously three sets of pyrimidine-purine parameters in the second column, and the four purine-purine differences in the third column. While there is some variation within columns, as there should be, there are striking patterns that are common within, and distinct between, each column. This observation suggests that the parameter fitting is successfully capturing sequence dependent effects of differing stacking interactions.

Figure 8.3 presents the logarithm of the (positive) eigenvalues of two independent 1-mer parameter stiffness matrices and ten independent 2-mer parameter stiffness matrices. Several eigenvalues are of order $10^{-6}$ or less. Depending on the desired level of error we observe that the 1-mer parameter matrices could be approximated to be of rank only 2, 3 or 4. Similarly several of the 2-mer parameter stiffness matrices could be approximated to not be of full rank.

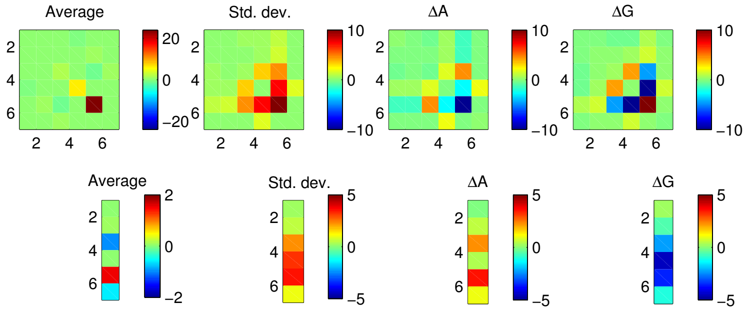


Figure 8.1: Averages and standard deviations over all four bases of 1-mer stiffness and weighted shape parameters along with differences $\Delta\mathtt{A}$ and $\Delta\mathtt{G}$ to two independent cases. Non dimensional units as described in Section 2.4 of Chapter 2. The indices $1, \ldots, 6$ correspond respectively to Buckle, Propeller, Opening, Shear, Stretch, Stagger.

The eigenvalues presented in Figure 8.3 are non dimensional, but in dimension full terms they have units $\mathring{A}^2 k_B T$ so that the eigenvalues are truly small. (Because the eigenvectors involve both translational and rotational degrees of freedom, this single unit for an eigenvalue depends on our scaling between translations and rotations in which an Å is taken to be equivalent to 1/5 radian, or approximately 11 degrees.) Interestingly when the dimer stiffness matrices are constructed from the 2-mer parameter stiffness matrices by adding 1-mer stiffness blocks, the resulting stiffness matrices, without exception, have a smallest eigenvalue of order $10^{-1}$. In other words the soft modes of the 2-mer parameter stiffnesses are substantially stiffened in the dimer stiffness matrices by the addition of the 1-mer $6 \times 6$ stiffness blocks, despite them each having four rather small eigenvalues.

The Euclidean averages of the parameter set illustrated in Figures 8.1 and 8.2 provide one way to construct a homogeneous model of DNA in which the occurrence of every sequence is assumed to be equally likely. The reconstructed oligomer shape vector $\widehat{\mathrm{w}}^*_{\mathrm{h,m}}$ for a homogeneous oligomer approaches the constant values shown in Table 8.1, where the values stabilise to the three significant digits shown only six base pairs away from the ends, with the values stabilising to one digit two base pairs from the ends. The “odd" parameters Buckle, Shear, Tilt and Shift all vanish exactly, because with the averaging procedure we adopt the homogeneous model is palindromic. A comparison with the sequence averaged MD values presented in Table 1 of [42] are also given. Both sets of values share a common origin in the underlying ABC MD data base [42], but the averaging procedures utilised to construct the homogeneous parameters are quite different. The two sets of sequence independent B-form parameters are rather close, so that in particular we can conclude that our coarse-grain parameter set is consistent with accepted B-form sequence-averaged values to within an acceptable tolerance. Interestingly our value of homogeneous Twist is slightly higher than that extracted directly from the MD data. In fact this Twist is of the base pair frame, so that its precise value is dependent upon the chosen embedding of the frame in the base pair; the partition between Roll and Twist could be altered by choosing a slightly modified frame (the partition between

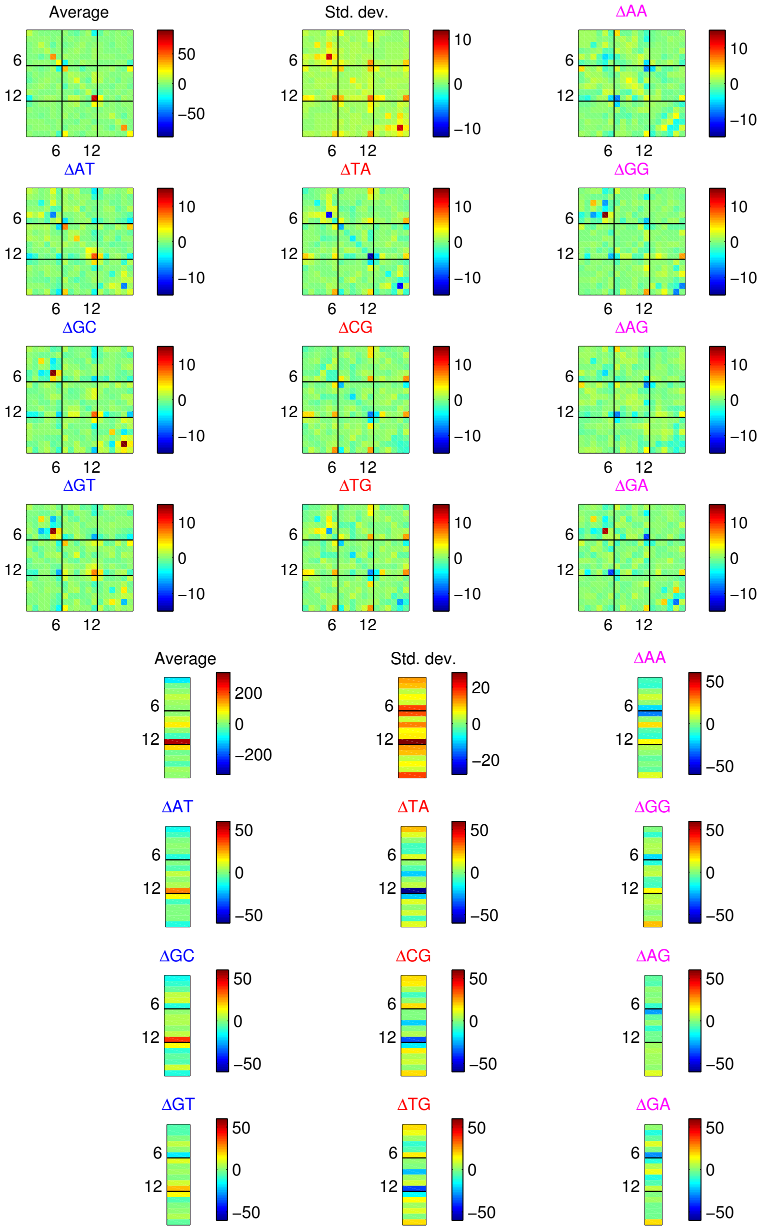


Figure 8.2: 2-mer weighted shape and stiffness parameters. Averages and standard deviations over the 16 possible dimer steps, then ten independent differences $\Delta$`AT` etc from the average with purine-pyrimidine, pyrimidine-purine, and purine-purine dimer steps in respectively columns one, two and three. While there are differences within columns as there should be, there are patterns common to each column that are quite distinct between columns. Non dimensional units as described in Section 2.4 of Chapter 2. The intra indices $1, \ldots, 6$ and $13, \ldots, 18$ are ordered as in Figure 8.1 with inter indices $7, \ldots, 12$ being in order Tilt, Roll, Twist, Shift, Slide Rise.

Slide and Rise would then also be altered). In a homogeneous model a Twist and Rise per base pair of the associated helical structure, which is independent of the precise choice of base pair framing, can be computed in the following way. For a homogeneous configuration the base pair frames lie on a circular helix, where, because we have adopted a Cayley vector parametrisation of rotation matrices, the coordinates in the base pair frame of the central axis of the helix are precisely the Tilt, Roll and Twist, so that the twist per base pair about the helix centreline is the norm of the Cayley vector, or, for the homogeneous parameters given in Table 8.1, 33.10 degrees. This is equivalent to a helical repeat of $360/33.10 = 10.9$ base pairs, which is within 5% of the experimentally reported value of $10.4 \pm 0.1$ [75]. The pitch or rise of the helical structure can be computed from the scalar product of the (Shift, Slide, Rise) vector with a unit vector parallel to the Cayley vector which yields 3.25Å per base pair, or 35.4Å per period. The radius of the circular helix on which the origins of the base pair frames lie is 1.49Å.

| | Average coarse-grain, $\frac{1}{5}$ rad/Å | Average coarse-grain, °/Å | Average MD °/Å |
|---|---|---|---|
| Buckle | 0 | 0 | 1.2 |
| Propeller | -1.089 | -12.43 | -11.0 |
| Opening | 0.109 | 1.24 | 2.1 |
| Shear | 0 | 0 | 0.02 |
| Stretch | 0.022 | 0.02 | 0.03 |
| Stagger | 0.168 | 0.17 | 0.09 |
| Tilt | 0 | 0 | -0.3 |
| Roll | 0.258 | 2.87 | 3.6 |
| Twist | 2.961 | 33.98 | 32.6 |
| Shift | 0 | 0 | -0.05 |
| Slide | -0.562 | -0.56 | -0.44 |
| Rise | 3.314 | 3.31 | 3.32 |

Table 8.1: Homogeneous, sequence averaged, nearest-neighbour rigid base DNA shape parameters compared to analogous sequence-averaged MD parameters. Column 1, coarse-grain values in the units of this work, Column 2 the same values in degrees and Å, which allows direct comparison with the MD average values shown in Column 3 and taken from Table 1 of [42].

To prepare for our discussion of sequence dependence, and to quantify the depth of penetration of end effects, properties of the reconstructed parameters for a homogeneous 18-mer can be plotted as a function of position along the oligomer. Figure 8.4 has eight panels. The top four panels plot the reconstructed shape parameter vector $\widehat{\mathrm{w}}^{*}_{\mathrm{h,m}}$ versus oligomer position, with discrete point values visualised using linear interpolation. Each of the four panels contains plots of three of the twelve internal variables, grouped by inter and intra and translation and rotation. All quantities use the non dimensional units introduced in Section 2.4 of Chapter 2, although the numerical scale on the ordinate is varied and indicated panel by panel to suit the scale of the pertinent data. Despite the sequence being homogeneous, the effect of ends means that the expected values vary from those given in Table 8.1 as the ends are approached, with the effects still being visible on this scale four base pairs from the end. The palindromic symmetry of the homogeneous model is evident, with oddness of Buckle, Shear, Tilt and Shift (all plotted in black), and evenness of the remaining variables. The values of the plotted shape parameters are also given in Table 8.2. The corresponding reconstructed stiffness matrix has many non zero entries, but the decay of end effects are captured in the diagonal entries, which are therefore plotted in the remaining four panels. The presentation is directly analogous to the shape parameters above, except that now the diagonal entries of the reconstructed stiffness matrix $\mathsf{K}^*_{\mathrm{h,m}}$ are plotted. In contrast to the nonlocal dependence of $\widehat{\mathsf{w}}^*_{\mathrm{h,m}}$, it is a consequence of our nearest-neighbour model that the only end effects in stiffnesses are localised precisely to the first and last intra $6 \times 6$ blocks. However these changes are significant, approaching 50%. Palindromy implies that all diagonal stiffnesses are even functions. The stiffness panels of Figure 8.4 are included both to emphasise the difference of end effects on shape and stiffness, and to provide a homogeneous reference point for the analogous sequence-dependent plots described in the next section.

| Buckle | Propeller | Opening | Shear | Stretch | Stagger | Tilt | Roll | Twist | Shift | Slide | Rise |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.20 | -0.97 | 0.03 | -0.00 | 0.03 | 0.13 | -0.02 | 0.26 | 2.98 | 0.03 | -0.58 | 3.35 |
| 0.09 | -1.12 | 0.11 | -0.00 | 0.02 | 0.16 | -0.01 | 0.26 | 2.98 | 0.00 | -0.57 | 3.33 |
| 0.02 | -1.11 | 0.11 | -0.00 | 0.02 | 0.17 | -0.00 | 0.26 | 2.96 | 0.00 | -0.56 | 3.32 |
| -0.00 | -1.09 | 0.11 | 0.00 | 0.02 | 0.17 | -0.00 | 0.26 | 2.96 | 0.00 | -0.56 | 3.31 |
| -0.00 | -1.09 | 0.11 | 0.00 | 0.02 | 0.17 | 0.00 | 0.26 | 2.96 | 0.00 | -0.56 | 3.31 |
| -0.00 | -1.09 | 0.11 | 0.00 | 0.02 | 0.17 | 0.00 | 0.26 | 2.96 | -0.00 | -0.56 | 3.31 |
| -0.00 | -1.09 | 0.11 | 0.00 | 0.02 | 0.17 | 0.00 | 0.26 | 2.96 | -0.00 | -0.56 | 3.31 |
| 0.00 | -1.09 | 0.11 | -0.00 | 0.02 | 0.17 | -0.00 | 0.26 | 2.96 | -0.00 | -0.56 | 3.31 |
| 0.00 | -1.09 | 0.11 | -0.00 | 0.02 | 0.17 | 0.00 | 0.26 | 2.96 | 0.00 | -0.56 | 3.31 |
| -0.00 | -1.09 | 0.11 | 0.00 | 0.02 | 0.17 | 0.00 | 0.26 | 2.96 | 0.00 | -0.56 | 3.31 |
| -0.00 | -1.09 | 0.11 | 0.00 | 0.02 | 0.17 | -0.00 | 0.26 | 2.96 | 0.00 | -0.56 | 3.31 |
| 0.00 | -1.09 | 0.11 | -0.00 | 0.02 | 0.17 | -0.00 | 0.26 | 2.96 | 0.00 | -0.56 | 3.31 |
| 0.00 | -1.09 | 0.11 | -0.00 | 0.02 | 0.17 | -0.00 | 0.26 | 2.96 | -0.00 | -0.56 | 3.31 |
| 0.00 | -1.09 | 0.11 | -0.00 | 0.02 | 0.17 | 0.00 | 0.26 | 2.96 | -0.00 | -0.56 | 3.31 |
| 0.00 | -1.09 | 0.11 | -0.00 | 0.02 | 0.17 | 0.00 | 0.26 | 2.96 | -0.00 | -0.56 | 3.32 |
| -0.02 | -1.11 | 0.11 | 0.00 | 0.02 | 0.17 | 0.01 | 0.26 | 2.98 | -0.00 | -0.57 | 3.33 |
| -0.09 | -1.12 | 0.11 | 0.00 | 0.02 | 0.16 | 0.02 | 0.26 | 2.98 | -0.03 | -0.58 | 3.35 |
| -0.20 | -0.97 | 0.03 | 0.00 | 0.03 | 0.13 |  |  |  |  |  |  |

Table 8.2: The elements of the reconstructed shape parameter vector $\widehat{\mathsf{w}}^*_{\mathrm{hom,m}}$ for a homogeneous 18-mer sequence, ordered along the sequence from top to bottom. Note that parameter values at the ends of the oligomer (top and bottom rows) are different from the ones in the middle, especially for the intra base pair parameters.

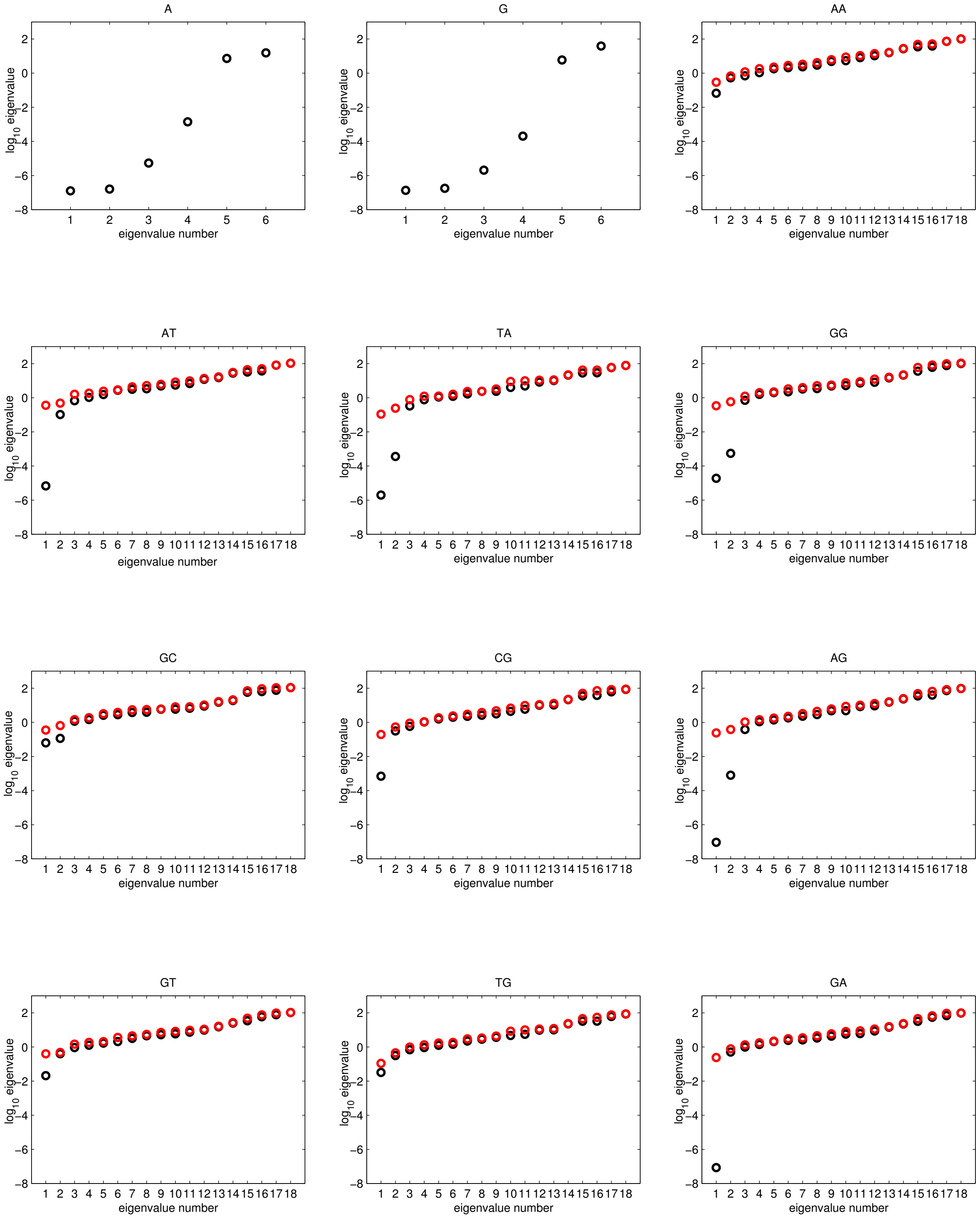


Figure 8.3: Logs of the eigenvalues of stiffness parameter matrices for two independent 1-mers, and of ten independent 2-mer steps (black circles). There are small eigenvalues of order $10^{-3}$ to $10^{-6}$ or less in several cases. However, when the 2-mer parameter stiffnesses are combined with 1-mer parameter stiffnesses, to construct dimer stiffness matrices, in all ten independent cases the smallest eigenvalues are of order $10^{-1}$ or larger (red circles).

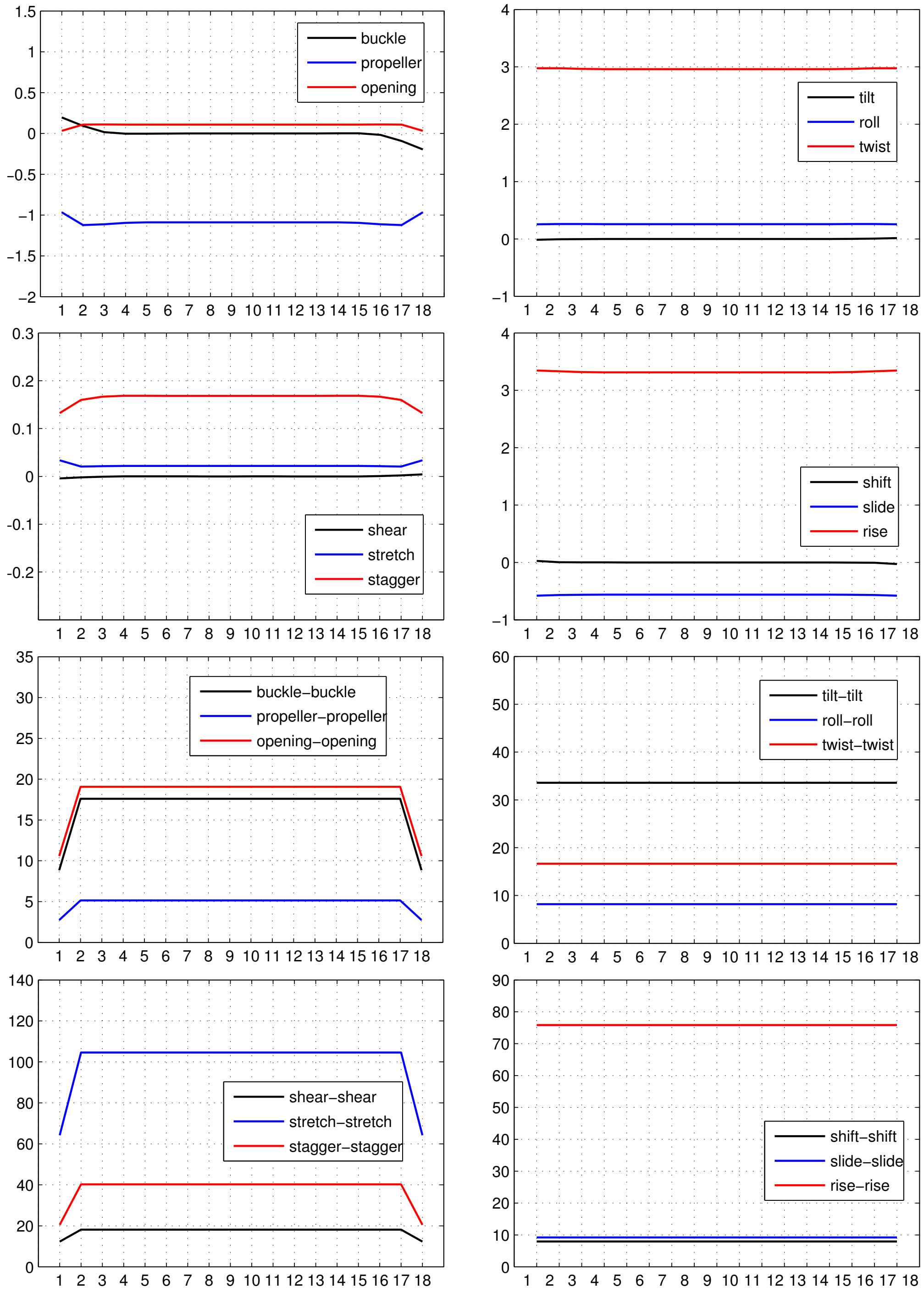


Figure 8.4: Reconstruction of model parameters for a homogeneous 18-mer. Top four panels: Plots of the reconstructed shape parameter vector $\widehat{\mathsf{w}}^*_{\mathrm{h,m}}$ along the sequence. Bottom four panels: Diagonal entries of the reconstructed stiffness parameter matrix $\mathsf{K}^*_{\mathrm{h,m}}$ along the sequence. Parameter values are interpolated by piecewise-linear curves. As a consequence of the reconstruction model, the intra base pair stiffnesses change only at the first and last base pair, while the inter base pair stiffnesses do not change at all. However the changes in oligomer shape parameters, both intra and inter, are not localised to the ends. The magnitude and decay rate of end effects are similar for both intra and inter parameters, but are more evident in the figure for the intra variables because of the differences in axes scale between panels.

## 8.2 Reconstructions of example oligomers from the training set

In this section we discuss in more detail the four training set sequences $\mathsf{S}_\mu$ with $\mu = 1, 3, 8, 42$ as detailed in Table 4.1. $\mathsf{S}_1$ is a palindrome with a period two sequence in the interior, $\mathsf{S}_3$ a palindrome with a period four sequence, and $\mathsf{S}_8$ a non-palindrome again of period four. All of these three sequences are 18-mers with $\mathtt{C}p\mathtt{G}$ end dimers. In contrast $\mathsf{S}_{42}$ is a 12-mer with $\mathtt{G}p\mathtt{G}$ ends. The diagonal blocks in Table 8.3 record two scaled reconstruction errors for each oligomer, namely the divergence $D(\rho_{i,\mathrm{M}}, \rho_{i,\mathrm{o}})/D_\mathrm{o}$ between the training set density $\rho_{i,\mathrm{o}}$ and the nearest-neighbour oligomer model $\rho_{i,\mathrm{M}}$, and the divergence $D(\rho_{i,\mathrm{m}}, \rho_{i,\mathrm{M}})/D_\mathrm{o}$ between the nearest-neighbour oligomer fit and the reconstructed nearest-neighbour dimer model $\rho_{i,\mathrm{m}}$. The off-diagonal blocks in Table 8.3 record the pairwise divergences $D(\rho_{i,\mathrm{o}}, \rho_{j,\mathrm{o}})/D_\mathrm{o}$ and $D(\rho_{i,\mathrm{m}}, \rho_{j,\mathrm{m}})/D_\mathrm{o}$ between pairs of training set and reconstructed probability density functions for each oligomer. It is apparent that the model reconstruction is sufficiently accurate to capture to a good approximation the variations due to differences in sequence in the overall sense of divergences between probability distributions for each oligomer.

| | $\mathsf{S}_1$ | $\mathsf{S}_3$ | $\mathsf{S}_8$ | $\mathsf{S}_{1'}$ |
|---|---|---|---|---|
| $\mathsf{S}_1$ | **0.087**<br>**0.076** | 0.676<br>0.619 | 0.567<br>0.407 | 0.116<br>0.053 |
| $\mathsf{S}_3$ | 0.616<br>0.489 | **0.095**<br>**0.075** | 0.541<br>0.297 | 0.621<br>0.466 |
| $\mathsf{S}_8$ | 0.423<br>0.310 | 0.422<br>0.314 | **0.091**<br>**0.088** | 0.418<br>0.286 |
| $\mathsf{S}_{1'}$ | 0.139<br>0.040 | 0.679<br>0.574 | 0.588<br>0.364 | **0.089**<br>**0.082** |

Table 8.3: A table of pairwise divergences, all scaled by the characteristic scale $D_\mathrm{o} \approx 85$, between observed and reconstructed probability densities for four 18-mer oligomers. $\mathsf{S}_1$, $\mathsf{S}_3$, and $\mathsf{S}_8$ are from the training set and are discussed in Section 8.2, while $\mathsf{S}_1'$ is a single nucleotide permutation of $\mathsf{S}_1$ as discussed in Section 8.3. In the off diagonal cells, the top entries are the divergences $D(\rho_{i,\mathrm{o}}, \rho_{j,\mathrm{o}})/D_\mathrm{o}$ between observed distributions, while the bottom entries are the divergences $D(\rho_{i,\mathrm{m}}, \rho_{j,\mathrm{m}})/D_\mathrm{o}$ between reconstructed distributions, where $i, j = 1, 3, 8 \,\mathrm{and}\, 1'$. The top entries in the diagonal cells are the model error $D(\rho_{i,\mathrm{M}}, \rho_{i,\mathrm{o}})/D_\mathrm{o}$, while the bottom entries are the model error $D(\rho_{i,\mathrm{m}}, \rho_{i,\mathrm{M}})/D_\mathrm{o}$.

We further discuss our reconstructions of the four oligomers in the context of a standard set of Figures prepared for each oligomer $\mathsf{S}_\mu$. The figures illustrate pointwise differences in the model parameters along the different oligomers. For $\mu = 1, 3, 8, 42$, Figures 8.5, 8.6, 8.7, and 8.8, are directly analogous to Figure 8.4, but now with information about the sequence dependence at various different levels of model approximation. The base sequence of the reference strand of each oligomer is indicated on the abscissa, with intra variables evaluated at each base pair and inter variables evaluated at each junction. Discrete point values are again visualised using linear interpolation, but now with different line styles for different model approximations to the same quantities. For each of the four sequences, the top four panels plot both the training set shape vector $\widehat{\mathsf{w}}_{\mu,\mathrm{o}}$ (which are the MD time series averages) in solid lines, and the reconstructed shape vector $\widehat{\mathsf{w}}_{\mu,\mathrm{m}}$, shown with dashed lines, each versus sequence position. The corresponding stiffness matrices have many non zero entries, as shown by the two examples of Figure 6.2, but much of the sequence dependence is captured by the diagonal entries, which are therefore plotted in the remaining four panels. The presentation is directly analogous to the shape parameters above, except now three plots of each diagonal stiffness is made corresponding to the diagonal entries of the training set stiffness matrix $\mathsf{K}_{\mu,\mathrm{o}}$ (solid), the best nearest-neighbour oligomer fit stiffness matrix $\mathsf{K}^*_{\mu,\mathrm{M}}$ (dash-dot), and the reconstructed nearest-neighbour, dimer dependent stiffness matrix $\mathsf{K}_{\mu,\mathrm{m}}$ (dashed).

Our conclusion is that the coarse-grain reconstructions are remarkably good. Frequently the different approximations are indistinguishable, and with very few exceptions the pointwise differences in the various model approximations are less than the variation with sequence. There is a tendency for the reconstruction errors to be larger at the ends, particularly for the oligomer $\mu = 42$ which may indicate a lack of sampling of `GG` end data in the training set. Visually the errors in the intra variables appear larger, but the scales in the intra and inter panels are of necessity different, although the units are identical. For both intra and inter shape parameters, and away from the ends, rather few errors are larger than 0.1 Å in translational variables or 2 degrees in rotational variables. All three model reconstructions clearly reflect the interior period two or four sequence dependence in oligomers $\mu = 1, 3, 8$.

By construction the nearest-neighbour, dimer dependent rigid base model exactly satisfies the palindromy symmetry present in oligomers $\mu = 1, 3$. The MD time series data expressed in the solid lines for the most part also closely satisfies the requisite palindromy symmetries. Note, however, that errors can arise from lack of convergence of the MD simulation of any given sequence, rather than the reconstruction. For example the breaking of evenness of the twist-twist stiffness plot of MD data shown in Figure 8.5 violates the palindromic symmetry for the $\mu = 1$ oligomer, and must reflect lack of convergence of the MD time series. Similarly the profile of the MD Stagger expectations shown in Figure 8.5 noticeably violates the evenness required by palindromy of the sequence.

Another way in which to judge quality of the training set data and of the reconstructions is to consider histograms, or $1D$ plots, of the marginal distributions of each of the 12 internal configuration variables along each oligomer according to each of the oligomer models. Figure 8.9 contains plots of the marginal distributions of all of the intra base pair internal coordinates at each base along the oligomer $\mathsf{S}_8$, where there are four plots superimposed. Solid lines correspond to a histogram of the data observed in the MD simulation, which in principle could be far from Gaussian, but in most cases are quite close, as discussed in Chapter 5. Dotted lines correspond to the Gaussian marginal determined from the Gaussian oligomer training set density $\rho_{\mu,\mathrm{o}}$, dashed-dotted lines correspond to the Gaussian marginal determined from the nearest-neighbour Gaussian distribution $\rho_{\mu,\mathrm{M}}$, and finally dashed lines represents the Gaussian marginal determined from reconstructed Gaussian density $\rho_{\mu,\mathrm{m}}$. For almost all of the histograms the four curves are practicably indistinguishable. For the other three example sequences the four curves are even closer than for this $\mu = 8$ case, so that we do not include those plots. Figures 8.10, 8.11, 8.12, and 8.13 provide the analogous plots of the marginal distributions of the inter base pair rotational coordinates. Now it can be seen that there are cases where there is a noticeable deviation from a Gaussian marginal even away from the ends, particularly for Twist and Slide, and particularly for the dimer step `C`$p$`G` whose bimodal properties are discussed at length in [42]. We also observe that the MD time series histograms for the 12-mer oligomer $\mathtt{S}_{42}$ with $\mathtt{G}p\mathtt{G}$ ends are particularly far from Gaussian.

Accordingly it is apparent that there are errors associated with our assumption of a Gaussian coarse-grain model. Nevertheless it seems that our Gaussian, nearest-neighbour, dimer dependent model captures the dominant features of sequence variation in a satisfactory way. Specifically the differences between marginals at different dimer steps is almost always qualitatively larger than the differences between the different model marginals at the same dimer step.

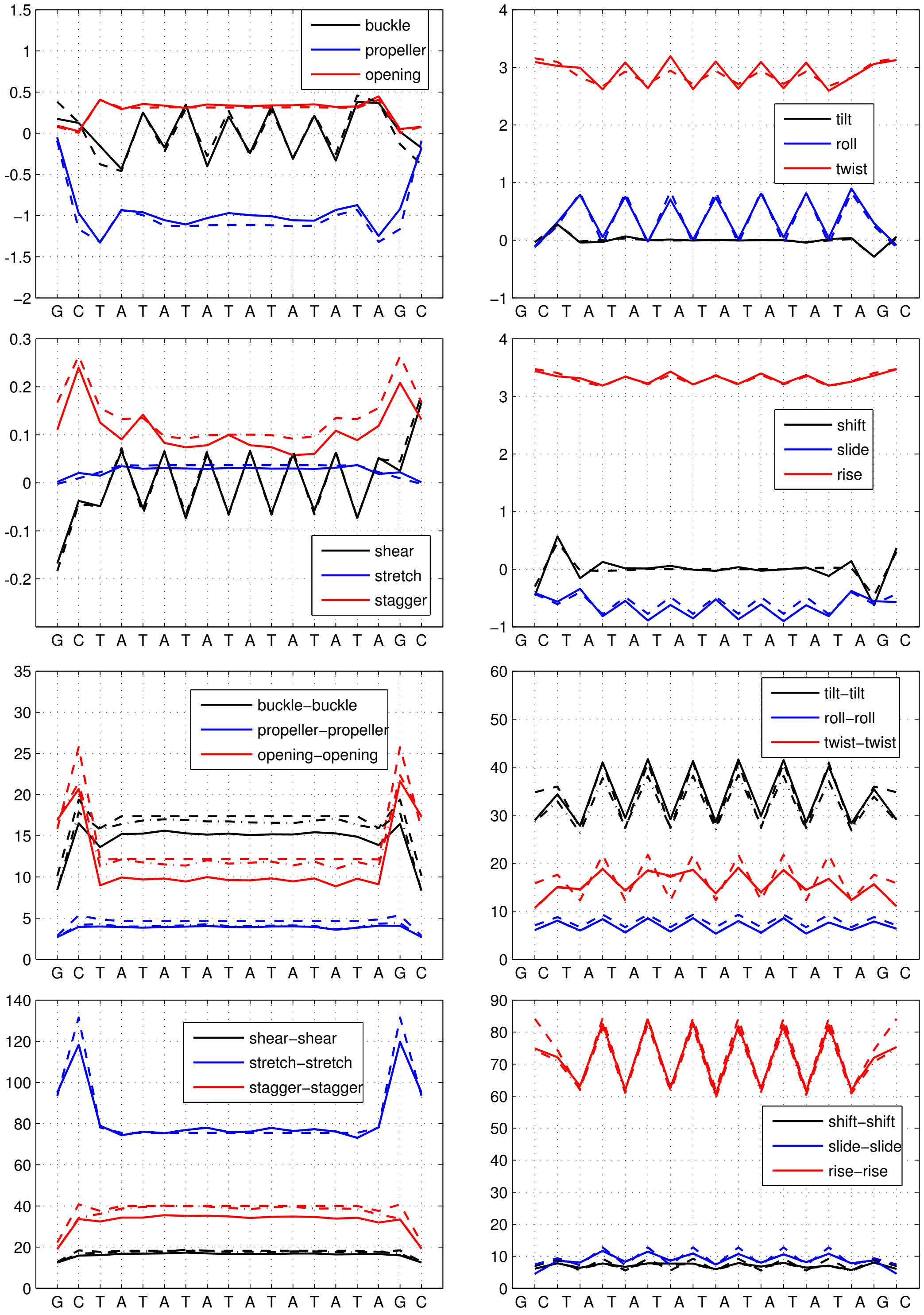


Figure 8.5: The palindromic, interior period two, training sequence $\mathsf{S}_1$. Top four panels: Observed shape parameters $\widehat{\mathrm{w}}_{\mu,\mathrm{o}}$ in solid lines, compared with reconstructed shape parameters $\widehat{\mathrm{w}}_{\mu,\mathrm{m}}$ in dashed lines. Bottom four panels: Diagonal entries of stiffness matrices; solid, observed, dash-dot oligomer fit, and dashed reconstructed. See text in Section 8.2.

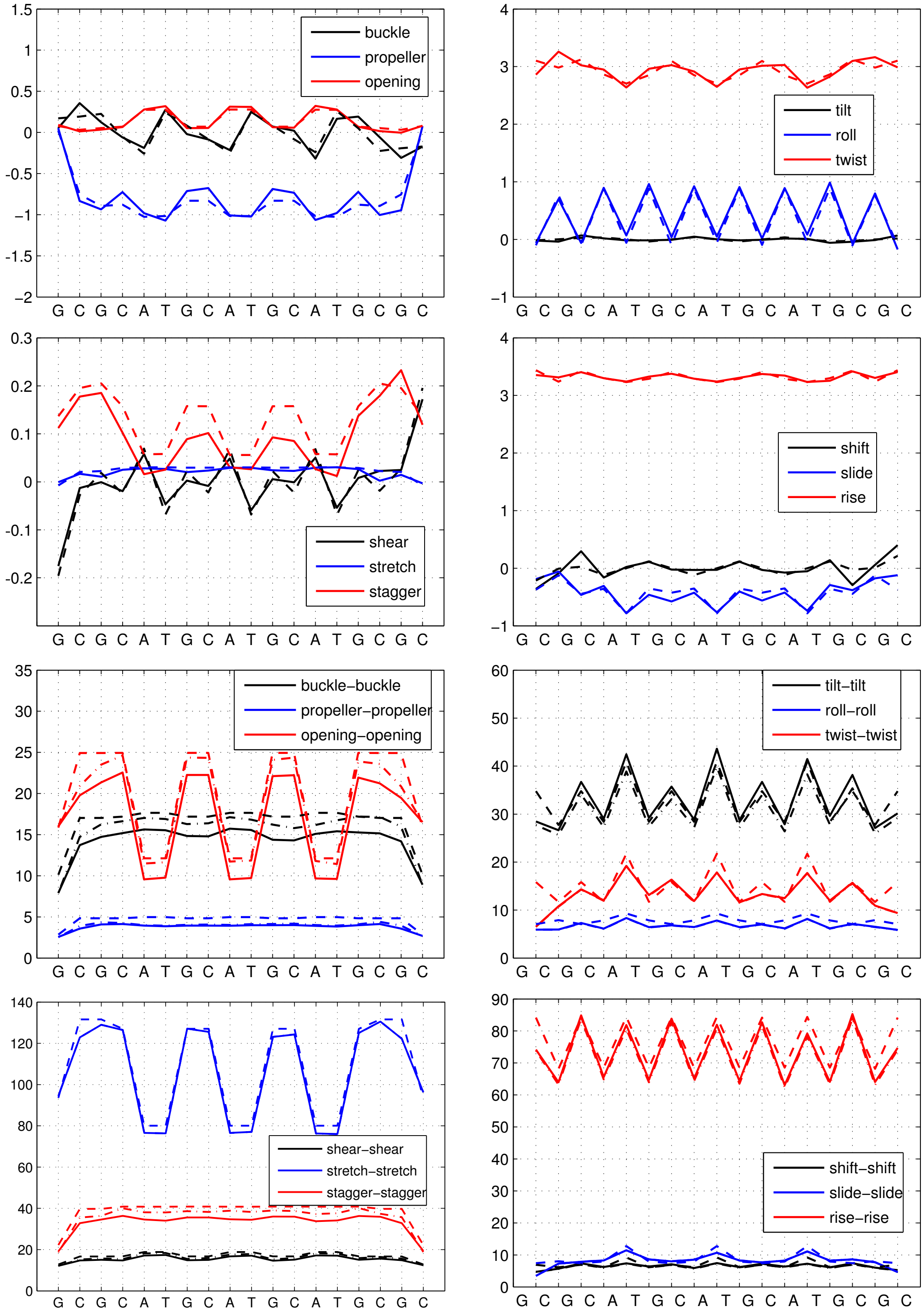


Figure 8.6: The palindromic, interior period four, training sequence $S_3$. Compare Figure 8.5 and text in Section 8.2.

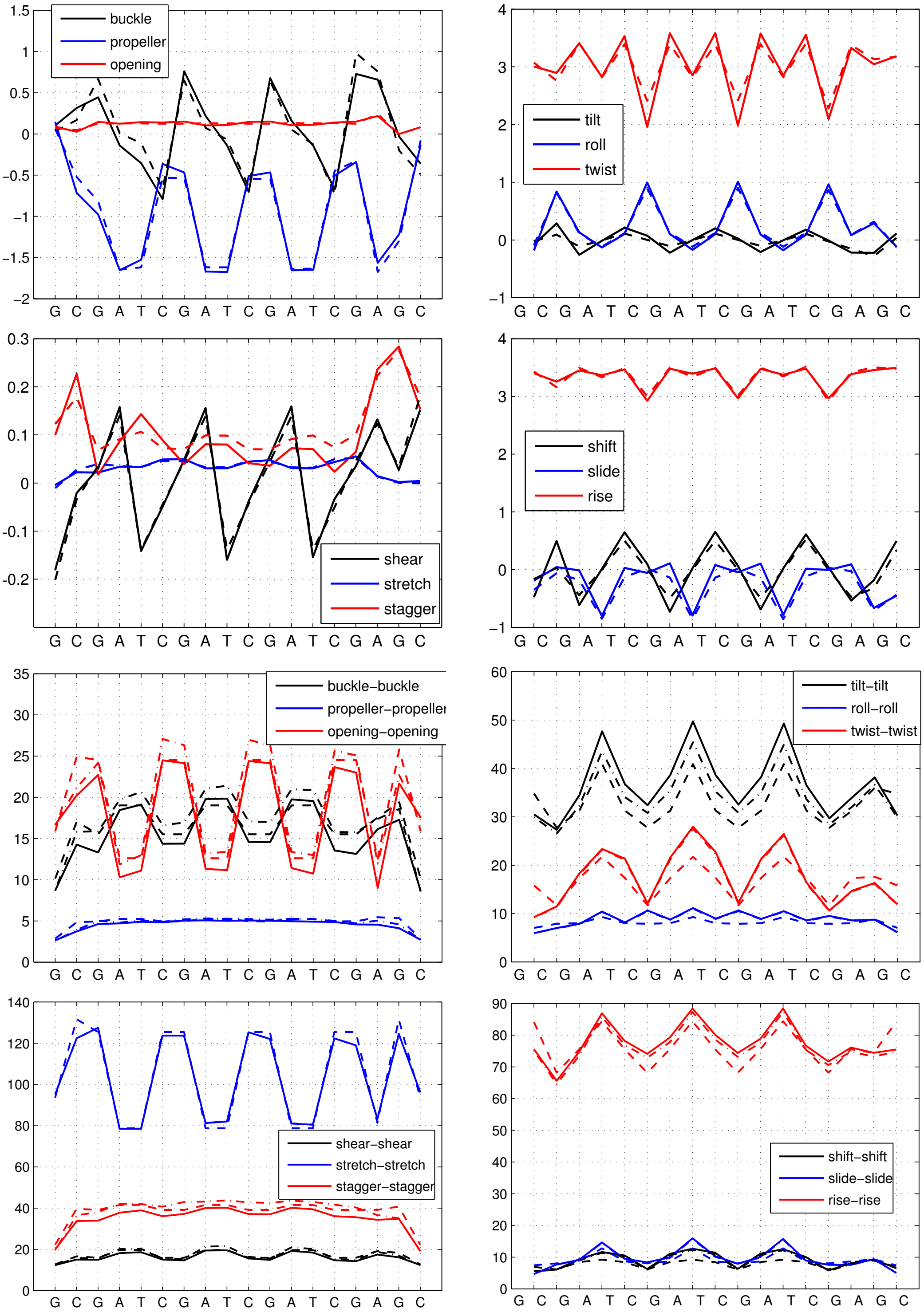


Figure 8.7: The non-palindromic, interior period four, training sequence $\mathsf{S}_8$. Compare caption of Figure 8.5 and text in Section 8.2.

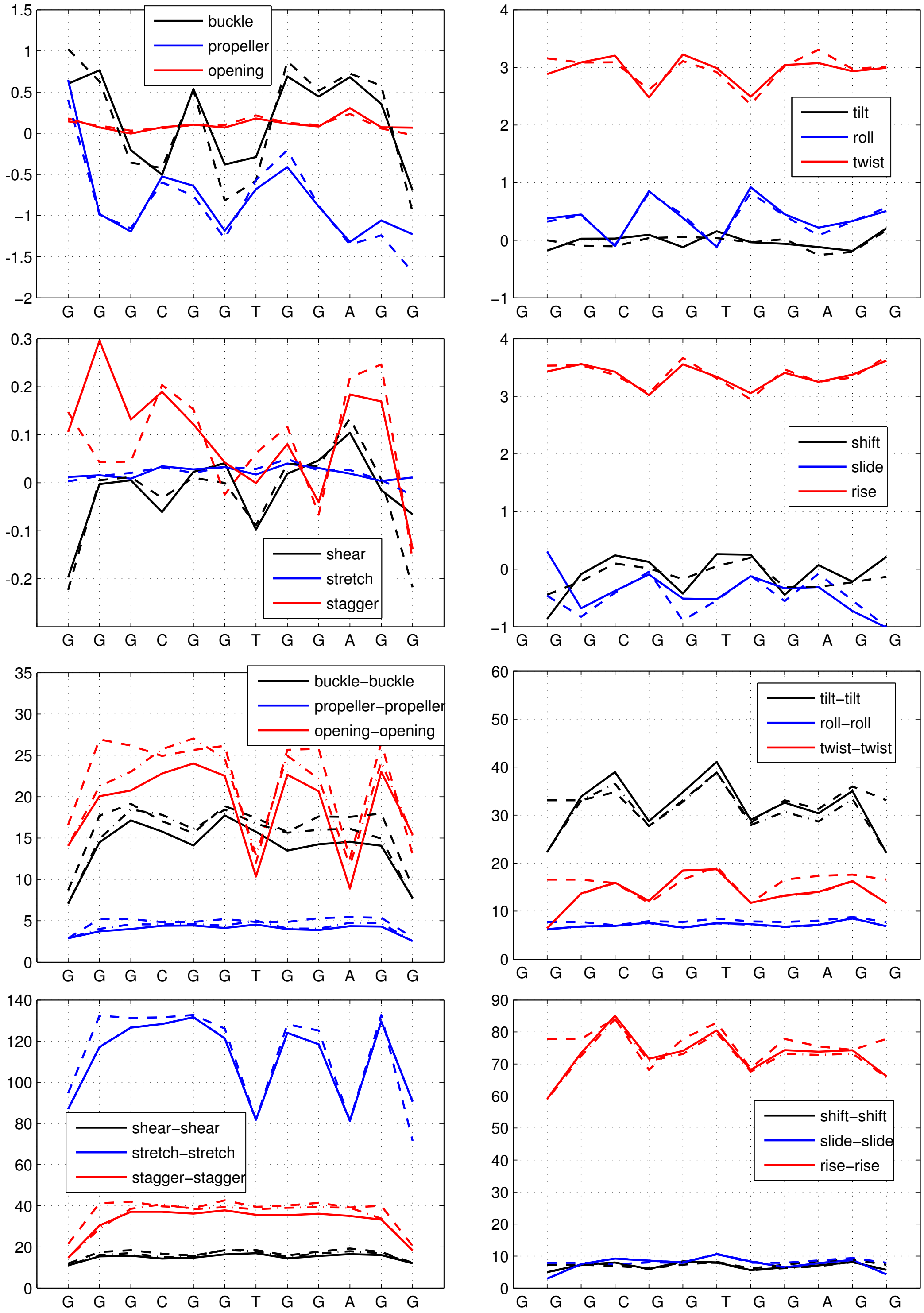


Figure 8.8: The non-palindromic 12-mer training sequence $\mathtt{S}_{42}$ with `GG` ends. Compare caption of Figure 8.5 and text in Section 8.2.

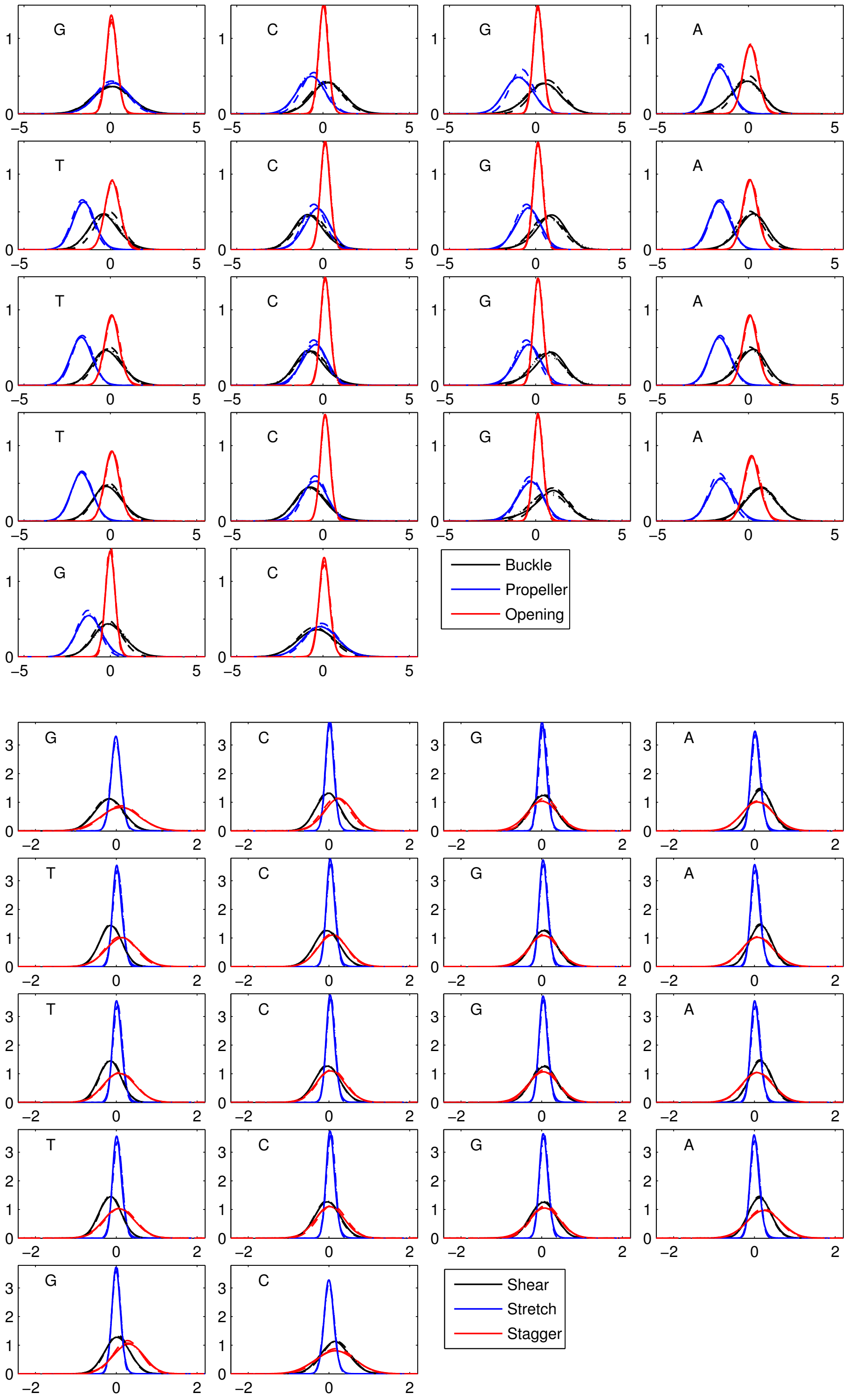


Figure 8.9: Plots of marginal distributions of the intra base pair parameters for the sequence $S_8$ to be read from right to left and top to bottom. The base on the reference strand is marked in each panel. Each colour plot in each panel is actually an overlay of four different line styles corresponding to histograms along the MD time series, which may be far from Gaussian, and three Gaussian marginals of our hierarchy of Gaussian oligomer models. See text in Section 8.2. For intra variables the superposed plots are virtually indistinguishable.

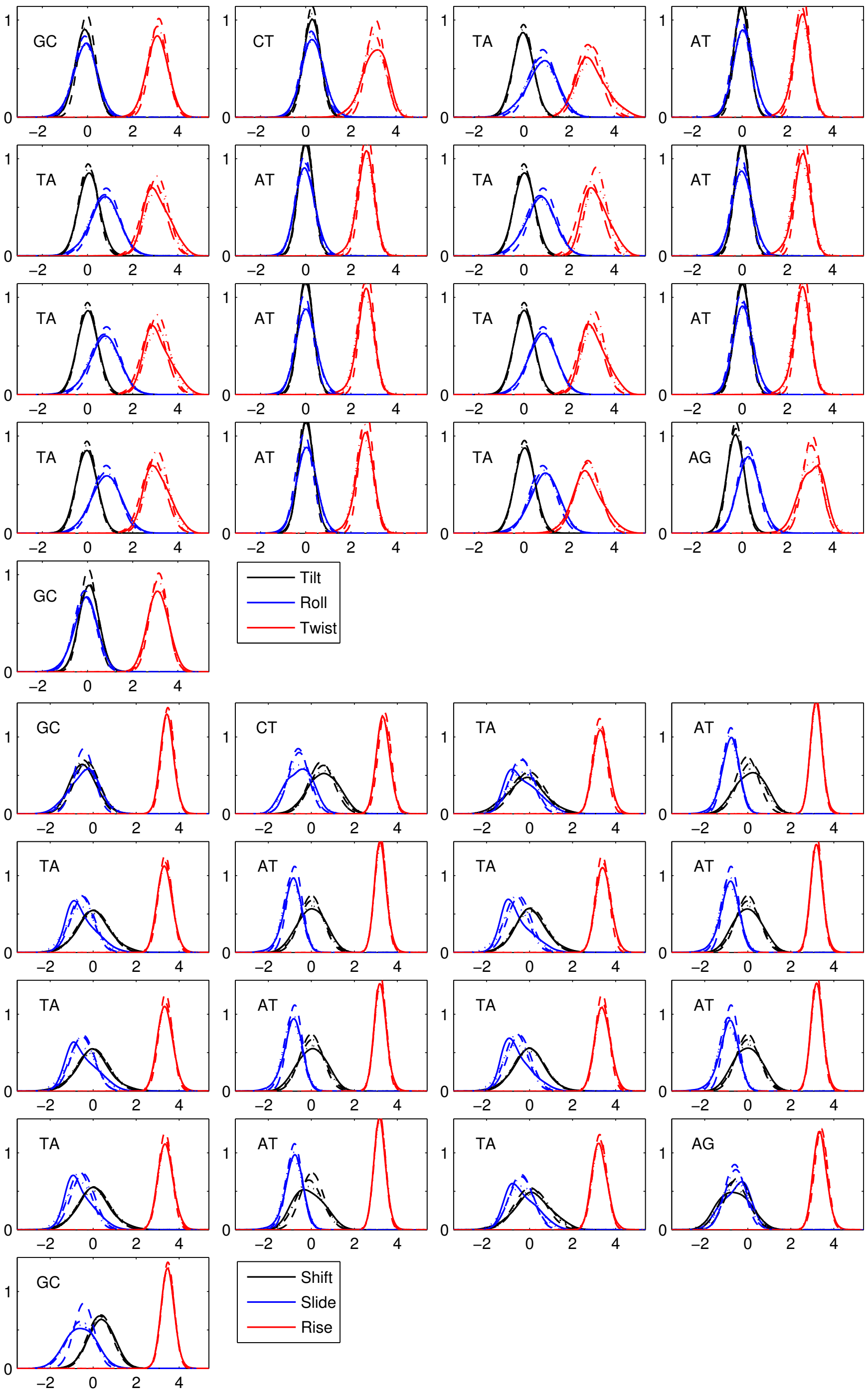


Figure 8.10: Plots of marginal distributions of the inter base pair parameters along the sequence $\mathsf{S}_1$ to be read from right to left and top to bottom. The dimer step on the reference strand at each junction is marked in each panel. Each colour plot in each panel is actually an overlay of four different line styles corresponding to histograms along the MD time series, which may be far from Gaussian, and three Gaussian marginals of our hierarchy of Gaussian oligomer models. In contrast to the intra variable marginals, the solid lines representing the MD histograms are now sometimes visibly far from Gaussian, with associated noticeable differences in the various Gaussian marginals at the same or adjacent junctions. See text in Section 8.2.

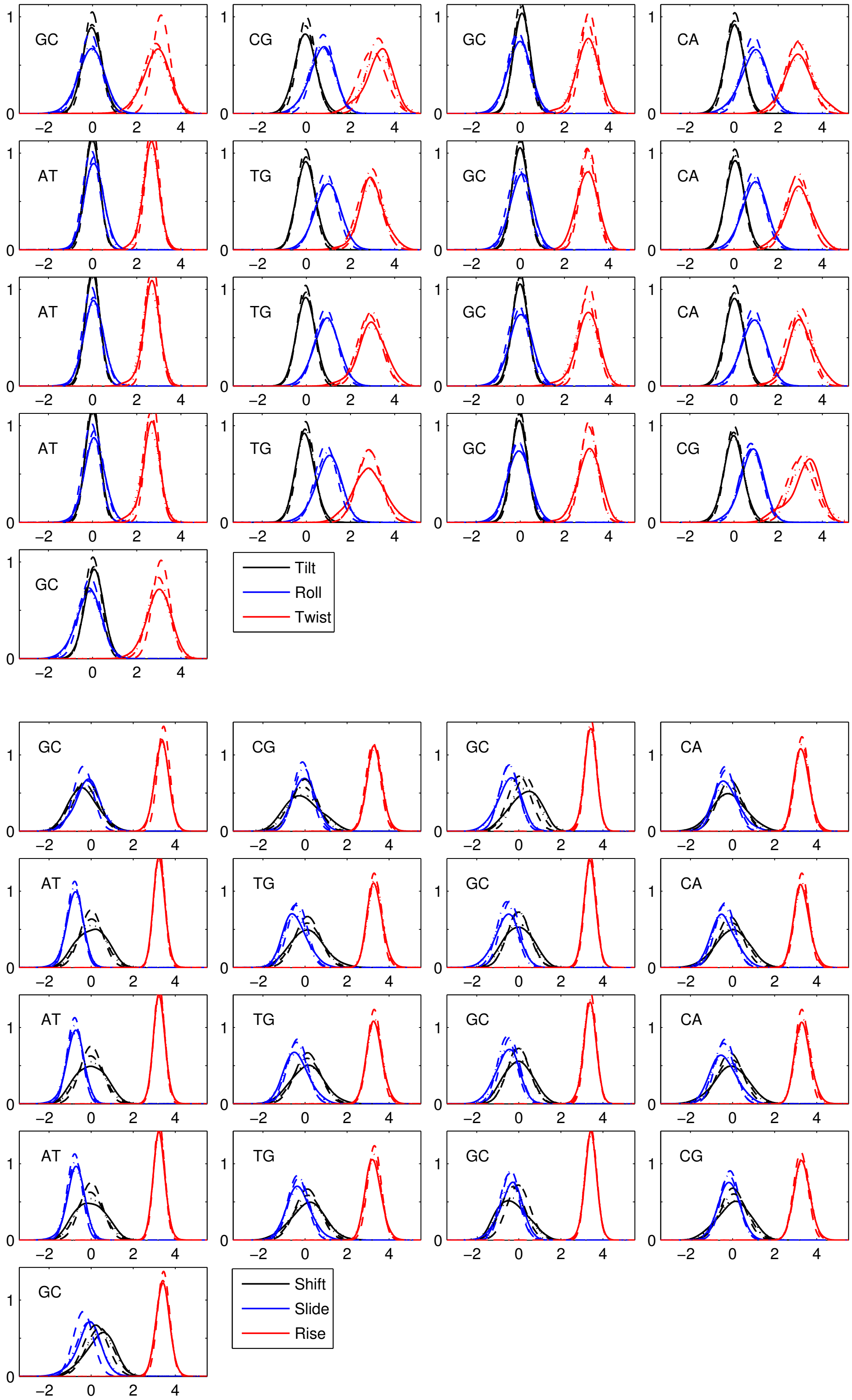


Figure 8.11: Plots of marginal distributions of the inter base pair parameters for the sequence $S_3$. See caption of Figure 8.10 and text in Section 8.2.

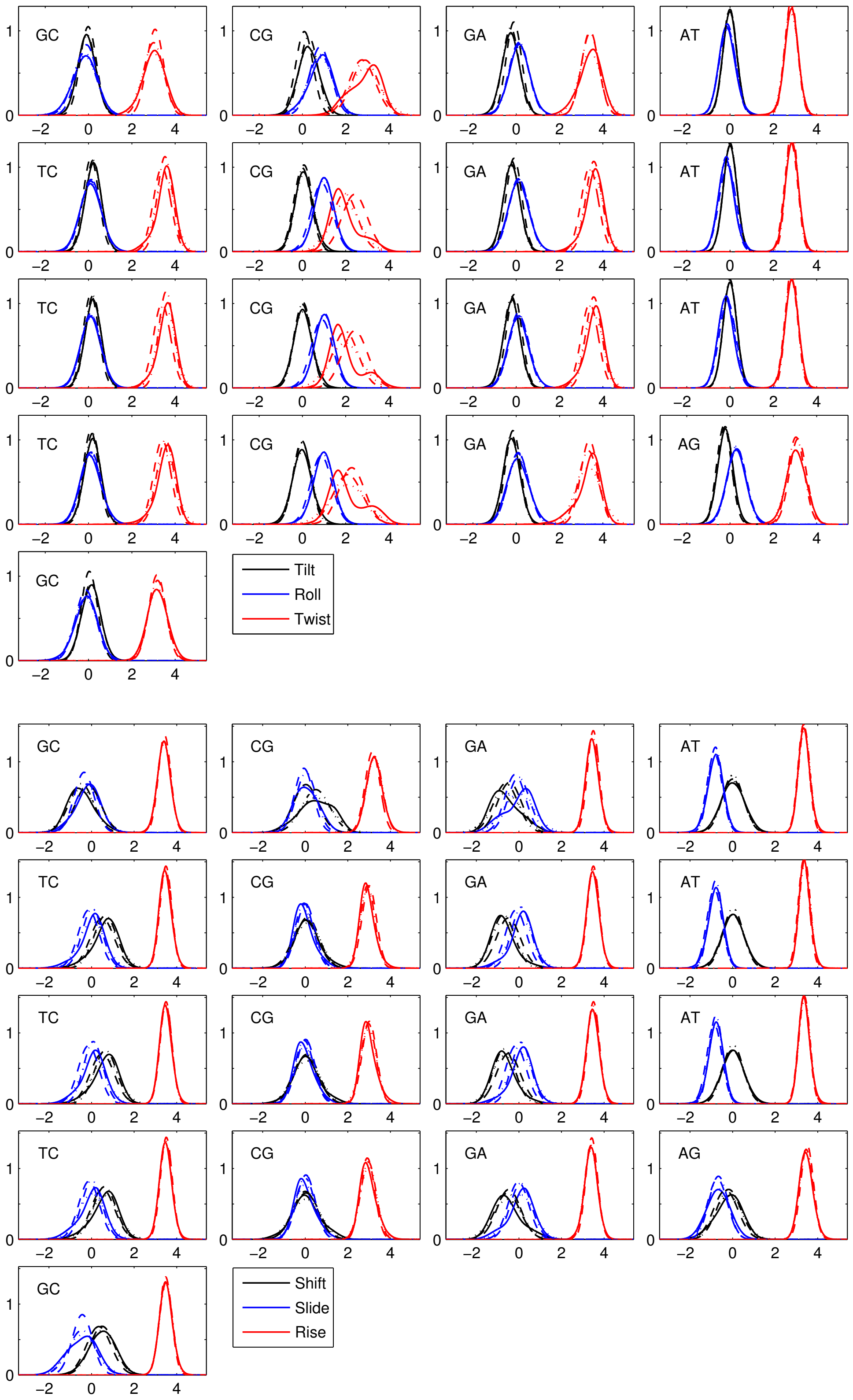


Figure 8.12: Plots of marginal distributions of the inter base pair parameters for the sequence $S_8$. See caption of Figure 8.10 and text in Section 8.2.

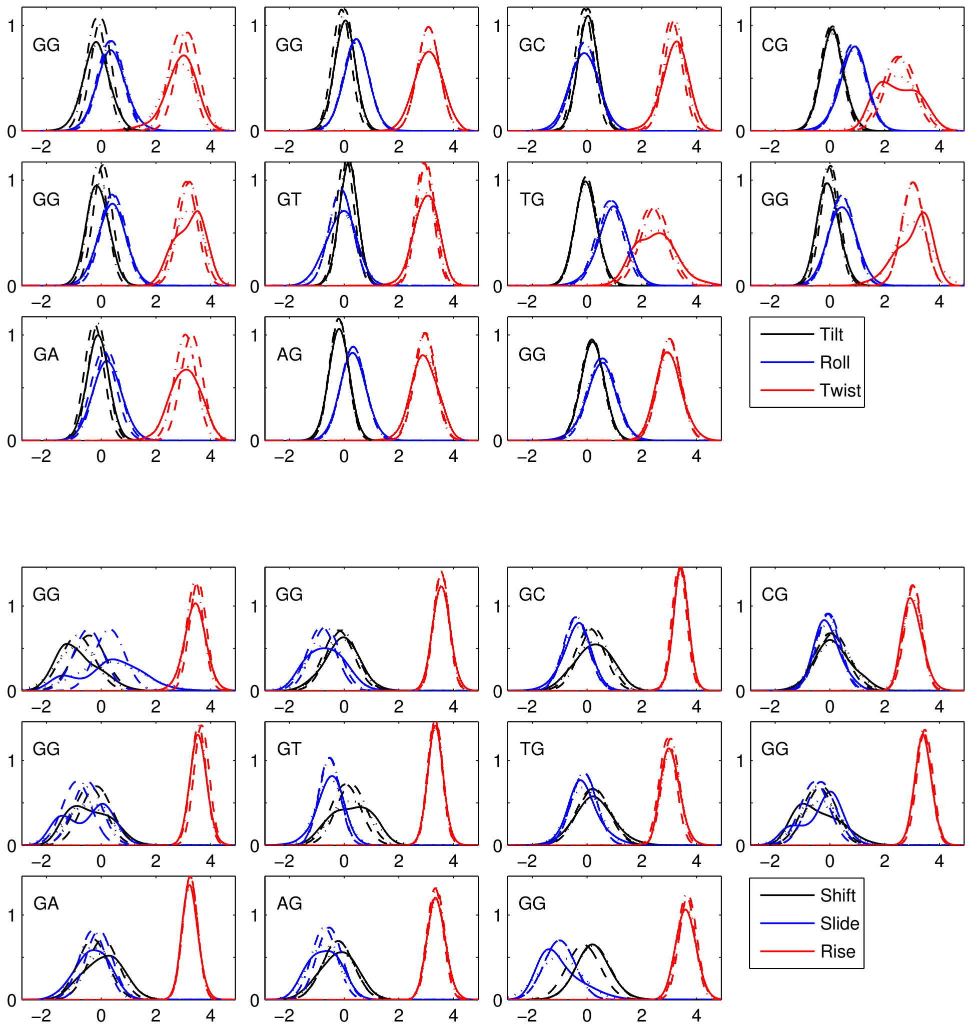


Figure 8.13: Plots of marginal distributions of the inter base pair parameters for the sequence $\mathtt{S}_{42}$. See caption of Figure 8.10 and text in Section 8.2. The Slide, Shift and Twist marginals at the $\mathtt{G}p\mathtt{G}$ steps both at the ends and in the interior of this oligomer have a pronounced deviation from Gaussian.

## 8.3 Example reconstruction of an oligomer not from the training set

As a further test of the discriminatory resolution of our sequence dependent model we made a reconstruction of an 18-mer oligomer not in the original training set. Specifically, sequence $\mathtt{S}_{1'}$, is a point mutation from the training set oligomer $\mathtt{S}_1$. The two sequences differ by only one base at position 6; $\mathtt{S}_{1'}$ has a `T` in this position, whereas $\mathtt{S}_1$ has an `A`. We reconstruct the distribution $\rho_{1',m}$ directly from the parameter set $\mathcal{P}^\dagger$ with no further MD simulation necessary. However to assess the accuracy of the reconstruction, we also carried out a full MD simulation of the oligomer $\mathtt{S}_{1'}$. This MD data was not added to the training set, but we do have available all quantities discussed for the training set oligomers.

As can be seen in Table 8.3 the scaled divergence $D(\rho_{1,\mathrm{m}},\ \rho_{1',\mathrm{m}})/D_\mathrm{o}$ between the two reconstructed distributions is smaller than the scaled reconstruction error $D(\rho_{1',\mathrm{m}},\ \rho_{1',\mathrm{M}})/D_\mathrm{o}$, so that this measure does not capture the effect of the point mutation in the sequence in a significant way. However our reconstruction of the point mutation does exhibit significant pointwise changes in the parameters. Figure 8.14 shows data for the sequence $\mathtt{S}_{1'}$, which should be directly compared with the analogous Figure 8.5 for the sequence $\mathtt{S}_1$. In addition to the plots for the sequence $\mathtt{S}_{1'}$ in Figure 8.14 that are directly analogous to those for the sequence $\mathtt{S}_1$ appearing in Figure 8.5, in order to enhance comparison, the observed vector $\widehat{\mathrm{w}}_{1,\mathrm{o}}$ for the original sequence $\mathtt{S}_1$ is also included in Figure 8.14 with a dotted line style. The top four panels of Figure 8.14 show that a change of one base in the sequence can have pronounced, non-local effects on the oligomer shape parameters, and that the coarse-grain, nearest-neighbour, dimer dependent reconstructions accurately mimic the nonlocal changes observed in the MD simulation. The non-local effects are most pronounced for Buckle, Propeller and Stagger, where the major effects are spread over approximately five bases, and are less pronounced, but still noticeable, for Shear, Twist, Shift and Slide, where the effects are spread over approximately three bases. For other quantities, for example Roll, the effect of the point mutation appears rather local, and for others, for example Opening, Stretch and Rise, any effect is not clearly visible on the scale of the plots. A comparison of the predicted vector $\widehat{\mathrm{w}}^*_{1',\mathrm{m}}$ with the observed vector $\widehat{\mathrm{w}}_{1',\mathrm{o}}$ shows that the nearest-neighbour, dimer-based model introduced here can predict the local and non-local effects of the mutation rather well. Figure 8.15 contains analogous plots for a sequence obtained by two point mutations from $\mathtt{S}_1$. The same conclusions hold for this simulation, namely that our model is able to predict non-local effects in shapes of the point mutations.

The nonlocal consequences of our nearest-neighbour, dimer dependent model on oligomer shape parameters can be further quantified using the data associated with $\mathtt{S}_1$ and $\mathtt{S}_{1'}$. In particular Table 8.4 provides the expected values of the intra parameters buckle and propeller of the central `AT` base pair in the two pentamer sub-sequences `ATATA` and `TTATA` of the oligomers $\mathtt{S}_1$ and $\mathtt{S}_{1'}$. In each case the expectations observed from the MD time series and reconstructed from our model are given. We conclude that our model can accurately predict changes in intra parameter expectations consequent upon changes in the sequence at a next to nearest-neighbour base pair. Similarly differences in the expectation of the Slide inter base pair junction variable at a `TA` step embedded in the tetramers `ATAT` and `TTAT` are accurately predicted by our model reconstruction.

| | `ATATA` | `TTATA` |
|---|---|---|
| Observed Buckle | -0.27 | -0.73 |
| Reconstructed Buckle | -0.25 | -0.53 |
| Observed Propeller | -1.03 | -0.72 |
| Reconstructed Propeller | -1.09 | -0.75 |

| | `ATAT` | `TTAT` |
|---|---|---|
| Observed Slide | 0.74 | -0.29 |
| Reconstructed Slide | 0.84 | -0.24 |

Table 8.4: Dependence of Buckle and Propeller of the central `AT` base pair of the `ATATA` and `TTATA` pentamer subsequences of the oligomers $\mathsf{S}_1$ and $\mathsf{S}_{1'}$, along with dependence of Slide of the central `TpA` junction on surrounding `ATAT` and `TTAT` tetramers. In each case the expected value of the parameter is both observed in an MD time series, and predicted by our reconstruction model.

The bottom four plots of Figure 8.14 illustrate the effects of the point mutation between the sequences $\mathsf{S}_1$ and $\mathsf{S}_{1'}$ on the diagonal entries of the stiffness parameter matrices $\mathsf{K}^*_{1',\mathrm{m}}$, $\mathsf{K}_{1',\mathrm{o}}$ and $\mathsf{K}_{1,\mathrm{o}}$. For stiffnesses, as predicted by our model, the effects of the mutation are local, and are again predicted rather well by our reconstruction.

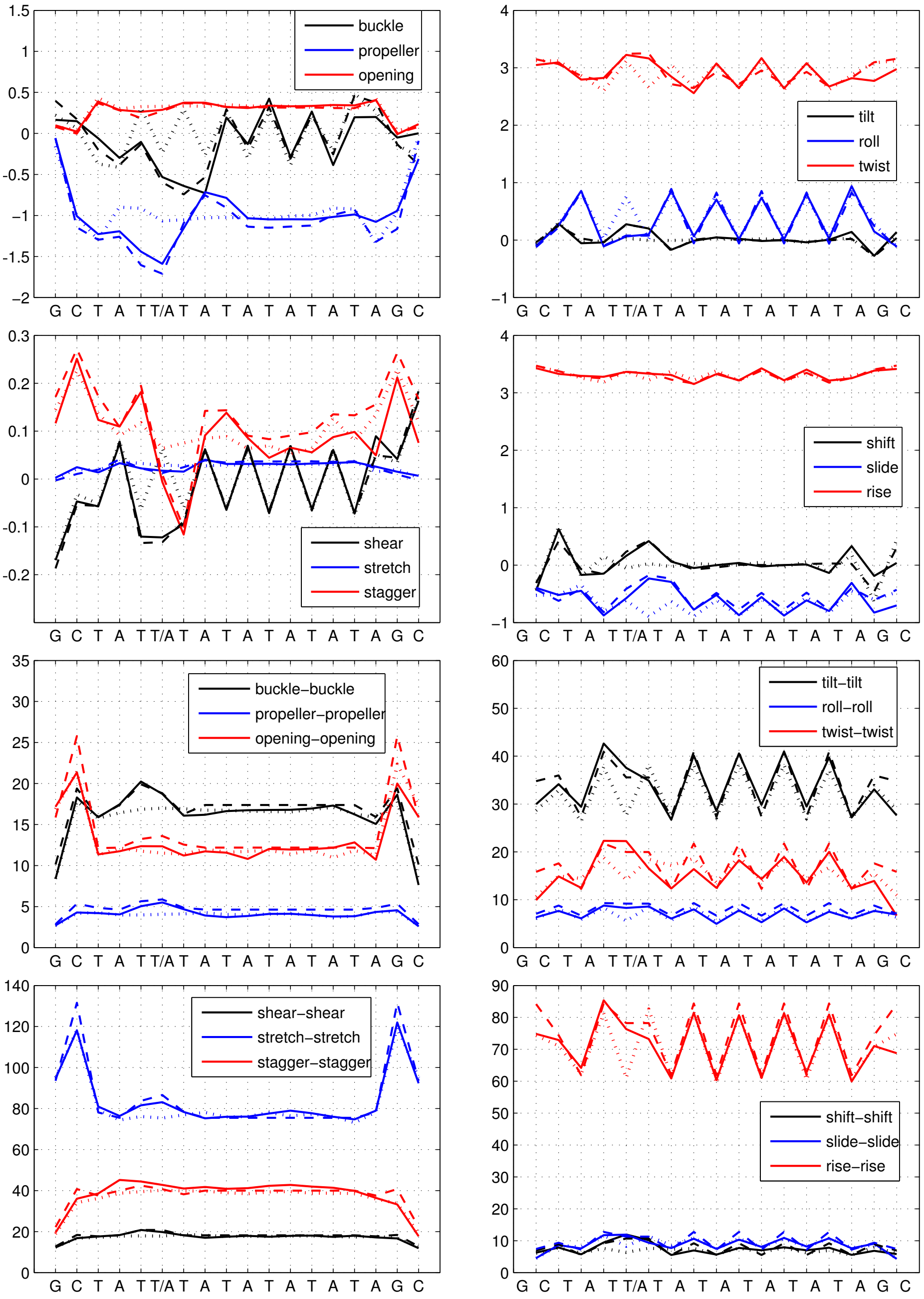


Figure 8.14: Data for the sequence $S_{1'}$, which is a point mutation of the training sequence $S_1$ described in Figure 8.5. The base sequence of the reference strand is indicated on the horizontal axis, where the symbol T/A denotes the mutation site. Top four panels: Observed shape parameters $\widehat{w}_{1',o}$ in solid lines, compared with reconstructed shape parameters $\widehat{w}_{1',m}$ in dashed lines. For reference the vector $\widehat{w}_{1,o}$ is also plotted in dotted lines. Bottom four panels: Diagonal entries of the stiffness matrices; observed $K_{1',o}$ in solid, reconstructed $K^*_{1',m}$ dashed, and for reference the observed $K_{1,o}$ for $S_1$ in dotted. As predicted the changes between $S_1$ and $S_{1'}$ are local for stiffness parameters, but nonlocal for shapes. See text in Section 8.3.

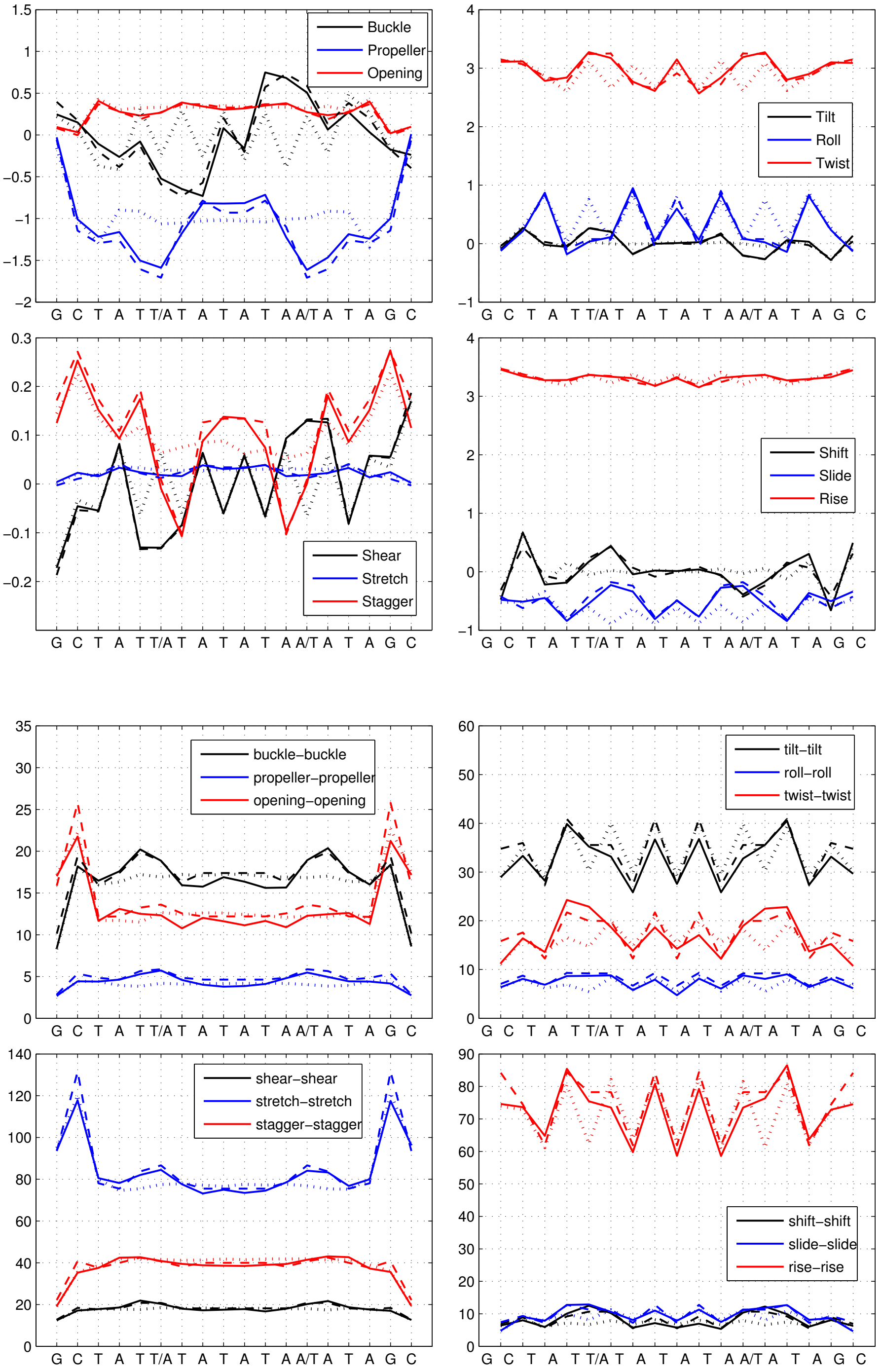


Figure 8.15: Data for a sequence that is obtained as a double point mutation of the training sequence $\mathsf{S}_1$ described in Figure 8.5. The base sequence of the reference strand is indicated on the horizontal axis, where the symbols `T/A` and `A/T` denote mutation sites. Top four panels: Observed shape parameters in solid lines, compared with reconstructed shape parameters in dashed lines. For reference the vector $\widehat{\mathsf{w}}_{1,\mathrm{o}}$ is also plotted in dotted lines. Bottom four panels: Diagonal entries of the stiffness matrices; observed in solid, reconstructed dashed, and for reference the observed $\mathsf{K}_{1,\mathrm{o}}$ for $\mathsf{S}_1$ in dotted. As predicted the changes between the two sequences are local for stiffness parameters, but nonlocal for shapes. See text in Section 8.3.

# Chapter 9

# Summary, conclusions and discussion

We have developed a sequence-dependent coarse-grain model of DNA that is able to predict the non local dependence of oligomer shape parameters, despite the local sequence dependence of the model parameters. The main novelty of the model is the concept of frustration energy, which, as described in Section 6.1, is the reason for this non local shape dependence. Therefore, the model can give more accurate oligomer shape and stiffness parameter predictions than the widely used local rigid base pair models.

We first give an overview of the development presented in this thesis, emphasising open questions. We model DNA as a system of interacting rigid bases. The configuration of a rigid base DNA oligomer is described by a vector $\mathsf{w}$ of relative coordinates, which can be partitioned into translational $v^i \in \mathbb{R}^3$ and rotational $u^i \in \mathbb{R}^3$ degrees of freedom for every base pair and every junction. For the relative rotations $u^i$ we choose a scaled Cayley vector, as in [36] and [41]. Following [36] and [74], we assume that the equilibrium distribution of $\mathsf{w} \in \mathbb{R}^{12n-6}$ for a DNA molecule of length $n$ base pairs is described by the probability density function

$$\rho(\mathsf{w}) = \frac{1}{Z_J} e^{-\mathsf{U}(\mathsf{w})}\, J(\mathsf{w})\, d\mathsf{w}. \tag{9.1}$$

Here $Z_J$ is a normalising constant, $\mathsf{U}(\mathsf{w}) \in \mathbb{R}$ is the non-dimensional free energy function, and $J$ is a Jacobian factor, which is present due to the non-Cartesian nature of the rotation group $SO(3)$, and which, for our choice of coordinates $u_i$, takes the explicit form

$$J(\mathsf{w}) = \prod_{i=1}^{2n-1} \frac{1}{(1 + \frac{1}{100}|u^i|^2)^2}. \tag{9.2}$$

We further assume that the free energy is a shifted quadratic of the form

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}(\mathsf{w} - \widehat{\mathsf{w}}) \cdot \mathsf{K}(\mathsf{w} - \widehat{\mathsf{w}}) + \widehat{\mathsf{U}}, \tag{9.3}$$

where $\widehat{\mathsf{w}} \in \mathbb{R}^{12n-6}$ is an equilibrium shape vector, $\mathsf{K} \in \mathbb{R}^{(12n-6)\times(12n-6)}$ is a symmetric positive definite stiffness matrix, and $\widehat{\mathsf{U}}$ is the energy of frustration, which is a novel concept and an important feature of our model. As discussed in [36] and [74], the equilibrium shape and stiffness parameters of the assumed potential can be estimated from MD time series using the discrete version of the extraction relations

$$\frac{\langle \mathsf{w}/J\rangle}{\langle 1/J\rangle} = \widehat{\mathsf{w}} \quad \text{and} \quad \frac{\langle \Delta\mathsf{w}\otimes\Delta\mathsf{w}/J\rangle}{\langle 1/J\rangle} = \mathsf{K}^{-1}, \tag{9.4}$$

where $\Delta\mathsf{w} = \mathsf{w} - \widehat{\mathsf{w}}$. Assuming ergodicity of simulations, data from Molecular Dynamics (MD) simulations can be used for this purpose. As our training data set, we used an ensemble of 50-200 ns MD trajectories for 53 DNA oligomers, the main part of which was produced by the ABC consortium [42], and the rest designed and simulated to get parameter estimates for different oligomer ends. Following [36], we exclude snapshots with broken hydrogen bonds from our analysis, as broken hydrogen bonds indicate possible structural defects of DNA, such as fraying oligomer ends, a phenomenon that can not be captured by a quadratic energy model.

For all 53 sequences of our training data set, we confirmed, as quantified in Figure 7.3, a conclusion of [36], made from MD data for a single different sequence and using a different MD force field, namely that a rigid base pair (in contrast to rigid base) DNA model with only nearest-neighbour interactions is a poor representation of DNA.

Having noticed that most of the marginal distributions of our oligomer configuration variables are bell-shaped curves, it was tempting to approximate the Jacobian $J$ by a constant, so that (9.1) becomes a Gaussian distribution. To examine the consequences of this approximation, we quantified the differences between shape and stiffness parameters, corresponding to both distributions, and concluded that those differences were small, as shown in Figure 5.4. Most importantly, we showed that the density function $\rho$ in (9.1) is logarithmically concave (i.e. that $\ln\rho$ is concave), which implies that the presence of the Jacobian factor can not predict bimodal marginal distributions of some parameters (mainly Twist of the $\mathtt{C}p\mathtt{G}$ junctions), which is observed from the MD data. Still, most of the observed marginal distributions were unimodal, hence the Gaussian approximation of the probability density function seems to be a reasonable first approximation.

At this point we have prepared the ground for introducing a Gaussian nearest-neighbour rigid base DNA model. The starting assumption of this model is that the oligomer potential energy $\mathsf{U}(\mathsf{w})$ in (9.3) can be expressed as a sum of dimer and of monomer energies along the oligomer:

$$\mathsf{U}(\mathsf{w}) = \frac{1}{2}\sum_{j=1}^{2}\sum_{a=1}^{n-j+1}(\mathsf{w}_j^a - \widehat{\mathsf{w}}_j^a)\cdot\mathsf{K}_j^a(\mathsf{w}_j^a - \widehat{\mathsf{w}}_j^a). \tag{9.5}$$

Here, for $j = 1, 2$, $\widehat{\mathsf{w}}_j^a \in \mathbb{R}^{12j-6}$ and $\mathsf{K}_j^a \in \mathbb{R}^{(12j-6)\times(12j-6)}$ are equilibrium shape and stiffness parameters of the base pair $(\mathtt{X},\overline{\mathtt{X}})_a$ when $j = 1$, and of the junction between base pairs $(\mathtt{X},\overline{\mathtt{X}})_a$ and $(\mathtt{X},\overline{\mathtt{X}})_{a+1}$ when $j = 2$. Each variable vector $\mathsf{w}_j^a \in \mathbb{R}^{12j-6}$ is a collection of corresponding (and overlapping) elements of the oligomer configuration vector $\mathsf{w}$. This model is actually a simple representative of a hierarchical family of models, where $\mathsf{U}(\mathsf{w})$ is the sum of energies of all overlapping $j$-mers $\mathtt{X}^a\ldots\mathtt{X}^{a+j-1}$, $j = 1,\ldots,k$, $1 < k < n$, and there are $n - j + 1$ such $j$-mers for each $j$.

We were also lead to consider a stress-like variable $\sigma = \mathsf{K}\widehat{\mathsf{w}}$, which we call a weighted shape vector. The nearest-neighbour assumption implies that the stiffness matrix has a special structure and can be obtained by summing the blocks $\mathsf{K}_j$, $j = 1, 2$. Analogously, the weighted shape vector can be obtained by summing the vectors $\sigma_j = \mathsf{K}_j \widehat{\mathsf{w}}_j$, $j = 1, 2$.

Simply by completing squares, the energy (9.5) can be expressed in the form (9.3), thus providing the expression of the parameters $\widehat{\mathsf{w}}$ and $\widehat{\mathsf{U}}$ in terms of monomer and dimer parameters. In particular, we find

$$\widehat{\mathsf{w}} = \mathsf{K}^{-1}\sigma. \tag{9.6}$$

Next we introduce sequence-dependence by assuming that the parameters $\{\widehat{\mathsf{w}}_1^a, \mathsf{K}_1^a\}$ and $\{\widehat{\mathsf{w}}_2^a, \mathsf{K}_2^a\}$ in (9.5) depend only on the local dimer sequence $\mathtt{X}_a\mathtt{X}_{a+1}$, and not explicitly on the oligomer length $n$ or the location $a$. This model is completely defined by the parameters $\{\sigma_1^\gamma, \mathsf{K}_1^\gamma\}$, $\gamma \in \mathtt{M}$ and $\{\sigma_2^{\alpha\beta}, \mathsf{K}_2^{\alpha\beta}\}$, $\alpha\beta \in \mathtt{D}$, where $\mathtt{M}$ is a set of any two independent monomers and $\mathtt{D}$ is a set of any ten independent dimers. Due to the anti-parallel double stranded structure of DNA, the parameters for the remaining two monomers and six dimers can be found using simple symmetry relations. Given such a parameter set, one can reconstruct the oligomer stiffness matrix $\mathsf{K}$ and shape vector $\widehat{\mathsf{w}}$ for any desired oligomer. Notice that the resulting stiffness matrix $\mathsf{K}$ has a local dependence on the base pair sequence, but, because of identity (9.6), the shape vector $\widehat{\mathsf{w}}$ does not. However, due to the banded structure of $\mathsf{K}$, the entries of $\mathsf{K}^{-1}$ decay away from the diagonal, and consequently so do the effects of the remote base pairs on the entries of $\widehat{\mathsf{w}}$. This non local sequence dependence of shapes can be explained in physical terms by the phenomenon of frustration, which arises due to the fact that all of the rigid bases of the oligomer can not minimise their energy simultaneously and so have to compromise. The minimum energy that can be achieved in a given oligomer is $\widehat{\mathsf{U}}$, which is in general non zero. In this case the oligomer can be described as pre-stressed.

Having a model, we next have to fit a parametrisation to a training data set. The available experimental data is insufficient for this purpose, so we instead use an extended version of the aforementioned ABC data set [42]. First, for each simulated oligomer, we enforce sparsity of the observed stiffness matrix, i.e. the range of interactions. We do this by minimising the Kullback-Leibler divergence between the oligomer dependent distributions with full and sparse stiffness matrices. Then we enforce the dimer sequence dependence of stiffness and weighted shape parameters, i.e. we fit a dimer-based model parameter set to our simulation data for all 53 oligomers. To achieve this goal, we first find all of the solutions in the null space of a least squares system without the equations for the oligomer ends, because we trust the MD data at the end of an oligomer less than in the middle. Then we choose a solution in this null space by adding equations for the oligomer ends with a vanishing weight. The resulting parameter set provides reasonably good reconstructions of oligomer shapes. However, and surprisingly, this solution, as well as all others in the null space that we could find, led to some indefinite stiffness parameter matrices. As a modelling decision, we choose to restrict stiffness parameter matrices to be at least positive semi-definite. Accordingly we construct a perturbation scheme that numerically raised the negative eigenvalues to yield a set of positive definite stiffness parameter matrices $\mathsf{K}_1^\alpha$ and $\mathsf{K}_2^{\alpha\beta}$. Finally, we numerically obtain our best-fit parameter set $\mathcal{P}^\dagger$ as a maximum relative entropy [49] optimiser of a Kullback-Leibler divergence. We note that the choice of the direction in the asymmetric Kullback-Leibler divergence does not have a big impact on the final parameter set, i.e. the two choices of an objective function for a fitting procedure give very similar results, consistently for each fitting step.

The first test of the parameter set $\mathcal{P}^\dagger$ was to compare the reconstructed oligomer parameters for the sequences from our training data set. Our conclusion was that these reconstructions were remarkably good, but that the reconstruction errors were larger at the ends, as can be seen, for example, in Figure 8.5. We also looked at the MD scalar marginal histograms of all of the individual observed configuration variables and compared with the corresponding marginals determined from three, related Gaussian probability densities, namely directly estimated from MD, nearest-neighbour fitted and dimer-based reconstructed (see, for example, Figure 8.10). Certainly, there are apparent errors associated with our assumption of a Gaussian coarse-grain model. However in general our Gaussian, nearest-neighbour, dimer dependent model captures the dominant features of sequence variation.

Another test was to reconstruct oligomer parameters for a sequence that was not used in the training set, and to compare these reconstructions with parameter estimations from a MD simulation. To be able to quantify the non local dependence of shape parameters, we carried out a simulation of two sequences, which were designed as point mutations from a training set oligomer. We observed that a point mutation indeed had a non local effect on some shape parameter values, as was well predicted by our model reconstruction, as shown in Figures 8.14 and 8.14. More plots of reconstructed parameters for oligomers both from and outside of the training set can be found on the webpage http://lcvmwww.epfl.ch/cgDNA, together with the full parameter set $\mathcal{P}^\dagger$ and the Matlab® scripts used to obtain these reconstructions.

Various limitations of our model arise from its simple origins. The first simplification is that DNA is modelled only as a system of rigid bases. The behaviour of the sugar-phosphate backbones is not explicitly taken into account, and the phosphate-base interactions only appear implicitly in the interactions between bases. Ignoring the backbones may certainly be a reason for some modelling errors. Another simplification is the linear approximation of interactions (quadratic energy). In general, some interactions between DNA atoms are highly nonlinear, such as van der Waals and electrostatic interactions. Also the contribution of the solvent caused entropy to the energy of DNA does not have a simple functional form. This nonlinearity could result in nonquadratic terms in the base interaction energy. One more simplification of reality is the assumption that only nearest-neighbour interactions of rigid bases contribute to the free energy of a DNA molecule.

The main strength of the model is that the parameter set is comparatively small, but, nevertheless, in the Gaussian regime, apparently sufficient for relatively good reconstructions of elastic parameters for any sequence of any reasonable length. One way to make use of our model is to explore the properties of DNA fragments that are too long to be investigated by Molecular Dynamics simulations, but still short enough to depend upon sequence. For example, R.S. Manning and J.S. Mitchell are developing a Monte Carlo code with the goal of computing sequence dependent DNA looping probabilities, which can then be compared to experimentally determined values. J.S. Mitchell also used our parameter set to construct the rigid base stiffness and shape parameters for eight 500 bp length sequences and then carried out Monte Carlo simulations of these sequences to compute sequence dependent persistence lengths. The persistence length is computed as a gradient of a line, fitted to the points $(n, -n/\ln\langle \boldsymbol{t}_1 \cdot \boldsymbol{t}_n\rangle)$, for $n = 2, \dots, 500$, where $\boldsymbol{t}_i$, is a tangent to the centreline of the oligomer at base pair $i$. The resulting persistence lengths were compared to those determined through cyclisation experiments [22]. This comparison revealed that the persistence lengths obtained from our model reconstructions were 1.3-1.8 times larger than the ones fit to cyclisation experimental data. A possible reason for this stiffness estimation error could be in the parametrisation of the MD force field, used for producing our training data set, or the discrepancy could lie in one of our coarse-grain modelling assumptions.

Another numerical experiment with our data set done by J.S. Mitchell was running Metropolis Monte Carlo simulations of the same sequences, but this time sampling the distribution with a Jacobian factor. The stiffness and shape parameters for these simulations were constructed from the same "Gaussian" parameter set $\mathcal{P}^\dagger$ as in the previous simulations. However, as shown in Chapter 5, the differences between the quadratic potential parameters in distributions with and without the Jacobian are negligible. His conclusion was that even though the Jacobian appears to soften DNA at short length scales, its overall effect on the persistence length is small. Thus our conclusion that the Gaussian approximation of a rigid base DNA probability density function is appropriate seems to hold. Using Metropolis Monte Carlo to compute Kullback-Leibler divergences between distributions with and without Jacobian would be another way to quantify errors of the Gaussian approximation.

As the method of parameter estimation for our dimer-based nearest-neighbour model from an MD data set is now established, it is possible to repeat this procedure on other data sets. The first new parameter extraction is planned from the latest ABC consortium data set, which was almost ready at the time of writing this thesis. This data set will contain simulations for the original 39 oligomers, but this time extended at least to one microsecond for each oligomer. We expect that these simulations will be better converged, i.e. closer to ergodic, thus hopefully resulting in a better dimer-based parameter set. It would also be interesting to study the dependence of our parameter set on the MD force field, solvent model, temperature, different ions and different salt concentrations. In particular, a dimer dependent parameter set is a compact concentrated set of values that can be used to quantify the elastic properties of DNA coming from any simulation, model or experiment. For example, there is an increasing interest in effects of methylation on DNA elastic properties [58]. To explore these effects, we plan to add parameters for the methylated `C`$p$`G` step to our parameter set. One limitation of using these parameters for comparing different systems is that the amount of data needed for their estimation is quite large.

Wider stencil (i.e. beyond nearest-neighbour) rigid base interactions could also be considered. Estimating a parameter set for the trimer based next to nearest-neighbour model and comparing the performance of this model with the one developed here would be a natural way of trying to improve the oligomer reconstructions. It is also possible (independently) to assume that a nearest-neighbour parameter set has a wider sequence dependence than dimers. Even though this choice appears to be not completely natural, it could still lead to better oligomer reconstructions.

The continuum analog of a rigid base DNA model is an elastic birod [53]. A. Grandchamp in his masters thesis [25] made a link between the parameters appearing in the Lagrangian system for a birod, and the discrete nearest-neighbour rigid base DNA parameters. Solving Lagrangian (or Hamiltonian) systems for a sequence dependent birod model could give interesting insights about the sequence dependent elastic behaviour of DNA, for example in the presence of external forces. Alternatively, using the rigid base pair (nonlocal) parameters obtained from a marginal distribution of our rigid base reconstruction, DNA could be modelled as a sequence dependent rod with nonlocal energy.

Non Gaussian deformations of DNA, such as strand separation (“melting") [77], [66], [4] and kinking [37], have also been widely investigated by various authors. One possible direction for future development of our rigid base nearest-neighbour model of DNA would be to try to encompass such deformations by including non quadratic terms in the free energy. However, sequence dependent parameter fitting in such an extension seems to be a non trivial task.

# Appendix A

# A parameter set for the nearest-neighbour model of DNA

Here we present tables with our model parameter set $\mathcal{P}^{\dagger}$, described in Sections 7.1 and 8.1. This parameter set is also available electronically (in Matlab® data format) on the webpage http://lcvmwww.epfl.ch/cgDNA.

## A.1 Monomer stiffness blocks $\mathsf{K}_1^{\gamma}$

| | | | | | |
|---|---|---|---|---|---|
| 0.107 | 0 | 0 | -0.268 | 0 | 0 |
| 0 | 0.328 | 0.206 | 0 | 1.389 | 0.346 |
| 0 | 0.206 | 2.085 | 0 | -1.435 | 0.850 |
| -0.268 | 0 | 0 | 6.470 | 0 | 0 |
| 0 | 1.389 | -1.435 | 0 | 23.810 | 0.531 |
| 0 | 0.346 | 0.850 | 0 | 0.531 | 0.675 |

Table A.1: Average monomer (symmetric) stiffness block $\mathsf{K}_1^{\mathrm{h}}$, computed over all four possible monomers ($\mathtt{A}, \mathtt{G}, \mathtt{T}$ and $\mathtt{C}$) and rounded to three decimal places. The ordering of parameters (from top to bottom and from left to right) is Buckle, Propeller, Opening, Shear, Stretch, Stagger. Zero entries corresponding to the couplings of the odd-even parameters (Buckle and Shear are odd and the rest are even) arise due to the palindromic symmetry, i.e. because $\mathsf{K}_1^{\mathrm{h}} = \mathsf{E}_1\, \mathsf{K}_1^{\mathrm{h}}\, \mathsf{E}_1$, where $\mathsf{E}_n$ is defined in (2.38).

| | | | | | |
|---|---|---|---|---|---|
| 0.141 | -0.000 | -0.353 | 0.149 | -1.381 | -0.074 |
| -0.000 | 0.031 | 0.138 | 0.436 | -0.087 | 0.145 |
| -0.353 | 0.138 | 1.511 | 1.566 | 3.117 | 0.835 |
| 0.149 | 0.436 | 1.566 | 6.269 | -2.640 | 1.956 |
| -1.381 | -0.087 | 3.117 | -2.640 | 13.845 | 0.344 |
| -0.074 | 0.145 | 0.835 | 1.956 | 0.344 | 0.717 |

Table A.2: $\mathsf{K}_1^{\gamma}$ for $\gamma = \mathtt{A}$. The ordering of parameters is as in Table A.1.

| | | | | | |
|---|---|---|---|---|---|
| 0.073 | -0.081 | -0.435 | -0.684 | 0.764 | -0.170 |
| -0.081 | 0.625 | 0.274 | 0.373 | 2.865 | 0.546 |
| -0.435 | 0.274 | 2.659 | 4.210 | -5.986 | 0.866 |
| -0.684 | 0.373 | 4.210 | 6.671 | -9.831 | 1.324 |
| 0.764 | 2.865 | -5.986 | -9.831 | 33.775 | 0.718 |
| -0.170 | 0.546 | 0.866 | 1.324 | 0.718 | 0.632 |

Table A.3: $\mathsf{K}_1^{\gamma}$ for $\gamma = \mathtt{G}$.

## A.2 Eigenvalues of monomer stiffness blocks $\mathsf{K}_1^\gamma$

| `A` | `G` | Average |
|---:|---:|---:|
| 1.3e-07 | 1.4e-07 | 4.8e-02 |
| 1.6e-07 | 1.8e-07 | 9.6e-02 |
| 5.3e-06 | 2.0e-06 | 3.5e-01 |
| 1.4e-03 | 2.0e-04 | 2.5e+00 |
| 7.2e+00 | 5.9e+00 | 6.5e+00 |
| 1.5e+01 | 3.9e+01 | 2.4e+01 |

Table A.4: Eigenvalues (sorted in increasing order) of monomer stiffness blocks $\mathsf{K}_1^\gamma$ for the two independent monomers $\gamma = \texttt{A}$ and $\gamma = \texttt{G}$. The last column contains eigenvalues of the average block $\mathsf{K}_1^{\mathrm{h}}$, given in Table A.1.

## A.3 Monomer weighted shape vectors $\sigma_1^\gamma$

| `A` | `G` | Average |
|---:|---:|---:|
| -0.117 | 0.233 | 0 |
| 0.743 | -0.478 | 0.132 |
| 1.287 | -3.189 | -0.951 |
| 0.386 | -4.495 | 0 |
| 5.011 | -1.963 | 1.524 |
| 0.369 | -1.579 | -0.605 |

Table A.5: Monomer weighted shape vectors $\sigma_1^\gamma$ for two independent monomers ($\gamma = \texttt{A}$ and $\gamma = \texttt{G}$), rounded to three decimal places. The last column is the average vector $\sigma_1^{\mathrm{h}}$, computed over all four monomers ($\texttt{A}, \texttt{G}, \texttt{T}$ and $\texttt{C}$). The ordering of parameters is as in Table A.2. Zeros in the average column appear due to the palindromic symmetry, i.e. because $\sigma_1^{\mathrm{h}} = \mathsf{E}_1\, \sigma_1^{\mathrm{h}}$, where $\mathsf{E}_n$ is defined in (2.38).

## A.4 Dimer stiffness blocks $\mathsf{K}_2^{\alpha\beta}$

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8.741 | 1.058 | -0.167 | -2.100 | -1.162 | -0.652 | 0.660 | -3.225 | -6.113 | -0.233 | 3.762 | -21.366 | -7.937 | 1.756 | -0.174 | 2.346 | -0.857 | 0.600 |
| 1.058 | 2.407 | 1.395 | 0.171 | -1.828 | 1.888 | -2.550 | 0.107 | -0.130 | 1.613 | 0.404 | -2.672 | -1.756 | -0.587 | -1.278 | 0.334 | 0.863 | -1.117 |
| -0.167 | 1.395 | 8.489 | 0.265 | -7.039 | 1.737 | -4.237 | -0.236 | 0.514 | 4.710 | -0.963 | 0.399 | 0.174 | -1.278 | -3.555 | 0.363 | 1.503 | -2.293 |
| -2.100 | 0.171 | 0.265 | 5.839 | -0.925 | 0.218 | -1.557 | 1.543 | 4.551 | 0.552 | -1.134 | 5.376 | 2.346 | -0.334 | -0.363 | -1.309 | 1.371 | -0.855 |
| -1.162 | -1.828 | -7.039 | -0.925 | 40.386 | -2.464 | 4.012 | 0.974 | 2.066 | -1.375 | -0.822 | 3.458 | 0.857 | 0.863 | 1.503 | -1.371 | -2.286 | 2.196 |
| -0.652 | 1.888 | 1.737 | 0.218 | -2.464 | 19.786 | -20.387 | 1.490 | -0.789 | 5.395 | 1.597 | 2.652 | -0.600 | -1.117 | -2.293 | 0.855 | 2.196 | -13.697 |
| 0.660 | -2.550 | -4.237 | -1.557 | 4.012 | -20.387 | 33.571 | 0 | 0 | -8.461 | 0 | 0 | 0.660 | 2.550 | 4.237 | -1.557 | -4.012 | 20.387 |
| -3.225 | 0.107 | -0.236 | 1.543 | 0.974 | 1.490 | 0 | 8.191 | 5.521 | 0 | -1.777 | 8.762 | 3.225 | 0.107 | -0.236 | -1.543 | 0.974 | 1.490 |
| -6.113 | -0.130 | 0.514 | 4.551 | 2.066 | -0.789 | 0 | 5.521 | 16.653 | 0 | -6.320 | 9.270 | 6.113 | -0.130 | 0.514 | -4.551 | 2.066 | -0.789 |
| -0.233 | 1.613 | 4.710 | 0.552 | -1.375 | 5.395 | -8.461 | 0 | 0 | 7.950 | 0 | 0 | -0.233 | -1.613 | -4.710 | 0.552 | 1.375 | -5.395 |
| 3.762 | 0.404 | -0.963 | -1.134 | -0.822 | 1.597 | 0 | -1.777 | -6.320 | 0 | 9.193 | -5.460 | -3.762 | 0.404 | -0.963 | 1.134 | -0.822 | 1.597 |
| -21.366 | -2.672 | 0.399 | 5.376 | 3.458 | 2.652 | 0 | 8.762 | 9.270 | 0 | -5.460 | 75.855 | 21.366 | -2.672 | 0.399 | -5.376 | 3.458 | 2.652 |
| -7.937 | -1.756 | 0.174 | 2.346 | 0.857 | -0.600 | 0.660 | 3.225 | 6.113 | -0.233 | -3.762 | 21.366 | 8.741 | -1.058 | 0.167 | -2.100 | 1.162 | 0.652 |
| 1.756 | -0.587 | -1.278 | -0.334 | 0.863 | -1.117 | 2.550 | 0.107 | -0.130 | -1.613 | 0.404 | -2.672 | -1.058 | 2.407 | 1.395 | -0.171 | -1.828 | 1.888 |
| -0.174 | -1.278 | -3.555 | -0.363 | 1.503 | -2.293 | 4.237 | -0.236 | 0.514 | -4.710 | -0.963 | 0.399 | 0.167 | 1.395 | 8.489 | -0.265 | -7.039 | 1.737 |
| 2.346 | 0.334 | 0.363 | -1.309 | -1.371 | 0.855 | -1.557 | -1.543 | -4.551 | 0.552 | 1.134 | -5.376 | -2.100 | -0.171 | -0.265 | 5.839 | 0.925 | -0.218 |
| -0.857 | 0.863 | 1.503 | 1.371 | -2.286 | 2.196 | -4.012 | 0.974 | 2.066 | 1.375 | -0.822 | 3.458 | 1.162 | -1.828 | -7.039 | 0.925 | 40.386 | -2.464 |
| 0.600 | -1.117 | -2.293 | -0.855 | 2.196 | -13.697 | 20.387 | 1.490 | -0.789 | -5.395 | 1.597 | 2.652 | 0.652 | 1.888 | 1.737 | -0.218 | -2.464 | 19.786 |

Table A.6: Average dimer (symmetric) stiffness block $\mathsf{K}_2^{\mathrm{h}}$, computed over all 16 possible dimer blocks and rounded to three decimal places. The ordering of intra parameters with indices $1, ..., 6$ and $13, ..., 18$ (entries separated by lines) is as in Table A.2, with inter indices $7, ..., 12$ being in order Tilt, Roll, Twist, Shift, Slide, Rise. Zero entries in the central sub-block corresponding to the couplings of the odd-even parameters (Tilt and Shift are odd and the rest are even) arise due to the palindromic symmetry, i.e. because $\mathsf{K}_2^{\mathrm{h}} = \mathsf{E}_2\,\mathsf{K}_2^{\mathrm{h}}\,\mathsf{E}_2$, where $\mathsf{E}_n$ is defined in (2.38). Note that, also due to the palindromic symmetry, the other $6 \times 6$ sub-blocks $\mathsf{K}_{2,ij}^{\mathrm{h}}$ also have symmetry relations, e.g. $\mathsf{K}_{2,11}^{\mathrm{h}} = \mathsf{E}_1\,\mathsf{K}_{2,33}^{\mathrm{h}}\,\mathsf{E}_1$ and $\mathsf{K}_{2,12}^{\mathrm{h}} = \mathsf{E}_1\,\mathsf{K}_{2,23}^{\mathrm{h}}\,\mathsf{E}_1$.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10.520 | 1.743 | -0.443 | -2.030 | 1.377 | -1.868 | 1.906 | -2.578 | -5.112 | -1.045 | 3.121 | -25.888 | -9.510 | 2.663 | 0.256 | 2.080 | -0.638 | 2.155 |
| 1.743 | 2.822 | 1.315 | 0.255 | -0.870 | 1.271 | -1.797 | -0.095 | 0.598 | 1.771 | 0.406 | -5.230 | -2.663 | -0.986 | -1.360 | -0.234 | 1.055 | -1.615 |
| -0.443 | 1.315 | 5.681 | 0.065 | -4.623 | 1.725 | -3.402 | -0.227 | 0.812 | 5.659 | -0.960 | 0.334 | -0.256 | -1.360 | -4.479 | -0.229 | 1.706 | -2.311 |
| -2.030 | 0.255 | 0.065 | 6.718 | -2.509 | -1.270 | -0.287 | 2.422 | 6.116 | -0.381 | -0.903 | 4.699 | 2.080 | 0.234 | 0.229 | -2.411 | 1.394 | 0.294 |
| 1.377 | -0.870 | -4.623 | -2.509 | 32.702 | -1.653 | 3.687 | 0.784 | 2.426 | -0.766 | -1.425 | 3.454 | 0.638 | 1.055 | 1.706 | -1.394 | -2.913 | 0.474 |
| -1.868 | 1.271 | 1.725 | -1.270 | -1.653 | 22.000 | -25.443 | 0.581 | -0.514 | 5.683 | 2.012 | -0.884 | -2.155 | -1.615 | -2.311 | -0.294 | 0.474 | -18.369 |
| 1.906 | -1.797 | -3.402 | -0.287 | 3.687 | -25.443 | 40.897 | 0 | 0 | -7.124 | 0 | 0 | 1.906 | 1.797 | 3.402 | -0.287 | -3.687 | 25.443 |
| -2.578 | -0.095 | -0.227 | 2.422 | 0.784 | 0.581 | 0 | 9.300 | 6.713 | 0 | -2.086 | 6.561 | 2.578 | -0.095 | -0.227 | -2.422 | 0.784 | 0.581 |
| -5.112 | 0.598 | 0.812 | 6.116 | 2.426 | -0.514 | 0 | 6.713 | 21.740 | 0 | -6.556 | 8.015 | 5.112 | 0.598 | 0.812 | -6.116 | 2.426 | -0.514 |
| -1.045 | 1.771 | 5.659 | -0.381 | -0.766 | 5.683 | -7.124 | 0 | 0 | 9.189 | 0 | 0 | -1.045 | -1.771 | -5.659 | -0.381 | 0.766 | -5.683 |
| 3.121 | 0.406 | -0.960 | -0.903 | -1.425 | 2.012 | 0 | -2.086 | -6.556 | 0 | 12.778 | -2.652 | -3.121 | 0.406 | -0.960 | 0.903 | -1.425 | 2.012 |
| -25.888 | -5.230 | 0.334 | 4.699 | 3.454 | -0.884 | 0 | 6.561 | 8.015 | 0 | -2.652 | 84.384 | 25.888 | -5.230 | 0.334 | -4.699 | 3.454 | -0.884 |
| -9.510 | -2.663 | -0.256 | 2.080 | 0.638 | -2.155 | 1.906 | 2.578 | 5.112 | -1.045 | -3.121 | 25.888 | 10.520 | -1.743 | 0.443 | -2.030 | -1.377 | 1.868 |
| 2.663 | -0.986 | -1.360 | 0.234 | 1.055 | -1.615 | 1.797 | -0.095 | 0.598 | -1.771 | 0.406 | -5.230 | -1.743 | 2.822 | 1.315 | -0.255 | -0.870 | 1.271 |
| 0.256 | -1.360 | -4.479 | 0.229 | 1.706 | -2.311 | 3.402 | -0.227 | 0.812 | -5.659 | -0.960 | 0.334 | 0.443 | 1.315 | 5.681 | -0.065 | -4.623 | 1.725 |
| 2.080 | -0.234 | -0.229 | -2.411 | -1.394 | -0.294 | -0.287 | -2.422 | -6.116 | -0.381 | 0.903 | -4.699 | -2.030 | -0.255 | -0.065 | 6.718 | 2.509 | 1.270 |
| -0.638 | 1.055 | 1.706 | 1.394 | -2.913 | 0.474 | -3.687 | 0.784 | 2.426 | 0.766 | -1.425 | 3.454 | -1.377 | -0.870 | -4.623 | 2.509 | 32.702 | -1.653 |
| 2.155 | -1.615 | -2.311 | 0.294 | 0.474 | -18.369 | 25.443 | 0.581 | -0.514 | -5.683 | 2.012 | -0.884 | 1.868 | 1.271 | 1.725 | 1.270 | -1.653 | 22.000 |

Table A.7: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \texttt{AT}$, rounded to three decimal places. The ordering of parameters is as in Table A.6. As the step `AT` is self-symmetric (i.e. the same reading along both strands), the matrix $\mathsf{K}_2^{\texttt{AT}} = \mathsf{E}_2\,\mathsf{K}_2^{\texttt{AT}}\,\mathsf{E}_2$ has the same block palindromic symmetries as $\mathsf{K}_2^{\mathrm{h}}$ (see the caption of Table A.6).

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10.052 | 1.359 | 0.535 | -1.923 | -3.214 | -2.031 | 1.516 | -3.763 | -6.442 | -0.524 | 3.351 | -24.843 | -8.809 | 2.831 | 0.496 | 2.906 | -1.298 | 1.810 |
| 1.359 | 2.229 | 0.550 | -0.741 | -2.196 | 2.851 | -3.582 | -0.689 | -0.236 | 2.004 | 0.110 | -6.032 | -2.831 | -0.703 | -0.920 | 0.138 | 0.456 | -2.943 |
| 0.535 | 0.550 | 13.202 | 0.816 | -11.659 | -0.076 | -3.952 | -0.449 | -0.193 | 4.246 | -0.346 | -0.852 | -0.496 | -0.920 | -3.416 | 0.205 | 1.902 | -1.987 |
| -1.923 | -0.741 | 0.816 | 6.102 | -4.066 | -1.761 | -0.283 | 1.715 | 4.643 | -0.093 | -1.218 | 6.133 | 2.906 | -0.138 | -0.205 | -1.798 | 2.188 | 0.170 |
| -3.214 | -2.196 | -11.659 | -4.066 | 59.764 | 1.732 | 4.063 | 0.925 | 3.064 | -1.304 | -1.035 | 4.252 | 1.298 | 0.456 | 1.902 | -2.188 | -3.580 | 1.723 |
| -2.031 | 2.851 | -0.076 | -1.761 | 1.732 | 21.569 | -22.113 | 0.497 | 0.166 | 4.379 | 0.454 | -2.579 | -1.810 | -2.943 | -1.987 | -0.170 | 1.723 | -16.197 |
| 1.516 | -3.582 | -3.952 | -0.283 | 4.063 | -22.113 | 34.772 | 0 | 0 | -8.033 | 0 | 0 | 1.516 | 3.582 | 3.952 | -0.283 | -4.063 | 22.113 |
| -3.763 | -0.689 | -0.449 | 1.715 | 0.925 | 0.497 | 0 | 7.057 | 4.659 | 0 | -0.544 | 11.491 | 3.763 | -0.689 | -0.449 | -1.715 | 0.925 | 0.497 |
| -6.442 | -0.236 | -0.193 | 4.643 | 3.064 | 0.166 | 0 | 4.659 | 15.831 | 0 | -5.308 | 11.672 | 6.442 | -0.236 | -0.193 | -4.643 | 3.064 | 0.166 |
| -0.524 | 2.004 | 4.246 | -0.093 | -1.304 | 4.379 | -8.033 | 0 | 0 | 6.933 | 0 | 0 | -0.524 | -2.004 | -4.246 | -0.093 | 1.304 | -4.379 |
| 3.351 | 0.110 | -0.346 | -1.218 | -1.035 | 0.454 | 0 | -0.544 | -5.308 | 0 | 7.419 | -5.145 | -3.351 | 0.110 | -0.346 | 1.218 | -1.035 | 0.454 |
| -24.843 | -6.032 | -0.852 | 6.133 | 4.252 | -2.579 | 0 | 11.491 | 11.672 | 0 | -5.145 | 84.189 | 24.843 | -6.032 | -0.852 | -6.133 | 4.252 | -2.579 |
| -8.809 | -2.831 | -0.496 | 2.906 | 1.298 | -1.810 | 1.516 | 3.763 | 6.442 | -0.524 | -3.351 | 24.843 | 10.052 | -1.359 | -0.535 | -1.923 | 3.214 | 2.031 |
| 2.831 | -0.703 | -0.920 | -0.138 | 0.456 | -2.943 | 3.582 | -0.689 | -0.236 | -2.004 | 0.110 | -6.032 | -1.359 | 2.229 | 0.550 | 0.741 | -2.196 | 2.851 |
| 0.496 | -0.920 | -3.416 | -0.205 | 1.902 | -1.987 | 3.952 | -0.449 | -0.193 | -4.246 | -0.346 | -0.852 | -0.535 | 0.550 | 13.202 | -0.816 | -11.659 | -0.076 |
| 2.906 | 0.138 | 0.205 | -1.798 | -2.188 | -0.170 | -0.283 | -1.715 | -4.643 | -0.093 | 1.218 | -6.133 | -1.923 | 0.741 | -0.816 | 6.102 | 4.066 | 1.761 |
| -1.298 | 0.456 | 1.902 | 2.188 | -3.580 | 1.723 | -4.063 | 0.925 | 3.064 | 1.304 | -1.035 | 4.252 | 3.214 | -2.196 | -11.659 | 4.066 | 59.764 | 1.732 |
| 1.810 | -2.943 | -1.987 | 0.170 | 1.723 | -16.197 | 22.113 | 0.497 | 0.166 | -4.379 | 0.454 | -2.579 | 2.031 | 2.851 | -0.076 | 1.761 | 1.732 | 21.569 |

Table A.8: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{GC}$. Because the step `GC` is self-symmetric, $\mathsf{K}_2^{\mathtt{GC}}$ has symmetries described in the caption of Figure A.6.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6.718 | 0.360 | -0.604 | -2.086 | -0.324 | 0.795 | -0.447 | -2.902 | -5.876 | 0.113 | 4.218 | -16.261 | -6.367 | 0.793 | -0.433 | 2.098 | -0.865 | 0.006 |
| 0.360 | 1.772 | 1.036 | 0.969 | 0.056 | 1.604 | -2.553 | 0.245 | -0.071 | 1.211 | 0.291 | 0.077 | -0.793 | -0.288 | -1.286 | 0.457 | 0.630 | -0.433 |
| -0.604 | 1.036 | 5.000 | 0.230 | -4.318 | 1.226 | -4.586 | -0.279 | 0.726 | 3.612 | -0.927 | 0.559 | 0.433 | -1.286 | -2.769 | 0.424 | 1.293 | -2.785 |
| -2.086 | 0.969 | 0.230 | 5.180 | 1.874 | 2.568 | -2.828 | 0.775 | 2.945 | 0.930 | -0.620 | 4.862 | 2.098 | -0.457 | -0.424 | -0.632 | 1.181 | -1.658 |
| -0.324 | 0.056 | -4.318 | 1.874 | 28.974 | -0.778 | 4.205 | 0.673 | 0.787 | -1.473 | -0.588 | 3.169 | 0.865 | 0.630 | 1.293 | -1.181 | -1.148 | 2.714 |
| 0.795 | 1.604 | 1.226 | 2.568 | -0.778 | 17.267 | -17.254 | 1.856 | -1.741 | 5.426 | 2.932 | 6.800 | -0.006 | -0.433 | -2.785 | 1.658 | 2.714 | -7.780 |
| -0.447 | -2.553 | -4.586 | -2.828 | 4.205 | -17.254 | 27.315 | 0 | 0 | -7.954 | 0 | 0 | -0.447 | 2.553 | 4.586 | -2.828 | -4.205 | 17.254 |
| -2.902 | 0.245 | -0.279 | 0.775 | 0.673 | 1.856 | 0 | 6.688 | 5.252 | 0 | -0.757 | 6.639 | 2.902 | 0.245 | -0.279 | -0.775 | 0.673 | 1.856 |
| -5.876 | -0.071 | 0.726 | 2.945 | 0.787 | -1.741 | 0 | 5.252 | 12.258 | 0 | -5.096 | 7.347 | 5.876 | -0.071 | 0.726 | -2.945 | 0.787 | -1.741 |
| 0.113 | 1.211 | 3.612 | 0.930 | -1.473 | 5.426 | -7.954 | 0 | 0 | 5.538 | 0 | 0 | 0.113 | -1.211 | -3.612 | 0.930 | 1.473 | -5.426 |
| 4.218 | 0.291 | -0.927 | -0.620 | -0.588 | 2.932 | 0 | -0.757 | -5.096 | 0 | 7.220 | -5.769 | -4.218 | 0.291 | -0.927 | 0.620 | -0.588 | 2.932 |
| -16.261 | 0.077 | 0.559 | 4.862 | 3.169 | 6.800 | 0 | 6.639 | 7.347 | 0 | -5.769 | 61.877 | 16.261 | 0.077 | 0.559 | -4.862 | 3.169 | 6.800 |
| -6.367 | -0.793 | 0.433 | 2.098 | 0.865 | -0.006 | -0.447 | 2.902 | 5.876 | 0.113 | -4.218 | 16.261 | 6.718 | -0.360 | 0.604 | -2.086 | 0.324 | -0.795 |
| 0.793 | -0.288 | -1.286 | -0.457 | 0.630 | -0.433 | 2.553 | 0.245 | -0.071 | -1.211 | 0.291 | 0.077 | -0.360 | 1.772 | 1.036 | -0.969 | 0.056 | 1.604 |
| -0.433 | -1.286 | -2.769 | -0.424 | 1.293 | -2.785 | 4.586 | -0.279 | 0.726 | -3.612 | -0.927 | 0.559 | 0.604 | 1.036 | 5.000 | -0.230 | -4.318 | 1.226 |
| 2.098 | 0.457 | 0.424 | -0.632 | -1.181 | 1.658 | -2.828 | -0.775 | -2.945 | 0.930 | 0.620 | -4.862 | -2.086 | -0.969 | -0.230 | 5.180 | -1.874 | -2.568 |
| -0.865 | 0.630 | 1.293 | 1.181 | -1.148 | 2.714 | -4.205 | 0.673 | 0.787 | 1.473 | -0.588 | 3.169 | 0.324 | 0.056 | -4.318 | -1.874 | 28.974 | -0.778 |
| 0.006 | -0.433 | -2.785 | -1.658 | 2.714 | -7.780 | 17.254 | 1.856 | -1.741 | -5.426 | 2.932 | 6.800 | -0.795 | 1.604 | 1.226 | -2.568 | -0.778 | 17.267 |

Table A.9: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{TA}$. Because the step `TA` is self-symmetric, $\mathsf{K}_2^{\mathtt{TA}}$ has symmetries described in the caption of Figure A.6.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6.914 | 0.294 | -0.540 | -2.163 | -1.771 | 0.346 | -0.079 | -3.049 | -6.247 | 0.088 | 3.902 | -16.803 | -6.273 | 0.589 | -0.690 | 2.306 | -0.805 | -0.265 |
| 0.294 | 1.983 | 1.645 | 0.107 | -2.543 | 2.132 | -2.326 | 1.248 | 0.135 | 1.000 | 0.424 | 1.554 | -0.589 | 0.102 | -1.392 | -0.162 | 0.667 | 0.639 |
| -0.540 | 1.645 | 9.062 | -0.391 | -8.187 | 3.245 | -4.737 | -0.228 | 1.217 | 3.619 | -1.083 | 1.180 | 0.690 | -1.392 | -2.558 | 0.145 | 0.689 | -2.046 |
| -2.163 | 0.107 | -0.391 | 3.910 | 0.736 | 1.003 | -2.126 | 1.199 | 2.863 | 0.431 | -0.673 | 4.970 | 2.306 | 0.162 | -0.145 | -1.225 | 1.050 | -1.451 |
| -1.771 | -2.543 | -8.187 | 0.736 | 38.062 | -4.523 | 4.034 | 1.235 | 0.942 | -0.993 | 0.263 | 4.074 | 0.805 | 0.667 | 0.689 | -1.050 | -1.648 | 3.376 |
| 0.346 | 2.132 | 3.245 | 1.003 | -4.523 | 17.530 | -16.473 | 2.813 | -1.848 | 4.564 | 1.791 | 8.745 | 0.265 | 0.639 | -2.046 | 1.451 | 3.376 | -6.976 |
| -0.079 | -2.326 | -4.737 | -2.126 | 4.034 | -16.473 | 27.735 | 0 | 0 | -8.563 | 0 | 0 | -0.079 | 2.326 | 4.737 | -2.126 | -4.034 | 16.473 |
| -3.049 | 1.248 | -0.228 | 1.199 | 1.235 | 2.813 | 0 | 7.896 | 4.484 | 0 | -1.191 | 9.236 | 3.049 | 1.248 | -0.228 | -1.199 | 1.235 | 2.813 |
| -6.247 | 0.135 | 1.217 | 2.863 | 0.942 | -1.848 | 0 | 4.484 | 11.701 | 0 | -5.065 | 7.935 | 6.247 | 0.135 | 1.217 | -2.863 | 0.942 | -1.848 |
| 0.088 | 1.000 | 3.619 | 0.431 | -0.993 | 4.564 | -8.563 | 0 | 0 | 6.134 | 0 | 0 | 0.088 | -1.000 | -3.619 | 0.431 | 0.993 | -4.564 |
| 3.902 | 0.424 | -1.083 | -0.673 | 0.263 | 1.791 | 0 | -1.191 | -5.065 | 0 | 8.035 | -6.318 | -3.902 | 0.424 | -1.083 | 0.673 | 0.263 | 1.791 |
| -16.803 | 1.554 | 1.180 | 4.970 | 4.074 | 8.745 | 0 | 9.236 | 7.935 | 0 | -6.318 | 68.167 | 16.803 | 1.554 | 1.180 | -4.970 | 4.074 | 8.745 |
| -6.273 | -0.589 | 0.690 | 2.306 | 0.805 | 0.265 | -0.079 | 3.049 | 6.247 | 0.088 | -3.902 | 16.803 | 6.914 | -0.294 | 0.540 | -2.163 | 1.771 | -0.346 |
| 0.589 | 0.102 | -1.392 | 0.162 | 0.667 | 0.639 | 2.326 | 1.248 | 0.135 | -1.000 | 0.424 | 1.554 | -0.294 | 1.983 | 1.645 | -0.107 | -2.543 | 2.132 |
| -0.690 | -1.392 | -2.558 | -0.145 | 0.689 | -2.046 | 4.737 | -0.228 | 1.217 | -3.619 | -1.083 | 1.180 | 0.540 | 1.645 | 9.062 | 0.391 | -8.187 | 3.245 |
| 2.306 | -0.162 | 0.145 | -1.225 | -1.050 | 1.451 | -2.126 | -1.199 | -2.863 | 0.431 | 0.673 | -4.970 | -2.163 | -0.107 | 0.391 | 3.910 | -0.736 | -1.003 |
| -0.805 | 0.667 | 0.689 | 1.050 | -1.648 | 3.376 | -4.034 | 1.235 | 0.942 | 0.993 | 0.263 | 4.074 | 1.771 | -2.543 | -8.187 | -0.736 | 38.062 | -4.523 |
| -0.265 | 0.639 | -2.046 | -1.451 | 3.376 | -6.976 | 16.473 | 2.813 | -1.848 | -4.564 | 1.791 | 8.745 | -0.346 | 2.132 | 3.245 | -1.003 | -4.523 | 17.530 |

Table A.10: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{CG}$. Because the step `CG` is self-symmetric, $\mathsf{K}_2^{\mathtt{CG}}$ has symmetries described in the caption of Figure A.6.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9.783 | 0.977 | -0.545 | -1.653 | -0.581 | -1.574 | 0.262 | -2.725 | -5.197 | -1.485 | 3.198 | -23.637 | -8.885 | 2.833 | 0.682 | 2.020 | -1.125 | 1.352 |
| 0.977 | 2.219 | 0.717 | -0.397 | -1.994 | 2.621 | -3.383 | -1.427 | -0.111 | 1.709 | 0.334 | -5.224 | -2.438 | -0.919 | -0.994 | 0.221 | 0.827 | -2.662 |
| -0.545 | 0.717 | 13.172 | 1.005 | -13.081 | 0.192 | -3.585 | -1.086 | -0.717 | 4.908 | -1.158 | 0.600 | 0.260 | -1.300 | -4.002 | 0.491 | 1.576 | -2.076 |
| -1.653 | -0.397 | 1.005 | 6.052 | -4.505 | -1.713 | -1.375 | 1.951 | 5.449 | -0.238 | -1.126 | 5.761 | 2.441 | -0.254 | 0.309 | -1.279 | 1.598 | -0.610 |
| -0.581 | -1.994 | -13.081 | -4.505 | 54.558 | 2.047 | 4.608 | 1.422 | 2.784 | -1.402 | -1.667 | 2.121 | 0.408 | 0.881 | 1.901 | -1.664 | -3.164 | 1.111 |
| -1.574 | 2.621 | 0.192 | -1.713 | 2.047 | 22.209 | -23.324 | -0.255 | 0.220 | 5.653 | 0.397 | -3.764 | -2.774 | -1.621 | -1.904 | -0.565 | 0.486 | -17.517 |
| 0.262 | -3.383 | -3.585 | -1.375 | 4.608 | -23.324 | 38.902 | 0.050 | -1.222 | -8.205 | 0.926 | 2.608 | 2.743 | 1.514 | 3.767 | 0.101 | -4.037 | 23.995 |
| -2.725 | -1.427 | -1.086 | 1.951 | 1.422 | -0.255 | 0.050 | 8.459 | 5.855 | -0.660 | -0.882 | 9.231 | 3.482 | 0.617 | 0.255 | -2.300 | 0.644 | 1.506 |
| -5.197 | -0.111 | -0.717 | 5.449 | 2.784 | 0.220 | -1.222 | 5.855 | 19.248 | -1.310 | -6.108 | 8.028 | 5.683 | 0.087 | 1.530 | -5.148 | 3.243 | -0.157 |
| -1.485 | 1.709 | 4.908 | -0.238 | -1.402 | 5.653 | -8.205 | -0.660 | -1.310 | 8.208 | -0.428 | 0.754 | 0.169 | -2.311 | -5.372 | -0.370 | 1.057 | -5.123 |
| 3.198 | 0.334 | -1.158 | -1.126 | -1.667 | 0.397 | 0.926 | -0.882 | -6.108 | -0.428 | 10.700 | -2.891 | -3.222 | -0.043 | -0.376 | 1.047 | -1.582 | 1.948 |
| -23.637 | -5.224 | 0.600 | 5.761 | 2.121 | -3.764 | 2.608 | 9.231 | 8.028 | 0.754 | -2.891 | 82.929 | 25.570 | -4.848 | -0.427 | -4.512 | 4.509 | 0.430 |
| -8.885 | -2.438 | 0.260 | 2.441 | 0.408 | -2.774 | 2.743 | 3.482 | 5.683 | 0.169 | -3.222 | 25.570 | 10.302 | -1.895 | -0.476 | -2.141 | 0.213 | 2.070 |
| 2.833 | -0.919 | -1.300 | -0.254 | 0.881 | -1.621 | 1.514 | 0.617 | 0.087 | -2.311 | -0.043 | -4.848 | -1.895 | 2.698 | 1.661 | -0.025 | -0.654 | 1.332 |
| 0.682 | -0.994 | -4.002 | 0.309 | 1.901 | -1.904 | 3.767 | 0.255 | 1.530 | -5.372 | -0.376 | -0.427 | -0.476 | 1.661 | 5.587 | -0.158 | -2.830 | 1.302 |
| 2.020 | 0.221 | 0.491 | -1.279 | -1.664 | -0.565 | 0.101 | -2.300 | -5.148 | -0.370 | 1.047 | -4.512 | -2.141 | -0.025 | -0.158 | 6.330 | 2.552 | 1.174 |
| -1.125 | 0.827 | 1.576 | 1.598 | -3.164 | 0.486 | -4.037 | 0.644 | 3.243 | 1.057 | -1.582 | 4.509 | 0.213 | -0.654 | -2.830 | 2.552 | 34.481 | -2.112 |
| 1.352 | -2.662 | -2.076 | -0.610 | 1.111 | -17.517 | 23.995 | 1.506 | -0.157 | -5.123 | 1.948 | 0.430 | 2.070 | 1.332 | 1.302 | 1.174 | -2.112 | 20.712 |

Table A.11: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{GT}$. Note that $\mathsf{K}_2^{\mathtt{AC}} = \mathsf{E}_2\,\mathsf{K}_2^{\mathtt{GT}}\,\mathsf{E}_2$, where $\mathsf{E}_n$ is defined in (2.38).

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6.985 | 0.234 | -0.842 | -1.813 | -0.233 | 0.586 | -0.695 | -3.270 | -5.953 | -0.085 | 3.838 | -17.279 | -6.416 | 0.419 | -0.720 | 2.229 | -0.691 | -0.886 |
| 0.234 | 2.131 | 0.949 | 0.996 | -0.286 | 1.850 | -2.889 | 0.371 | 0.083 | 0.932 | 0.485 | 0.658 | -0.882 | -0.289 | -1.404 | 0.191 | 0.666 | 0.173 |
| -0.842 | 0.949 | 4.928 | 0.486 | -3.231 | 1.016 | -4.519 | -0.207 | 1.187 | 3.477 | -0.940 | 0.666 | 0.665 | -1.240 | -2.548 | 0.266 | 1.088 | -2.233 |
| -1.813 | 0.996 | 0.486 | 5.838 | 2.479 | 2.534 | -3.308 | 0.312 | 2.629 | 2.096 | -0.790 | 4.811 | 2.012 | -0.580 | -0.435 | 0.050 | 1.189 | -1.998 |
| -0.233 | -0.286 | -3.231 | 2.479 | 33.547 | -1.534 | 4.099 | 1.085 | 0.405 | -0.924 | 0.382 | 3.728 | 0.708 | 0.820 | 1.011 | -0.739 | -1.524 | 3.481 |
| 0.586 | 1.850 | 1.016 | 2.534 | -1.534 | 18.068 | -17.525 | 2.459 | -2.260 | 5.097 | 1.566 | 8.567 | 0.015 | 0.431 | -2.712 | 1.651 | 3.592 | -7.224 |
| -0.695 | -2.889 | -4.519 | -3.308 | 4.099 | -17.525 | 28.391 | 0.425 | -0.135 | -8.093 | 0.680 | 0.446 | 0.198 | 2.295 | 4.908 | -2.375 | -4.368 | 16.710 |
| -3.270 | 0.371 | -0.207 | 0.312 | 1.085 | 2.459 | 0.425 | 7.847 | 4.660 | -0.269 | -0.979 | 8.444 | 3.085 | 1.152 | -0.097 | -1.291 | 1.024 | 2.667 |
| -5.953 | 0.083 | 1.187 | 2.629 | 0.405 | -2.260 | -0.135 | 4.660 | 11.713 | 0.488 | -4.936 | 6.991 | 6.036 | -0.094 | 0.936 | -2.704 | 0.832 | -2.139 |
| -0.085 | 0.932 | 3.477 | 2.096 | -0.924 | 5.097 | -8.093 | -0.269 | 0.488 | 6.160 | -0.335 | 0.103 | 0.362 | -1.177 | -3.369 | 0.751 | 1.406 | -4.868 |
| 3.838 | 0.485 | -0.940 | -0.790 | 0.382 | 1.566 | 0.680 | -0.979 | -4.936 | -0.335 | 7.922 | -5.332 | -4.189 | 0.293 | -1.213 | 0.342 | -0.504 | 3.655 |
| -17.279 | 0.658 | 0.666 | 4.811 | 3.728 | 8.567 | 0.446 | 8.444 | 6.991 | 0.103 | -5.332 | 68.531 | 17.157 | 1.567 | 1.334 | -4.983 | 3.969 | 9.503 |
| -6.416 | -0.882 | 0.665 | 2.012 | 0.708 | 0.015 | 0.198 | 3.085 | 6.036 | 0.362 | -4.189 | 17.157 | 7.060 | -0.278 | 0.471 | -2.096 | 1.882 | -0.282 |
| 0.419 | -0.289 | -1.240 | -0.580 | 0.820 | 0.431 | 2.295 | 1.152 | -0.094 | -1.177 | 0.293 | 1.567 | -0.278 | 1.981 | 1.854 | -0.249 | -2.728 | 2.063 |
| -0.720 | -1.404 | -2.548 | -0.435 | 1.011 | -2.712 | 4.908 | -0.097 | 0.936 | -3.369 | -1.213 | 1.334 | 0.471 | 1.854 | 9.074 | 0.308 | -8.014 | 3.437 |
| 2.229 | 0.191 | 0.266 | 0.050 | -0.739 | 1.651 | -2.375 | -1.291 | -2.704 | 0.751 | 0.342 | -4.983 | -2.096 | -0.249 | 0.308 | 3.945 | 0.132 | -1.256 |
| -0.691 | 0.666 | 1.088 | 1.189 | -1.524 | 3.592 | -4.368 | 1.024 | 0.832 | 1.406 | -0.504 | 3.969 | 1.882 | -2.728 | -8.014 | 0.132 | 33.500 | -5.399 |
| -0.886 | 0.173 | -2.233 | -1.998 | 3.481 | -7.224 | 16.710 | 2.667 | -2.139 | -4.868 | 3.655 | 9.503 | -0.282 | 2.063 | 3.437 | -1.256 | -5.399 | 18.677 |

Table A.12: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{TG}$. Note that $\mathsf{K}_2^{\mathtt{CC}} = \mathsf{E}_2\,\mathsf{K}_2^{\mathtt{TG}}\,\mathsf{E}_2$.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9.059 | 1.705 | 0.161 | -2.525 | 0.295 | -2.081 | 4.955 | -3.268 | -6.050 | 1.032 | 3.544 | -22.689 | -8.563 | 0.977 | -1.140 | 2.388 | -0.460 | 1.191 |
| 1.705 | 3.025 | 2.088 | 0.074 | -1.368 | 1.407 | -1.677 | 1.580 | 0.238 | 2.400 | -0.394 | -4.867 | -2.580 | -0.479 | -2.029 | 0.248 | 0.825 | -0.079 |
| 0.161 | 2.088 | 5.619 | 0.206 | -6.410 | 1.718 | -4.102 | 0.925 | 1.786 | 5.722 | -1.010 | -0.126 | -0.250 | -0.994 | -4.118 | 0.518 | 2.344 | -1.169 |
| -2.525 | 0.074 | 0.206 | 6.249 | -2.558 | -1.583 | 0.226 | 1.401 | 5.125 | -0.568 | -1.765 | 4.752 | 2.307 | 0.155 | -0.279 | -2.194 | 1.301 | -0.003 |
| 0.295 | -1.368 | -6.410 | -2.558 | 35.258 | -3.290 | 4.231 | 0.406 | 2.642 | -1.696 | -0.516 | 4.170 | 1.018 | 0.926 | 1.749 | -1.928 | -2.308 | 1.822 |
| -2.081 | 1.407 | 1.718 | -1.583 | -3.290 | 19.598 | -22.364 | 0.492 | -1.137 | 5.558 | 1.956 | -0.307 | 0.600 | -0.609 | -2.568 | 1.627 | 2.767 | -14.246 |
| 4.955 | -1.677 | -4.102 | 0.226 | 4.231 | -22.364 | 35.975 | 0.086 | 1.182 | -8.120 | 0.676 | -7.538 | -3.326 | 2.649 | 3.859 | -3.061 | -4.458 | 19.521 |
| -3.268 | 1.580 | 0.925 | 1.401 | 0.406 | 0.492 | 0.086 | 8.769 | 5.575 | 0.804 | -2.802 | 7.236 | 2.480 | -0.271 | -1.070 | -1.351 | 1.875 | 3.126 |
| -6.050 | 0.238 | 1.786 | 5.125 | 2.642 | -1.137 | 1.182 | 5.575 | 17.600 | 1.576 | -6.279 | 9.141 | 5.949 | 0.102 | -0.433 | -4.878 | 1.821 | -0.364 |
| 1.032 | 2.400 | 5.722 | -0.568 | -1.696 | 5.558 | -8.120 | 0.804 | 1.576 | 8.922 | 1.080 | -3.673 | -1.567 | -0.826 | -5.031 | 0.437 | 1.868 | -4.463 |
| 3.544 | -0.394 | -1.010 | -1.765 | -0.516 | 1.956 | 0.676 | -2.802 | -6.279 | 1.080 | 9.352 | -5.667 | -4.332 | 1.551 | -1.303 | 0.106 | -2.472 | 0.563 |
| -22.689 | -4.867 | -0.126 | 4.752 | 4.170 | -0.307 | -7.538 | 7.236 | 9.141 | -3.673 | -5.667 | 74.441 | 21.562 | -0.434 | 2.009 | -5.345 | 2.450 | 6.048 |
| -8.563 | -2.580 | -0.250 | 2.307 | 1.018 | 0.600 | -3.326 | 2.480 | 5.949 | -1.567 | -4.332 | 21.562 | 9.266 | -0.962 | 1.082 | -1.979 | 1.258 | -1.177 |
| 0.977 | -0.479 | -0.994 | 0.155 | 0.926 | -0.609 | 2.649 | -0.271 | 0.102 | -0.826 | 1.551 | -0.434 | -0.962 | 2.483 | 2.027 | -0.275 | -4.287 | 1.590 |
| -1.140 | -2.029 | -4.118 | -0.279 | 1.749 | -2.568 | 3.859 | -1.070 | -0.433 | -5.031 | -1.303 | 2.009 | 1.082 | 2.027 | 9.925 | 0.837 | -8.115 | 3.823 |
| 2.388 | 0.248 | 0.518 | -2.194 | -1.928 | 1.627 | -3.061 | -1.351 | -4.878 | 0.437 | 0.106 | -5.345 | -1.979 | -0.275 | 0.837 | 5.617 | -1.487 | -1.533 |
| -0.460 | 0.825 | 2.344 | 1.301 | -2.308 | 2.767 | -4.458 | 1.875 | 1.821 | 1.868 | -2.472 | 2.450 | 1.258 | -4.287 | -8.115 | -1.487 | 38.052 | -5.626 |
| 1.191 | -0.079 | -1.169 | -0.003 | 1.822 | -14.246 | 19.521 | 3.126 | -0.364 | -4.463 | 0.563 | 6.048 | -1.177 | 1.590 | 3.823 | -1.533 | -5.626 | 18.558 |

Table A.13: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{AG}$. Note that $\mathsf{K}_2^{\mathtt{CT}} = \mathsf{E}_2\,\mathsf{K}_2^{\mathtt{AG}}\,\mathsf{E}_2$.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8.521 | 1.370 | 0.737 | -2.056 | -3.340 | -1.726 | 2.603 | -3.403 | -7.034 | 0.269 | 3.621 | -20.142 | -7.687 | 1.528 | 0.177 | 2.578 | -1.272 | 0.775 |
| 1.370 | 2.338 | 1.354 | -0.381 | -2.409 | 2.299 | -2.752 | -0.825 | -0.470 | 2.115 | -0.268 | -6.682 | -2.200 | -0.712 | -1.168 | 0.924 | 0.904 | -2.383 |
| 0.737 | 1.354 | 12.800 | 0.639 | -11.114 | 1.048 | -4.621 | 0.721 | -1.028 | 5.121 | -0.739 | 0.102 | 0.013 | -0.916 | -3.716 | 1.244 | 1.520 | -2.444 |
| -2.056 | -0.381 | 0.639 | 5.383 | -3.338 | -1.298 | -0.287 | 2.358 | 4.907 | -0.044 | -1.990 | 6.591 | 2.606 | -0.309 | -0.267 | -1.950 | 1.327 | 0.744 |
| -3.340 | -2.409 | -11.114 | -3.338 | 53.597 | -1.491 | 2.352 | 0.332 | 2.933 | -1.354 | 0.058 | 1.955 | 0.404 | 0.922 | 1.502 | -1.822 | -2.241 | 1.452 |
| -1.726 | 2.299 | 1.048 | -1.298 | -1.491 | 20.988 | -19.296 | -0.976 | 0.317 | 6.477 | 1.336 | -6.922 | -0.194 | -1.875 | -2.382 | 1.971 | 2.086 | -15.182 |
| 2.603 | -2.752 | -4.621 | -0.287 | 2.352 | -19.296 | 31.288 | -0.145 | -0.595 | -9.507 | 1.050 | -2.902 | -1.960 | 2.698 | 4.401 | -2.539 | -4.246 | 19.610 |
| -3.403 | -0.825 | 0.721 | 2.358 | 0.332 | -0.976 | -0.145 | 8.029 | 5.903 | 0.108 | -2.272 | 9.485 | 3.474 | -0.401 | -1.132 | -1.225 | 0.972 | 3.151 |
| -7.034 | -0.470 | -1.028 | 4.907 | 2.933 | 0.317 | -0.595 | 5.903 | 17.338 | -0.986 | -7.190 | 10.896 | 6.416 | -0.992 | 0.492 | -5.135 | 1.881 | -0.385 |
| 0.269 | 2.115 | 5.121 | -0.044 | -1.354 | 6.477 | -9.507 | 0.108 | -0.986 | 8.435 | 0.968 | -1.994 | -0.457 | -1.196 | -4.714 | 2.115 | 1.184 | -6.285 |
| 3.621 | -0.268 | -0.739 | -1.990 | 0.058 | 1.336 | 1.050 | -2.272 | -7.190 | 0.968 | 8.642 | -6.784 | -4.239 | 0.998 | -0.609 | 1.175 | -1.130 | 1.527 |
| -20.142 | -6.682 | 0.102 | 6.591 | 1.955 | -6.922 | -2.902 | 9.485 | 10.896 | -1.994 | -6.784 | 75.542 | 20.964 | -0.669 | -0.199 | -5.761 | 3.136 | 7.243 |
| -7.687 | -2.200 | 0.013 | 2.606 | 0.404 | -0.194 | -1.960 | 3.474 | 6.416 | -0.457 | -4.239 | 20.964 | 8.365 | -0.670 | 0.209 | -2.356 | 0.496 | -0.807 |
| 1.528 | -0.712 | -0.916 | -0.309 | 0.922 | -1.875 | 2.698 | -0.401 | -0.992 | -1.196 | 0.998 | -0.669 | -0.670 | 2.386 | 0.936 | -0.688 | -0.734 | 1.434 |
| 0.177 | -1.168 | -3.716 | -0.267 | 1.502 | -2.382 | 4.401 | -1.132 | 0.492 | -4.714 | -0.609 | -0.199 | 0.209 | 0.936 | 5.416 | -0.634 | -2.160 | 1.558 |
| 2.578 | 0.924 | 1.244 | -1.950 | -1.822 | 1.971 | -2.539 | -1.225 | -5.135 | 2.115 | 1.175 | -5.761 | -2.356 | -0.688 | -0.634 | 6.695 | -2.086 | -2.682 |
| -1.272 | 0.904 | 1.520 | 1.327 | -2.241 | 2.086 | -4.246 | 0.972 | 1.881 | 1.184 | -1.130 | 3.136 | 0.496 | -0.734 | -2.160 | -2.086 | 32.254 | -0.861 |
| 0.775 | -2.383 | -2.444 | 0.744 | 1.452 | -15.182 | 19.610 | 3.151 | -0.385 | -6.285 | 1.527 | 7.243 | -0.807 | 1.434 | 1.558 | -2.682 | -0.861 | 18.859 |

Table A.14: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{GA}$. Note that $\mathsf{K}_2^{\mathtt{TC}} = \mathsf{E}_2\,\mathsf{K}_2^{\mathtt{GA}}\,\mathsf{E}_2$.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9.426 | 1.792 | 0.362 | -2.812 | 0.277 | -2.165 | 5.066 | -3.998 | -6.572 | 0.430 | 3.033 | -23.381 | -8.755 | 1.459 | 0.071 | 2.279 | -1.140 | 1.219 |
| 1.792 | 3.045 | 2.175 | 0.122 | -1.501 | 1.355 | -1.736 | 1.076 | -0.013 | 3.145 | -0.336 | -5.493 | -2.464 | -0.854 | -1.679 | 1.134 | 0.998 | -0.909 |
| 0.362 | 2.175 | 6.043 | 0.359 | -4.996 | 1.472 | -3.937 | 0.607 | 2.040 | 5.962 | -1.613 | -0.498 | -0.185 | -1.180 | -3.667 | 0.621 | 1.526 | -2.425 |
| -2.812 | 0.122 | 0.359 | 6.720 | -2.394 | -1.216 | -0.130 | 2.376 | 5.763 | 0.431 | -2.331 | 6.158 | 2.493 | -0.257 | -0.404 | -1.045 | 1.035 | 0.324 |
| 0.277 | -1.501 | -4.996 | -2.394 | 35.605 | -5.447 | 1.984 | 0.824 | 2.730 | -0.706 | 1.077 | 4.264 | 1.063 | 1.186 | 0.560 | -1.500 | -2.067 | 1.198 |
| -2.165 | 1.355 | 1.472 | -1.216 | -5.447 | 20.213 | -21.645 | 0.510 | -0.999 | 5.912 | 3.259 | -2.126 | 1.109 | -1.342 | -2.247 | 1.941 | 2.639 | -15.695 |
| 5.066 | -1.736 | -3.937 | -0.130 | 1.984 | -21.645 | 35.555 | -0.831 | 0.299 | -8.676 | 0.284 | -7.934 | -4.080 | 3.313 | 3.155 | -2.945 | -4.069 | 20.934 |
| -3.998 | 1.076 | 0.607 | 2.376 | 0.824 | 0.510 | -0.831 | 9.201 | 6.535 | 0.914 | -3.027 | 8.554 | 3.218 | -1.010 | -1.498 | -0.995 | 1.572 | 2.281 |
| -6.572 | -0.013 | 2.040 | 5.763 | 2.730 | -0.999 | 0.299 | 6.535 | 19.981 | 1.538 | -7.990 | 10.825 | 6.443 | -0.345 | -0.931 | -5.354 | 2.016 | -0.608 |
| 0.430 | 3.145 | 5.962 | 0.431 | -0.706 | 5.912 | -8.676 | 0.914 | 1.538 | 10.681 | 1.283 | -2.390 | -0.911 | -1.256 | -4.962 | 1.458 | 1.396 | -6.071 |
| 3.033 | -0.336 | -1.613 | -2.331 | 1.077 | 3.259 | 0.284 | -3.027 | -7.990 | 1.283 | 11.304 | -5.841 | -3.986 | 1.760 | -1.492 | 0.656 | -1.185 | 0.802 |
| -23.381 | -5.493 | -0.498 | 6.158 | 4.264 | -2.126 | -7.934 | 8.554 | 10.825 | -2.390 | -5.841 | 78.231 | 22.137 | -1.305 | -0.145 | -5.659 | 2.870 | 5.943 |
| -8.755 | -2.464 | -0.185 | 2.493 | 1.063 | 1.109 | -4.080 | 3.218 | 6.443 | -0.911 | -3.986 | 22.137 | 9.246 | -0.945 | 0.509 | -2.130 | 0.073 | -1.710 |
| 1.459 | -0.854 | -1.180 | -0.257 | 1.186 | -1.342 | 3.313 | -1.010 | -0.345 | -1.256 | 1.760 | -1.305 | -0.945 | 2.796 | 0.616 | -1.184 | -0.727 | 1.676 |
| 0.071 | -1.679 | -3.667 | -0.404 | 0.560 | -2.247 | 3.155 | -1.498 | -0.931 | -4.962 | -1.492 | -0.145 | 0.509 | 0.616 | 6.049 | 0.272 | -3.646 | 0.733 |
| 2.279 | 1.134 | 0.621 | -1.045 | -1.500 | 1.941 | -2.945 | -0.995 | -5.354 | 1.458 | 0.656 | -5.659 | -2.130 | -1.184 | 0.272 | 7.824 | -2.589 | -2.813 |
| -1.140 | 0.998 | 1.526 | 1.035 | -2.067 | 2.639 | -4.069 | 1.572 | 2.016 | 1.396 | -1.185 | 2.870 | 0.073 | -0.727 | -3.646 | -2.589 | 37.188 | -0.498 |
| 1.219 | -0.909 | -2.425 | 0.324 | 1.198 | -15.695 | 20.934 | 2.281 | -0.608 | -6.071 | 0.802 | 5.943 | -1.710 | 1.676 | 0.733 | -2.813 | -0.498 | 19.752 |

Table A.15: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{AA}$ Note that $\mathsf{K}_2^{\mathtt{TT}} = \mathsf{E}_2\,\mathsf{K}_2^{\mathtt{AA}}\,\mathsf{E}_2$.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8.616 | 1.508 | 0.583 | -1.459 | -4.329 | -2.259 | 2.958 | -3.445 | -6.346 | 0.221 | 3.825 | -21.345 | -7.709 | 0.967 | -0.631 | 2.350 | -1.156 | 0.182 |
| 1.508 | 2.255 | 0.893 | -0.478 | -2.801 | 2.315 | -3.144 | -0.371 | -0.461 | 1.644 | -0.227 | -6.095 | -2.476 | -0.507 | -1.391 | 0.734 | 0.833 | -1.789 |
| 0.583 | 0.893 | 13.938 | 1.354 | -13.701 | 0.496 | -5.381 | 0.507 | 0.113 | 5.261 | -0.953 | 0.454 | 0.349 | -1.200 | -3.779 | 1.226 | 1.748 | -2.612 |
| -1.459 | -0.478 | 1.354 | 5.214 | -4.756 | -1.325 | -1.071 | 1.384 | 4.476 | 0.553 | -2.322 | 5.391 | 2.440 | -0.451 | 0.184 | -1.020 | 0.410 | -1.106 |
| -4.329 | -2.801 | -13.701 | -4.756 | 60.874 | 0.187 | 4.433 | 0.874 | 2.727 | -2.403 | -0.865 | 3.900 | 0.663 | 1.220 | 1.938 | -1.615 | -2.335 | 3.124 |
| -2.259 | 2.315 | 0.496 | -1.325 | 0.187 | 20.776 | -20.799 | -0.642 | -0.628 | 5.791 | 0.170 | -3.433 | -0.820 | -0.859 | -2.794 | 1.755 | 3.082 | -15.050 |
| 2.958 | -3.144 | -5.381 | -1.071 | 4.433 | -20.799 | 33.102 | 0.413 | -0.498 | -9.253 | 1.585 | -3.021 | -1.054 | 2.491 | 4.885 | -2.621 | -5.325 | 19.180 |
| -3.445 | -0.371 | 0.507 | 1.384 | 0.874 | -0.642 | 0.413 | 7.749 | 5.086 | -0.070 | -1.967 | 10.182 | 3.459 | 0.505 | -0.524 | -1.636 | 0.930 | 3.770 |
| -6.346 | -0.461 | 0.113 | 4.476 | 2.727 | -0.628 | -0.498 | 5.086 | 16.578 | 0.153 | -7.045 | 10.796 | 6.451 | -0.524 | 0.690 | -4.676 | 1.819 | -0.544 |
| 0.221 | 1.644 | 5.261 | 0.553 | -2.403 | 5.791 | -9.253 | -0.070 | 0.153 | 7.296 | 0.508 | -1.599 | -0.343 | -1.117 | -4.326 | 1.325 | 2.064 | -4.972 |
| 3.825 | -0.227 | -0.953 | -2.322 | -0.865 | 0.170 | 1.585 | -1.967 | -7.045 | 0.508 | 7.894 | -7.222 | -4.573 | 1.076 | -0.680 | 1.078 | -1.964 | 1.174 |
| -21.345 | -6.095 | 0.454 | 5.391 | 3.900 | -3.433 | -3.021 | 10.182 | 10.796 | -1.599 | -7.222 | 77.859 | 22.204 | 0.277 | 1.387 | -5.620 | 3.300 | 9.164 |
| -7.709 | -2.476 | 0.349 | 2.440 | 0.663 | -0.820 | -1.054 | 3.459 | 6.451 | -0.343 | -4.573 | 22.204 | 9.018 | -0.839 | 0.276 | -2.373 | 2.827 | 0.363 |
| 0.967 | -0.507 | -1.200 | -0.451 | 1.220 | -0.859 | 2.491 | 0.505 | -0.524 | -1.117 | 1.076 | 0.277 | -0.839 | 2.354 | 2.508 | 0.213 | -4.212 | 2.401 |
| -0.631 | -1.391 | -3.779 | 0.184 | 1.938 | -2.794 | 4.885 | -0.524 | 0.690 | -4.326 | -0.680 | 1.387 | 0.276 | 2.508 | 10.336 | -0.089 | -6.545 | 4.867 |
| 2.350 | 0.734 | 1.226 | -1.020 | -1.615 | 1.755 | -2.621 | -1.636 | -4.676 | 1.325 | 1.078 | -5.620 | -2.373 | 0.213 | -0.089 | 5.642 | -0.753 | -0.432 |
| -1.156 | 0.833 | 1.748 | 0.410 | -2.335 | 3.082 | -5.325 | 0.930 | 1.819 | 2.064 | -1.964 | 3.300 | 2.827 | -4.212 | -6.545 | -0.753 | 37.765 | -10.180 |
| 0.182 | -1.789 | -2.612 | -1.106 | 3.124 | -15.050 | 19.180 | 3.770 | -0.544 | -4.972 | 1.174 | 9.164 | 0.363 | 2.401 | 4.867 | -0.432 | -10.180 | 19.808 |

Table A.16: $\mathsf{K}_2^{\alpha\beta}$ for $\alpha\beta = \mathtt{GG}$. Note that $\mathsf{K}_2^{\mathtt{CC}} = \mathsf{E}_2\,\mathsf{K}_2^{\mathtt{GG}}\,\mathsf{E}_2$.

## A.5 Eigenvalues of dimer stiffness blocks $\mathsf{K}_2^{\alpha\beta}$

| `AT` | `GC` | `TA` | `CG` | `GT` | `TG` | `AG` | `GA` | `AA` | `GG` | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| 6.8e-06 | 6.3e-02 | 2.0e-06 | 6.9e-04 | 2.1e-02 | 3.2e-02 | 9.2e-08 | 8.5e-08 | 6.6e-02 | 1.9e-05 | 4.9e-01 |
| 1.0e-01 | 1.1e-01 | 3.6e-04 | 3.1e-01 | 4.0e-01 | 3.1e-01 | 8.0e-04 | 5.0e-01 | 5.3e-01 | 5.5e-04 | 6.4e-01 |
| 6.7e-01 | 1.2e+00 | 3.3e-01 | 5.8e-01 | 9.1e-01 | 6.9e-01 | 3.8e-01 | 9.8e-01 | 7.0e-01 | 7.1e-01 | 2.0e+00 |
| 1.0e+00 | 1.5e+00 | 7.7e-01 | 1.0e+00 | 1.3e+00 | 9.2e-01 | 1.1e+00 | 1.4e+00 | 1.1e+00 | 1.6e+00 | 2.3e+00 |
| 1.5e+00 | 2.6e+00 | 1.1e+00 | 1.6e+00 | 1.7e+00 | 1.2e+00 | 1.4e+00 | 2.1e+00 | 1.8e+00 | 2.0e+00 | 2.4e+00 |
| 2.7e+00 | 2.8e+00 | 1.2e+00 | 2.0e+00 | 2.1e+00 | 1.4e+00 | 1.8e+00 | 2.4e+00 | 2.1e+00 | 2.2e+00 | 3.1e+00 |
| 3.1e+00 | 3.8e+00 | 1.7e+00 | 2.2e+00 | 3.1e+00 | 2.2e+00 | 2.3e+00 | 2.6e+00 | 2.3e+00 | 3.3e+00 | 3.3e+00 |
| 3.3e+00 | 3.9e+00 | 2.4e+00 | 2.6e+00 | 4.5e+00 | 2.8e+00 | 2.9e+00 | 3.3e+00 | 2.9e+00 | 3.4e+00 | 3.8e+00 |
| 4.9e+00 | 5.8e+00 | 2.4e+00 | 3.1e+00 | 5.2e+00 | 3.6e+00 | 4.7e+00 | 4.3e+00 | 4.8e+00 | 4.9e+00 | 4.0e+00 |
| 5.5e+00 | 5.9e+00 | 4.0e+00 | 4.4e+00 | 5.9e+00 | 4.7e+00 | 4.8e+00 | 5.7e+00 | 5.4e+00 | 5.1e+00 | 5.2e+00 |
| 6.7e+00 | 6.7e+00 | 4.8e+00 | 5.8e+00 | 7.4e+00 | 5.5e+00 | 8.3e+00 | 6.0e+00 | 8.0e+00 | 7.4e+00 | 6.3e+00 |
| 1.2e+01 | 8.9e+00 | 8.2e+00 | 1.0e+01 | 9.7e+00 | 9.6e+00 | 9.3e+00 | 8.6e+00 | 1.1e+01 | 8.1e+00 | 9.7e+00 |
| 1.5e+01 | 1.5e+01 | 1.0e+01 | 1.0e+01 | 1.5e+01 | 1.0e+01 | 1.5e+01 | 1.4e+01 | 1.6e+01 | 1.5e+01 | 1.3e+01 |
| 2.8e+01 | 1.9e+01 | 2.1e+01 | 2.1e+01 | 2.5e+01 | 2.2e+01 | 2.3e+01 | 2.2e+01 | 2.7e+01 | 2.1e+01 | 2.2e+01 |
| 3.2e+01 | 5.8e+01 | 2.8e+01 | 3.5e+01 | 3.4e+01 | 3.2e+01 | 3.5e+01 | 3.2e+01 | 3.5e+01 | 3.6e+01 | 3.9e+01 |
| 3.7e+01 | 6.4e+01 | 2.9e+01 | 3.8e+01 | 5.9e+01 | 3.3e+01 | 4.0e+01 | 5.5e+01 | 3.9e+01 | 6.1e+01 | 4.1e+01 |
| 8.1e+01 | 7.5e+01 | 5.8e+01 | 6.1e+01 | 7.7e+01 | 6.1e+01 | 7.2e+01 | 6.9e+01 | 7.2e+01 | 7.6e+01 | 7.0e+01 |
| 1.1e+02 | 1.1e+02 | 7.7e+01 | 8.5e+01 | 1.0e+02 | 8.5e+01 | 9.6e+01 | 9.7e+01 | 1.0e+02 | 1.0e+02 | 9.5e+01 |

Table A.17: Eigenvalues (sorted in increasing order) of dimer stiffness blocks $\mathsf{K}_2^{\alpha\beta}$ for ten independent dimers $\alpha\beta$ . The last column contains eigenvalues of the average block $\mathsf{K}_2^{\mathrm{h}}$, given in Table A.6.

## A.6 Dimer weighted shape vectors $\sigma_2^{\alpha\beta}$

| `AT` | `GC` | `TA` | `CG` | `GT` | `TG` | `AG` | `GA` | `AA` | `GG` | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| -109.629 | -108.010 | -73.962 | -77.076 | -101.586 | -77.696 | -99.122 | -93.619 | -103.117 | -94.929 | -94.742 |
| -19.496 | -21.077 | -1.102 | 3.946 | -18.457 | 1.326 | -18.316 | -24.267 | -21.710 | -21.254 | -11.486 |
| 4.012 | -0.363 | 4.450 | 8.058 | 3.569 | 6.704 | 4.779 | -0.546 | 4.111 | 6.342 | 4.052 |
| 33.137 | 38.560 | 23.717 | 23.551 | 38.916 | 22.757 | 32.156 | 42.193 | 40.010 | 36.782 | 32.385 |
| 17.659 | 27.485 | 10.236 | 18.249 | 18.848 | 11.267 | 18.512 | 17.626 | 18.986 | 22.763 | 18.229 |
| -5.459 | -6.034 | 17.306 | 24.334 | -11.855 | 22.175 | -4.152 | -23.112 | -12.499 | -13.833 | 6.284 |
| 0 | 0 | 0 | 0 | 9.870 | 0.964 | -27.386 | -12.138 | -30.365 | -12.586 | 0 |
| 42.260 | 53.883 | 41.794 | 48.633 | 49.147 | 46.916 | 43.349 | 53.753 | 50.316 | 53.267 | 48.754 |
| 92.184 | 90.369 | 64.334 | 63.916 | 87.248 | 61.057 | 86.802 | 96.564 | 102.454 | 92.196 | 85.215 |
| 0 | 0 | 0 | 0 | -1.774 | 2.825 | -9.577 | -12.534 | -7.061 | -5.949 | 0 |
| -39.106 | -36.725 | -36.787 | -37.628 | -34.545 | -35.089 | -45.253 | -49.498 | -52.574 | -52.023 | -43.013 |
| 320.593 | 329.521 | 230.517 | 253.891 | 313.598 | 252.000 | 288.089 | 297.558 | 305.582 | 304.792 | 291.110 |
| 109.629 | 108.010 | 73.962 | 77.076 | 108.172 | 78.118 | 97.065 | 94.368 | 100.052 | 99.356 | 94.742 |
| -19.496 | -21.077 | -1.102 | 3.946 | -18.421 | 3.356 | -6.017 | -8.144 | -9.985 | -4.160 | -11.486 |
| 4.012 | -0.363 | 4.450 | 8.058 | 3.200 | 7.391 | 6.382 | 1.561 | -1.132 | 6.310 | 4.052 |
| -33.137 | -38.560 | -23.717 | -23.551 | -31.357 | -22.939 | -29.868 | -35.422 | -34.069 | -32.731 | -32.385 |
| 17.659 | 27.485 | 10.236 | 18.249 | 22.688 | 17.912 | 18.957 | 15.581 | 13.747 | 21.156 | 18.229 |
| -5.459 | -6.034 | 17.306 | 24.334 | 2.272 | 24.787 | 17.216 | 24.806 | 15.982 | 28.608 | 6.284 |

Table A.18: Dimer weighted shape vectors $\sigma_2^{\alpha\beta}$ for ten independent dimers, rounded to three decimal places. The last column is the average vector $\sigma_2^{\mathrm{h}}$, computed over all 16 different dimers. The ordering of parameters is as in Table A.7. Zeros in the first four columns and in the average column appear due to the palindromic symmetry, i.e. because $\sigma_2^{\alpha\beta} = \mathsf{E}_2\, \sigma_2^{\alpha\beta}$ for $\alpha\beta \in \{\mathtt{AT}, \mathtt{GC}, \mathtt{TA}, \mathtt{CG}\}$ and $\sigma_2^{\mathrm{h}} = \mathsf{E}_2\, \sigma_2^{\mathrm{h}}$, where $\mathsf{E}_n$ is defined in (2.38). Note that $\sigma_2^{\alpha\beta}$ for $\alpha\beta \in \{\mathtt{AC}, \mathtt{CA}, \mathtt{CT}, \mathtt{TC}, \mathtt{TT}, \mathtt{CC}\}$ can be found using the identity $\sigma_2^{\alpha\beta} = \mathsf{E}_2\, \sigma_2^{\overline{\beta\alpha}}$, where $\overline{\mathtt{X}}$ is a complementary base to `X`.

# Appendix B

# Plots of sequence averaged nearest-neighbour stiffness blocks

Here we present plots of averages and standard deviations of nearest-neighbour stiffness matrix blocks, that are assumed to depend on a dimer (Figure B.3) and on a trimer (Figures B.4 and B.5). See discussion in Section 7.3. The positions of dimer and trimer dependent blocks in a matrix are shown in Figures B.1 and B.2. Plots of averaged weighted shape vector elements, although ordered by parameter name, can be found in Section 7.3 of Chapter 7. We conclude that our model assumptions are compatible with the MD data.

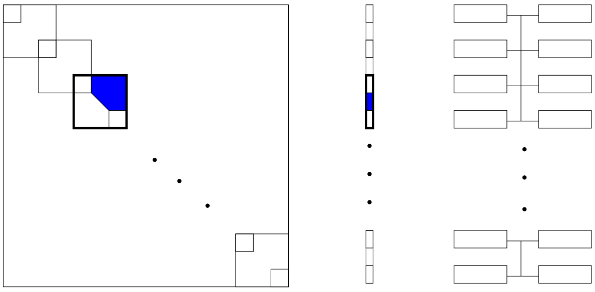


Figure B.1: The structure of a nearest-neighbour stiffness matrix $\mathsf{K}$ and a weighted shape vector $\sigma$. Elements coloured in blue are assumed to have a dimer dependence.

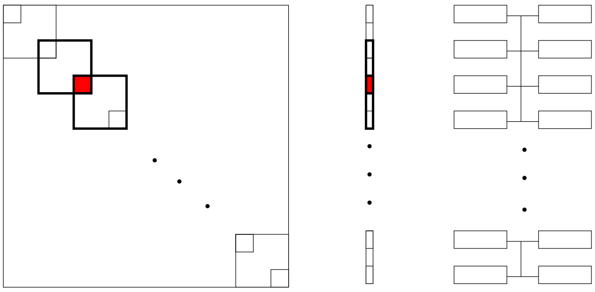


Figure B.2: The structure of a nearest-neighbour stiffness matrix $\mathsf{K}$ and a weighted shape vector $\sigma$. Elements coloured in red (intra base pair variables) are assumed to have a trimer dependence.

## B.1 Dimer averages of the nearest-neighbour stiffness matrix blocks with an assumed dimer dependence

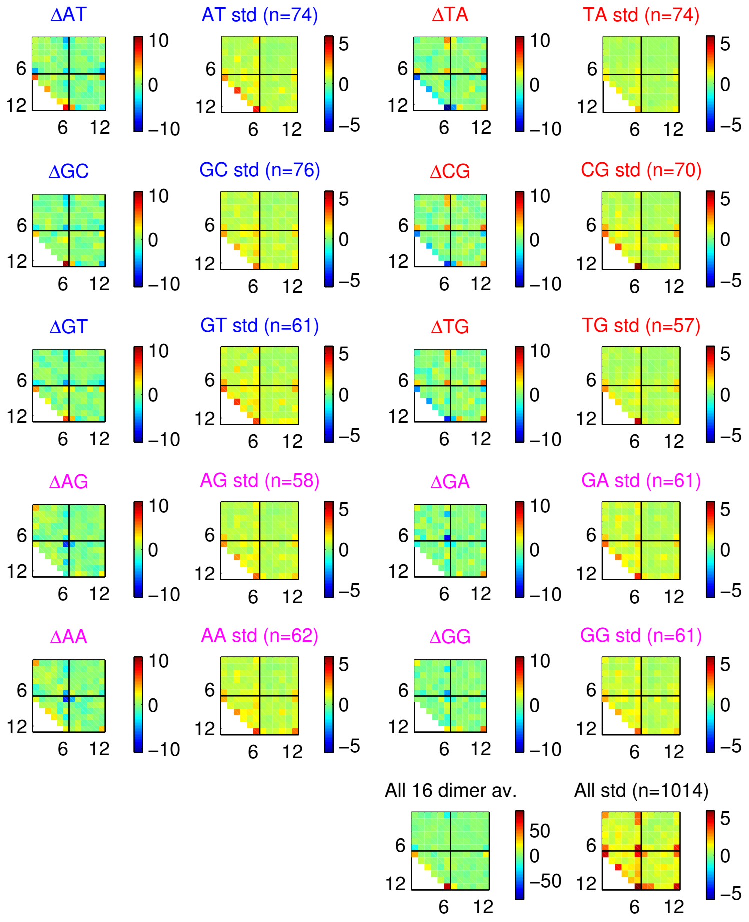


Figure B.3: Centred averages and standard deviations of the nearest-neighbour stiffness blocks with an assumed dimer dependence, computed over all instances of each independent dimer on both strands of all ABC oligomers. The dimers in each plot are ordered in three groups: `RY` (blue titles), `YR` (red titles) and `RR` (magenta titles), where `R` denotes a purine and `Y` a pyrimidine. The last two plots on the bottom right are averages and standard deviations, computed over all instances of all 16 dimers. Centred averages mean that averages over all dimers were subtracted from averages over a particular dimer. One can notice that the observed stiffness parameters does depend on a sequence. The standard deviations within the dimer groups are smaller than the total standard deviation, indicating that considering a dimer sequence dependence of these stiffnesses seems a reasonable choice.

## B.2 Trimer averages of the diagonal intra base pair stiffness matrix blocks

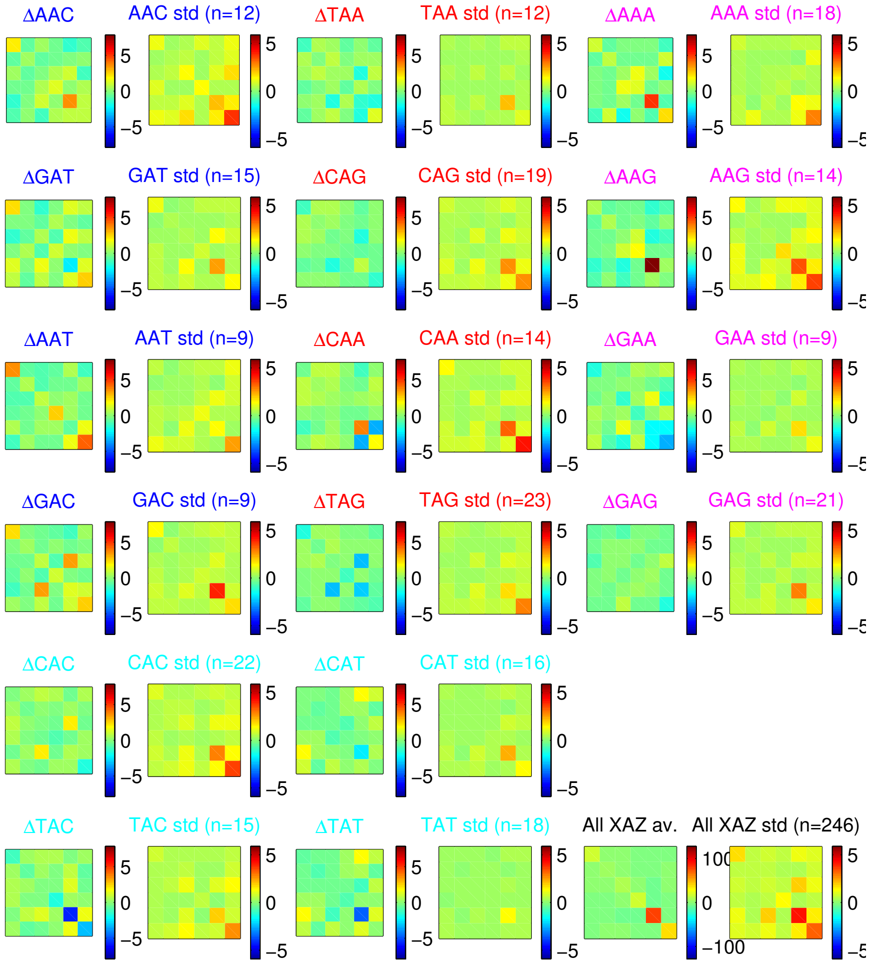


Figure B.4: Intra base pair stiffness block centred averages and standard deviations, computed over all instances of each independent trimer `XAZ`, where $\mathtt{X}, \mathtt{Z} \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$, on both strands of all ABC oligomers. The trimers in each plot are ordered in groups of four: `RRY` (blue titles), `YRR` (red titles), `RRR` (magenta titles) and `YRY` (cyan titles), where `R` denotes a purine and `Y` a pyrimidine. The last two plots on the bottom right are averages and standard deviations, computed over all instances of all `XAZ` trimers. Centred averages mean that averages over all trimers were subtracted from averages over a particular trimer. Notice that the average stiffness values for `XAZ` trimers are very different from the ones for `XGZ` trimers, shown in Figure B.5. Even though there is still some variation within each of the 32 independent trimer groups, considering a trimer dependence of these stiffness parameters seems a reasonable choice.

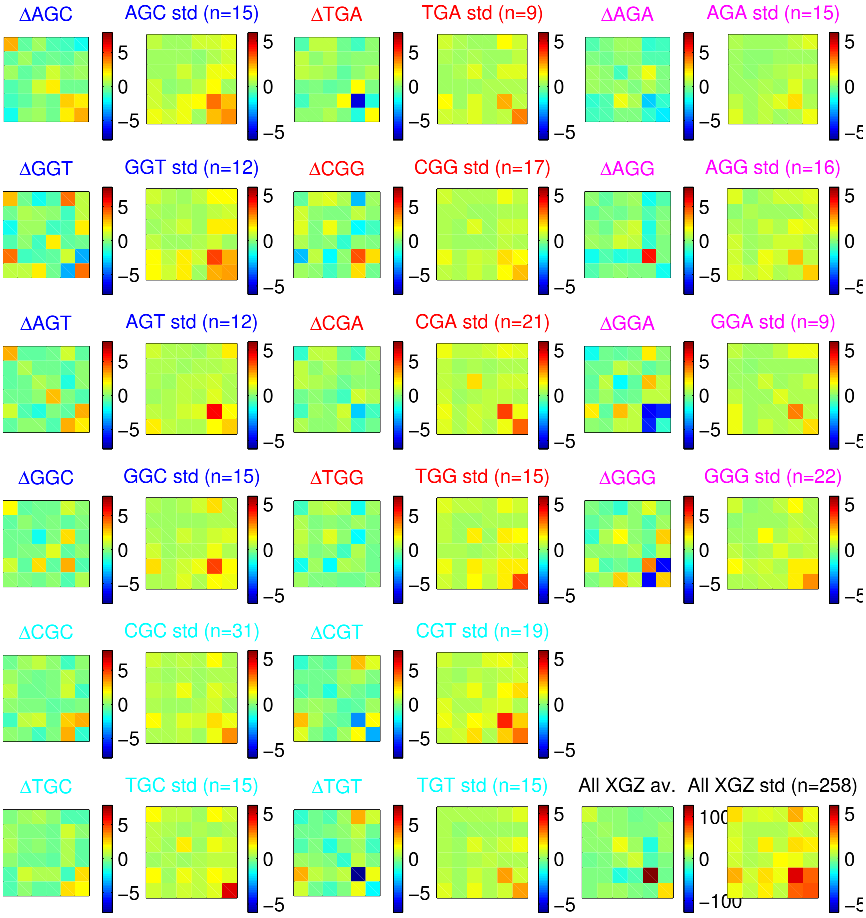


Figure B.5: Intra base pair stiffness block centred averages and standard deviations, computed over all instances of each independent trimer XGZ, where X, Z ∈ {A, G, C, T}, on both strands of all ABC oligomers. The trimers in each plot are ordered in groups of four: RRY (blue titles), YRR (red titles), RRR (magenta titles) and YRY (cyan titles), where R denotes a purine and Y a pyrimidine. The last two plots on the bottom right are averages and standard deviations, computed over all instances of all XGZ trimers. Centred averages mean that averages over all trimers were subtracted from averages over a particular trimer. Also see the caption of Figure B.4.

# Bibliography

[1] W.K. Olson A.A. Gorin, V.B. Zhurkin. B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, 247:34–48, 1995.

[2] N. B. Becker and R. Everaers. From rigid base pairs to semiflexible polymers: Coarse-graining DNA. *Phys. Rev. E*, 76:021923, 2007.

[3] C.J. Benham. An elastic model of the large-scale structure of duplex DNA. *Biopolymers*, 18:609–623, 1979.

[4] C.J. Benham. Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *J. Mol. Biol.*, 255:425–434, 1996.

[5] C.J. Benham and S. P. Mielke. DNA mechanics. *Annu. Rev. Biomed. Eng.*, 7:21–53, 2005.

[6] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[7] D.L. Beveridge, G. Barreiro, K.S. Byun, D.A. Case, T.E. Cheatham III, S.B. Dixit, E. Giudice, F. Lankas, R. Lavery, J.H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K.M. Thayer, P. Varnai, and M.A. Young. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I: Research design and results on d(CpG) steps. *Biophys. J.*, 87:3799–3813, 2004.

[8] D.L. Beveridge, S.B. Dixit, G. Barreiro, and K.M. Thayer. Molecular dynamics simulations of DNA curvature and flexibility: helix phasing and premelting. *Biopolymers*, 73(3):380–403, 2004.

[9] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30:1545–1614, 2009.

[10] B.R. Brooks, R.E. Bruccoleri, D.J. Olafson, D.J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[11] C.R. Calladine, H.R. Drew, B. F. Luisi, and A. A. Travers. *Understanding DNA*. Elsevier Academic Press, 2004. 3rd edn.

[12] D. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26:1668–1688, 2005.

[13] T.E. Cheatham III, P. Cieplak, and P. Kollman. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, 16:845–862, 1999.

[14] D. M. Chenoweth and P. B. Dervan. Allosteric modulation of DNA by small molecules. *Proc. Natl. Acad. Sci. USA*, 106(32):13175–13179, 2009.

[15] G. Chirikjian. The stochastic elastica and excluded-volume perturbations of DNA conformational ensembles. *Int J Non Linear Mech.*, 43(10):1108–1120, 2008.

[16] B.D. Coleman, W.K. Olson, and D. Swigon. Theory of sequence-dependent DNA elasticity. *J. Chem. Phys.*, 118(15):7127–7140, 2003.

[17] L.X. Dang. Mechanism and thermodynamics of ion selectivity in aqueous-solutions of 18-crown-6 ether - a molecular dynamics study. *J. Am. Chem. Soc.*, 117:6954–6960, 1995.

[18] S. Demko, W. F. Moss, and Ph. W. Smith. Decay rates for inverses of band matrices. *Mathematics of Computation*, 43(168):491–499, 1984.

[19] R.E. Dickerson, M. Bansal, C.R. Calladine, S. Diekmann, W.N. Hunter, O. Kennard, R. Lavery, H.C.M. Nelson, W.A. Olson, W. Saenger, Z. Shakked, H. Sklenar, D.M. Soumpasis, C.-S. Tung, E. von Kitzing, A. Wang, and V. Zhurkin. Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, 205:787–791, 1989.

[20] S.B. Dixit, D.L. Beveridge, D.A. Case, T.E. Cheatham III, E. Giudice, F. Lankas, R. Lavery, J.H. Maddocks, R. Osman, H. Sklenar, K.M. Thayer, and P. Varnai. Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, 89:3721–3740, 2005.

[21] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee., and L.G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995.

[22] S. Geggier and A. Vologodskii. Sequence dependence of DNA bending rigidity. *Proc. Natl Acad. Sci. USA 107*, 107:15421–15426, 2010.

[23] O. Gonzalez and J.H. Maddocks. Extracting parameters for base-pair level models of DNA from molecular dynamics simulations. *Theor. Chem. Acc.*, 106:76–82, 2001.

[24] O. Gonzalez, D. Petkevičiūtė, and J.H. Maddocks. A sequence-dependent rigid-base model of DNA. *Submitted.*

[25] A. Grandchamp. Equilibrium birod theory, high twist and coarse grain dna models. Master's thesis, EPFL, 2011.

[26] A. Griswold. Genome packaging in prokaryotes: the circular chromosome of e. coli. *Nature Education*, 1(1), 2008.

[27] J.E. Hearst. Elastic model of DNA supercoiling in the infinite-length limit. *J. Chem. Phys.*, 95:9329–9338, 1991.

[28] J.R. Grigera H.J.C. Berendsen and T.P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91:6269–6271, 1987.

[29] R.V. Hogg and A.T. Craig. *Introduction to Mathematical Statistics.* Macmillan, New York, 3rd edition, 1970.

[30] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[31] A.S. Lewis J.V. Burke and M.L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM J. Optimization*, 15:751–779, 2005.

[32] C.L. Kielkopf, S. Ding, P. Kuhn, and D.C. Rees. Conformational flexibility of B-DNA at 0.74 Angstrom resolution: d(CCAGTACTGG)$_2$. *J. Mol. Biol.*, 296(3):787–801, 2000.

[33] S. Kullback. *Information Theory and Statistics.* John Wiley and Sons, New York, 1959.

[34] K.J. Kwak, H. Kudo, and M. Fujihira. Imaging stretched single DNA molecules by pulsed-force-mode atomic force microscopy. *Ultramicroscopy*, 97(1-4):249–255], 2003.

[35] F. Lankaš. Modelling nucleic acid structure and flexibility: From atomic to mesoscopic scale. In *Chapter 1 of Innovations in Biomolecular Modeling and Simulations*, volume 2, pages 3–32. The Royal Society of Chemistry, 2012.

[36] F. Lankaš, O. Gonzalez, L.M. Heffler, G. Stoll, M. Moakher, and J.H. Maddocks. On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Physical Chemistry Chemical Physics*, 11:10565–10588, 2009.

[37] F. Lankas, R. Lavery, and J.H. Maddocks. Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure*, 14:1527–1534, 2006.

[38] F. Lankaš, J. Šponer, P. Hobza, and J. Langowski. Sequence-dependent elastic properties of DNA. *J. Mol. Biol.*, 299:695–709, 2000.

[39] F. Lankaš, J. Šponer, J. Langowski, and T.E. Cheatham III. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, 85:2872–2883, 2003.

[40] F. Lankaš, J. Šponer, J. Langowski, and T.E. Cheatham III. DNA deformability at the base pair level. *J. Am. Chem. Soc.*, 126:4124–4125, 2004.

[41] R. Lavery, M. Moakher, J.H. Maddocks, D. Petkeviciute, and K. Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, 37:5917–5929, 2009.

[42] R. Lavery, K. Zakrzewska, D. Beveridge, T. Bishop, D. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, J. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Spackova, and J. Sponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, 38(1):299–313, 2010.

[43] A. R. Leach. *Molecular Modelling: Principles and Applications.* Pearson Education Limited, 2nd edition, 2001.

[44] M. Levitt. Computer simulation of DNA double-helix dynamics. *Cold Spring Harb. Symp. Quant. Biol.*, 47:251–262, 1983.

[45] A.S. Lewis and M.L. Overton. Nonsmooth optimization via quasi-Newton methods. *Math. Programming*, pages 1–29, 2012.

[46] X.-J. Lu and W.K. Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, 31(17):5108–5121, 2003.

[47] A.D. MacKerel Jr., C.L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus. *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*, volume 1 of *The Encyclopedia of Computational Chemistry*, pages 271–277. John Wiley & Sons: Chichester, 1998.

[48] T. Maehigashi, C. Hsiao, K. K. Woods, T. Moulaei, N. V. Hud, and L. D. Williams. B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.*, 40(8):3714–3722, 2012.

[49] A.J. Majda and X. Wang. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows.* Cambridge University Press, Cambridge, 2006.

[50] F. Michor, J. Liphardt, M. Ferrari, and J. Widom. What does physics have to do with cancer? *Nature Reviews Cancer 11*, 11:657–670, 2011.

[51] V. Miele, C. Vaillant, Y. d'Aubenton-Carafa, C. Thermes, and T. Grange. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, 36(11):3746–3756, 2008.

[52] M. Moakher. On the averaging of symmetric positive-definite tensors. *Journal of Elasticity*, 82:273–296, 2006.

[53] M. Moakher and J. H. Maddocks. A double-strand elastic rod theory. *Arch. Rational Mech. Anal.*, 177:53 – 91, 2005.

[54] W.K. Olson, M. Bansal, S.K. Burley, R.E. Dickerson, M. Gerstein, S.C. Harvey, U. Heinemann, X.-J. Lu, S. Neidle, Z. Shakked, H. Sklenar, M. Suzuki, C.-S. Tung, E. Westhof, C. Wolberger, and H.M. Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, 313:229–237, 2001.

[55] W.K. Olson, A.A. Gorin, X.-J. Lu, L.M. Hock, and V.B. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, 95:11163–11168, 1998.

[56] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye. Structural, mechanical, and thermodynamic properties of a coarse-grained dna model. *J. Chem. Phys.*, 134:085101, 2011.

[57] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham III, S. DeBolt, D. Ferguson, G.L. Seibel, and P.A. Kollman. 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. *Comp. Phys. Commun.*, 91:1–41, 1995.

[58] A. Pérez, C. L. Castellazzi, F. Battistini, K. Collinet, O. Flores, O. Deniz, M. L. Ruiz, D. Torrents, R. Eritja, M. Soler-López, and M. Orozco. Impact of methylation on the physical properties of DNA. *Biophys. J.*, 102(9):2140–2148, 2012.

[59] A. Pérez, F. Lankas, F.J. Luque, and M. Orozco. Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.*, 36(7):2379–2394, 2008.

[60] A. Pérez, F. J. Luque, and M. Orozco. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.*, 129(47):14739–14745, 2007.

[61] A. Pérez, F.J. Luque, and M. Orozco. Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res*, 45(2):196–205, 2012.

[62] A. Pérez, I. Marchan, D. Svozil, J. Sponer, T.E. Cheatham III, C.A. Laughton, and M. Orozco. Refinement of the AMBER force field for nucleic acids: Improving the description of $\alpha/\gamma$ conformers. *Biophys. J.*, 92:3817–3829, 2007.

[63] J. P. Peters and L. J. Maher. DNA curvature and flexibility in vitro and in vivo. *Q Rev Biophys.*, 43:22–63, 2010.

[64] A. Prékopa. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343, 1973.

[65] J.P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.*, 23:327–341, 1977.

[66] G. Saccomandi and I. Sgura. The relevance of nonlinear stacking interactions in simple models of double-stranded DNA. *J. R. Soc. Interface*, 3(10):655–667, 2006.

[67] T. Schleif. DNA looping. *Annu. Rev. Biochem.*, 61:199–223, 1992.

[68] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. Moore, J. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442:772–778, 2006.

[69] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J.P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, 2008.

[70] D. Shore and R. L. Baldwin. Energetics of DNA twisting: I. Relation between twist and cyclization probability. *Journal of Molecular Biology*, 170(4):957–981, 1983.

[71] D Shore, J Langowski, and R L Baldwin. DNA flexibility studied by covalent closure of short fragments into circles. *Proc Natl Acad Sci USA*, 78(8):4833–4837, 1981.

[72] D. Swigon, B.D. Coleman, and I. Tobias. The elastic rod model for DNA and its application to the tertiary structure of DNA minicircles in mononucleosomes. *Biophys. J.*, 74:2515–2530, 1998.

[73] A. A. Travers and J. M. T. Thompson. An introduction to the mechanics of DNA. *Phil. Trans. R. Soc. Lond.*, 362:1265–1279, 2004.

[74] J. Walter, O. Gonzalez, and J.H. Maddocks. On the stochastic modeling of rigid body systems with application to polymer dynamics. *SIAM Multiscale Modeling and Simulation*, 8:1018–1053, 2010.

[75] J. C. Wang. Helical repeat of DNA in solution. *Proc Natl Acad Sci U S A*, 76(1):200–203, 1971.

[76] J. D. Watson and F. H. C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171:964–967, 1953.

[77] A. Wildes, N. Theodorakopoulos, J. Valle-Orero, S. Cuesta-Lopez, J-L Garden, and M. Peyrard. The thermal denaturation of DNA studied with neutron scattering. *Physical Review Letters*, 106:048101–1–4, 2011.

# Curriculum vitae

Daiva Petkevičiūtė was born in Kaunas, Lithuania in 1983. She got her bachelor and master degrees in applied mathematics in Kaunas University of Technology. Afterwards she pursued with doctoral studies in mathematics at Ecole Polytechnique Fédérale de Lausanne, Switzerland.