Sequence Dependent Elasticity of DNA

Nils Becker

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
vorgelegt der

Fakultät Mathematik und Naturwissenschaften
Technische Universität Dresden

Abstract

The DNA contained in every living cell not only stores the genetic information; it functions in a complex molecular network that can condense, transcribe, replicate and repair genes. The essential role played by the sequence dependent structure and deformability of DNA in these basic processes of life, has received increasing attention over the past years.

The present work aims at better understanding sequence dependent elasticity of double stranded DNA elasticity, across biologically relevant length scales. A theoretical description is developed that makes is possible to relate structural, biochemical and biophysical experiments and simulation. It is based on the rigid base–pair chain (rbc) model which captures all basic deformation modes on the scale of individual base–pair (bp) steps.

Existing microscopic parametrizations of the rbc model rely on indirect methods. A way to relate them to biochemical experiments is provided by the indirect readout mechanism, where DNA elasticity determines protein–DNA complexation affinities. By correlating theoretical affinity predictions with *in vitro* measurements in a well–studied test case, different parameter sets were evaluated. As a result a new, hybrid parameter set is proposed which greatly reduces prediction errors. Indirect readout occurs mostly at particular binding subsites in a complex. A statistical marker is developed which localizes indirect readout subsites, by detecting elastically optimized sub-sequences.

By a systematic coarse–graining of the rbc to the well–characterized worm–like chain (wlc) model, a quantitative connection between microscopic and kbp scale elasticity is established. The general helical rbc geometry is mapped to an effective, linear 'on-axis' version, yielding the full set of wlc elastic parameters for any given sequence repeat. In the random sequence case, structural variability adds conformational fluctuations which are correlated by sequence continuity. The sequence disorder correction to entropic elasticity in the rbc model is shown to coincide with the conformational correction. The results show remarkable overall agreement of the coarse–grained with the mesoscale wlc parameters, lending support to the model and to the microscopic parameter sets.

A continuum version of the rbc is formulated as Brownian motion on the rigid motion group. Analytic expressions for angular correlation functions and moments of the end–to–end distance distribution are given. In an equivalent Lagrangian approach, conserved quantities along, and the linear response around, a general equilibrium shape are explored.

Zusammenfassung

Die in jeder lebenden Zelle enthaltene DNS speichert nicht nur die genetische Information; Sie funktioniert innerhalb eines komplexen molekularen Netzwerks, das in der Lage ist, Gene zu kondensieren, transkribieren, replizieren und reparieren. Die zentrale Rolle, welche der sequenzabhängigen Struktur und Deformierbarkeit von DNS in diesen grundlegenden Lebensprozessen zukommt, erregte in den letzten Jahren zunehmendes Interesse.

Die vorliegende Arbeit hat ein besseres Verständnis der sequenzabhängigen elastischen Eigenschaften von DNS auf biologisch relevanten Längenskalen zum Ziel. Es wird eine theoretische Beschreibung entwickelt, die es ermöglicht, strukturbiologische, biochemische und biophysikalische Experimente und Simulationen in Beziehung zu setzen. Diese baut auf dem Modell einer Kette aus starren Basenpaaren (rbc) auf, das alle wichtigen Deformationsmoden von DNS auf der Ebene von einzelnen Basenpaar (bp)–Schritten abbildet.

Bestehende Parametersätze des rbc-Modells beruhen auf indirekten Methoden. Eine direkte Beziehung zu biochemischen Experimenten kann mithilfe des indirekten Auslese-Mechanismus hergestellt werden. Hierbei bestimmt die DNS–Elastizität Komplexierungsaffinitäten von Protein–DNS–Komplexen. Durch eine Korrelation von theoretischen Vorhersagen mit *in vitro* Messungen in einem gut untersuchten Beispielfall werden verschiedene Parametersätze bewertet. Als Resultat wird ein neuer Hybrid–Parametersatz vorgeschlagen, der die Vorhersagefehler stark reduziert. Indirektes Auslesen tritt meistens an speziellen Teilbindungsstellen innerhalb eines Komplexes auf. Es wird eine statistische Kenngröße entwickelt, die indirektes Auslesen durch Detektion elastisch optimierter Subsequenzen erkennt.

Durch ein systematisches Coarse–Graining des rbc-Modells auf das gut charakterisierte Modell der wurmartigen Kette (wlc) wird eine quantitative Beziehung zwischen der mikroskopischen und der Elastizität auf einer kbp-Skala hergestellt. Die allgemeine helikale Geometrie wird auf eine effektive, lineare Version der Kette ‘auf der Achse’ abgebildet. Dies führt zur Berechnung des vollen Satzes von wlc-elastischen Parameters für eine beliebig vorgegebene periodische Sequenz. Im Fall zufälliger Sequenz führt die Strukturvariabilität zu zusätzlichen Konformationsfluktuationen, die durch die Kontinuität der Sequenz kurzreichweitig korreliert sind. Es wird gezeigt, daß die Sequenzunordnungs-Korrektur zur entropischen Elastizität im rbc-Modell identisch ist zur Korrektur der Konformationsstatistik. Die Ergebnisse zeigen eine bemerkenswerte Übereinstimmung der hochskalierten mikroskopischen mit den mesoskopischen wlc-Parameter und bestätigen so die Wahl des Modells und seiner mikroskopischen Parametrisierung.

Eine Kontinuumsversion des rbc-Modells wird formuliert als Brownsche Bewegung auf der Gruppe der Starrkörpertransformationen. Analytische Ausdrücke für Winkelkorrelationsfunktionen und Momente der Verteilung des End-zu-End–Vektors werden angegeben. In einem äquivalenten Lagrange-Formalismus werden Erhaltungsgrößen entlang von Gleichgewichtskonformationen und die lineare Antwort in ihrer Umgebung untersucht.

Acknowledgments

First of all, I would like Ralf Everaers for his supervision and his encouragement during this work. His unerring physical intuition was invaluable and he never failed to remind me that theoretical physics is about: the real world, and real data.

Frank Jülicher kindly admitted me into his group. I am very grateful for his teaching and for the extremely pleasant and stimulating scientific crowd he has gathered at the Max-Planck-Institute for Complex Systems.

My fellow students have helped me in many different ways. I would like to thank the 'first generation', Andreas Hilfinger, Gernot Klein, Peter Borowski, Frank Pollmann[1] and Tobias Bollenbach for the team spirit and also Christian Simm, Thomas Bittig, Eva-Maria Schötz, Elisabeth Fischer, Benjamin Friedrich, Lars Wolff and Kai Dierkes for their interest, their support and their encouragement.

It has been a great learning experience and a lot of fun to interact with Ben Lindner, Karsten Kruse, Simon Tolic-Norrelykke, Eric Galburt and John Maddocks.

Like everyone in the biological physics department, I am very much indebted to Nadine Baldes who has 'run the place' and still had the time for all my administrative problems.

My deepest gratitude belongs to my parents for their love and constant support in these last three years, again.

[1] honorary group member

Contents

Introduction

The implementation of the genome

When asked to name the most important biomolecule, one would probably say it's DNA, deoxyribonucleic acid. DNA is present in every living cell, with a chemical structure that has been conserved over billions of years. It functions as the physical implementation of the genome, preserving the genetic information of any living organism with unmatched storage density and reliability. Our DNA base sequence defines if not who we are, so at least what we are, by encoding for the protein components all cells are made of.

After completion of the Human Genome Project [Int03b, Int03a], the genetic information of man is readily available, and more and more species are being sequenced. Given the rapid progress in efficiency, it will soon be possible to sequence entire genomes of individuals for an affordable price. So in a way, one could think that all secrets that have surrounded DNA are finally resolved, and one should move on to study something else.

However, neither the complete genome sequence nor the atomic structure of the double helix discovered 50 years earlier [Wat53b] can explain how the molecule really works. How exactly is DNA able to perform the enormous tasks of storing gigabytes of genetic information in an error–tolerant way, repairing inevitable damage? How can the appropriate bits of that information be read out with appropriate frequency? How does the machinery work that allows DNA to replicate itself faithfully, then to condense and separate before cell division and finally to de-condense in the nucleus afterward?

Like any component of a complex system, DNA does not function on its own. Understanding DNA means understanding its interactions with a multitude of co-evolved proteins, whose intricate biochemical network performs essential molecular processes of life collectively.

DNA as a physical object

In all of these interactions, the *physical* properties of the DNA molecule as a complex polymer are essential. Here, thinking in terms of physics can give insight of the constraints under which the biological system works. Some examples follow.

In a stereotyped eucaryotic cell 10 μm in length, between divisions, DNA is concentrated in the nucleus of 1 μm radius. The total contour length of DNA is of the order of 1 cm, less than could be fit into the nucleus by tight packing. So is DNA really compressed at all? From polymer physics one knows that the bending persistence length of 50 nm sets the scale for the extension of a coil of DNA free in solution. The result is at least 50 μm radius for 1 cm of DNA, indicating that confinement into the nucleus does require work.

Separating such a highly condensed coil of threadlike polymer for cell division is a nontrivial task, since the inevitable entanglement of strands poses topological constraints [Sch04]. Cells deal with them on one hand by a set of enzymes that can actively change the linking of DNA coils, and on the other hand by a whole hierarchy of organized packing structures which compact DNA and limit entanglement at the same time (see e.g. [Sin94, Alb02]).

Of this packing hierarchy, the lowest level is best understood. The basic packing motif is called the nucleosome core particle. It consists of about 50 nm of DNA wrapped in 1.7 turns around a cylindrical spool with about 10 nm diameter [Ric03]. The histones that form the spool and other DNA–associated proteins actually make up more than half of the material in the cell nucleus. The tight bending of DNA onto the 5 nm outer radius of the histone spool costs energy, and there exists a free energy balance between chemical bonds of DNA with the histone surface, and its wrapping. In effect, histones are bound strongly enough to occupy DNA almost densely but still not too strongly to block transcription [Sch03].

Protein levels in the cell are regulated in response to cell fate and to environmental conditions. One of the involved feedback mechanisms works at the level of transcription of DNA to RNA (ribonucleic acid). Here, depending on protein product concentration, a regulatory protein binds DNA at a specific sequence of several base–pairs, close to the transcription initiation site, thereby modifying the rate of transcription. In crystal structures of such complexes, DNA is often deformed from its equilibrium shape. As a result, the base–sequence dependent deformability of DNA affects the binding strength of the complex and thus also the resulting protein levels [Kou06, Kou87, Heg02].

In fact, sequence–dependent packing and transcription regulation do have an overlap: Nucleosome core particles are known to form much [Clo04] more stably with DNA sequences whose specific structure and bendability match well with the prescribed wrapping path of DNA around the spool [Dre85, Shr90]. Stable nucleosomes suppress transcription [Kne86], so the detailed positioning of nucleosomes has a regulatory effect.

Even regulatory proteins that bind at a specific location hundreds of base–pairs away from the actual gene, have been observed to influence transcription rates [Sch75]. In order for this to happen, DNA loops back onto itself to allow direct contact of the 'distant' regulatory protein and the transcription initiation site of the gene [Sch92]. The free energy associated with such a loop depends on its size and on the stiffness of the looped DNA and thus plays a role in the final expression levels of the gene [Vil03, Sai06].

DNA elasticity across scales

In all of these examples, the elastic properties of the intact double helical B-DNA structure are important for the functioning of a biological process. It is not surprising that DNA elasticity and macromolecular structure is a long–standing field of research, see e.g. [Gar07]. A large variety of experimental techniques are sensitive to some combination of the intrinsic conformation and deformability of DNA on different length scales.

On a μm scale, the topological constraint that the molecule cannot pass through itself is most important, and has been studied, e.g, using the gel electrophoretic mobility of different knot types of circular DNA [Sta96]. On shorter length scales, this constraint becomes weaker since bending persistence of the molecule suppresses self–intersections. At around 50 nm contour length, DNA behaves on average like a thin homogeneous elastic rod that can resist thermal bending and twisting forces so that its contour looks only 'mildly curved'.

The elasticity of DNA on this scale can be measured comparatively well. The basic idea of a widely used biochemical method is to observe the cyclization reaction of short pieces of DNA that have 'sticky ends'. The stiffness and structure of the molecule can be reconstructed from the reaction kinetics [Clo05, Vol02, Du05]. A biophysical technique consists in tracing the thermally randomized conformation of DNA molecules adsorbed on a surface, either by electron microscopy [Bed95] or by atomic force microscopy [Wig06]. Finally, micro-manipulation experiments

allow to probe the stiffness of individual molecules in solution by recording force–extension or twist–extension relations of DNA tethers [Str00, Lio06, Gor06].

At 5 nm contour length, the scale of one turn of the double helix, thermal forces cannot bend the molecule very much, but interactions with proteins can and typically do. Also, the deformation free energy is strongly sequence–dependent on this short scale. Experimental data on sequence–dependent DNA elasticity on the scale of individual bases is rather indirect. From the distortions of DNA observed in crystal structures of protein–DNA complexes or in oligonucleotide structures, empirical sequence–dependent elastic potentials can be constructed [Ols98]. In a related approach, molecular dynamics simulations have been used to characterize DNA deformability on the scale of base–pairs [Lan03].

Answers and questions

The aim of this work is to understand better how the elastic properties of DNA influence its biological function. A general strategy in pursuing this goal will be to combine a range of available experimental data from biochemistry, single–molecule biophysics, and structural biology, as well as atomistic simulations. To make this possible, a new theoretical framework is developed that is able to quantitatively connect DNA statistical mechanics on different length scales.

The first part of the thesis concentrates on the question

- How in detail can DNA elasticity influence gene regulation?

Besides the chemical features of individual base–pairs, the short–scale deformability of DNA is another property specific to certain sequences, providing another 'interface' that connects DNA to the network of proteins in which it functions. In this way, binding affinities and eventually, the expression levels of proteins in the cell are influenced by the short–scale, sequence–dependent structure and deformability of DNA. These properties are captured by the rigid base–pair model, introduced in chapter 1. The combined statistical mechanics of sequence and deformation of this model allow predictions for biochemical competitive binding experiments. This comparison of structural and biochemical data is carried out in chapter 2 for a well–studied test case. In this way, the rather indirect parametrizations of the rigid base–pair model are directly compared to experiment. A further application is presented in 3: A new statistical marker allows the local detection of elastically optimized subsites in a given protein–DNA crystal structure.

The state of the art of parametrization of the rigid base–pair model is based

solely on indirect methods. To improve on this point, in the second part of the thesis, a connection is established between the microscopic parametrization of the rigid base–pair model and direct stiffness measurements on larger scales:

- What is the relation between sequence–dependent base–pair elasticity and effective mesoscopic elasticity of the molecule?

Here, a quantitative answer is possible. The discussion begins with a new theoretical framework for chains of elastically coupled rigid base–pairs. Large parts of chapter 4 are concerned with details of the mathematical formalism and may be skipped by the reader interested primarily in the physics; the main ideas are summarized in the last section. The main virtue of the formalism is that it allows a convenient description of the combined elasticity of groups of coupled base–pairs. This is put to work in chapter 5, where a systematic coarse–graining procedure is presented that links the rigid base–pair model to the worm–like chain, which is the established model of DNA elasticity on scales of hundreds of base–pairs. Chapter 6 extends the procedure to the case of irregular DNA sequence. This closes the gap between the microscopic parametrizations of the rigid base–pair model on one hand, and direct measurements of the conformational statistics of DNA, as well as single–molecule experiments of DNA stiffness on the other hand. Detailed quantitative comparisons are given.

The third part of the work is more theoretical in nature. It revolves around the question

- What is the appropriate continuum description of DNA elasticity with all rigid body degrees of freedom?

Generalizing the worm–like chain, a continuous model for the conformation of a chain of rigid base–pairs with a total of six, translational and rotational, local degrees of freedom is constructed in chapter 7. The resulting ‘continuous rigid body chain’ is motivated by DNA, but the model is more general; it may be applied for other macromolecules that exhibit coupled shear and bending deformation modes as well as for the diffusion of self–propelled particles. The formulation in terms of a Brownian path evolving on the Lie group of rigid motions allows explicit evaluation of several interesting moments of the end–to–end frame distribution. In chapter 8, the continuous rigid base–pair chain is treated in a Lagrangian formalism; the associated equations of motion govern the equilibrium shape of the chain, and allow to identify a set of conserved quantities. Finally, the linear response around an arbitrary, known equilibrium shape is computed.

The set of methods developed to address the questions above suggest a number of other exciting topics of research. They could be barely scratched in this work. Nevertheless, in the outlook chapter 9, some preliminary results are presented to give a hint how some of the following points could be investigated in the future:

- Do different classes of DNA–binding proteins have characteristic ways of recognizing DNA deformability?
- What local forces and torques does DNA experience in a given complex structure?
- Can the histone positioning observed throughout eucaryotic genomes be quantitatively explained by elastic effects?

1 DNA at the base pair level

The rigid base–pair model for DNA elasticity is introduced, and its basic assumptions as well as microscopic parametrization methods are discussed. This sets the ground for applications and further theoretical development in the later chapters.

1.1 Sequence dependent DNA elasticity

What is the best description of DNA elasticity? Judging by the sheer number of different models for DNA deformability that are in use, ranging from atomistic molecular dynamics (MD) interaction potentials to continuous semiflexible polymer models [Wig05, Gol00, Yam97, Kam97, O'H98, Kra49, Col03, Moa05, Mar94, Win03, Leb96], this is not a simple question. The answer, as usual, depends on what aspect and which length scale of the problem are most interesting.

1.1.1 Basic structure of the molecule

The structure of the DNA molecule has been known for more than 50 years [Wat53b, Wat53a]. Free DNA in physiological conditions occurs as a right–handed double helix in which two sugar–phosphate chains, the backbones, wind around a core of stacked base–pairs, see fig. 1.1.

The bases form planar pairs that are held together by hydrogen bonds. A two–cycle purine combines with its complementary single–cycle pyrimidine to form the Watson–Crick pairs Adenosine–Thymine (A·T) or Guanine–Cytosine (G·C), fig. 1.2.

Since the bases are covalently bound to the backbone sugar rings in an asymmetric manner, the two backbones are unevenly spaced, so that their double helical path around the base core leaves a small (minor) and a large (major) groove. The backbones are strongly negatively charged due to the presence of one phosphate group per base–pair step (bps). They also carry a structural asymmetry that allows to assign a direction; base sequences are conventionally read from the end where the phosphate is bound to the carbon at the 5' position, to the end where it is bound to the 3' carbon, see fig. 1.3.

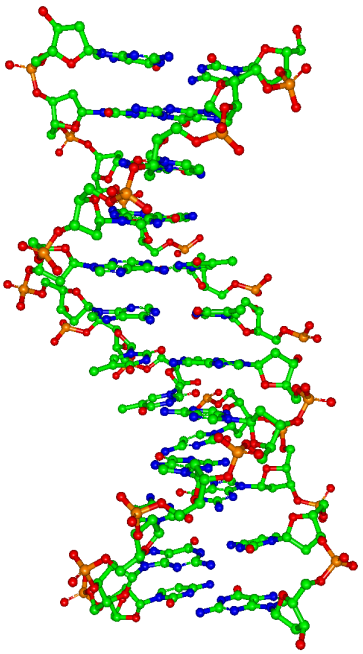

Figure 1.1 | B-form DNA oligonucleotide structure.[Dre81]

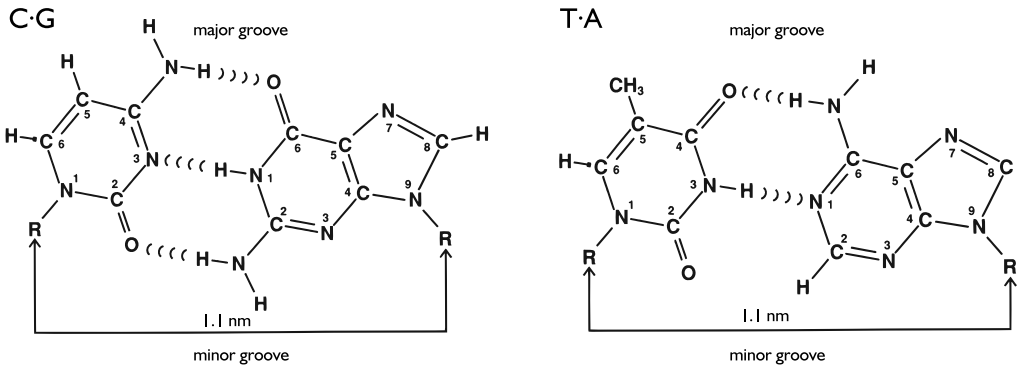

Figure 1.2 | The canonical Watson–Crick base pairs. Adapted from [Sin94].

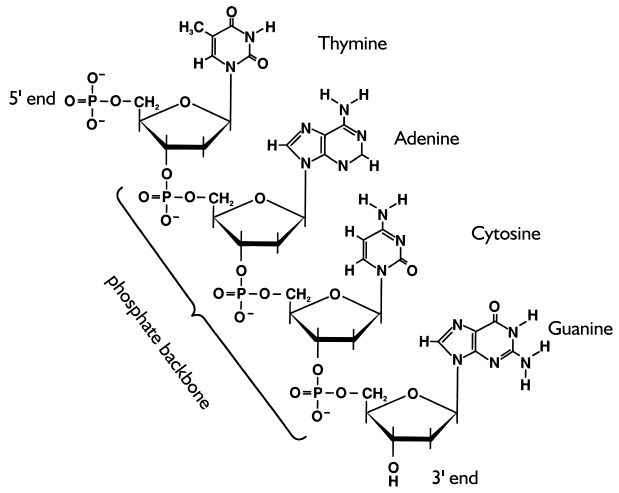

Figure 1.3 | A single stranded tetranucleotide. Adapted from [Sin94] (not to scale).

The two strands are paired together in opposite directions, so that the structure has a strand–change symmetry: Interchange of complementary bases combined with reversal of the base sequence, transforms the molecule to itself, except for a rotation by 180°. When the base sequence is disregarded, the molecule has no preferred direction but it is chiral: spatial inversion changes the handedness of the helix.

Apart from the B-form just described, DNA exists in a variety of other helical geometries (A, Z, etc.), depending on salt concentrations, humidity and on the tension in the molecule.

1.1.2 Quantum effects?

The deformations of the DNA double helix on the base–pair level will be treated throughout as a purely classical system, disregarding all effects of quantum interference. An estimate to justify this approximation follows.

The energy scale of thermal excitations is $k_B T \simeq 4\times10^{-21}\,\mathrm{J}$ at room temperature. An internal energy scale for the deformation of a bp step is given by a quantum mechanical energy level spacing $\hbar\omega$. Here $\omega = \sqrt{\kappa/m}$ is the characteristic frequency of a harmonic oscillator describing small deformations of one bp step. The stiffness can be estimated from the known deformability of DNA, to be discussed below. For the extension and shear modes, a typical value is $\kappa = 2\,\frac{\mathrm{N}}{\mathrm{m}}$. The mass of a naked bp is around $m \simeq 700\,\mathrm{a.u.} = 1.2\times10^{-21}\,\mathrm{g}$, so that $\hbar\omega \simeq 1.3\times10^{-22}\,\mathrm{J}$.

Thus the mean thermal energy is bigger than the quantum level spacing by at least one order of magnitude. This is not as much as one might have expected. Still, thermal excitations over many levels are possible, and the strong coupling to the surrounding heat bath is expected to destroy quantum correlations along the DNA molecule. In conclusion, a classical treatment of DNA conformations, starting from single base–pairs, appears justified.

1.2 Rigid base–pair elasticity

In this work, we focus on the sequence dependent linear elastic response of DNA ranging from a single bps to DNA loops several hundred base–pairs long. The natural discretization for sequence–dependent properties is one bps, about 0.34 nm. We consider a level of detail that captures all deformation modes at this scale of double–stranded DNA. In the corresponding model, each bp constitutes one basic unit without internal structure, and the DNA molecule is built up as a helical stack of base–pairs. Any two base–pairs are related through a rigid body transformation, i.e. by a three–dimensional rotation and translation, which specify their relative orientation and position in space. This widely used description is called the rigid base–pair (rbp) model [Cal04, Cal84].

The name is somewhat misleading: In contrast to the relatively rigid and planar aromatic rings of the individual bases, the hydrogen bonds that connect complementary bases are flexible, so that in atomic structures of DNA, base–pairs often deviate considerably from the coplanar equilibrium shape. Figure 1.4 illustrates the internal deformations of a bp as well as the rbp parameters which relate different base–pairs. In the rbp model, internal bp deformations are effectively averaged out, and each rigid bp represents the *mean* structure of a flexible real bp.

1.2.1 Basic mechanics

Many features of the elastic response of double–stranded DNA (dsDNA) can be understood by looking at a brick representation [Cal04] of the rbp model, fig. 1.5. Each brick has the spatial dimensions of a bp, about $0.3 \times 1 \times 1.8$ nm. The sugar–phosphate backbones of the double helix are approximated by inextensible sticks that are attached to the minor groove edges of each bp via flexible hinges.

Two basic physical effects govern the response of a bps to deformation. The van der Waals–like stacking interaction has an energy minimum for base–pairs

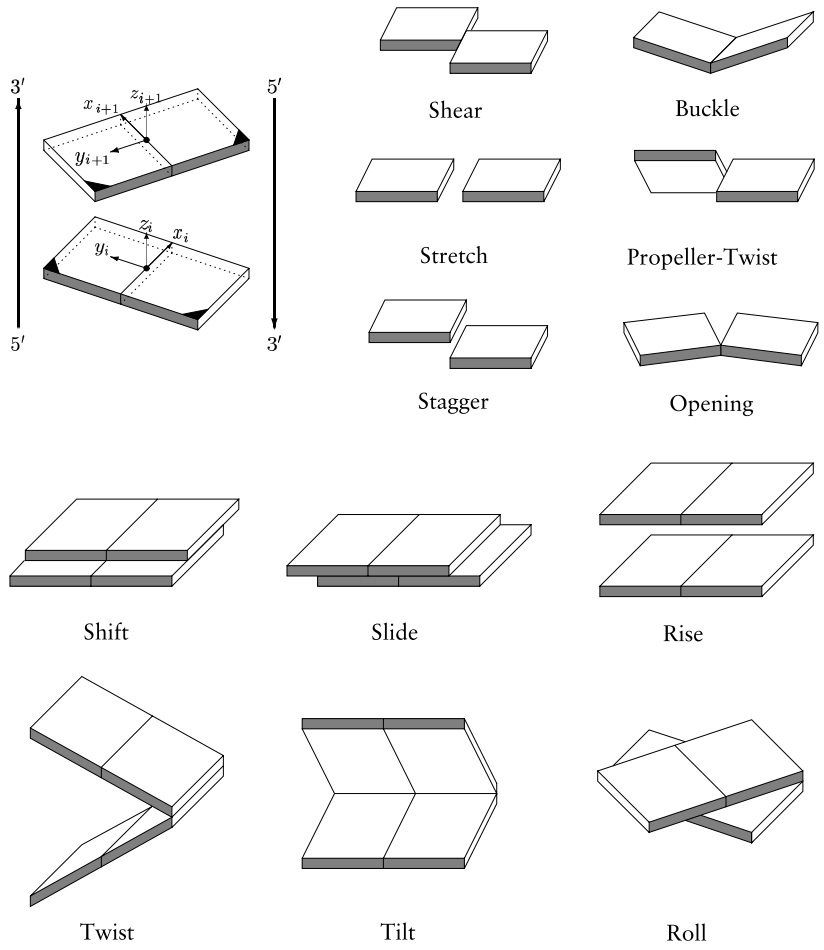

Figure 1.4 | Internal bp deformations (Shear, Stretch, Stagger, Buckle, Propeller–Twist, Opening) and rbp parameters (Shift, Slide, Rise, Tilt, Roll, Twist). Bases are represented as bricks, the minor groove face is shaded. Adapted from [Dic89].

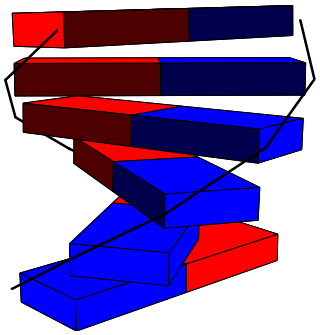

Figure 1.5 | Brick representation of double–stranded DNA. Base–pairs are represented as bricks, backbones as lines. The minor groove is shaded.

aligned on top of each other, i.e. when Rise $\simeq$ 0.34 nm and all other rbp parameters vanish. On the other hand, the backbone linkers with a fixed length of about 5.5 nm impose a conformational constraint that forbids aligned stacking. The helical shape with Twist $\simeq 33^\circ$ shown in fig. 1.5 (but not the handedness) then appears naturally as a ground state of the system. On the basis of this picture one can also guess some of the main features of the elastic response:

- Bending into the grooves (Roll) is easier than towards the backbones (Tilt).
- Stretching (ΔRise > 0) is coupled to untwisting (ΔTwist < 0)
- Slide and Shift are coupled to Tilt.

This intuition is correct, at least for fairly *large* deformations of dsDNA. A physical model expanding on this idea was developed in [Eve03, Mer03].

In this work, we will be concerned with *small* rbp deformations within the regime of linear response. These are strongly sequence dependent, and internal deformations lead to an elastic response that differs in some aspects from that expected by the arguments above. A notable example is the *anti*correlation of Twist and Stretch for small deformations.

1.2.2 State space

Within the rbp model, the conformation of a bps is described by a set of 3+3 variables specifying the relative position and orientation of the two base pairs. We combine these variables into a vector $q = (\mathrm{Ti}, \mathrm{Ro}, \mathrm{Tw}, \mathrm{Sh}, \mathrm{Sl}, \mathrm{Ri})^\mathsf{T}$. Here Ti, Ro, Tw are the rotation angles Tilt, Roll and Twist around the x, y and z-axes of the material frame. Correspondingly Sh, Sl, Ri are the translations Shift, Slide and

Rise along the axes, as depicted in fig. 1.4.[1]

To characterize a bps completely, one also needs to specify the identity of the bases b_1, b_2 along one preferred strand in 5' to 3' direction, that is their two–base sequence σ, e.g. $\sigma = b_1 b_2 = \mathrm{AC}$. A rbp step is then fully specified by (q, σ).

1.2.3 Strand change symmetry

When specifying the state (q, σ) of some rbp step, the choice of preferred strand is arbitrary. Therefore, physical quantities have to be invariant under the symmetry operation of switching strands. When changing from 'Watson' to 'Crick', the reading sense has to be reversed simultaneously, to keep the 5' to 3' convention, cf. fig. 1.4. Referring now to the Crick strand, one will describe the *same* physical bps by the complementary sequence $\overline{\sigma} = \overline{b_2}\,\overline{b_1}$, e.g. $\overline{\mathrm{AC}} = \overline{\mathrm{C}}\,\overline{\mathrm{A}} = \mathrm{GT}$, and by a new conformation $\overline{q}$. Conventionally [Dic89], the conformation variables are defined such that $\overline{q}$ has entries with a definite parity under strand change. Specifically $\overline{q} = \mathcal{J} q$, where $\mathcal{J} = \mathrm{diag}(-1, 1, 1, -1, 1, 1)$.[2]

Due to this symmetry, only ten out of the sixteen possible step sequences are physically different, and symmetry relations exist between the energy functions of complementary steps. For a detailed account thereof see [Col03].

1.3 Fluctuations of rigid base–pair steps

We now discuss the equilibrium statistical mechanics of uncoupled rigid base–pair steps, taking a probabilistic approach.

1.3.1 Joint distribution

Suppose we have by some means collected an ensemble $\{(q_i, \sigma_i)\}_{1 \leqslant i \leqslant N}$ of elastically fluctuating, independent rbp steps. Their conformations and sequences are jointly distributed according to some normalized probability density function (pdf)[3] $p(q, \sigma)$, which contains all available statistical information. This pdf is given with respect to the measure $dV_q d\sigma$, which reflects an unbiased distribution on the

[1] There are many different ways to define the material frame as well as the rbp parameters in detail, see sec. 4.2.8. In chapters 2 and 3 we will adhere to the definition used in the 3DNA program [Lu03].

[2] I.e, the body x-axis vector is even and the y, z-axes are odd under strand change.

[3] As is customary, the various pdfs are always written with the same symbol p and can be distinguished by their arguments.

state space. The sequence measure $d\sigma \equiv 1$ just assigns unit weight to each step sequence and will be omitted. The conformation measure dV_q depends on the choice of curvilinear coordinates and is generally *different* from $dq^1 \cdots dq^6$, see chapter 4.

At inverse temperature $\beta = (k_B T)^{-1}$ we associate to the joint pdf a free energy K, where

$$\beta K(q, \sigma) = -\ln[\nu\, p(q, \sigma)]. \tag{1.1}$$

The constant ν is a volume scale in q space needed to fix dimensions, and will drop out in all free energy differences. Log–relative probabilities of bps that differ in sequence and structure, are K–differences:

$$\ln\left[\frac{p(q', \sigma')}{p(q, \sigma)}\right] = \beta\left(K(q, \sigma) - K(q', \sigma')\right). \tag{1.2}$$

Taking partial averages, we get the marginal pdfs: $p(\sigma) = \int p(q, \sigma) dV_q$ gives the frequency of a sequence σ in the ensemble while $p(q) = \sum_\sigma p(q, \sigma)$ is the pdf to find the conformation q in any sequence step. Using the notation of a dot $\cdot$ for an empty slot in a function, one can also write the them as expectation values: $p(\sigma) = \langle \delta_{\sigma\cdot} \rangle$ and $p(q) = \langle \delta(q - \cdot) \rangle$.

1.3.2 Conformation distribution

Sequence–dependent elasticity determines the conformation probabilities for fixed sequence. They follow the normalized conditional pdf to find q given σ,

$$p(q|\sigma) = \frac{p(q, \sigma)}{p(\sigma)}. \tag{1.3}$$

We associate a conformation free energy,

$$\beta F_\sigma(q) = -\ln[\nu\, p(q|\sigma)] = \beta K(q, \sigma) + \ln[p(\sigma)]. \tag{1.4}$$

A free energy difference $F_\sigma(q) - F_\sigma(q')$ expresses the relative probability to observe the conformation q' rather than q in the data, given that one is looking at a fixed sequence σ. F differs from K only by a sequence–dependent normalization offset.

1.3.3 Sequence distribution

Similarly, we may ask for the probability to find the sequence step σ among all steps at fixed conformation q in the ensemble. It is given by the (discrete) normalized

conditional pdf

$$p(\sigma|q) = \frac{p(q,\sigma)}{p(q)}, \tag{1.5}$$

and we associate a sequence potential

$$\beta G_q(\sigma) = -\ln p(\sigma|q) = \beta K(q,\sigma) + \ln[\nu\, p(q)]. \tag{1.6}$$

A potential difference $G_q(\sigma) - G_q(\sigma')$ expresses the relative probability to find the sequence σ' rather than σ, at a fixed conformation q. By normalization, when $G_q(\sigma) = 0$, the sequence σ occurs with certainty among steps with conformation q. G differs from K by a conformation–dependent normalization offset.

1.3.4 Relations between free energies

Quite generally, differences in K can be split up into ΔF and ΔG terms :

$$K(q,\sigma) - K(q',\sigma') = F_\sigma(q) - F_\sigma(q') + G_{q'}(\sigma) - G_{q'}(\sigma') \tag{1.7}$$

Often, it is interesting to compare sequences in an unbiased ensemble where each sequence step is equally probable, so $p(\sigma) = const$. In this special situation, the formulas look simpler. E.g,

$$\beta G_q(\sigma) = \beta F_\sigma(q) + \ln \sum_{\sigma'} e^{-\beta F_{\sigma'}(q)}. \tag{1.8}$$

From (1.7) or (1.8), also $G_q(\sigma) - G_q(\sigma') = F_\sigma(q) - F_{\sigma'}(q)$, so the relative probabilities of sequences are in this case expressed by their F differences.

1.3.5 Thermodynamic analogy

A analogy to basic thermodynamics may help clarify how the different free energies are related. Note first that deformation and sequence are not conjugate variables, so F and G are not related by a Legendre transformation. Consider a thermodynamic system at constant temperature consisting of some gas in a box. We let the deformation of a bps correspond to a change of the volume V of the box. Further, the step sequence is analogous to the chemical composition of the gas in the box, given by the particle numbers N_i.

In this setup, fixed σ corresponds to a closed box with a certain gas species. Since the N_i cannot change, the Helmholtz free energy A of the system is then a function of the volume only, $dA = p\, dV$. It corresponds to the conformation free

energy F which is a function of the deformation. Indeed, statistical mechanics tells us that the Helmholtz free energy is the log of the normalized canonical partition function, which in the bp setting, corresponds to $p(q|\sigma)$.

On the other hand, fixing q while allowing σ to vary, corresponds to an open box allowing particle exchange, at fixed volume. In this situation, the Helmholtz free energy is a function of the particle numbers only, $dA = \sum \mu_i dN_i$. Consequently, the sequence potential G corresponds to the Helmholtz free energy, in the grand canonical ensemble with the constraint of fixed volume. Taking the analogy a step further, when considering single bp steps, we can index the set of particle numbers by the sequence; $\{N_i\} = \sigma$ corresponds to one particle of type σ, and no particles of other types. Then, a sequence free energy difference $G_q(\sigma) - G_q(\sigma')$ can be identified with a difference in chemical potential $\mu_\sigma - \mu_{\sigma'}$ of the two species.

Releasing the volume constraint, the joint pdf $p(q, \sigma)$ corresponds to the grand canonical partition function of the gas mixture in the box, and the grand potential K is the Helmholtz free energy without constraint; $dA = p\, dV + \sum \mu_i dN_i$.

In summary, the different free energies arise by imposing different kinds of constraint on the system.

1.4 Fluctuations of rigid base–pair chains

We build up a statistical model for DNA by combining independently fluctuating rbp steps into a chain.

1.4.1 Basic assumptions

A piece of DNA in solution undergoes thermal fluctuations. We describe it as a chain of rbp steps or short, a rbc. The main basic assumption of the model is that *conformational fluctuations* of any two rbp steps along the chain *are independent*. In other words, coupling terms between steps in the conformational free energy are neglected. This also means that the *internal* base–pair deformations are treated on a mean–field level: fluctuations within a bps are averaged, and correlations in internal fluctuations between steps are discarded.

The assumption of independence is motivated by mathematical simplicity but also by the fact that microscopic parameter sets for the conformation free energy are available only without coupling of neighboring steps, see sec. 1.6. It is worth mentioning that independent step conformations also imply that no repulsive self–

contact interactions in a looped rbc are included in the model. We are therefore considering only 'ideal' chains. For chains shorter than a few bending persistence lengths, this is not a serious limitation.

On the other hand, there is no assumption of linear elasticity inherent the model. The functional form of the conformation free energy is in principle completely arbitrary. Again, microscopic parameters are available only for the regime of linear elasticity, so only this case will be considered in detail later on.

1.4.2 Free energies

We now extend the free energies introduced above for single steps, to a chain of consecutive steps. By the basic assumption of the rbp model, bps conformations are independent random variables. However, we have to make sure that consecutive steps form a meaningful sequence, e.g. AC can only be followed by CN where $N = A, C, G, T$. This requirement of sequence continuity correlates the sequences of neighboring steps. Clearly, the correlation is just a result of considering the bp steps as the basic objects rather than the individual base pairs.

Extending previous notation, we now denote a rbp *chain* made of l bps by $(q, \sigma) = ((q_j, \sigma_j))_{1 \leqslant j \leqslant l}$. We additionally require that the sequence steps match up, $\sigma_j = b_j b_{j+1}$ where $\sigma = b_1 \dots b_{l+1}$ is some sequence of $l + 1$ bases.

We now compute the free energies for chains. To start with, we immediately have $p(q|\sigma) = \prod_j p(q_j|\sigma_j)$ since the conformations are independent. Consequently the chain conformation free energy, $F_\sigma(q) = \sum_j F_{\sigma_j}(q_j)$ is the sum of step free energies.

Chain free energies depending on the sequence argument σ are generally not additive. This is because the sequence pdf $p(\sigma)$ has to be renormalized so that its sum over all *matching* sequences $\sum'_\sigma = \sum_{b_1,\dots,b_{l+1}}$ is unity. If the normalization factor

$$W_l = \sum_{\sigma'}{}' \prod_{i=1}^{l} p(\sigma'_i), \tag{1.9}$$

then clearly $p(\sigma) = W_l^{-1} \prod_j p(\sigma_j)$ is the properly normalized sequence distribution. In the special case where all sequences are equally likely, one can check that indeed $p(\sigma) = 4^{-(l+1)}$.

Likewise one finds that the joint distribution $p(q, \sigma) = W_l^{-1} \prod_j p(q_j, \sigma_j)$. This

renormalization makes the joint free energy K non-additive,

$$\beta K(q,\sigma) = -\ln[\nu^l p(q,\sigma)] = \beta \sum_{j=1}^{l} K(q_j,\sigma_j) + \ln W_l. \tag{1.10}$$

Finally, we compute the sequence distribution for given chain conformation as

$$p(\sigma|q) = \frac{p(q,\sigma)}{\sum'_{\sigma'} p(q,\sigma')}. \tag{1.11}$$

The resulting sequence free energy,

$$\beta G_q(\sigma) = -\ln[p(\sigma|q)] = \beta K(q,\sigma) + \ln[\nu^l \sum_{\sigma'}{}' p(q,\sigma')] \tag{1.12}$$

can be written in a more compact form. We first note that

$$\nu^l \sum_{\sigma'}{}' p(q,\sigma') = W_l^{-1} \prod_{j=1}^{l} \nu\, p(q_j) \sum_{b'_1,\ldots,b'_{l+1}} p(b'_j b'_{j+1}|q_j). \tag{1.13}$$

We now introduce the 4×4 transfer matrix $T(q_j)$ with entries

$$\big(T(q_j)\big)_{b',b''} = p(b'b''|q_j) = e^{-\beta G_{q_j}(b'b'')} \tag{1.14}$$

and rewrite the primed sum as a matrix multiplication. With $\mathbb{1}^T = (1,1,1,1)$ and using (1.10) and (1.6), the sequence free energy of a rbc can be rearranged as

$$\beta G_q(\sigma) = \beta \sum_{j=1}^{l} G_{q_j}(\sigma_j) + \ln[\mathbb{1}^T T(q_1)\cdots T(q_l)\mathbb{1}]. \tag{1.15}$$

Note that for $l = 1$, the formula does reduce to the single step result since $p(\sigma|q)$ is normalized. G is not stepwise additive, and is defined as an average over an exponentially growing set of 4^{l+1} sequences (1.12). Still when using the transfer matrix approach, the computational cost of evaluating it is only $O(l)$! No approximation by an additive quantity (as used in [Mor05] in a related context) is necessary for efficient computation in longer chains.

Finally, from eqns. (1.10) and (1.12) the basic relation $\Delta K = \Delta F + \Delta G$ (1.7) follows also for chains of bps. Whenever the sequences are equidistributed, the chain free energies reduce to simpler expressions. In particular, one can see from eqns. ((1.14),(1.15)) that like for single steps, $G_q(\sigma) - G_q(\sigma') = F_\sigma(q) - F_{\sigma'}(q)$ whenever $p(\sigma) = const.$

1.5 Linear elasticity of rigid base–pair steps

How are the elastic free energies introduced above related to the more familiar concepts of *elastic energy* and *linear elasticity*?

1.5.1 Linear response

Consider a rbp step with fixed sequence in thermal equilibrium, so that its conformation is a random variable. When an external generalized force[4] μ is exerted on the step (by global bending of the chain, protein contacts etc.) the conformation distribution is modified; we can write this as $p(q|\sigma;\mu)$ or $p(q|\mu)$, suppressing sequence notation in the remainder of this section. The response of its first moment, the average step conformation $\langle q|\mu\rangle$, may be written as

$$\langle q|\mu\rangle - \langle q|0\rangle = \beta C\mu + o(\mu). \tag{1.16}$$

Here, βC is the linear response coefficient. The force and torque μ is a six-dimensional vector, so C is a 6×6 matrix. To parametrize it, it is not necessary to actually measure the linear response in experiment, since C is related to the *equilibrium* fluctuations of q. From linear response theory one knows that C is identical to the covariance matrix of deformations at zero force:

$$C^{ij} = \left\langle (q - \langle q|0\rangle)^i \, (q - \langle q|0\rangle)^j \,\middle|\, 0 \right\rangle. \tag{1.17}$$

1.5.2 Linear elasticity

From a slightly different perspective, one may view the rbp step as an elastic element and write an expansion of its elastic internal energy to second order in the strains $(q - q_0)$ as

$$E(q) = \frac{1}{2} (q - q_0)^T S (q - q_0) + O(q - q_0)^3, \tag{1.18}$$

where S is the 6×6 stiffness matrix. In a thermal environment with no external force and at inverse temperature β, one obtains a Boltzmann distribution of step conformations,

$$p(q|0) = Z^{-1} e^{-\beta E(q)}, \tag{1.19}$$

[4] μ is the variable conjugate to q, as discussed in detail in chapter 4.

where the partition sum $Z = \int e^{-\beta E(q)} dV_q$. At this point, the choice of curvilinear coordinates matters, since the volume element $dV_q = A(q) d^6 q$ makes the integral non-Gaussian [Gon01]. However, for typical rbp steps in DNA, the distribution $p(q|0)$ is sharply peaked around q_0. Then to a good approximation, the metric factor $A(q)$ is constant, and one can also neglect the finite integration boundaries of the angular part of q. By doing a Gaussian integral,

$$Z = \det(2\pi\beta S)^{-1/2}, \tag{1.20}$$

$$\langle q|0\rangle = q_0, \text{ and} \tag{1.21}$$

$$\left\langle (q-q_0)^i (q-q_0)^j \,\middle|\, 0 \right\rangle = \left((\beta S)^{-1}\right)^{ij}. \tag{1.22}$$

If we let $\beta S = C^{-1}$, this is consistent with (1.16), bringing the two views into agreement. For a somewhat more accurate version of these small–angle relations for particular choices of coordinates, see section 4.2.8 and appendix A.7.

Conceptually, it is misleading to think of a rbp as a macroscopic elastic element, since the thermal environment is inherent in this microscopic system; the elastic response as well as the structure of DNA depend heavily on the solution conditions and temperature, and an elastic response of DNA without fluctuations is an abstraction that does not correspond to a realizable experiment. So when talking about linear elasticity in the following, this is always to be understood in the sense of linear response theory of average quantities as outlined in section 1.5.1.

In particular, q_0 and S are *defined* by eqns. (1.21) and (1.22), and the quadratic elastic energy $\frac{1}{2}(q-q_0)^T S (q-q_0)$ is by definition the second order term in the expansion of the conformation free energy F around its minimum. Using ((1.19),(1.4)) we can write

$$F_\sigma(q) = E_\sigma(q) + \beta^{-1} \ln Z(\sigma) + \beta^{-1} \ln \nu \tag{1.23}$$

Disregarding the irrelevant global constant $\beta^{-1} \ln \nu$, we can rewrite this as $F = E - T\Sigma$ where the term $\Sigma(\sigma) = -k_B \ln Z(\sigma)$ has the form of a *sequence–dependent entropy* of the harmonic rbp step. Clearly, Σ is not the thermodynamic entropy, $\Sigma \neq -\frac{\partial F}{\partial T}$ since E as defined here is *not* the true internal energy of the system. Rather, E is a version of the free energy F, with a sequence–dependent offset which ensures that $E(q_0) = 0$. The true entropic and enthalpic parts of F are not separable just from data at constant temperature.

By construction, C is a positive definite, symmetric matrix, therefore also S has

these properties, so that equilibrium (or spontaneous) configuration q_0 is stable . Both q_0 and S depend on the sequence σ of the step.

1.5.3 Boltzmann inversion

Whenever the conformational distribution $p(q|\sigma)$ is given as a result from experiment, eqns. ((1.21),(1.22)) can be used to extract $q_0(\sigma)$ and $S(\sigma)$. This is equivalent to fitting the Boltzmann distribution (1.19) with a six-dimensional Gaussian, which is the maximum entropy distribution with the mean and covariance of the data. In the following, we will always use the linearized versions of the elastic free energies K, F and G which result from such a fitting procedure.[5]

The basic assumption here is that the observed pdf is indeed an equilibrium distribution at a certain fixed temperature T, so that thermal energy is equally distributed among hard and soft modes: $\frac{1}{2}C^{ij}S_{jk} = \frac{1}{2}k_B T\delta^i_k$.

It is worth mentioning that when external constraints are present, this assumption can easily break down. Although the total system may have a Gaussian distribution, constraints destroy internal equipartition of energy. As an example, imagine a toy model in which two springs with stiffness constants $k_1 > k_2$, are arranged in series, so that they share the common tension f. Consequently, their elongations have the constant ratio $\frac{x_2}{x_1} = \frac{fk_1}{fk_2} = \frac{k_1}{k_2}$. If now the combined spring elongation $x = x_1 + x_2$ has a Boltzmann–like distribution with a certain variance $\langle x^2 \rangle$, one calculates immediately that the ratio of mean energies is $\frac{1/2\, k_2 \langle x_2^2 \rangle}{1/2\, k_1 \langle x_1^2 \rangle} = \frac{k_1}{k_2} > 1$, so equipartition is violated. Similar considerations apply to elongation rather than force constraints, or to springs in parallel.

1.6 Microscopic parametrization of rbp potentials

To completely specify the linear elasticity of the rbp model, for each step there are 6 (for q_0) + 21 (for S) parameters required, which gives a total of 270 (!) for the 10 different steps. It is not surprising that even the most detailed rbp parameter sets are only given in linear approximation. Two different approaches [Ols98, Lan03] have been utilized to tackle the problem of parametrization, detailed below. Both start from a set of atomistic bps structures which are interpreted as representing a sample from a thermal equilibrium distribution. The recipe then consists of first

[5] With more detailed (multi-modal) free energy functions, parametrization would become even more difficult and conformation space integrals would get more involved but the formalism would not change.

extracting the corresponding ensemble of rbp parameters and then computing the first moments.

1.6.1 Extraction of rbp parameters

In order to compute the rbp parameters of a particular bps in a given all-atom structure, the standard procedure is to perform a least-squares fit of an ideal, coplanar model base–pair to each of the deformed base–pairs that make up the step. The rbp parameters of the step are then defined by the rigid body transformation between reference frames fixed to the two best–fit model base–pairs.[6]

This procedure involves a 'physical' choice of details of the fitting procedure and reference frame used and a 'mathematical' choice of parameters used to describe the rigid body transformation between the reference frames. Both of these choices have been partially fixed in the community by agreements on basic symmetry properties of the parameters [Dic89] and on a reference frame [Ols01]. For a comparison of several different extraction schemes as implemented in various computer programs see [Lu99a, Lu99b]. In the following we will use the 3DNA program by Lu and Olson [Lu03].

1.6.2 Molecular dynamics simulation

Lankaš et al. [Lan03] obtained an ensemble of fluctuating base pair steps at temperature $T = 300K$ from MD simulation of oligonucleotides. Under the assumption that the MD trajectories are equilibrated sufficiently, the bps ensemble is Boltzmann distributed, and the equilibrium values [Lan06b] and stiffness matrices [Lan03], $\{q_{0,MD}, S_{MD}\}$ can be extracted as described above. The partition sum $Z(\sigma, T) = \det(2\pi\beta S(\sigma))^{-1/2}$ gives a natural measure for the overall strength of fluctuations, counting all six degrees of freedom.

1.6.3 Crystal structure analysis

In an experimental, but rather indirect approach, Olson et al. [Ols98] used ensembles of statically deformed bps, obtained from high–resolution DNA crystallographic structures. Their 'B-DNA' ensemble consists of B-form DNA oligonu-

[6] In this sense, it is the least–squares fitting procedure that defines exactly which deformations are internal to the bp, and which are step parameters.

cleotide structures, while their 'P•DNA' ensemble is obtained from protein–DNA co-crystals.

The means $q_0(\sigma)$ and covariance matrices $\hat{C}$ describe the ensemble on a Gaussian level. Stiffness matrices can be extracted under the additional assumption that equipartition of energy holds also in crystal ensembles at some, yet undetermined *effective* temperature. External constraints such as force balance make this a problematic assumption, see section 1.5.3. However, this strategy is the only way of gaining experimental access to rbp flexibility.

To fix the energy scale given by the effective temperature, we require that the fluctuation strength of the MD ensemble and that of the crystal ensembles X = B, P be equal on sequence average, i.e. $\langle Z_X(\sigma, T)\rangle = \langle Z_{MD}(\sigma, T)\rangle$ [Bec07]. If we define the crystal stiffness matrices by $S_X(\sigma) = k_B T_X \hat{C}_X(\sigma)^{-1}$, then the effective temperature definition

$$T_X = 300\,\mathrm{K}\Big\langle \Big(\frac{\det \hat{C}_X(\sigma)}{\det \hat{C}_{MD}(\sigma)}\Big)^{\frac{1}{6}} \Big\rangle \tag{1.24}$$

satisfies this requirement. Performing the calculation, we obtain $T_B = 107\,\mathrm{K}$ and $T_P = 233\,\mathrm{K}$. Our resulting B and P ensembles then have equilibrium values and stiffness matrices $\{q_{0,B}, S_B\}$ and $\{q_{0,P}, S_P\}$, and each ensemble has by construction the same overall stiffness as the MD simulations [Lan03]. After the effective temperature is set, we replace the observed distribution of deformations, $\hat{p}_X(q|\sigma)$, by the corresponding Boltzmann distribution $p_X(q|\sigma)$ at $T = 300\,\mathrm{K}$. This distribution has covariance $C_X(\sigma) = (k_B 300\mathrm{K}) S_X^{-1}(\sigma)$.

Also in [Lan03], the effective temperature for the P-DNA ensemble [Ols98] was computed, by comparing the persistence lengths for DNA oligomers as extrapolated from a normal mode analysis of oligomers without temperature scale [Mat02], to experimental values for B-DNA in solution. This yielded a value of $T_{P,\,La} = 295\,\mathrm{K}$. While our microscopic approach matches fluctuations of all six rigid bp degrees of freedom to an MD simulation, this mesoscopic method effectively matches the *bending* fluctuations only, to experimental data. For comparison, we have repeated our fixing of effective temperatures, eqn. (1.24), using only the bending (i.e, Roll and Tilt) stiffness submatrices. This gives effective temperatures of $T_{B'} = 166\,\mathrm{K}$ and $T_{P'} = 232\,\mathrm{K}$, the latter value surprisingly unchanged from T_P. We denote the resulting crystal ensembles by B' and P'.

1.6.4 Hybrid potential parametrizations

In rescaling the crystal stiffness matrices with a single parameter to match MD simulation we have, strictly speaking, constructed a hybrid parametrization. One could extend the procedure, introducing multiple effective temperatures that match all sequences or even all deformation degrees of freedom separately to the MD stiffness matrices. At the extreme, one ends up with the B and P equilibrium values combined with the pure MD stiffness matrices. We also include these hybrid combinations $\{q_{0,B}, S_{MD}\}$ and $\{q_{0,B}, S_{MD}\}$ in the analysis, denoted MB and MP, respectively.

Although this combination of data from different sources seems somewhat artificial, it does avoid some of the weaknesses of the 'pure' approaches:

- The equilibrium values obtained from MD using the parm94 force field [Cor95] are known to have Twist and Rise values that are lower than commonly accepted on the basis of structural data [Bev04].
- The stiffness matrices (even if rescaled) obtained from the crystal ensembles suffer from the unjustified assumption of equipartition of energy.

In chapter 2, we compare the success of different parametrizations MD, B, P, MB and MP in predicting binding affinities.

2 Indirect Readout in Protein-DNA complexes

Sequence–dependent elasticity of DNA plays an important role in regulating specific protein–DNA interactions. The formalism developed in the previous chapter can give some insight into the mechanism of regulation. A test of its validity using biochemical data on bacteriophage 434 repressor–DNA affinities is presented.

2.1 DNA-protein recognition

The DNA base sequence together with the genetic code as a dictionary encodes for the amino acid sequence of all proteins that a living cell can produce. However, the set of expressed proteins is not nearly enough information to keep a cell running. At the very least, the expression levels of proteins have to be regulated in response to environmental conditions, cell fate, cell cycle phase etc. Also, the DNA molecule has to be physically handled; packing and replication need to be spatially organized to allow the separation of genetic material at cytokinesis. For all of these processes, targeted binding of a host of specialized proteins to their specific sequence motifs on DNA is essential.

2.1.1 Another code in DNA?

It is an appealing idea is that besides the genetic code, an additional sequence code is used at the binding sites of regulatory proteins on DNA, the operators. This 'recognition code' would be used to store information required for gene regulation and DNA management, and read out by the DNA–binding proteins. Despite much effort to understand the mechanisms of protein–DNA recognition, it has proved impossible to decipher a simple sequence code that can explain specific protein–DNA interactions, based on direct chemical contacts between amino acid side chains and bases [Mat88]. Refined versions of a recognition code include adapted weights for each combination of residue and base, which may also depend on their spatial arrangement [Suz95, Pab00, Cho97]. Their applicability is however restricted to certain geometries or certain protein classes. It seems that the recognition code resembles an industrial–strength encryption algorithm

much more than a simple look-up table for codons. A possible explanation of this complexity lies in the fact that DNA–protein recognition involves the *coupled* elastic properties of DNA along the operator site.

2.1.2 Mechanisms of specific binding

The non-covalent binding (or complexation) of a protein to a specific stretch of DNA is driven by the free energy gained in bringing the amino residues into contact with their base counterparts. This involves enthalpic contributions, e.g. from the formation of hydrogen bonds and salt bridges, as well as entropic parts, mainly due to changes in the solvent entropy [JJ00], see also [Bru02] for a lucid review of the basic physics involved. The complexation free energy thus depends on chemical properties of the bases, which allows for specific binding according to the chemical base identity. Thus, the DNA operator sequence is directly read out by the protein, which usually binds in the major groove. An example in which binding specificity is dominated by *direct readout*, is the zinc–finger class of proteins, see e.g. [Dau99, Cho97].

However, complexation necessarily also depends on the *elastic* free energy required to distort both the protein and the operator into their three–dimensional structure in the complex. This extra term always disfavors complexation, but it does depend on DNA sequence. In this way, sequence–dependent structure and deformability can contribute to binding specificity. This effect is called *indirect readout*, and has been found to be important in the transcription factors bacteriophage 434 repressor [Kou87], lac repressor [Sas90a, Sas90b] and papillomavirus E2 protein [Hin98], among many other examples. It is also important in nucleosome positioning [Tha99, Wid97]. See [Kou06] for a recent review of indirect readout.

The relative importance of direct and indirect readout for protein–DNA binding affinities have also been addressed computationally. The considered elastic models range from a combination of fixed coarse–grained protein structure with DNA rigid rod [Gro97, Gro05], rigid base–pair [Ste02, Gro04] and rigid base [Mor05] models, to all-atom force fields with partial protein structure relaxation [Pai04a, Pai04b] and more recently, with residue–dependent protein sidechain relaxation [End06, Ash06], leading to numerically demanding algorithms.

In the following the focus is on the *indirect* readout part of the complexation free energy and its relation to the elastic free energies within the rbp model.

2.1.3 Competitive binding

In an idealized experiment, consider a protein that can bind different operator sequences σ , deforming them into corresponding conformations q. This idealized 'indirect readout–only' protein has no *intrinsic* sequence preference, i.e. we assume that direct readout which drives complex formation has the same strength for all operators. It may be put in contact with a solution containing an ensemble of fluctuating rbp chains, perhaps viral DNA, containing operators with relative frequencies $p(\sigma)$. Further assume that the free energy required to deform the *protein* into its structure in the complex does not depend on the DNA sequence. In this situation, the relative occupancies of the protein with the different operators are entirely determined by elastic free energy differences.

Fixing an operator with sequence σ into a structure q costs a deformation free energy $F_\sigma(q)$. We multiply the probability $p(\sigma)$ to find σ at all in the ensemble and get the relative occupancy Q of (q, σ) compared to (q', σ') ,

$$Q = \frac{p(q, \sigma)\, dV_q}{p(q', \sigma')\, dV_{q'}} = e^{-\beta(K(q,\sigma)-K(q',\sigma'))}. \tag{2.1}$$

Using (1.7), this expression simplifies in the following two special cases:

On one hand, whenever all steps in the bps ensemble are equally frequent ($p(\sigma) = const$), those sequences will bind best whose bound structures are the most relaxed. Here,

$$Q = \frac{p(q|\sigma)}{p(q'|\sigma')} = e^{-\beta(F_\sigma(q)-F_{\sigma'}(q'))}. \tag{2.2}$$

F accounts for the entropic cost of fixing the rbp deformation fluctuations to a value q in the complex; softer steps acquire a higher entropic penalty of complexation. Note that substituting the elastic energy E in (2.2) would give different results, since E does not capture this sequence–dependent physical effect. For the parameter sets we used, the term $E - F = T\Sigma(\sigma)$ varies by up to 2 $k_B T$.

On the other hand, the protein may be very stiff, forcing all sequences into one fixed deformation q, combined with an arbitrary sequence distribution $p(\sigma)$. In that situation

$$Q = \frac{p(\sigma|q)}{p(\sigma'|q)} = e^{-\beta(G_q(\sigma)-G_q(\sigma'))}. \tag{2.3}$$

The sequence that minimizes G is the one that fits best with the prescribed structure, when weighted with its frequency in the ensemble.

These two special cases coincide when q is fixed and $p(\sigma)$ is uniform. Then

indeed F and G differ only by a constant, giving identical relative occupancies, see (1.8).

How realistic is the idealized experiment discussed above? Neglecting direct readout is far off most of the time, but it can be justified for appropriate *subsites* of the operators, see chapter 3. For the sake of argument we have also treated the bound conformation of the step as non-fluctuating. This is not a necessary assumption. In fact, when treating the bound fluctuations as finite but *independent* of (q, σ), all free energy differences are unchanged from the values given above. Thus, the approximation made here is effectively that of weak dependence of bound fluctuation strength on sequence and conformation.

2.1.4 Sequence–structure threading

In the sequence–structure threading approach (see e.g. [Pai04b, Mor05, Gro97]), a set of different operator sequences is threaded through a single, fixed complex structure, usually obtained from x-ray scattering. The resulting free energies are then used to predict binding affinities in solution. This can clearly work well only if the crystal structure is representative of the protein in solution. Further, the protein needs to be much stiffer than the operator, enforcing a single, sequence–independent bound conformation also in solution. At the same time, a stiff protein stores little deformation energy, which justifies the assumption of sequence–independent protein conformational energy. Also, the remaining fluctuations of the bound rbp are suppressed by a stiff protein, so that their sequence dependence can be neglected.[1]

In summary, the stiff protein limit justifies the idealizations made in the previous paragraph, and coincides with the special case (2.3). The sequence–structure threading approach will next be applied to a specific test case to evaluate its performance.

[1] A toy model to clarify the limit: We represent a protein by a linear spring with stiffness k_{pr} and resting position x_{pr}. The DNA operator sequence is represented by a spring with stiffness $k_{op} = \eta k_{pr}$ and resting position x_{op}. Upon 'complexation', driven by some external binding energy, the springs are connected in parallel, summing up to k. Then one calculates the resting position of the complex to be $\frac{k_{pr}}{k} x_{pr} + \frac{k_{op}}{k} x_{op} = x_{pr} + O(\eta)$. At this position the ratio of stored elastic energies becomes $\frac{E_{pr}}{E_{op}} = \eta$. In this sense, in the stiff protein limit $\eta \to 0$, the bound DNA conformation and protein deformation energy are sequence–independent. Deviations occur in first order in the relative stiffness η.

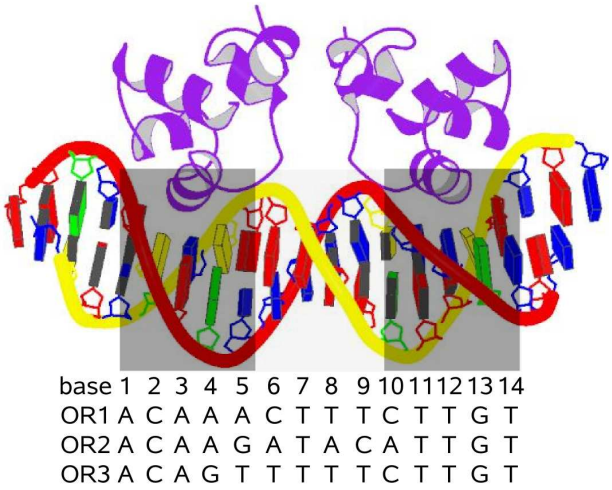

Figure 2.1 | Representation of 434 repressor–O_R3 complex structure [Rod93]. The outer 5+5 and the inner 4 base pairs are shaded differently. Together they form the 14 base pair binding site. The O_R sequences are also shown.

2.2 Indirect readout in 434 repressor

The bacteriophage 434 repressor, fig. 2.1, is a well-studied example of indirect readout. Mutations of the *non-contacted* region of the protein were surprisingly found to affect DNA binding affinities. This was one of the first pieces of evidence for indirect readout [Kou87].

2.2.1 Structure of the complex

The 434 repressor is a viral transcription factor that forms part of a genetic switch between the lytic and lysogenic states in the bacteriophage 434 virus. There exist two operator regions O_R, O_L in the bacteriophage genome with three binding sites of 14 base pairs in each region [Kou87]. High–resolution x-ray crystal structures have been solved for the three operators $O_R1,2,3$ [Agg88, Shi93, Rod93]. The protein binds in dimeric form in a so-called helix–turn–helix motif, making the complex approximately two-fold rotationally symmetric. The outermost 5+5 bases on each binding site are directly contacted by two α-helices of the protein dimer. The sequence of the outermost 4+4 bases is conserved in all six $O_{R,L}$ binding sites, with a single base exception. The consensus sequence of the contacted outer 5+5 bases shows the two-fold symmetry that can be expected from the

structural symmetry.

In contrast, the inner four bases are *not* contacted directly. Their sequence is neither conserved nor rotationally symmetric. Interestingly, binding affinities of the native binding sites vary 40-fold, and those of synthetic binding sites can vary as much as 200-fold, depending only on the sequence of the inner four bases [Agg88, Kou87]. This is true even though in the existing structures none of the individual bps is kinked strongly, and the overall bend is moderate, between 25 and 40 degrees. In gel shift experiments [Kou91], the overall bend was estimated to be small and sequence-independent, supporting the idea that the protein is indeed stiffer than DNA also in physiological solution conditions.

A correlation of affinity to the twisting rigidity and intrinsic twist of these mutations was found in further biochemical studies, such as insertion or deletion of a bp in the central region [Kou92, Kou88, Kou98].

Together these facts establish that indirect readout in the central region of the complex is important in tuning the relative affinities of 434 repressor for different operators. On the other hand, for the contacted outer $5+5$ base pairs we expect no particular elastic specificity. At these positions, protein–DNA contacts necessarily dominate interaction energies since they drive the complexation. DNA distortion is moderate and the protein is reasonably stiff, so quadratic bps potentials should reflect this behavior. Moreover, the existence of three different structures allows an evaluation of the sequence–structure threading approach. The 434 repressor is thus an ideal candidate for a test of our formulation of the elastic free energies within the rbp model, to be described next.

2.2.2 Relative binding affinities in 434 repressor

Experimental evidence for indirect readout in 434 repressor comes from the dependence of binding affinity on the sequence of the central, non-contacted bases [Kou87]. Can DNA elasticity alone already explain the observed affinities? If it can and if in addition the protein forces all of the equally probable artificial sequences into a common structure q, one expects that

$$\beta(G_q(\sigma) - G_q(\sigma')) = \beta(F_\sigma(q) - F_{\sigma'}(q)) = \ln\left[\frac{c_{1/2}(\sigma)}{c_{1/2}(\sigma')}\right]. \tag{2.4}$$

Here the affinity $c_{1/2}(\sigma)$ is the (normalized) repressor concentration needed to occupy half of the operators σ, which is proportional to $p(\sigma|q)$ in dilute solution.

But which version of F is the correct one? There exist three different structures that may serve as template, and a total of seven different ways to parametrize the elastic energy, see section 1.6. An overview of the affinity predictions for all 21 different resulting combinations is given in fig. 2.2 [Bec06]. In each of the panels, the left hand side vs. right hand side of eqn. (2.4) is shown.[2] They exhibit widely varying root mean square (rms) deviation, ranging from 1.5 $k_B T$ to 26 $k_B T$ depending on the parametrization and structure used. This variation occurs even though the global energy scale agrees for all potentials, with the sole exception of the rescaled B', see sec 1.6.3.

Possible reasons are:

1. Failure of basic model assumptions: independent bps, stiff protein, elasticity dominates binding free energy differences in the central region of 434;

2. The crystal structures do not correspond to the relevant structures in solution closely enough;

3. The parametrizations of the potential are inexact.

2.2.3 Choice of a preferred parameter set

A posteriori, we can now check whether one combination of parametrization and crystal structure stands out as the best model for the measured solution affinities.

The linear correlation coefficients shown in each panel vary between -0.52 and 0.64. They measure the quality of a linear regression of the data points with *arbitrary* slope. Although a negative correlation coefficient does identify bad correspondence, the correlation coefficients are insufficient as indicators of fit quality. E.g, B' has higher correlation than B but is far off the correct energy scale. Indeed, the theoretical model to be compared with the data is a line of *fixed* slope equal to one. The rms deviations from this model together with the linear correlation indicate clearly that overall, the combination of MP potential and O_R3

[2] We used affinity data of ten 14-bp artificial sequences, which differ only in the central base pairs [Kou87]. The experimental affinities for the R1-R69 subdomain of the repressor given in this paper were used, since this eliminates cooperative binding effects and corresponds to the domain that was crystallized [Agg88]. All F differences are computed using the total 14-bp deformation free energies for the same sequences in each of the three O_R crystal structures. Out of the two possible orientations in which the repressor can bind, we used the one with lower F value. This makes a difference only for those three artificial sequences that are not self-complementary. All possible combinations of σ and σ' are shown, so the plots are inversion symmetric.

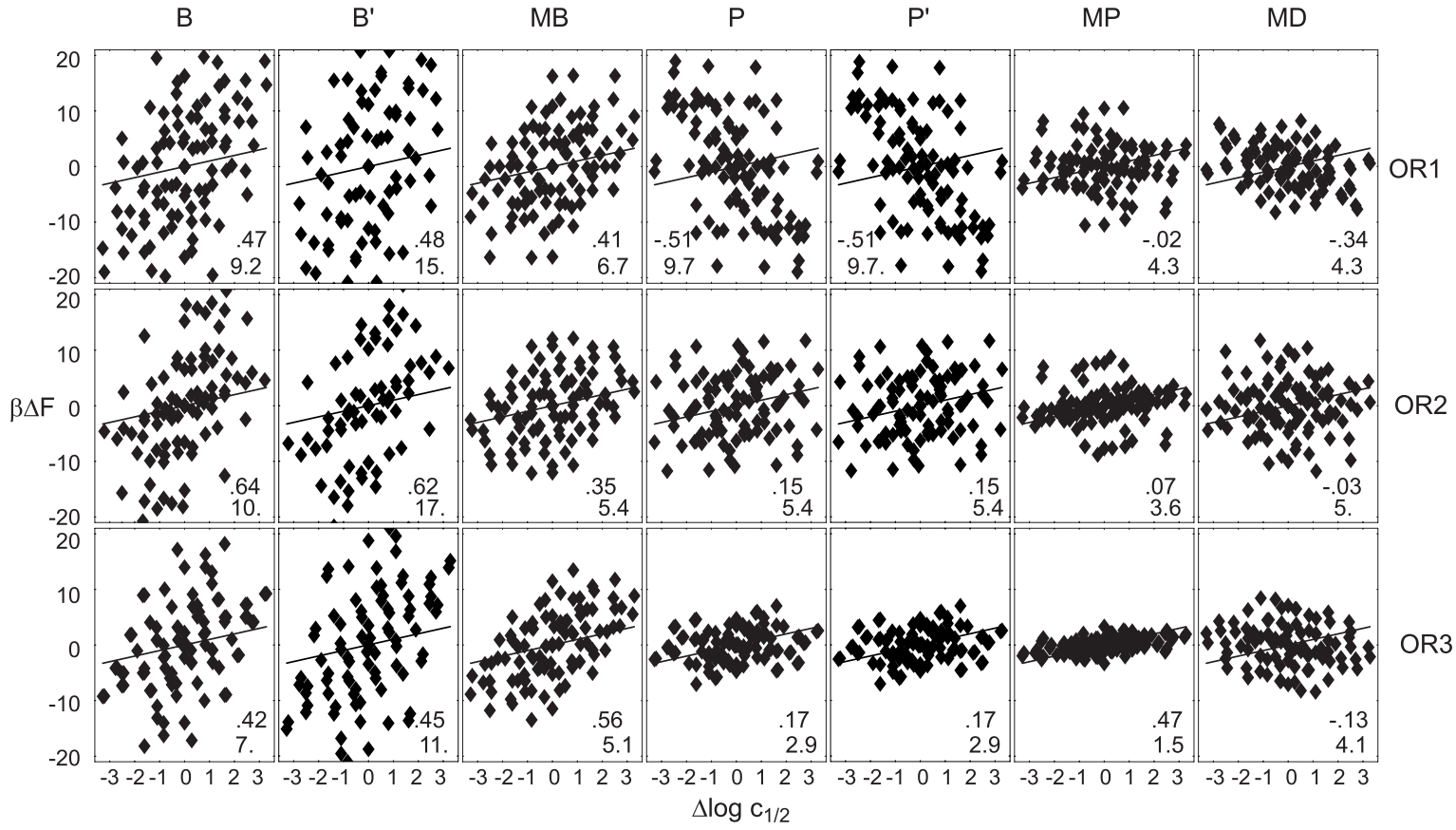

Figure 2.2 | Computed deformation free energy differences vs. measured log affinity differences, for all combinations of crystal structure and employed parametrization. Inset: linear correlation coefficients (upper number) and root mean square deviation from the line $\beta\Delta F = \Delta \log c_{1/2}$ (lower number).

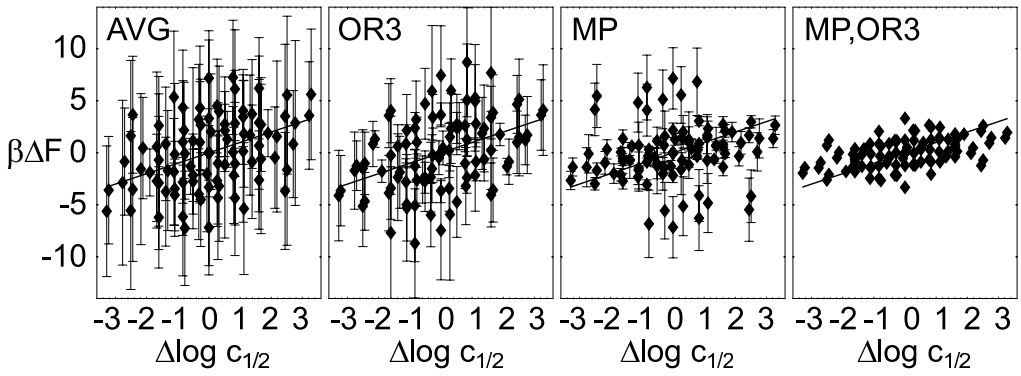

Figure 2.3 | Computed deformation free energy differences vs. measured log affinity differences. From left to right, we used ΔF values for all structures and parametrizations (AVG), the O_R3 structure and all parametrizations (O_R3), all structures and the MP parametrization (MP), and O_R3 together with MP (MP, O_R3). Error bars indicate the spread in ΔF.

structure gives the best agreement with measured affinities, at a comparatively very low rms error 1.5 k_BT and acceptable correlation 0.47. Although this is clearly not enough for quantitative predictions, it should be mentioned that no fitting parameter is involved here; in comparison to related, knowledge–based potential approaches (e.g. in [Mor05]) that use a learning set of complexes with known affinities as input, the quality of correlation appears surprisingly good.

Fig. 2.3 shows the same data as fig. 2.2, in summarized form. One can see that the variation among parametrizations within the best structure (O_R3 panel), is greater than that among structures for the MP parametrization (MP). A standard χ^2-test using the respective error bars reveals that at a 5% confidence level, the model $\beta\Delta F = \Delta \log c_{1/2}$ is compatible with the averaged data in the panel (O_R3), but is rejected for those in (MP). This is in accord with the observation that MP together with O_R1,2 give no positive correlation, while O_R3 together with B, MB, P and MP results in acceptable correlation coefficients.

These observations give some indication that the parametrization error 3. is more important than the failure of basic approximations 1. made in the model, and that improvements in the determination of a harmonic base pair potential will eventually lead to quantitative affinity predictions. If we accept the MP potential as a valid representation of solution DNA elasticity based on its small rms deviation, we can then identify the O_R3 structure as the template that is the best representative of the affinities in solution.

This also points to the basic limitation 2. of sequence–structure threading: it is not clear *a priori* that the given structural template is at all suitable to calculate solution binding affinities. In fact, if the protein structure is not rigid enough, there may not even exist a single structural template that is able to account for indirect readout.

Table 2.1 | Computed free energy differences for mutations of the inner four bases of the sequence ACAATNNNNATTGT. Sequences used in [Kou87] are shown with the experimental log affinity difference $\Delta \log c$. In addition to these, the five highest and lowest affinity random sequences are shown. For complementary sequences, the lower F value was used.

rank	$\beta\Delta F$	$\Delta \log c$	NNNN	rank	$\beta\Delta F$	$\Delta \log c$	NNNN
1.	−1.9		AAAA	39.	0.9	2.7	ACGT
2.	−1.5		AAAG	51.	1.3	1.1	GTAC
3.	−1.4		ATAA	55.	1.5	2.8	AGCT
4.	−1.2	0.3	TTAA	75.	2.2	0.3	AATT
5.	−1.		ATAG	132.	5.7		CATA
8.	−0.5	−0.5	AAAT	133.	6.2		TGCA
17.	0.	0.	ATAT	134.	6.8		CACA
21.	0.1	1.1	CTAG	135.	7.		CATC
25.	0.3	0.6	GTAT	136.	8.6		CATG
37.	0.9	1.4	AGAT				

Table 2.1 lists some of the binding affinity predictions made with the MP hybrid potential–O_R3 template combination [Bec06]. One can see that the range of computed free energies is bigger than that of the measured ones, and that the measured affinities are generally higher. The highest affinity sequence AAAA coincides with the central part of the native sequence of O_R3, which however differs slightly from the consensus in the non-contacted region, see fig. 2.1. These observations underscore the importance of ongoing efforts to improve DNA elastic potentials [Bev04]. The quantitative prediction of indirect readout–mediated relative affinities is suggested as a method to benchmark them. To test improved rbp potentials, it appears helpful to extend this kind of experiments to the sequences with extreme predicted affinities.

In conclusion of these results, for the discussion of localized specificity in chapter 3, the MP hybrid potential parametrization will be used chosen on the basis of its superior performance in affinity prediction. To give an idea of the sensitivity of results to the choice of parametrization, plots with error bars showing the variation

among parameter sets are included in appendix A.1.

3 Local elastic optimization

Although sequence–dependent elasticity is important for specific binding of DNA operator sequences to proteins, there are always other important mechanisms for specificity. In this chapter, we discuss how to identify binding subsites *in which elastic effects are dominant for specificity, by searching for local elastic sequence optimization.*

3.1 Local elasticity in 434 repressor

From the 434 repressor crystal structures, it is evident that the central region of the operator is not contacted by the protein, so that direct chemical interactions cannot provide a recognition mechanism there. This fact allowed the conclusion that the central base–pair sequence is read out according to elastic free energy differences. It was shown in chapter 2 that calculated elastic free energies of sequences that differ in the central region, do indeed correlate with the experimentally observed affinities.

It is now interesting to ask conversely: What is special about the detailed structure of the central stretch of the bound operator that produces this specificity? Is the structure somehow optimized to perform indirect readout? Is it possible to quantify such a feature more rigorously than by referring to DNA–protein distances in a crystal structure?

3.1.1 Elastic free energy profiles

To start addressing these questions, consider the distribution of elastic energy along the available 434 repressor–O_R1,2,3 structures. In fig. 3.1 elastic energies $E_\sigma(q)$ vs. base pair number are plotted in a moving window of length 3 steps around each bps. Partial energies for bend, twist, shear and stretch are calculated by replacing the full covariance matrix $C = S^{-1}$ (see eqn. (1.22)) by its (Ti,Ro), (Tw), (Sh,Sl) and (Ri) submatrices, respectively.[1]

[1] In each case, the covariance 1×1 or 2×2 submatrix is inverted to give the partial energy stiffness matrix. It is easily checked that this is equivalent to integrating out the other variables. Since all

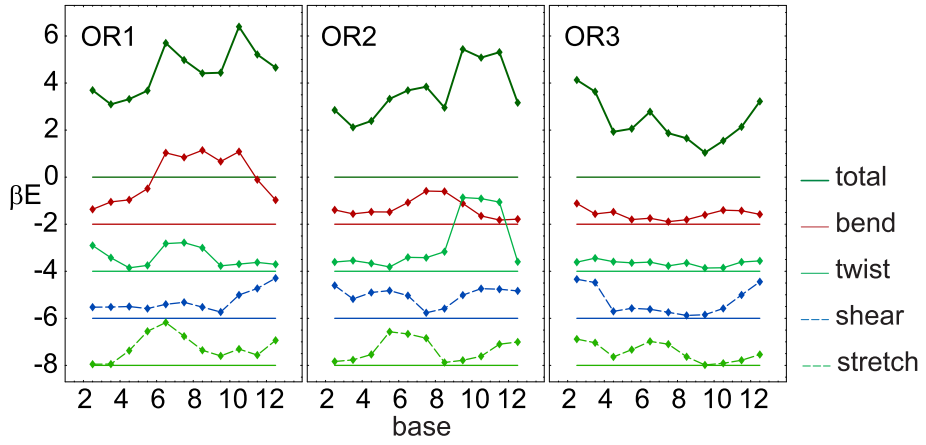

Figure 3.1 | Elastic energy E per bps along O_R1,2,3, shown in units of k_BT. A 3 bps window and the MP parameter set were used. The top curve shows the full energy; partial contributions are successively shifted down, see the legend.

The full and partial energies for the three crystal structures show significant variation along the structure. However, curves for different structural templates look remarkably different and have no common features at the central four base positions. E.g, the increase in bending energy in the center visible in O_R1,2 is absent in O_R3.

The overall bending angles for O_R1,2,3 are around 25, 40 and 30 degrees, respectively. Although O_R1 has the lowest overall bend, O_R3 clearly is the most relaxed structure.

Also, the elastic energy is not strongly dominated by any one of the partial energies. Rather, the identity of the most important partial energy varies between the structures and even within each structure. In O_R1 and O_R3, bend and stretch appear most important, respectively. In O_R2 there is a balance between all four partial energies.

The main contributions to the twist energy at bases 6 to 10 result from *over*-twisting, in accord with experimental results that indicate overtwisting of the central region [Kou92]. However, twist does not appear more important than other partial energies.

The reason to show E here instead of the free energy F here is that the normal-

coupling stiffnesses are averaged out, the partial energies obtained in this way do *not* sum up to the full energy.

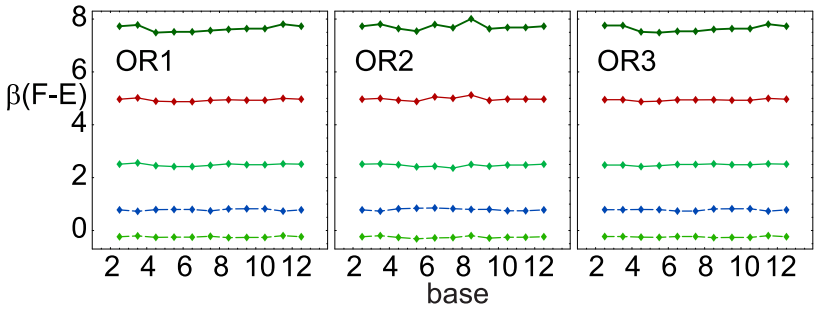

Figure 3.2 | Difference of free energy F to elastic energy E. The normalization–dependent constant offsets are superimposed with a sequence–dependent variation. MP, 3 bps average, partial energies as in fig. 3.1.

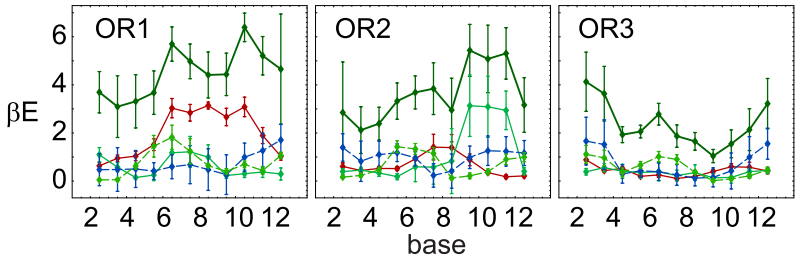

Figure 3.3 | Elastic energy as in fig. 3.1 but including parametrization error bars. No offset, MP, 3 bps average.

ization $E(q_0) = 0$ allows a direct comparison of partial energies with different dimensionality. When using F, the choice of volume scale in the angular vs. linear dimensions of q space adds a constant offset, see (1.23). This is clearly visible in fig. 3.2, where $F - E$ is shown. Apart from the uninteresting constant, the two versions of the deformation energy also differ by a sequence–dependent term $-T\Sigma(\sigma)$.

To complete the picture, the parametrization uncertainty of elastic energies is shown in fig. 3.3. The error bars summarize the variation due to different choices of parametrization. Their size is ± the standard deviation of E values computed with the full set of {MP, MB, P, B, MD} parametrizations. While single values do vary by up to 2 $k_B T$, the global shape of the curves is well above this 'noise'. Even the finer details of the *partial* energies are significant.

3.2 Elastic optimization

A first guess about optimization in protein–DNA complexes could be that the elastic energy should be low ('optimized') at those base positions where indirect readout takes place. As shown in the previous section, this is incorrect: The elastic energy in itself does not exhibit any special features at the central four bases of the 434 complex. It seems worthwhile to think more carefully about how sequence specificity and elastic optimization are related.

3.2.1 Optimal subsequences and indirect readout

In thermal equilibrium, when some protein binds specifically to a certain sequence, this happens because the sequence has optimal binding free energy. It is then interesting to ask, which part of the binding free energy is most important for specificity. Certainly, if DNA elasticity is the dominant part, the operator must be optimized with respect to DNA elasticity.

Our working hypothesis is the converse: We assume that if the sequence is elastically optimized at a certain position along the operator, then DNA elasticity is the dominant part of the binding free energy. Otherwise, elastic optimization would occur just by coincidence. It is natural to call that position an *indirect readout position*.

The strategy is then to detect elastic optimization as a marker for indirect readout. This may lead to false positive detections. To systematically exclude these false positives, one would have to make a comparison of the indirect to the direct readout part of the free energy. This additional information requires much more detailed modeling, which is not attempted here. One can nevertheless reduce the probability of false positives by a reasonably high threshold for detection and by considering simultaneous optimization of multiple–base subsites, see below.

The question whether an operator is elastically optimal can be given two different precise meanings. Consider a known structure of some stretch of DNA in a co-crystal. We may ask

1. Is the structure optimal for the observed sequence? I.e, is the given structure the most relaxed one for that sequence?

2. Is the sequence optimal for the observed structure? In other words, is it more relaxed than other sequences?

The respective answers were already formulated in chapter 1: Question 1. corresponds to the q with minimal $F_\sigma(q)$, question 2. to the σ with minimal $G_q(\sigma)$.

One important objection[2] can be made at this point: It may be true that the elastically optimized sequences bind most strongly in an idealized test tube experiment as introduced in section 2.1.3. However in Nature, the direct readout part of the free energy drives binding and at the same time restricts the set of possible binding sites to just a few cognate DNA operators in the relevant genome. There is no reason why the indirect readout contribution should further increase binding strength, if its biological function is just to fine–tune the affinities in a certain range. On the contrary, overly strong binding must be avoided, since there has to be a way to remove the protein from DNA at some point.

In the 434 repressor example, the $5+5$ outer base pair sequence could provide enough specificity to allow binding just to the $O_{R,L}$ sites, and the inner four positions, even though they exhibit indirect readout, would not appear optimized in the sense of question 2.

In response to this objection, one can point out that while there may be no reason why biological function requires optimization of elastic energy in the complex, it is true that the observed crystal structures represent states where the total free energy is minimized. The structural templates used to calculate free energies, are co-crystallized with their respective native operator sequences. Therefore the measured co-crystal structure is adapted to the native sequence.

The amount to which it is adapted to its *elastic* properties can be measured by trying to fit other sequences into the same structure and checking if the elastic free energy drops. At positions where this can be done, the complex structure is obviously not accommodating very well the elastic properties of the native sequence. Turning this around, at positions where the native sequence has the lowest elastic free energy among all sequences, we postulate that indirect readout dominates.

In this line of thought, the adaptation of the crystal structure to the native operator sequence introduces a bias which justifies the assumption that the native sequence must have the lowest free energy. In summary it is fair to say that the local elastic optimization of the native sequence is an interesting observable, pointing to dominant indirect readout. In accord with this picture, table 2.1 lists the native O_R3 free energy as the lowest possible one, and the O_R1,2 free energies still in the

[2] Thanks to J. Widom for pointing this out.

low 20% of the trial set of mutated sequences.

3.2.2 Measures of elastic optimization

For some stretch (q, σ) of DNA in a given co-crystal structure, we would like to tell whether it is specifically bound because of DNA elasticity. Naively, one might assume that this is the case if it carries a small elastic energy, but this not correct. We are really asking: Compared to all mutated sequences, is σ elastically optimal? In general, this is the case if $K(q, \sigma) < K(q', \sigma')$ for all other (q', σ'), as explained in section 1.3.

The typical situation is that there is only one crystal structure q available as a model for the solution complex. When threading sequences through that particular structure, one automatically makes an additional simplification. The assumption is that the (experimentally inaccessible) complexes (q', σ') of the protein with any other DNA sequence σ' will force the DNA into essentially the same structure $q' \simeq q$, which is valid in the stiff protein limit. Considering eqn. (1.7), the approximation is that $|F_{\sigma'}(q) - F_{\sigma'}(q')| \ll |G_q(\sigma) - G_q(\sigma')|$, so that F difference between *structures* can be neglected. The same approximation is effectively made in [Pai04a], where after an initial partial structure relaxation the structure was kept fixed, and in the static model of [Mor05]. The validity of the stiff protein limit depends on the protein in question. However, when only one structure is known, it is a reasonable strategy to see what the known part of the free energy difference can explain.

Whenever all possible mutated sequences occur with equal probability $p(\sigma') = const$, G differences coincide with F differences between *sequences*, see sec. 1.3. An example of an F histogram of all sequence mutations is shown in fig. 3.5, discussed in more detail below. A widely used [Gro04, AB05] way to quantify optimization of the native sequence based on such histograms is the Z-score. In our case, it is given by

$$Z_{mean} = \frac{\langle F_\sigma \rangle - F_{\sigma_{nat}}}{\left\langle (F_\sigma - \langle F_\sigma \rangle)^2 \right\rangle^{1/2}}, \tag{3.1}$$

i.e. the difference of the mean F to the native F, normalized by the width of histogram. Since we are interested in the low F (optimized) tail of the histogram, we consider also a modified score: the normalized difference of the native F to the minimal F leads to

$$Z_{min} = \frac{F_{\sigma_{nat}} - \min_\sigma\{F_\sigma\}}{\left\langle (F_\sigma - \langle F_\sigma \rangle)^2 \right\rangle^{1/2}} > 0. \tag{3.2}$$

A shortcoming of any Z-score is that information on the global scale of free energy differences in a histogram is disregarded by the normalization. No quantitative connection to competitive binding experiments is impossible.

An alternative way to quantify optimization is to consider just the free energy $G_q(\sigma_{nat})$ of the native sequence. G is the logarithm of a normalized pdf, so a sequence σ has a higher–than–random probability of occurring if and only if $G_q(\sigma)$ is lower than that of an ensemble with $p(\sigma'|q) = const$. A value $G_q(\sigma) = 0$ means that σ occurs with certainty at that deformation, $G_q(\sigma) \leqslant \ln 2$ means that σ has half of the total probability, and $G_q(\sigma) = (l+1)\ln 4$ is the random value for a rbc with l steps.

Considering these properties, clearly the value of $G_q(\sigma)$ already summarizes information about how low the corresponding F_σ lies in the F histogram of all sequences in the ensemble. This observation can be made more precise: In the case $p(\sigma') = const$, from the definitions (1.4) and (1.6), we get $G_q(\sigma) = F_\sigma(q) - \overline{F}(q)$, where $\beta\overline{F}(q) = -\ln \sum_{\sigma'} e^{-\beta F_{\sigma'}(q)}$. This can be interpreted as the difference of F_σ to an 'exponential mean' $\overline{F}$ which is taken over the Boltzmann factors of all mutations. In this respect, the sequence potential $G_q(\sigma)$ is similar to a Z-score, but one that is computed for the histogram of Boltzmann weights.

Since G is a true free energy, it can be directly related to relative affinities in a competitive binding experiment, unlike the Z-scores. By normalizing G to the length of the considered window, an unbiased comparison of specificity for different subsequence window lengths is also possible. The expected dependence of a Z-score on window length is less clear [AB05].

3.2.3 Quantifying elastic optimization in 434 repressor

What can be learned by applying the different measures of elastic optimization of the native sequence to the 434 repressor test case? Fig. 3.4 gives an overview [Bec06]. The free energies F, G and the Z-scores are calculated with respect to sequences in a centered moving window of length 3 bps, which gives sufficient spatial resolution to distinguish the central from the outer base pairs.

As emphasized before, the deformation free energy shows no features special to the inner four bases (6 to 9). What matters for sequence optimization is only the native value of F compared to the whole F distribution of mutated sequences.

As an example of such a distribution, in fig. 3.5 the F histograms of sets of mutated sequences in three consecutive 5 bps windows along the O_R2 structure are

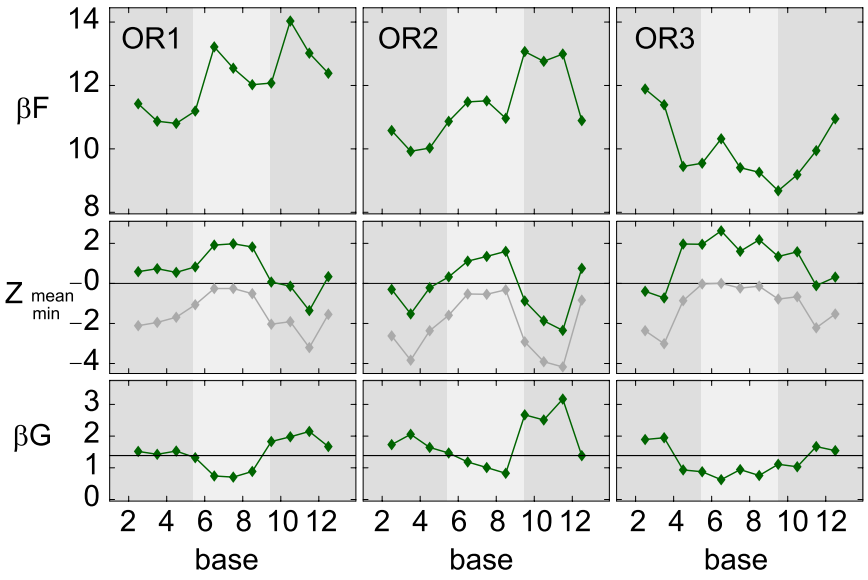

Figure 3.4 | Elastic optimization in 434 repressor structures $O_R1,2,3$. Deformation free energy F, first row. Z-scores of mean (green) and minimum (gray), second row. The third row shows the sequence potential G together with the random level. F is given per bps while G is per bp. Again, a moving window of 3 bps was used. Lighter shading highlights the inner 4 bp, see also 2.1. MP parameters.

shown. One sees that the free energies follow a skewed, Gamma–like distribution which varies in both mean and width. The native value is lowest in the left window position, but only in the *central* window does the native sequence lie below the mean and close to the minimum of the distribution. So the low F value in the left window does *not* correspond to an optimal sequence!

Quantifying these observations, consider again fig. 3.4. The second row shows the Z-scores Z_{mean} and Z_{min}, computed from F histograms of all mutated sequences in the same moving windows as in the rest of the panel. In correspondence with fig. 3.5, the O_R2 plot shows a maximum in the central region. Similar bumps in the other Z-score plots show that also in $O_R1,3$ the native sequence is particularly low–lying only at the *central* base positions. The constant difference $Z_{mean} - Z_{min}$ indicates that the shape (not the width!) of the histograms stays the same. Therefore the two Z-scores carry essentially the same information.

The third row of fig. 3.4 shows the sequence potential G, given per bp, together with the random G level. It is computed in a 3 bps (i.e. 4 bp) window and for a uniform sequence probability $p(\sigma)$. In contrast to the deformation energy, G

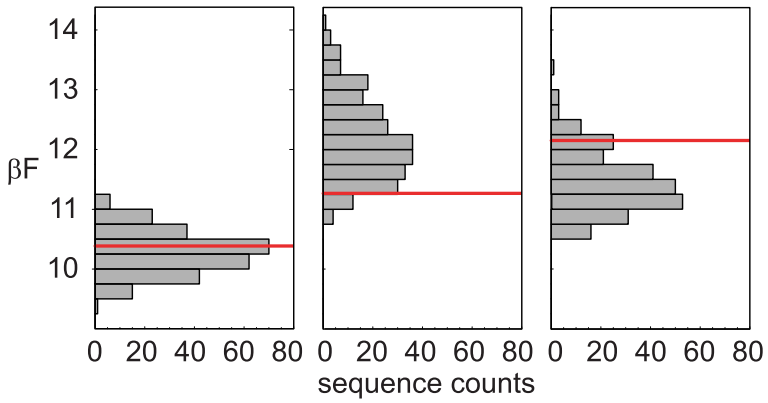

Figure 3.5 | Histograms of the free energy per bps of mutated sequences, around bps 3, 7 and 11 from left to right. The red line indicates the F value of the native sequence. All possible mutations inside a 5 bps window were generated. O_R2 structure.

shows a significant dip below the random value close to the center, in *all* structures. Since G is normalized per bp, a value $G = .5$ corresponds to 8% probability of the native 4 bp subsequence in the unbiased ensemble, which is 20 times the random value of $4^{-4} \simeq 0.4\%$.

The G dip shows that subsequences around the central, but not the outer, base pairs of the binding site occur with a probability above chance, when accounting only for DNA elasticity. In this sense the respective native sequences of the central base pairs are optimized in each of the three available structures. The minimum in G agrees well with the maximum of Z_{min}, which can be explained with the exponentially high weight of the sequences with low F.

Following the reasoning in section 3.2.1, these measures give a clear indication for indirect readout mediated by DNA elasticity in the central region of 434 repressor. The fact that all available structures show the same feature, lends support to the method of inferring the presence of indirect readout from one representative crystal structure in general.

The moving window used in the profiles smoothes the results, and provides better defined histograms in the case of the Z-score. In the above results, the moving window length is not crucial for the central dip. While any window from 1 to 5 bps will show the same trend, there is a tradeoff between spatial resolution and noise. Importantly, the feature of a central dip is robust with respect to parametrization errors, see Appendix A.1.

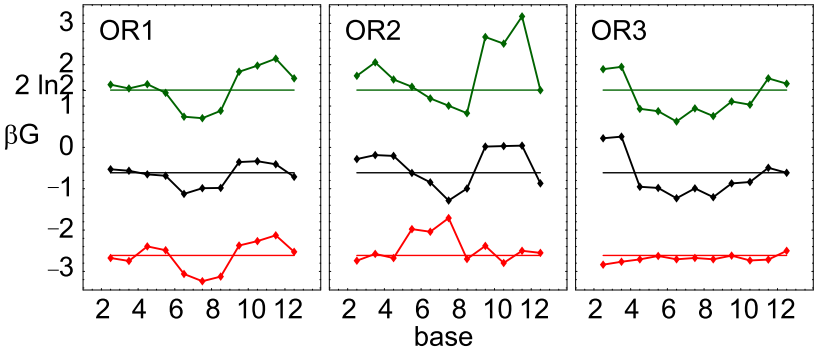

Figure 3.6 | Sequence potential G for O_R1,2,3. From top to bottom, shifted in 2 k_BT steps: Full sequence–dependence, averaged equilibrium values, averaged stiffness. The random baseline is at $2 \ln 2\, k_BT$.

3.3 Origins of specificity

3.3.1 Structure vs. stiffness

Indirect readout is caused by the sequence dependence of both DNA structure and DNA stiffness.Which dependence is stronger? Either one can be selectively switched off: By sequence–averaging the equilibrium values, the structural effect is eliminated; by averaging the covariance matrices, the effect of stiffness is suppressed.

The profiles of the resulting partially averaged sequence free energies in 434 repressor are shown in fig. 3.6 [Bec06]. Interestingly, the characteristic G dip at the central bases *persists* when stiffness matrices are averaged, in fact the G curve roughly traces the fully sequence–dependent one. In contrast, averaging equilibrium values and retaining sequence dependent stiffness, does alter the shape of the curves, and the central G dip is lost in O_R2,3. This indicates that sequence dependent structure is more important for indirect readout than sequence dependent stiffness, at least in the only moderately deformed example of the 434 repressor.

3.3.2 Bending vs. twisting

Is it possible to explain sequence specificity by a reduced set of variables? E.g. can twisting alone explain indirect readout in the 434 repressor, as suggested by the experimental fact [Kou92] that operators with higher twist in the central region have higher affinity for 434 repressor than those with lower twist? This question

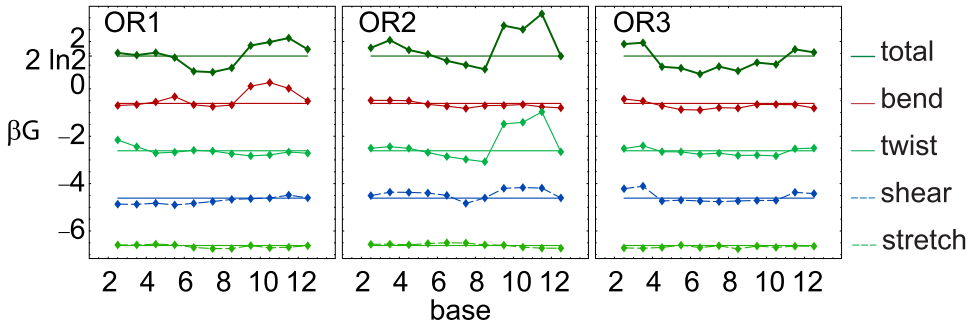

Figure 3.7 | Sequence potential G along $O_R1,2,3$, analogous to fig. 3.1. The partial free energies are shifted down by 2 k_BT successively, and each one is shown together with the level of random probability. A 3 bps moving window was used.

can be addressed using partial sequence free energies, defined in the same way as G but with subsets of the conformation variables. In fig. 3.7 we show both full and partial sequence free energies (compare fig. 3.1). The result is ambiguous. In O_R2, twist can account for the characteristic G minimum in the center. In the other structures, sequence specificity appears to arise from an interplay between all deformation modes, and and thus cannot be generally attributed to the twisting mode only.

3.4 Elastic consensus sequences

The central 4 bp native subsequences are elastically optimized in the 434 repressor structures. But how strongly is the identity of each individual base of the native sequence preferred? One can try using very short subsequences to calculate G, but then the results get noisy. A typical example of the tradeoff between spatial resolution and noise is shown in fig. 3.8.

The most widely used way of describing localized specificity is not a free energy profile; instead it is based on the concept of a consensus sequence. Here, usually on the basis of a biochemical competitive binding assay, a set of operators for the protein question is the data. Aligning these sequences, one can then look for base identities that are 'conserved'. E.g, it may happen that $b_4 = G$ is required for binding in any sequence. Combining this kind of information for different base position, the consensus sequence might look something like to ACNGNNA, with undetermined bases that are denoted N.

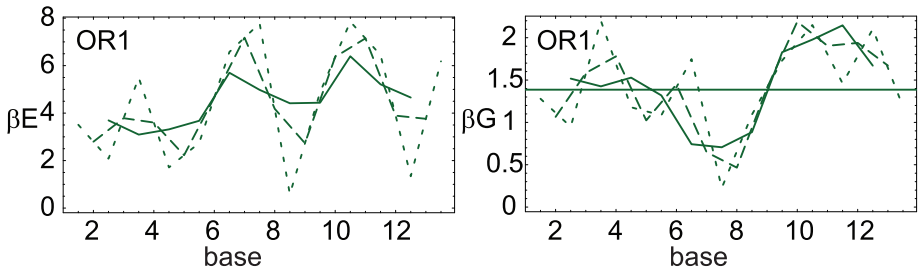

Figure 3.8 | Elastic energy (E, per bps) and sequence free energy (G, per bp) in the O_R1 structure. The moving window lengths 1, 2 and 3 bps are shown with short, long and no dashes, respectively. While using a moving window for E amounts to a simple moving average, this is not the case for G.

3.4.1 Single–base elastic consensus sequences

This approach has been made quantitative for the case where an exhaustive set of binding sequences is available [Sch90]: Instead of considering only whether a certain base identity is required, one can incorporate the *strength* of this requirement. This is done by scaling the height letters in the consensus sequence by the relative frequency of that base. This is a preliminary version of the so-called sequence logos [Sch90] which show the four base letters with varying height stacked on top of each other at operator position.

This kind of idea can be applied in the context of elastic sequence specificity. Moreover, we will extend the approach to include correlated preferences for short *subsequences* instead of isolated bases only [Bec06].

Assuming a stiff protein with structure q, and regarding only DNA elasticity, mutated operator sequences σ' of length l bind with a probability $p(\sigma'|q) = e^{-\beta G_q(\sigma')}$. Instead of looking at an entire sequence one can first ask for the probability $p_i(b)$ to find just the i-th base $b'_i = b$ in all length l subsequences. It is given by the expectation

$$p_i(b) = {\sum_{\sigma'}}' \delta_{bb'_i} e^{-\beta G_q(\sigma')} =: \langle \delta_{b\cdot} |_i \rangle \tag{3.3}$$

where δ is the Kronecker delta. See 1.4 for the notation. Using eqn. (1.15),we have [Bec06]

$$p_i(b) = \frac{1^T T(q_1)\cdots T(q_{i-1}) P_b T(q_i)\cdots T(q_l)1}{1^T T(q_1)\cdots T(q_l)1}. \tag{3.4}$$

Here, the projection matrix onto the base b, $(P_b)_{b'b''} = \delta_{b'b}\delta_{b\,b''}$ has to be

inserted at position i. Note that if the transfer matrices T were scalars, this would reduce to a purely local expression depending only on b_{i-1}, b_i, b_{i+1}. However the elastic *step* energy in conformation space induces short–range correlations in sequence space, embodied in the non-commuting transfer matrices. Note that it is not necessary to make an approximation of additive free energy in sequence space (as done in [Mor05] in a related context).

Calculating $p_i(b)$ for all bases b =A,T,C,G along a given structure, using centered windows of constant length, gives a base–per–base picture of elastic preference in the structure. This information is often called a weight matrix. To check for elastic preference for the *native* sequence, one can just pick out $b = b_{nat}$, the native base at every position.

3.4.2 Multiple–base correlated consensus

We have seen above that the step deformation energy introduces correlations in the sequence. It is therefore natural to extend the approach to correlation functions of $k+1$ bases (where $k = 0$ is the case in sec. 3.4.1). The joint probability $p_{i,i+k}(\sigma)$ to find $k+1$ specific bases $b_i \dots b_{i+k} = \sigma$ at positions $(i, \dots, i+k)$ is not hard to write down using the transfer matrix formulation. One just has to insert projectors at all of these base positions [Bec06],

$$p_{i,i+k}(\sigma) = \left\langle \delta_{b_i \cdot}\big|_i \cdots \delta_{b_{i+k} \cdot}\big|_{i+k} \right\rangle = \frac{\mathbf{1}^T T(q_1) \cdots T(q_{i-1}) P_{b_i} T(q_i) P_{b_{i+1}} T(q_{i+1}) \cdots P_{b_{i+k}} T(q_{i+k}) \cdots T(q_l) \mathbf{1}}{\mathbf{1}^T T(q_1) \cdots T(q_l) \mathbf{1}}. \quad (3.5)$$

Again, if the 'tails' of non-projected transition operators to the left and right were scalars and canceled with the denominator, one would end up with the probability $p(\sigma|q)$ of the $k+1$-bp sequence σ alone, see sec. 1.4. The difference to the full expression $p_{i,i+k}(\sigma)$ is that the latter includes sequence correlations extending left and right from the subsequence σ. When $k = l$, both probabilities agree. In practice, one can simply choose for l the whole binding site length, since the computational cost is $O(l)$ only.

It has been pointed out [Sch90] that different shapes of distributions $p_i(b)$ contain varying amounts of information, and that this gives a measure of sequence specificity at that position. E.g, any position i at which all bases are equally probable has zero Shannon information and it clearly carries *no* elastic specificity whatsoever.

The entropy of the distribution $p_{i,i+k}(\sigma)$ is

$$\Sigma_{i,i+k} = -\sum_{\sigma'}{}' p_{i,i+k}(\sigma') \ln[p_{i,i+k}(\sigma')] \leqslant (k+1)\ln 4. \tag{3.6}$$

From this a measure for the information content of the distribution that ranges from 0 to 1 is derived as

$$I_{i,i+k} = 1 - \Sigma_{i,i+k}/((k+1)\ln 4). \tag{3.7}$$

The method [Sch90] of scaling the base frequencies with the information content of the distribution at each position to get a compact representation of sequence preferences can be transferred to our situation: $I_{i,i+k}p_{i,i+k}(\sigma)$ gives the relative frequency of σ, scaled with the information content which indicates the overall strength of base preference at that position. However this quantity cannot be represented as a sequence logo with the usual letter scaling notation whenever $k > 0$, since the neighboring subsites in a moving window overlap with each other. However, the most interesting information can be shown by plotting $I_{i,i+k}p_{i,i+k}(\sigma_{nat})$ for the *native* subsequences only, see fig. 3.9 below. Such a plot shows directly whether a strong *elastic* sequence preference exists and how well the native sequence coincides with it. Thus $I_{i,i+k}p_{i,i+k}(\sigma_{nat})$ gives a local marker for which significantly nonzero values point to elastic specificity [Bec06]. Again, since the subsequence length of interest is usually just a few base pairs, computation is cheap.

3.4.3 Native vs. elastic consensus sequences in 434 repressor

How do the elastic consensus sequences look in our test case?

Figure 3.9 shows the similarity of the native sequence to elastic consensus. The plots of the scaled native probability Ip indicate elastic specificity of the native sequence on the level of single base–pairs, dimers, and tetramers, from top to bottom. Interestingly, in the $O_R1,2$ complex structures, elastic specificity is concentrated on two central bases at positions 7 and 8, while the O_R3 structure shows a more distributed preference, mainly at positions 5 and 7. Going from from base–pairs to dimers, the peak for $O_R1,2$ at bps 7/8 stays sharp. For the tetramers, the distribution shows still a concentration to a preferred set of sequences indicated by the maximum in I but the native sequence cannot claim a significant part of the total weight among its 4^4 competitors anymore. In summary, the O_R1 structure exhibits the strongest elastic specificity, localized to a dinucleotide, and O_R3 has a

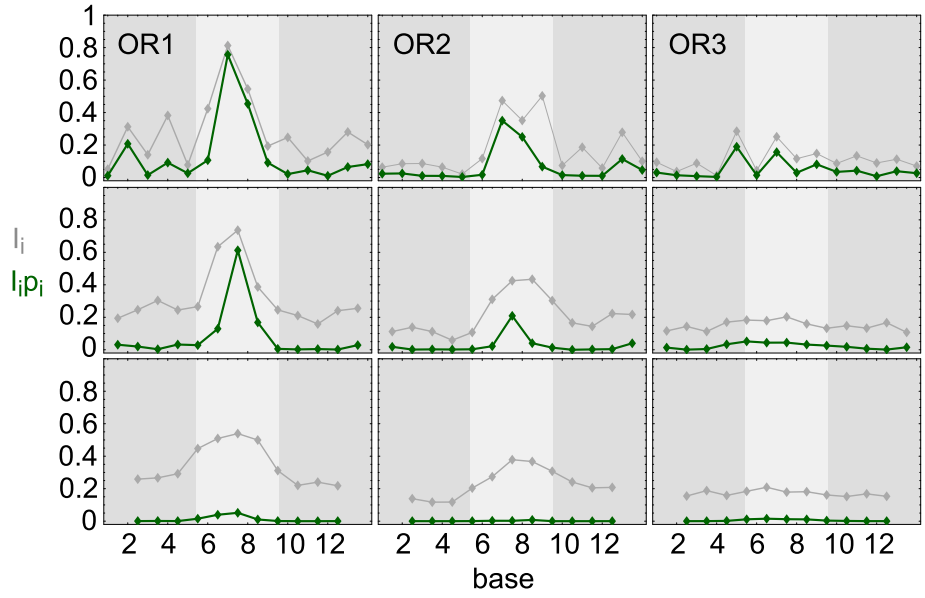

Figure 3.9 | Similarity to elastic consensus for native subsequences in the O_R complexes. Information (gray) and scaled native probability (green) are shown for 1, 2 and 4 bp subsequences, from top to bottom, centered on the subsequence window. MP parameters.

more distributed specificity.

3.5 Summary

A theoretical framework for modeling indirect readout based on appropriate elastic free energies was introduced in chapter 1. These describe affinities in idealized competitive binding experiments, and compare favorably to experimental affinities, see chapter 2. Starting form the elastic free energies, statistical markers were developed that can detect sites of dominant indirect readout by locating elastically optimized subsequences in protein–DNA co-crystals. They are linked to experimentally measurable ensemble properties of relative binding affinities of operators mutated at these sites, as detailed in section 3.2.1.

The success of this approach depends on the applicability of the particular model used to describe DNA elasticity, as well as on the quality of the parametrization. The description on the rigid base–pair level appears as a sensible compromise between computationally much more expensive all-atom models on one hand and coarser rigid rod representations on the other. State–of–the–art parametrizations

from MD simulation and from structural data analysis were combined using a new, microscopic method of adapting the effective temperature scale [Bec06].

Quantitative predictions for relative binding affinities depend quite sensitively on the choice of parametrization. In the case of the 434 repressor, results averaged over the available elastic potentials and structural templates are compatible with measured binding affinities, but the margins of error are too wide to allow quantitative predictions. Closer inspection showed that the new MP hybrid potential [Bec06] performs significantly better than alternative parametrizations.

Qualitative observations appear much more robust with respect to the parametrization uncertainty, as can be appreciated by plotting parametrization error bars. Examples are the location of indirect readout sites, the relative importance of structure and elasticity for specificity, or the distinction of contributions from different elastic degrees of freedom, see also appendix A.1.

The deformation fluctuations of base pairs in the model are taken to be independent, which is an oversimplification. Since adjacent rbp steps are coupled through the DNA sugar–phosphate backbones, their fluctuations are expected to be correlated to some extent. To overcome this limitation, two different ways to refine of the model can be considered.

One is the inclusion of nearest-neighbor step cross–correlation terms in the rbp elastic energy, leading to tetranucleotide stiffness matrices. The corrections to a dinucleotide model due to flanking base sequence were recently investigated [AB05] using MD simulation. In most cases these are much smaller than the difference between the dinucleotide potentials we used for the same step. At the precision of parametrization available today, these correlations are still a secondary effect.

Another possible refinement is to consider rigid bases instead of rigid base–pairs. There are indications from MD simulation that this improves the quality of a purely local description [Mad]. However, a corresponding parameter set is not available, so an experimental test of this extended model cannot yet be performed.

In the bacteriophage 434 repressor complex, the elastic energy (fig. 3.1) and specificity (fig. 3.7) profiles of the O_R1,2,3 co-crystal structures reveal differences in detail. However, in all three cases, agreement between the native and the elastic consensus sequence is confined to the central, not directly contacted part of the operator. On a qualitative level, this supports the working hypothesis that strong elastic optimization in protein–DNA co-crystals is an indicator for dominant indirect readout in real protein-DNA solution complexes, which is at heart of the

proposed marker for indirect readout [Bec06].

While the numerical complexity of the present analysis is negligible, DNA deformation (free) energies in protein–DNA co-crystals substantially extend the insights that can be gained from structural data.

4 Rigid base–pair chains

Starting with this chapter, the perspective on DNA elasticity is shifted; Instead of single base–pair steps that are individually constrained to some conformation, we now consider rigid–body chains that fluctuate as a whole. Deformations occurring at different steps of the chain can be conveniently related using group language which is successively developed in this chapter in a rather formal way. The final section then relates the introduced mathematical notions back to the physics.

While none of the mathematical tools presented in the section is new, their systematic collection and application constitutes a novel approach to DNA mechanics.

4.1 Linear elastic response of a rigid base–pair chain

In order to give a physical motivation for the formalism to be presented in the following, let's consider a stretch of homogeneous DNA, modeled by a rbc, in a thermal environment. We maintain the basic assumption that its step conformations fluctuate independently. Each rbp step then obeys a thermal equilibrium distribution, carrying the mean elastic energy of $\frac{6}{2}k_BT$.

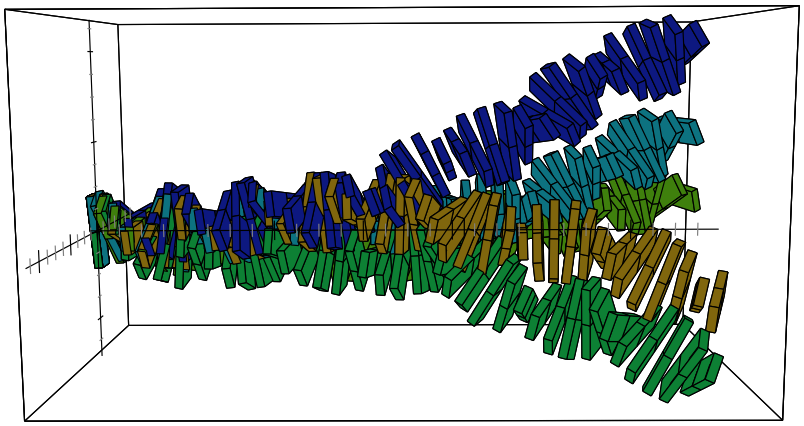

Figure 4.1 | Snapshots of a thermally fluctuating, 42-bp chain, aligned on, or clamped at, the first bp (far left).

Fig. 4.1 shows a collection of random conformations, as seen from from a frame of reference that is fixed to rbp frame 1. Clearly, the deviation from the regular helical equilibrium conformation grows with the bp number. More importantly, a *bending* deformation close to frame 1 results in a large *lateral displacement* of e.g. frame 42, while bending at frame 40 doesn't displace frame 42 much. When seen from a fixed material frame, all rbp steps are not the same; they differ by their respective *lever arm* with respect to the fixed frame.

The same observation applies to the linear response of a rbc to external forces and torques (cf. section 1.5.1): Fig. 4.1 can also be interpreted as representing a rbc which is clamped at the first bp frame. Any transversal *force* acting at frame 42 then induces a *torque* on the other bp steps by lever action. The lever arm is longest at frame 1, which will therefore feel the highest torque. The response of the chain is a sum of the responses of the individual steps, weighted with appropriate leverage terms.

The basic quantities that enter the description can be summarized as follows:

Step conformations (ξ) are the six degrees of freedom of every step, i.e. the configuration space of the model. Rotations are converted into lateral displacement at distant steps by the connecting chain segment.

Elastic energies (E) are naturally given with respect to the local material frame. Apart from sequence dependence, the functional form $E(\xi)$ in the material frame is the same for all steps.

Forces (μ) are the differential change of elastic energy when varying the conformation, $\mu = -dE$. The generalized force μ includes a linear force and a torque component. Lateral linear forces at distant steps induce torque by lever action of the intermediate chain segment.

To understand the collective mechanics of a rbc, it is essential to relate the different frames of reference along the chain. They are connected by the group $SE(3)$ of Euclidean frame transformations, or rigid motions. $SE(3)$ constitutes both the configuration space and the basic transformation group of the rbp model. We will consider its mathematical structure in some detail in section 4.2 below; the link to rbp elasticity will be made in section 4.3.

4.2 Basic properties of the rigid motion group

Some Lie group theory basics are put into the context of our specific example $SE(3)$, setting up the tools and notation used later on. The basic approach is adapted from the robot kinematics literature [Mur93, Zef02]; for more mathematical background, see [Sat86, Lee02].

4.2.1 Homogeneous representation

The position and orientation of a right–handed, orthonormal frame in three–dimensional Euclidean space is determined by six real parameters. The position is specified by a vector p from the lab frame to its origin. The three orientational parameters can be given in many possible ways, such as (some choice of) Euler angles, the orientation and magnitude of the rotation axis vector etc, each of which has its advantages and limitations. We avoid a choice of coordinates here by using the whole 3×3 rotation matrix which has the frame's body axes as its columns, $R = (e_1, e_2, e_3)$. To write the frame as a so-called homogeneous matrix, we add an extra row to obtain a 4×4 matrix. In block form[1]:

$$g = \begin{bmatrix} e_1 & e_2 & e_3 & p \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} R & p \\ 0 & 1 \end{bmatrix}. \tag{4.1}$$

The main advantage of this notation is that frame transformations can be written as a matrix multiplication. If $\cdot_x$ denotes a point's Cartesian coordinates with respect to some frame $x = 1, 2$, and R_{12}, p_{12} specify the frame 2 given with respect to frame 1, then

$$\begin{bmatrix} q_1 \\ 1 \end{bmatrix} = \begin{bmatrix} R_{12} q_2 + p_{12} \\ 1 \end{bmatrix} = g_{12} \begin{bmatrix} q_2 \\ 1 \end{bmatrix}. \tag{4.2}$$

Concatenating frame transformations, one sees immediately that

$$g_{13} = g_{12} g_{23} \text{ , and} \tag{4.3}$$

$$g_{21} = g_{12}^{-1} = \begin{bmatrix} R_{12}^T & -R_{12}^T p_{12} \\ 0 & 1 \end{bmatrix}. \tag{4.4}$$

The identity transformation e is given by the 4×4 identity matrix I_4.

This shows that the g-matrices form a faithful matrix representation of the group of frame transformations in three–dimensional Euclidean space. This so-

[1] As a notational convention, block matrices are always written with square brackets.

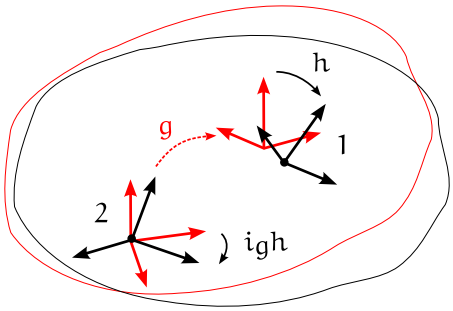

Figure 4.2 | Conjugation. A rigid motion transforming the body frames from their initial (red) to their final positions (black). In this particular example, it looks like a rotation and translation h when using the frames '1', or like a pure rotation $\mathrm{i}_\mathrm{g}\mathrm{h}$, when using the frames '2'.

called special Euclidean group $SE(3)$ is the semidirect product $SO(3) \ltimes \mathbb{R}^3$ of the three–dimensional rotations and translations. $SE(3)$ (or shorter, SE) is a six-dimensional Lie group, i.e. its group space is a smooth manifold, on which the group multiplication and inverse operations are smooth maps.

4.2.2 Left and right translations

Group multiplication by g from the left is called left translation[2], $\mathrm{l}_\mathrm{g} : SE \to SE, \mathrm{h} \mapsto \mathrm{gh}$. The complementary notion is right translation $\mathrm{r}_\mathrm{g} : \mathrm{h} \mapsto \mathrm{hg}$. The map $\mathrm{i}_\mathrm{g} = \mathrm{l}_\mathrm{g} \circ \mathrm{r}_{\mathrm{g}^{-1}} = \mathrm{r}_{\mathrm{g}^{-1}} \circ \mathrm{l}_\mathrm{g} : \mathrm{h} \mapsto \mathrm{ghg}^{-1}$ is called conjugation by g. As can be seen from eqns. (4.2) and (4.3), l_g implements a change of the (non-moving) *lab* frame of reference by an amount g^{-1}. In a similar way, r_g can be seen to correspond to a change of the (moving) *body* frame of reference from e to g. All of $\mathrm{l}_\mathrm{g}, \mathrm{r}_\mathrm{g}, \mathrm{i}_\mathrm{g}$ are bijective with inverse maps $\mathrm{l}_{\mathrm{g}^{-1}}, \mathrm{r}_{\mathrm{g}^{-1}}, \mathrm{i}_{\mathrm{g}^{-1}}$, respectively.

Consider for fixed h, the map $\mathrm{g} \mapsto \mathrm{i}_\mathrm{g}\mathrm{h}$. This amounts to changing both the lab and the body frame together, by an amount of g^{-1}, see fig. 4.2. It is shown readily that if h_p is a pure translation, $\mathrm{i}_\mathrm{g}\mathrm{h}_\mathrm{p}$ is a translation by the same distance but in a direction rotated by R. [3] Rotations on the other hand become *mixed* with translations when changing reference frames with i_g.

[2] This is not to be confused with "translations" which are the group elements without rotation part, i.e. with $\mathrm{R} = \mathrm{I}_3$.

[3] This makes the translations a 'normal subgroup' of SE.

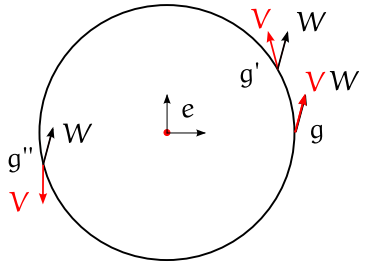

Figure 4.3 | Left (V, red) and right (W, black) invariant vector fields visualized on a subgroup of pure rotations around a fixed axis.

4.2.3 Vector fields

A tangent vector V based at a point g in the manifold SE can be defined as the velocity vector $V = \dot{g}(0)$ to some smooth curve $g(s)$ with $g(0) = g$. It acts on functions $f : SE \to \mathbb{R}$ by taking the directional derivative $Vf := \frac{d}{ds} f(g(s))$ (a real number). At each point g on the group manifold, the tangent vectors span the six–dimensional tangent space $T_g SE$. A smooth map $V : g \mapsto V|_g \in T_g SE$ is called a vector field. As a special case, coordinate vector fields ∂_{q^i} take the derivative in the direction of the local coordinate q^i.

The chain rule reads as follows: For a smooth map $\varphi : SE \to SE$, we have $V(f \circ \varphi) = (\varphi_* V)f$. Here $\varphi_* : T_g SE \to T_{\varphi(g)} SE$ is variously known as tangent map, differential map, or pushforward of φ. In local coordinates $\{q^i\}$ and $\{\varphi^i\}$, it is the Jacobian matrix $\left(\frac{\partial \varphi^i}{\partial q^j}\right)$. Writing also $V = V^i \partial_{q^i}$ in local coordinates, we get the usual chain rule: $V(f \circ \varphi) = \frac{\partial \varphi^i}{\partial q^j} V^j \partial_{\varphi^i} f$.

Using left translation in place of φ, we can move any vector to a different base point: If V is based at h, $l_{g*} V$ is based at gh. An important special class of vector fields is in some sense 'parallel to the group operation': V is called *left invariant* if $V|_{gh} = l_{g*} V|_h$ for all g, h. As a consequence, any left invariant field V is completely determined by its values at e; to evaluate it at other points, just left translate it over using l_{g*}. Since left invariance means invariance under changes of the lab frame (sec. 4.2.2), local material properties are necessarily left invariant.

In the same way, W is *right invariant* if $W|_{hg} = r_{g*} W|_h$ for all g, h, i.e. W does not change when using a different material frame. Therefore, external forces are expected to be right invariant.

4.2.4 The Lie algebra

The left invariant vector fields span a six-dimensional vector space $se(3)$ (or short, *se*), which can be identified with T_eSE. Its elements are the infinitesimal generators of the group. Moreover, the commutator of left invariant vector fields $[U, V]f = UVf - VUf$ is again left invariant. *se* with the commutator bracket $[\cdot,\cdot] : se \times se \to se$ is the Lie algebra of *SE*.

We calculate the homogeneous matrix representation of the infinitesimal generators. Any vector is the tangent vector of some curve, so $V\big|_g = \dot{g}$ for some choice of curve $g(s)$. Since the group operation is linear in the matrix representation, it coincides with its own tangent map: $l_{g*} = g\,\cdot$. We then get explicitly[4] $V\big|_e = l_{g^{-1}*}\dot{g} = g^{-1}\dot{g} = \begin{bmatrix} R^T\dot{R} & R^T\dot{p} \\ 0 & 0 \end{bmatrix}$. Since $R^T\dot{R}$ is an antisymmetric matrix, a matrix basis for *se* is given by

$$\begin{aligned} X_i &= \begin{bmatrix} \epsilon_i & 0 \\ 0 & 0 \end{bmatrix}, \text{ with } (\epsilon_i)_{jk} = \epsilon_{jik} \text{ and} \\ X_{i+3} &= \begin{bmatrix} 0 & d_i \\ 0 & 0 \end{bmatrix}, \text{ with } (d_i)_j = \delta_{ij}. \end{aligned} \tag{4.5}$$

Here, ϵ_{ijk} and δ_{ij} are the antisymmetric and symmetric tensors, respectively, and $1 \leqslant i, j, k \leqslant 3$. We can write any infinitesimal generator uniquely in terms of this basis as $V\big|_e = V^i X_i$.

For $1 \leqslant i \leqslant 3$, X_i generates a rotation around the d_i axis, while X_{i+3} generates a translation along d_i. The generators satisfy a real version of the usual commutation relations of angular and linear momentum in quantum mechanics:

$$[X_i, X_j] = \epsilon^k{}_{ij} X_k,\ [X_i, X_{j+3}] = \epsilon^k{}_{ij} X_{k+3},\ [X_{i+3}, X_{j+3}] = 0; \quad 1 \leqslant i, j, k \leqslant 3. \tag{4.6}$$

One sees that $\{X_i\}_{1\leqslant i\leqslant 3}$ span the subalgebra $so(3)$ (or short, *so*) of three–dimensional rotations, while $\{X_{i+3}\}_{1\leqslant i\leqslant 3}$ span the commutative subalgebra of translations. The commutation relations are tabulated in the structure constants $c^k{}_{ij}$ via

$$[X_i, X_j] = c^k{}_{ij} X_k, \quad 1 \leqslant i, j, k \leqslant 6. \tag{4.7}$$

A complementary set of relations are given by the anticommutators $\{X_i, X_j\} =$

[4] The notation does not distinguish between a tangent vector and its homogeneous matrix representation. Both are completely equivalent but suggest different a different viewpoint: A tangent vector carries the association of a direction, which is not what comes to mind first when one thinks about a matrix.

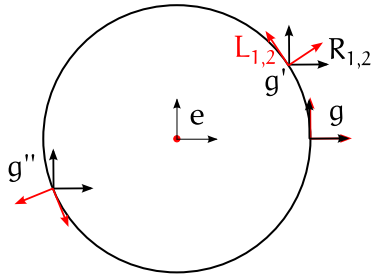

Figure 4.4 | Left and right invariant frames. The group elements $g, \ldots$ are pure rotations around the lab z axis. The vector fields $L_{1,2}$ rotate along, while $R_{1,2}$ do not change.

$(X_i X_j + X_j X_i)$. Unlike the commutators, the resulting matrices are *not* elements of the Lie algebra and thus cannot be represented in terms of the basis $\{X_i\}_i$. We add the symmetric 4×4 basis matrices Δ_{ij} with entries $(\delta^i{}_j + \delta^j{}_i)$. The anticommutation relations then are

$$\{X_i, X_j\} = \Delta_{ij} - \delta_{ij}\delta^{kl}\Delta_{kl},\ \{X_i, X_{j+3}\} = \epsilon^k{}_{ij}X_{k+3},\ \{X_{i+3}, X_{j+3}\} = 0;$$
$$1 \leqslant i, j, k, l \leqslant 3. \quad (4.8)$$

4.2.5 Invariant frames

The left invariant vector fields provide a basis for all tangent spaces $T_g SE$, i.e. a 'moving frame'. We will denote this left invariant frame[5] by $\{L_i\}_{1\leqslant i\leqslant 6}$, where $L_i\big|_g = l_{g*}X_i$ has the matrix representation gX_i. Any vector has unique components in this basis: $V = V^i L_i$. However, it is impossible to find local coordinates so that the L_i coincide everywhere with the partial derivatives in these coordinates; $\{L_i\}$ is *not* a coordinate frame.[6]

In the same way, a right invariant frame $\{R_i\}_{1\leqslant i\leqslant 6}$, $R_i\big|_g = r_{g*}X_i$ can be built. It has the matrix representation $X_i g$. Interestingly, left invariant fields do commute with right invariant ones, which follows from the fact that even for finite group operations, $l_g r_{g'} h = r_{g'} l_g h = ghg'$. Summarizing

$$[L_i, L_j] = c^k{}_{ij}L_k; \quad [R_i, R_j] = -c^k{}_{ij}R_k; \quad [L_i, R_j] = 0. \quad (4.9)$$

[5] This is not to be confused with a 'frame' which is an element of *SE*.
[6] Proof: for any coordinate frame $[\partial_i, \partial_j] = 0$, in contradiction to (4.6).

The action of any left invariant vector field V on a function f by differentiation can be written in matrix form as

$$V\big|_g f = V^i L_i\big|_g f = V^i \left.\frac{d}{ds}\right|_0 f\,(g(e + s\,X_i))\,. \tag{4.10}$$

In a similar way, the action of a right invariant vector field W on a function is given by

$$W\big|_g f = W^i R_i\big|_g f = W^i \left.\frac{d}{ds}\right|_0 f\,((e + s\,X_i)g)\,. \tag{4.11}$$

It is useful to group the invariant components in rotational and translational parts: $(V^i)_{1\leqslant i\leqslant 6} = (\omega, v)$. Here, ω and v are three–dimensional component vectors. The matrix representation becomes

$$V\big|_e = V^i X_i = \begin{bmatrix} \hat{\omega} & v \\ 0 & 0 \end{bmatrix},$$

where $\hat{\omega} = \omega^i \epsilon_i$ is the 3×3 matrix that implements the cross product: $\hat{\omega} = \omega \times \cdot$. Both parts of V have direct physical meaning: ω is the angular velocity and v is the linear velocity of the infinitesimal motion generated by V.

We will generally use the letter ξ to denote a column vector of left invariant components of the velocity of a curve $g(s)$, and ζ for right invariant components in the following. To get the velocity components ξ and ζ of some curve $g(s)$, one needs to solve the linear equations

$$\xi^i(s) X_i = g^{-1}(s)\dot{g}(s) \quad \text{and} \quad \zeta^i(s) X_i = \dot{g}(s) g^{-1}(s), \tag{4.12}$$

respectively.

4.2.6 The adjoint representation

To be able to switch between left and right invariant components, we calculate the g dependent transition matrix. Note that $L_i\big|_e = R_i\big|_e = X_i$ so

$$L_i\big|_g = l_{g*} R_i\big|_e = l_{g*} r_{g^{-1}*} R_i\big|_g = i_{g*} R_i\big|_g. \tag{4.13}$$

The matrix representation of i_{g*} is called the adjoint matrix $\mathrm{Ad}\, g$. If a vector field $V\big|_g = \xi^i(g) L_i\big|_g$ in left invariant components, its right invariant components are $\zeta = \mathrm{Ad}\, g\, \xi$. Explicitly, we get from (4.12) or (4.13),

$$g X_i g^{-1} = \mathrm{Ad}\, g^j{}_i X_j, \tag{4.14}$$

and the $(3+3)\times(3+3)$ block matrix comes out to be

$$\mathrm{Ad}\, g = \begin{bmatrix} R & 0 \\ \hat{p}R & R \end{bmatrix}. \tag{4.15}$$

The Ad matrices form an alternative faithful matrix representation of the group, isomorphic to the homogeneous representation. Specifically, we have the relations

$$\mathrm{Ad}^{-1}\, g := (\mathrm{Ad}\, g)^{-1} = \mathrm{Ad}(g^{-1}), \quad \mathrm{Ad}\, gh = \mathrm{Ad}\, g\, \mathrm{Ad}\, h. \tag{4.16}$$

They are also compatible with the commutators:

$$[\mathrm{Ad}\, g\, V, \mathrm{Ad}\, g\, W] = \mathrm{Ad}\, g\, [V, W]. \tag{4.17}$$

Thinking again of a smooth curve $g(s)$, if $\dot{g}(0)$'s left invariant components are $\xi = (\omega_l, v_l)$ we can get the corresponding right invariant components as

$$\zeta = \begin{bmatrix} \omega_r \\ v_r \end{bmatrix} = \mathrm{Ad}\, g(0) \begin{bmatrix} \omega_l \\ v_l \end{bmatrix} = \begin{bmatrix} R\omega_l \\ Rv_l + p \times R\omega_l \end{bmatrix}. \tag{4.18}$$

Not surprisingly, this is the composition rule for linear and angular velocities: The angular velocity vector is merely rotated, see also fig. 4.4. In contrast, the linear velocity v_r attains an extra contribution due to the axis offset.

Differentiating one step further, one defines $\mathrm{ad}\, V = V_i\big|_{g=e} \mathrm{Ad}\, g$, acting on the matrix entries, so $\mathrm{ad}\, V$ is again a 6×6 matrix. Using the definition,

$$\mathrm{ad}\, X_i\, X_j = X_i\big|_{g=e} g X_j g^{-1} = X_i X_j - X_j X_i, \tag{4.19}$$

so the adjoint matrix ad is the matrix representation of the commutator: $\mathrm{ad}\, V = [V, \cdot]$, from which follows $(\mathrm{ad}\, X_i)^j{}_k = c^j{}_{ik}$. In block matrix notation,

$$\mathrm{ad}\, V = \begin{bmatrix} \hat{\omega} & 0 \\ \hat{v} & \hat{\omega} \end{bmatrix}. \tag{4.20}$$

From (4.17) it follows that

$$\mathrm{ad}[V, W] = [\mathrm{ad}\, V, \mathrm{ad}\, W], \tag{4.21}$$

which is the Jacobi identity for the commutators.

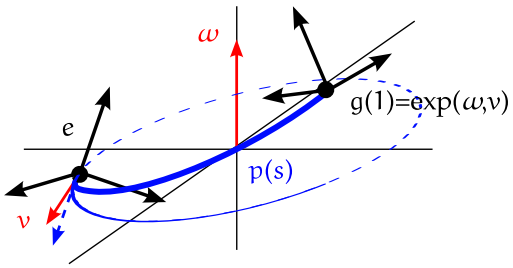

Figure 4.5 | Exponential coordinates on SE. The components ω and v give angular and linear velocity of a screw motion ending at $\exp(\omega, v)$.

4.2.7 The exponential map

Any Lie group G can be parametrized by its infinitesimal generators in a neighborhood of the identity. This is achieved by the exponential map $\exp : T_e G \to G$, which is defined by integration along left invariant vector fields: $\exp V = g(1)$, where $g(s)$ is the solution of $g(0) = e$, $\dot{g}(s) = l_{g*} V$.

In our case $G = SE$, the path $g(s)$ corresponds to a screw motion, i.e. a simultaneous rotation and translation about a common axis. A classical theorem of Chasles states that all rigid body motions can be expressed in this way.

In a matrix representation, the exponential map is the ordinary matrix exponential, defined by its series. For $V \in T_e SE$,

$$\exp V = \exp(V^i X_i) = \sum_{n=0}^{\infty} \frac{1}{n!} (V^i X_i)^n. \tag{4.22}$$

The exponential series of a conjugated group element is

$$g \exp V \, g^{-1} = \sum_{n=0}^{\infty} \frac{1}{n!} (g V g^{-1})^n = \exp(\mathrm{Ad}_g V), \tag{4.23}$$

This relation may be written as $i_g \circ \exp = \exp \circ i_{g*} : se \to SE$ in short. This means that exponential coordinates transform just like tangent vectors under simultaneous changes of lab and body frame.

In practice, it convenient to reduce (4.22) to a finite sum of matrix powers, with nonconstant coefficients, see appendix A.3.

The fact that the homogeneous and adjoint representations are isomorphic leads to the basic relation

$$\mathrm{Ad} \exp V = \exp \mathrm{ad} \, V \tag{4.24}$$

Here, exp on the right hand side is the matrix exponential of the 6×6 square matrix $\operatorname{ad} V$.[7] This correspondence makes it possible to switch to the most convenient representation when needed. As an instructive example, we use it to make explicit all screw motions that commute with a given one $g = \exp(\omega, v)$. Noting that

$$g \exp(\omega', v') g^{-1} = \exp \operatorname{Ad} g\,(\omega', v') = \exp\Big(\big(\exp \operatorname{ad}(\omega, v)\big)\,(\omega', v')\Big), \qquad (4.25)$$

one sees that $g \exp(\omega', v') g^{-1} = \exp(\omega', v') \Leftrightarrow (\omega', v') \in \ker \operatorname{ad}(\omega, v)$. The kernel of $\operatorname{ad}(\omega, v)$ is just the set of infinitesimal motions that commute with (ω, v). It is computed in appendix A.2. The result is that the commuting tangent vectors generate exactly either those screw motions that have the same axis: $\omega' = \omega$, or pure translations in that direction. In short screw motions commute iff they have the same screw axis.

4.2.8 Coordinate charts for *SE*

There are many different ways to represent a rigid body motion $g = (R, p)$ by a set of six parameters. We give a short, non-exhaustive overview of the possibilities that are relevant in the following.

Exponential coordinates use $\log = \exp^{-1}$ as the coordinate chart[8]. The linear order expansion is $g = \exp q = e + q^i X_i + o(q)$. They have a direct geometrical interpretation in terms of screw motions: ω gives the angular and v the linear velocity, which are constant in the instantaneous body frame $g(s)$, see fig. 4.5. Therefore, $\|\omega\|$ is the total angle of rotation, and v is the initial linear velocity $\dot{p}(0)$, in the lab frame, see fig. 4.5.

The exponential coordinates of g and its inverse sum up to 0. It is however *not* true that $\exp(q + q')$ equals the product $\exp q \exp q'$ because of non-commutativity.

The partial derivatives $\{\partial_{q^i}\}$ provide a coordinate frame that coincides with both invariant frames at e but nowhere else. The transformation relating invariant and exponential coordinate frames at other points is detailed in appendix A.4.

[7] This relation can be verified directly by noting that conjugation (a similarity transformation) and taking a matrix power, commute in each term of the exponential series.

[8] This function has multiple branches. However, the set $\{\exp(\omega, v) | \|\omega\| < \pi\}$ covers almost all of *SE*, except for the set of zero measure where $\operatorname{tr} R = -1$, i.e. rotations by π. The branch of log with $\|\omega\| < \pi$ is invertible. See appendix A.3 for more details.

Exponential coordinates based at g_0 are a left translated variant of exponential coordinates. Explicitly, $\tilde{q} = \log(g_0^{-1}g)$. In linear order, $g = g_0 + \tilde{q}^i g_0 X_i + o(\tilde{q})$, so that at the base point, the coordinate frame coincides with the left invariant frame, $\partial_{\tilde{q}^i}\big|_{g_0} = L_i\big|_{g_0}$. The relation to the exponential coordinate frame is $\partial_{\tilde{q}^i}\big|_g = l_{g_0 *}\partial_{q^i}\big|_{g_0^{-1}g}$.

Product coordinates result when some parametrization of $R \in SO$ is combined with separate coordinates for p. Of the many possibilities, we already used in chapter 2 the coordinate system implemented in the 3DNA program [Lu03]. Here, the R is described by a certain choice of Euler angles adapted to the geometry of B-DNA, and p is given in Cartesian coordinates with respect to the mid–frame. For details and conversion formulas to exponential coordinates see appendix A.8.

4.2.9 Invariant coframes

Every tangent space T_gSE has an associated six–dimensional, dual vector space T_g^*SE of linear maps $m : T_gSE \to \mathbb{R}$, called covectors. A smooth map $g \mapsto m\big|_g \in T_g^*SE$ is called a covector field or one-form. The space se^* of left invariant covector fields is spanned by the basis $\{\lambda^j\}_{1\leqslant j\leqslant 6}$, dual to $\{L_i\}$, so that $\langle \lambda^j, L_i\rangle = \lambda^j(L_i) = \delta^j_i$ everywhere. These covector fields form a basis of the cotangent spaces T_g^*SE at every point, i.e. a left invariant coframe. In the same way, a right invariant coframe is defined by $\langle \rho^j, R_i\rangle = \delta^j_i$.

The natural pairing with the coframe elements projects out vector components, similar to a scalar product on a vector space:$V = \langle \lambda^i, V\rangle L_i = \langle \rho^j, V\rangle R_j$, from which follows $\mathrm{Ad}\, g^i{}_j = \langle \rho^i, L_j\rangle$ and $\mathrm{Ad}^{-1}\, g^i{}_j = \langle \lambda^i, R_j\rangle$. The same also works for a covector: $m = \langle m, L_i\rangle\lambda^i = \langle m, R_j\rangle\rho^j$.

Clearly, V (m) is a left invariant (co)vector field exactly if all pairings $\langle \lambda^i, V\rangle$ ($\langle m, L_i\rangle$) are constant on the group.

4.2.10 The coadjoint representation

The components of a covector with respect to right and left invariant coframes are related in much the same way as the vector components. Changing the frame, a covector field $m\big|_g = \nu_i\rho^i\big|_g = \nu_i\langle \rho^i, L_j\rangle\lambda^j\big|_g = \nu_i \mathrm{Ad}\, g^i{}_j\lambda^j\big|_g$. We conclude that if a covector field has left invariant components $\mu(g) = (\tau_l, f_l)$, its right invariant

components are $\nu(g) = \mathrm{Ad}^{-\mathsf{T}}_g \mu(g)$. In block form, the transformation matrix is

$$\mathrm{Ad}^{-\mathsf{T}}_g = \begin{bmatrix} R & \hat{p}R \\ 0 & R \end{bmatrix}. \tag{4.26}$$

Separating this into three–vector components,

$$\nu = \begin{bmatrix} \tau_r \\ f_r \end{bmatrix} = \mathrm{Ad}^{-\mathsf{T}}_g \mu = \begin{bmatrix} R\tau_l + \hat{p}Rf_l \\ Rf_l \end{bmatrix}. \tag{4.27}$$

The τ component attains an extra leverage term, namely the cross product $p \times Rf_l$, see fig. 4.6. These are exactly the transformation rules for forces and torques, see sec. 4.3.

By differentiation of $\mathrm{Ad}^{\mathsf{T}}_g$ at the identity, one gets the map $\mathrm{ad}^{\mathsf{T}}_V : se^* \to se^*$,

$$\mathrm{ad}^{\mathsf{T}}_V \mu = V\big|_{g=e} \mathrm{Ad}^{\mathsf{T}}_g \mu = V\big|_{g=e} \langle \mu, \mathrm{Ad}_g \cdot \rangle. \tag{4.28}$$

Note that ad^{T} does not correspond to a commutator since it depends on one vector and one covector. A relation analogous to (4.24) holds, and $[\mathrm{ad}^{\mathsf{T}}_V, \mathrm{ad}^{\mathsf{T}}_W] = -\mathrm{ad}^{\mathsf{T}}_{[V,W]}$ (note the extra minus sign here).

4.2.11 Tensor fields

A vector V can be regarded as a linear map $V \equiv \langle \cdot, V \rangle : T^*_g SE \to \mathbb{R}$. More generally, a tensor of type (k, l) on some vector space T is a linear map $t : T^{*k} \times T^l \to \mathbb{R}$. A tensor field t on SE can be written in left invariant components as $t = t^{i_1 \dots i_k}{}_{j_1 \dots j_l} L_{i_1} \otimes \dots \otimes L_{i_k} \otimes \lambda^{j_1} \otimes \dots \otimes \lambda^{j_l}$. When changing from left invariant to the right invariant (co)frames $\{\rho^i\}$ and $\{R_i\}$, the k contravariant indices are transformed with Ad while the l covariant indices are transformed with $\mathrm{Ad}^{-\mathsf{T}}$. Clearly, vectors are $(1, 0)$ tensors and covectors are $(0, 1)$ tensors.

Symmetric tensors of types $(2, 0)$ and $(0, 2)$ play a major role in the context of DNA elasticity. We therefore introduce additional abbreviated notation. If C^{ij} are the left invariant components of a $(2, 0)$ tensor at g, its right invariant components can be written as a matrix by

$$C'^{kl} = (\mathrm{AD}_g C)^{kl} := (\mathrm{Ad}_g C \, \mathrm{Ad}^{\mathsf{T}}_g)^{kl} = \mathrm{Ad}_g{}^k{}_i C^{ij} \mathrm{Ad}_g{}^l{}_j, \tag{4.29}$$

where we have introduced the new representation AD of linear operators acting on the space of symmetric $(2, 0)$ tensors. The relations $\mathrm{AD}_g \mathrm{AD}_{g^{-1}} = \mathrm{AD}^{-1}_g$ and

$\mathrm{AD}\, gh = \mathrm{AD}\, g\, \mathrm{AD}\, h$, follow from the analogous properties of the Ad matrices.

If $S = C^{-1}$, then the inverse S' of C' in (4.29) is the component matrix of a $(0,2)$ tensor field,

$$S'_{kl} = (\mathrm{AD}^{-\mathsf{T}}\, g\, S)_{kl} := ((\mathrm{AD}\, g\, C)^{-\mathsf{T}})_{kl} = \mathrm{Ad}^{-1}\, g^{i}{}_{k} S_{ij} \mathrm{Ad}^{-1}\, g^{j}{}_{l}. \tag{4.30}$$

The $\mathrm{AD}^{-\mathsf{T}}$ representation introduced here thus acts on $(0,2)$ tensors.

After choosing a matrix basis for the 6×6 symmetric matrices (a 21-dimensional space), $\mathrm{AD}\, g$ can be written as an invertible 21×21 matrix[9]. In complete analogy to (4.19), one can further define the aD representation of the algebra *se*, by $\mathrm{aD}\, V = V\big|_{g=e}\, \mathrm{AD}\, g$. Then from ((4.24)) we immediately have the basic relations $\exp \mathrm{aD}\, V = \mathrm{AD} \exp V$.

Although the AD matrices may seem huge, they contain no more or less information than a group element itself. They can be seen as an abbreviated matrix notation for simultaneously transforming both tensor indices.[10]

4.3 Rigid base–pair elasticity revisited

Hoping to justify the rather dry collection of mathematical definitions in the previous section, we now give an interpretation of rbc elasticity within this framework. We will see that many of the quantities defined above have a natural physical meaning.

4.3.1 Rigid base–pair deformations

A bp step conformation in a rbc is naturally represented as an element of *SE*. Having found the equilibrium conformation of a step g_0, one may parametrize deformations from it by choosing exponential coordinates based at frame g_0. Then the total conformation

$$g = g_0 \exp(\xi^{i} X_{i}) = g_0 (e + \xi^{i} X_{i} + o(\xi)). \tag{4.31}$$

In thermally fluctuating B-DNA, step deformation angles are so small that the step conformation is well approximated by its linear order expansion. E.g, in the ensembles from [Lan03, Ols98], the angular width of the conformation distributions

[9] The reader is kindly asked to excuse the lack of explicit formulas in this section...

[10] Mathematically speaking, all of the properties of AD are an immediate consequence of the fact that C' is the component matrix of the *pullback* $\mathrm{i}_g^* \tau$ of a 2-tensor τ by the conjugation map i_g.

is below 8°, where $1 - \cos\alpha < 0.01$. Thus in the thermal regime, it is sufficient to think of single step deformations as elements of the tangent space $T_{g_0}SE$, with left invariant vector components ξ. Therefore, ξ is the *strain* of the elastic medium, given in the material frame.[11]

4.3.2 Forces

Generalized forces are defined as the conjugate variables of the step deformations, so that an increment of the step elastic energy can be written as $dE = \mu_i d\xi^i = \langle \mu, d\xi \rangle$ where μ is the force acting on the bps given in the material frame. Forces are therefore covectors. Explicitly, if $\mu = (\tau, f)$,

$$dE = \langle \mu, d\xi \rangle = \tau \cdot d\omega + f \cdot dv, \tag{4.32}$$

so the three–dimensional components of the generalized force μ are the torque τ with respect to the origin of the material frame, and the linear force f. In other words, μ is the *stress*, given in the material frame.

By choosing invariant frames to describe the components of deformations, we have obtained a formulation in which the generalized forces have a straightforward physical interpretation as the usual torque and linear force. An equally intuitive interpretation would have been impossible had we used a parametrization such as Euler angles for the rotation.

4.3.3 Elastic energy

In the rbc model with purely local elasticity, the elastic deformation energy E, quadratic or not, can depend only on the deformation in the material frame. If a chain has the configuration $g_{1k} g_0 \exp(\xi^i X_i)$, the elastic energy of the last step g_{kk+1} is thus a function of ξ only, independent of $g_{1k} g_0$. In other words, *the function* $E : T_g SE \to \mathbb{R}$ *is left invariant*.

The step deformations fluctuate around a mean value $\langle \xi \rangle = 0$. Their positive definite, symmetric covariance matrix $C^{ij} = \langle \xi^i \xi^j \rangle$ determines the linear response $d\xi$ to some external force $d\mu$, via

$$d\xi^i = \beta C^{ij} d\mu_j, \tag{4.33}$$

[11] The choice of reference frame within a single step is somewhat arbitrary. Instead of referring to g_0, one could have taken the start frame e of the step, which gives a component vector $\xi_{pre} = \mathrm{Ad}\, g_0 \xi$, or the frame at half of the equilibrium step leading to $\xi_{mid} = \mathrm{Ad}\, g_0^{1/2} \xi$. We will stay with ξ in the following.

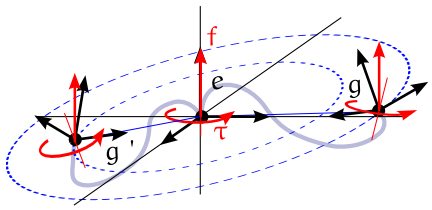

Figure 4.6 | Transformation of a force with right invariant components (τ, f). While the linear component is merely rotated, the torque gets an extra leverage contribution proportional to the offset radius, see (4.27).

which can be written as $d\xi = \beta C(d\mu, \cdot)$. The total elastic work done up to a deformation ξ within the linear response regime is then

$$E = \int \langle \mu, d\xi \rangle = \int \beta C^{ij} \mu_i d\mu_j = \frac{\beta}{2} C(\mu, \mu). \tag{4.34}$$

The covariance matrix C is thus a symmetric, positive definite, left invariant $(2,0)$ tensor field. Substituting $\beta S = C^{-1}$,

$$E = \tfrac{1}{2} S(\xi, \xi), \tag{4.35}$$

where S is a symmetric, positive definite, left invariant $(0,2)$ tensor field. Its left invariant components are the stiffness matrix. The linear stress–strain relation of the chain, written in the material frame, is thus $\mu = S\xi$. The left invariant components of C and S in general still depend on the base sequence, but not on chain conformation.

4.3.4 Change of frame along the chain

The local elastic response of a rbc is naturally described in the material frame, reflected in the fact that the stiffness is left invariant. However, external forces acting on the chain are typically not left invariant covectors. Consider for example a rbc, clamped to a fixed support at bp 1 and subject to an external linear force f applied at the end bp $k+1$, as may occur in an optical tweezers experiment. Here, the external force is naturally given in lab frame components, $\nu = (0, f)$. We can compute the material frame elastic response ξ of the last step deformation by transforming the external force ν to the material frame: If $g = g_{1k} g_0$ is the total

transformation from the fixed base to the mean material frame, then

$$\xi = C\ \mathrm{Ad}^{\mathsf{T}}{}_g\, \nu, \tag{4.36}$$

see sec. 4.2.10. The same relation holds for a general right invariant force/torque combination.[12]

Since there is nothing special about bp 1 as a reference point, we can also express forces or deformations given with respect to some frame along the chain, in any other frame. We just need to replace g in (4.36) by the interjacent transformation which connects the two frames.

Is there a restriction to what local forces can be achieved by an external right invariant force field? To start with, for fixed g, every material force/torque can be produced by varying ν, since $\mathrm{Ad}^{\mathsf{T}}{}_g$ is an invertible matrix. On the other hand, if the external force ν is fixed, by varying the orientation and position g of the end frame, one can only reach a certain subset of material frame forces $\mathrm{orb}^{\mathsf{T}}\nu = \{\mathrm{Ad}^{\mathsf{T}}{}_g\,\nu | g \in SE\}$, the so-called coadjoint orbit of $\nu = (\tau_r, f_r)$. From the block matrix representation one readily proves that $\mathrm{orb}^{\mathsf{T}}\nu = \{(\tau_l, f_l) | f_l \cdot f_l = f_r \cdot f_r, \tau_l \cdot f_l = \tau_r \cdot f_r\}$, i.e. the linear force magnitude and the torque projection in the linear force direction are conserved quantities, restricting the accessible force range.

4.3.5 Compound steps

The ability to express step deformations relative to an arbitrary bp reference frame, makes it possible to combine deformations occurring at *different* base pairs. This is done by first transforming all deformations to a common frame and then adding them up. Recalling sec. 4.2.6, from the basic relation $hgh' = g\mathrm{i}_{g^{-1}}(h)h'$ one derives the first–order relation for small deformations $h = e + V$, $h' = e + V'$,

$$hgh' = g(e + \mathrm{i}_{g^{-1}*}V + V') + o(V + V'). \tag{4.37}$$

Consider now a k-step rbc with equilibrium conformations $g_{0,l\,l+1}$,

$$g_{1\,k+1} = g_{0,12}(e + \xi^i_{12}X_i)g_{0,23}(e + \xi^i_{23}X_i)\cdots g_{0,k\,k+1}(e + \xi^i_{k\,k+1}X_i). \tag{4.38}$$

Applying (4.37), one can commute all of the first order deformations to the right and equilibrium steps to the left. The error made due to non-commutativity is

[12] In a typical magnetic tweezers setup, the torque τ is constant with respect to the bead center, not the base point. Such an external torque is *not* right invariant.

compensated by the appropriate Ad matrices in first order. Explicitly, the result is

$$g_{1\,k+1} = g_{0,1\,k+1}\Big(e + \sum_{l=1}^{k} \big(\mathrm{Ad}\, g^{-1}_{0,l+1\,k+1}\xi_{l\,l+1}\big)^{i} X_i\Big) + o\big(\sum \xi_{l\,l+1}\big), \qquad (4.39)$$

where $g_{0,lm} = g_{0,l\,l+1}\cdots g_{0,m-1\,m}$ and $g_{0,mm} = e$. This formula may look more complicated than it really is; essentially, each ξ acquires one Ad^{-1} factor per commutation.

4.3.6 Diagonalization of the stiffness matrix

We chose the basis $\{L_i\}$ of left invariant deformations in *se* without making reference to the stiffness matrix. An idea that comes to mind is that one should first diagonalize the symmetric, positive definite matrix S by choosing an appropriate basis of eigenvectors, before calculating anything else.

A problem in this approach comes from the fact that the 3×3 blocks of the stiffness matrix correspond to subalgebras of basis vectors with distinct commutation relations, see (4.6). If the new basis is to be interpreted as again consisting of infinitesimal pure rotations and pure translations, the transformation $S' = U^{\mathsf{T}} S U$ must preserve the commutation relations, $U([V, V']) = [U(V), U(V')]$.[13]

The Ad matrices have this property by equation (4.17). It is proved explicitly in appendix A.5 that in fact they are the only commutator–preserving transformations. They correspond to a physical rotation and offset of the frame of reference, just as the usual principal axis transformation of an inertia tensor in mechanics corresponds to a pure physical rotation. The larger set of transformations $SO(6)$ which allows full diagonalization of the stiffness matrix does *not* have an equally simple physical interpretation.

So what is the simplest form of S attainable by a physical change of reference frame? In other words, by transforming

$$S' = \mathrm{Ad}^{\mathsf{T}}_{g}\, S\, \mathrm{Ad}_{g} = \mathrm{AD}^{\mathsf{T}}_{g}\, S, \qquad (4.40)$$

what kind of partial diagonalization can be reached at best? A discussion of this question in the generic case is given in appendix A.6. The following alternative partial diagonal forms for the block matrix $S' = \left[\begin{smallmatrix} S'_{11} & S'_{12} \\ S'^{\mathsf{T}}_{12} & S'_{22} \end{smallmatrix}\right]$ result:

- S'_{11}, S'_{22} both diagonal,

[13] I.e, U is a Lie algebra automorphism.

- S'_{12} diagonal,
- S'_{11} diagonal and S'_{12} symmetric
- S'_{22} diagonal and S'_{12} symmetric.

In particular, the coupling of rotation and translation in the stiffness matrix cannot be eliminated by a change of reference frame. In general, the stiffness matrix depends also on the step sequence σ. Any diagonalization procedure would have to be carried out for each σ separately, which further limits its use. In the following chapters, we will therefore retain the full stiffness matrices, staying in our original original frame of reference determined by the choice of basis $\{X_i\}$.

5 Coarse graining of helical DNA

In this chapter, we relate descriptions of DNA elasticity on different length scales. While the rigid base–pair model captures sequence–dependent elasticity on a microscopic length scale of a few bps, the mesoscopic elastic properties of B-DNA over hundreds of bps are described by sequence–averaged, semiflexible polymer models. By coarse–graining the rigid base–pair chain to a semiflexible polymer model, experiments on the microscopic and mesoscopic scale are made comparable.

5.1 DNA elasticity is scale dependent

Local elastic properties of DNA on a nm length scale play a vital role in basic biological processes such as chromatin organization [Wid01, Seg06] and gene regulation, via indirect readout [Kou87, Hin98, Heg02, Pre93] or via DNA looping [Sch72, Sch92, Rip01].

On a mesoscopic length scale, it is possible to directly measure force–extension relations for DNA in single–molecule experiments [Cha04]. For small external forces, DNA behaves as a worm–like chain (wlc) [Bus94], i.e. an inextensible semiflexible polymer with a single parameter, the bending persistence length, and no explicit sequence dependence. An extension of the classical wlc model, reflecting the chiral symmetry of the DNA double helix, includes coupled twisting and stretching degrees of freedom [Str96, Mar97, Kam97, Mor97]. These become important in a force regime where the DNA molecule is already pulled straight but not yet overstretched [Clu96]. Recent measurements indicate that DNA overtwists when stretched in the linear response regime [Lio06, Gor06].

The issue of relating atomistic and mesoscopic descriptions of DNA elasticity has been addressed mainly by simulation of oligonucleotides. Normal mode analysis using atomistic [Mat99] or knowledge–based rigid base–pair chain (rbc) potentials [Mat02] can give an impression of global bending and twisting modes but disregards viscous damping. In a MD simulation study in explicit solvent, a full set of elastic constants of a fluctuating global helical axis were determined [Lan00]. A recent study [Maz06] extends this approach, elaborating on technical difficulties

of the global axis definition and on convergence criteria.

In this chapter, a systematic coarse–graining of the rbc model down to the wlc scale is described. Here, the average helical geometry of the chain is taken into account exactly. As a result, we obtain exact expressions for the average helical parameters and the full set of stiffnesses for bend, twist, stretch, as well as twist–stretch coupling that characterize an extended wlc elastic model.

5.2 Thermal fluctuations in a rigid base–pair chain

Consider a base pair step $g = g_{k\,k+1}$ in a rbc that fluctuates around a mean or equilibrium value g_0. Deformations can be conveniently expressed in exponential coordinates based at g_0; small deformations are well approximated just by the linear order expansion, i.e. as a tangent vector ξ in left invariant components, see chapter 4. We determine the mean g_0 such that the expectation over all fluctuations, $\langle \xi \rangle = 0$. This is always possible for not too wide step distributions [Ken90], and can be implemented by a gradient search with no numerical problems. The covariance matrix is $C^{ij} = \langle \xi^i \xi^j \rangle$.

Note that we have not specified the source of fluctuations yet. In this chapter, we will consider steps fluctuating thermally. The thermal mean values as well as the thermal covariance matrices depend on the sequence of the step; $g_0 = g_0(\sigma)$, $C = C(\sigma)$. In the next chapter, the effects of random sequence will be added as another independent source of randomness.

5.2.1 Compound steps

Using the matrix formalism described in 4, we can combine a chain of m consecutive steps into one compound step, which in turn is described in terms of its mean and covariance matrix. This is possible as long as the *combined* fluctuations stay small. In other words, the short chain must be well approximated by a (helical) rigid rod.

Consider a rbc with k steps as in eq. (4.38),

$$g_{1\,k+1} = g_{0,12}(e + \xi^i_{12} X_i) g_{0,23}(e + \xi^i_{23} X_i) \cdots g_{0,k\,k+1}(e + \xi^i_{k\,k+1} X_i), \qquad (5.1)$$

where $g_{0,l\,l+1} = g_0(\sigma_{l\,l+1})$ are the equilibrium steps.

Successively commuting the equilibrium steps g_0 to the left, using (4.39), the

compound step takes on the form

$$\xi_{1\,k+1} = \sum_{l=1}^{k} \left(\mathrm{Ad}\, g^{-1}_{0,l+1\,k+1} \xi_{l\,l+1}\right), \tag{5.2}$$

where $g_{0,lm} = g_{0,l\,l+1} \cdots g_{0,m-1\,m}$ and $g_{0,1\,k+1}$ is the equilibrium compound step. Since the Admatrices in this expression are non-random and all single step deformations are assumed independent, the compound covariance $C(\sigma_{1\,m+1})$ equals [Bec07]

$$\sum_{k=1}^{m} \mathrm{Ad}\, g^{-1}_{0,k+1\,m+1} C(\sigma_{k\,k+1}) \,\mathrm{Ad}^{\mathsf{T}}\, g^{-1}_{0k+1\,m+1} = \sum_{k=1}^{m} \mathrm{AD}\, g_0^{-1}(\sigma_{k+1\,m+1}) C(\sigma_{k\,k+1}). \tag{5.3}$$

We have now characterized the compound step in terms of its mean and covariance. This will allow us to treat repetitive, poly-$(\sigma_{1\,m})$ DNA on the same footing as homogeneous DNA. The validity of this combination of steps is limited by the first order approximation for the deformations. For combining, it is necessary that the *compound* step angles stay small, $\|\omega_{1\,m}\| \ll 1$.

5.3 Effective semiflexible polymer for homogeneous chains

What is the effective wlc model that corresponds to a given rigid base–pair chain? We address this question first for a homogeneous (or repetitive, see above) sequence.

Up to this point, step deformations and therefore also the covariance matrices were given with respect to a reference frame equal to the equilibrium base–pair frame g_0, which in general is offset and tilted relative to its own local helical axis. To relate the rbc deformations to a coarse–grained wlc model, we are much more interested in the elastic properties of the *centerline* of the chain.

Once a covariance matrix for deformations of centerline segments is known, the large–scale elastic properties of the wlc are determined. E.g, the bending persistence length of the wlc is defined as the decay length of bend angle correlations and thus depends only on the second moment of the centerline bend angle distribution.

5.3.1 Helical centerline

In the case of a non-fluctuating chain with identical steps, the centerline can be conveniently described using the matrix formalism introduced in chapter 4.

The screw motion $s \mapsto \exp[s\xi^i X_i]$ joins the identity frame e with g as s increases from 0 to 1, see fig. 4.5. Its screw axis is determined by a vector from the origin of e to a point on the axis, given by $p_{ax} = \|\omega\|^{-2}\omega \times v$, and by its direction, ω. It is the 'local helical axis' [Lav89] associated with the base pair step g. When concatenating many *identical* steps g one generates a rbc with frame origins lying on a regular helix with this axis.

In addition to p_{ax} we can define a matrix R_{ax} which rotates e such that ω becomes its third direction vector. One choice is to take p_{ax} as the second new direction. In combination, we then get [Bec07]

$$g_{ax} = \begin{bmatrix} R_{ax} & p_{ax} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{(\omega\times v)\times\omega}{\|(\omega\times v)\times\omega\|} & \frac{\omega\times v}{\|\omega\times v\|} & \frac{\omega}{\|\omega\|} & \frac{\omega\times v}{\|\omega\|^2} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{5.4}$$

which takes e to a frame $e' = eg_{ax} = g_{ax}$ sitting on the helix axis with its third direction pointing along it. One can check that $g' = gg_{ax}$ also has these properties. The primed, on-axis frames are 'local helical axis systems' in the terminology of [Lav89].

Under the influence of thermal fluctuations, the helical structure of the chain becomes irregular. It turns out that in this case the definition of a centerline is problematic in itself. One could try to define it as the local helical axis for each individual base–pair step, cf. fig. 4.5. This has the disadvantage that for a fluctuating chain, the local centerline pieces of consecutive steps do not form a continuous curve, since they are laterally offset. An alternative approach is to fit a continuous centerline globally to a stretch of a rbc, using the *Curves* algorithm [Lav89], as carried out in [Lan00]. The fitting procedure involves a free parameter, namely the relative weight of translational and rotational deviations from an ideal helix shape. By a reasonable choice of this relative weight *a posteriori*, periodic artifacts in the analysis can be reduced but not eliminated [Maz06]. Also, the fact that the resulting centerline depends non-locally on the base pair step conformations introduces artificial correlations on the length scale over which the fitting procedure extends.

We circumvent these problems in three steps. First we transform all rigid base–pairs of the chain to new frames of reference. These are chosen such that *without fluctuations*, all new bp frames lie exactly on, and point in the direction of, a single straight helical axis. We can then identify and average over the unwanted shear degrees of freedom. In a last step, this reduced model is averaged over the helical

phase angle and mapped to the wlc models.

5.3.2 On-axis rbc

The first task is to transform small deviations from an equilibrium conformation g_0 into small deviations from an equivalent on-axis version of g_0. Consider first a regular helix composed of identical g_0 steps. As explained in section 5.3.2, the on-axis step between the k-th and $k+1$-th on-axis frames is

$$g_{0\shortparallel} = (g_0^{k-1} g_{ax})^{-1} g_0^k g_{ax} = g_{ax}^{-1} g_0 g_{ax}, \tag{5.5}$$

where g_{ax} is the on-axis transformation (5.4) corresponding to g_0. Since $g_{0\shortparallel}$ is a transformation between on-axis frames, its rotation and displacement vectors point along the d_3 axis, $\omega_{0\shortparallel} = \|\omega_{0\shortparallel}\| d_3$ and $p_{0\shortparallel} = \|p_{0\shortparallel}\| d_3$.

For a step $g_{k\,k+1} = g_0 \exp[\xi^i X_i]$ of a *fluctuating* rbc we calculate an on-axis version as

$$(g_{1k} g_{ax})^{-1} g_{1k+1} g_{ax} = g_{ax}^{-1} g_{k\,k+1} g_{ax} = g_{0\shortparallel} g_{ax}^{-1} \exp[\xi^i X_i] g_{ax}. \tag{5.6}$$

The three rightmost factors in (5.6) clearly represent the deviation from the on-axis equilibrium step $g_{0\shortparallel}$. Using the property (4.23) we can rewrite

$$g_{0\shortparallel} g_{ax}^{-1} \exp[\xi^i X_i] g_{ax} = g_{0\shortparallel} \exp[\xi^i_{\shortparallel} X_i], \tag{5.7}$$

where the deviation from the on-axis equilibrium step $\xi_{\shortparallel} = \mathrm{Ad}\, g_{ax}^{-1} \xi$. $\xi_{\shortparallel}$ has zero mean and covariance matrix $C^{ij}_{\shortparallel} = \left\langle \xi^i_{\shortparallel} \xi^j_{\shortparallel} \right\rangle$,

$$C_{\shortparallel} = \mathrm{Ad}\, g_{ax}^{-1} C\, \mathrm{Ad}^{\mathsf{T}}\, g_{ax}^{-1} = \mathrm{AD}\, g_{ax}^{-1} C. \tag{5.8}$$

The rbc composed of steps (5.7) is an equivalent description of the original chain, which one may call its on-axis version [Bec07]. Intuitively, to each fluctuating frame g_{1k} of the original chain, we rigidly connected a frame g'_{1k} in such a way that the primed, on-axis chain fluctuates about a straight, but still twisted, equilibrium conformation. This is illustrated in fig. 5.1: The equilibrium conformations generate a tilted helix that is offset from the helical centerline. Thermal fluctuations distort it, producing an irregular helix. However, on average, the on-axis configuration is exactly lined up on a straight helical axis. Note that we had no need to compute a fluctuating axis explicitly, nor choose a weighting factor [Maz06].

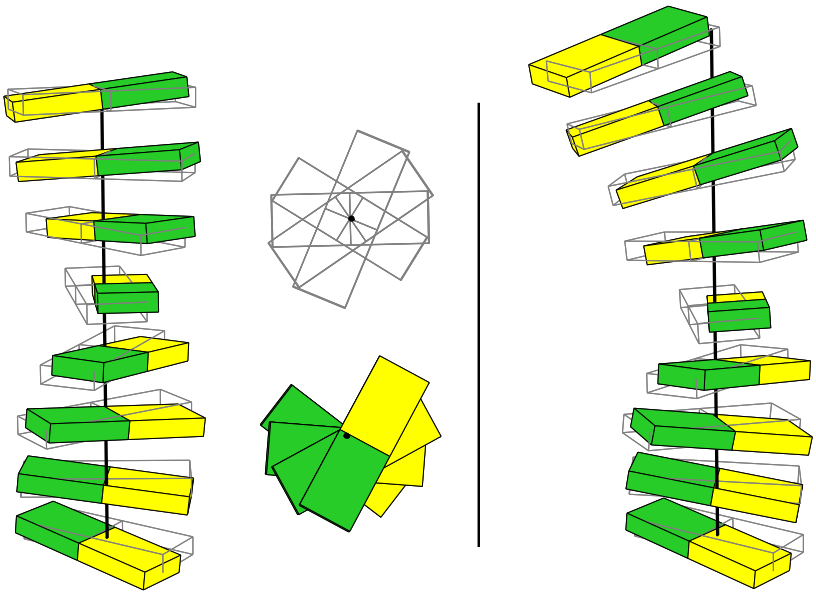

Figure 5.1 | Equivalent descriptions of a poly-G rbc. Left: Colored blocks represent base pairs in their equilibrium conformations. Wireframe blocks represent their on–axis counterparts. Right: Thermal fluctuations distort the helix. (MP parameter set, base pair size scaled down by 1/2 for clarity.)

5.3.3 Averaging over shear variables

The on-axis rbc has the nice property that the translational fluctuations $(\xi^4_{\shortparallel}, \xi^5_{\shortparallel}) = (\nu^1_{\shortparallel}, \nu^2_{\shortparallel})$ are now exactly transversal to the equilibrium helix axis. They are pure shear modes and do not contribute to compression fluctuations along the chain. Let $\eta = (\omega_{\shortparallel}, \nu^3_{\shortparallel})$ be the vector of the four remaining variables. Noting that the invariant volume element dV_ξ in exponential coordinates depends only the angular part (see appendix A.7), one has

$$\langle \eta^i \eta^j \rangle = \int \underbrace{d^3\omega_{\shortparallel} d\nu^3_{\shortparallel} A(\omega_{\shortparallel})}_{dV_\eta} \underbrace{\int d\nu^1_{\shortparallel} d\nu^2_{\shortparallel} p(\xi_{\shortparallel})}_{p(\eta)} \eta^i \eta^j, \tag{5.9}$$

from which one can see that the 4×4 covariance matrix $\widetilde{C}^{ij} = \langle \eta^i \eta^j \rangle$ is in fact the same as $C_{\shortparallel}$ with its $\nu^1_{\shortparallel}, \nu^2_{\shortparallel}$ rows and columns deleted. Thus, η has a distribution around 0 with covariance matrix $\widetilde{C}$. Here and in the following, $\widetilde{\cdot}$ indicates deletion of the shear rows and columns in an on-axis, 6×6 matrix. E.g, $\widetilde{Ad}$ is the 4×4 adjoint matrix. Its on-axis version $\widetilde{Ad}\, g_{0_{\shortparallel}}$ has a particularly simple form. Using

(4.15) and noting that $p_{0\shortparallel} \propto \omega_{0\shortparallel} \propto d_3$ we obtain

$$\widetilde{\mathrm{Ad}}\, g_{0\shortparallel} = \begin{pmatrix} \cos\|\omega_0\| & \sin\|\omega_0\| & 0 & 0 \\ -\sin\|\omega_0\| & \cos\|\omega_0\| & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5.10}$$

5.3.4 Averaging over the helical phase

A shear–averaged, on-axis rbc still has a finite equilibrium twist and anisotropic bending stiffness. To relate it to a wlc with isotropic bending rigidity, one can perform an average over a continuous helical phase angle rotation of the reference frame [Mar94]. An on-axis covariance matrix which is rotated by a helical phase angle φ around the average local helical axis (see (6.7)), is

$$\widehat{C}(\varphi) = \widetilde{\mathrm{Ad}}\, g_\varphi \widetilde{C} \widetilde{\mathrm{Ad}}^{\,T} g_\varphi, \tag{5.11}$$

where $g_\varphi = \exp[\varphi X_3]$ is a pure rotation by an angle φ around d_3. Since $\widetilde{\mathrm{Ad}}\, g_\varphi$ has the form (5.10), the helical phase average comes out as [Bec07]

$$\bar{C} = \frac{1}{2\pi}\int_0^{2\pi} \widehat{C}(\varphi)\mathrm{d}\varphi = \begin{pmatrix} \frac{\widetilde{C}^{11}+\widetilde{C}^{22}}{2} & 0 & 0 & 0 \\ 0 & \frac{\widetilde{C}^{11}+\widetilde{C}^{22}}{2} & 0 & 0 \\ 0 & 0 & \widetilde{C}^{33} & \widetilde{C}^{34} \\ 0 & 0 & \widetilde{C}^{34} & \widetilde{C}^{44} \end{pmatrix}. \tag{5.12}$$

From $\bar{C}$ one can read off the bend persistence length as $l_b = h_{\shortparallel}/\bar{C}^{11}$. E.g, the mean square end–to–end distance of a homogeneous chain $\langle R^2 \rangle \propto 2 l_b l$ for contour lengths $l \gg l_b$. The torsional modulus, normalized to units of length is called the twist persistence length $l_t = h_{\shortparallel}/\bar{C}^{33}$ [1] (see e.g.[Mar94]). Here, the on-axis helical rise $h_{\shortparallel} = \|p_{0\shortparallel}\|$. The wlc stiffness matrix $\beta\bar{S} = \bar{C}^{-1}$ can be found by inversion and has the same block structure as $\bar{C}$, see also appendix A.7.

When the considered covariance matrix actually belongs to a *compound* step, $\bar{C} = \bar{C}_{1\,m+1}$, all of the elastic parameters can be extracted in the same way, the only difference being that $h_{\shortparallel}$ has to be taken as the total helical rise on the compound step. Also, $\bar{S}$ will be the compound step stiffness, which can be renormalized to one bp step by multiplying with m.

[1] This is the modulus for unconstrained stretching degree of freedom.

5.4 Coarse–graining relations

We have derived all wlc elastic parameters starting from an arbitrarily oriented and offset homogeneous rbc. We now discuss in some detail how these coarse–grained parameters are related to the microscopic rbc parameters.

5.4.1 Equilibrium step

The transformation of the equilibrium step onto the helical axis (5.5) leaves the total rotation angle invariant. Therefore the equilibrium twist of $g_{0\shortparallel}$ is $\theta_{\shortparallel} = \|\omega_{0\shortparallel}\| = \|\omega_0\| \geqslant |\omega_0^3|$. I.e, the twist per base pair of the wlc equals the total angle of rotation, not the Tw angle of the off-axis step. The equilibrium rise on axis is $h_{\shortparallel} = \|p_{0\shortparallel}\| = \omega_0^T p_0 / \|\omega_0\|$ which is different from both off-axis quantities $\|p_0\|$ and p_0^3. These differences are of order $O(\omega_0^1 + \omega_0^2)^2$ so they become important only when the equilibrium rotation axis ω_0 has significant roll and tilt with respect to the material frame, i.e. when the local helical parameters Inclination and Tip [Dic89] are not negligible.

5.4.2 Fluctuations

Unlike the equilibrium step, the covariance matrix is changed not only by the rotation R_{ax} but also by the shift p_{ax} onto the average local helix axis. Intuitively, the on-axis frame g' is rigidly connected to g, cf. fig. 5.1. Therefore, a rotational fluctuation of g with rotation vector $\delta\omega$ will result in an additional *translational* fluctuations of g' equal to $\delta\omega \times p_{ax}$.

A familiar example of this geometrical effect is the stretching of an ordinary coil spring along its helix axis, see fig. 5.2. In the wire material, this deformation corresponds mainly to torsion, i.e. a rotational deformation of consecutive wire segments. On a larger scale, the same deformation is levered into a translation of one coil end along the helix axis. The transformation (6.7) captures exactly this lever arm effect, which is proportional to the total axial displacement $\|p_{ax}\|$ and so becomes relevant if the chain deviates from an idealized B-DNA form.

We calculate explicitly the 3×3 blocks $C_{\shortparallel}^{(ab)}$ of $C_{\shortparallel}$, (5.7), in terms of the corresponding blocks $C^{(ab)}$ of C, using (6.7) and (4.15). Here $a, b \in \{\omega, \nu\}$ stand for the set of rotational or translational components, respectively. Further, we let $C^{(ab)\prime} = R_{ax}^T C^{(ab)} R_{ax}$, and $P'_{ax} = R_{ax}{}^i{}_j p_{ax}^j \epsilon_i$, which is an antisymmetric matrix.

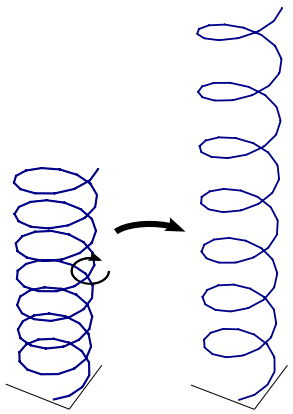

Figure 5.2 | What looks like linear extension of a coil spring on the "mesoscale" is almost exclusively due to torsion when described in the "microscopic" material frame of the wire.

Using this notation,

$$C_{\shortparallel} = \begin{pmatrix} C^{(\omega\omega)\prime} & C^{(\omega\nu)\prime} + C^{(\omega\omega)\prime}P'_{ax} \\ C^{(\nu\omega)\prime} - P'_{ax}C^{(\omega\omega)\prime} & C^{(\nu\nu)\prime} - P'_{ax}C^{(\omega\omega)\prime}P'_{ax} + \\ & +C^{(\nu\omega)\prime}P'_{ax} - P'_{ax}C^{(\omega\nu)\prime} \end{pmatrix}. \tag{5.13}$$

Inspecting this expression, the rotational block $C_{\shortparallel}^{(\omega\omega)}$ is merely a rotated version of the off-axis rotational block $C^{(\omega\omega)}$. In contrast, the translational block $C_{\shortparallel}^{(\nu\nu)}$ and the coupling block $C_{\shortparallel}^{(\omega\nu)}$ have 'leverage terms', since rotational fluctuations about directions perpendicular to the offset vector contribute through a cross product with p_{ax}. For $C_{\shortparallel}^{(\nu\nu)}$, these involve the off-axis coupling $C^{(\nu\omega)}$ in first order and rotational fluctuations $C^{(\omega\omega)}$ in second order in $\|p_{ax}\|$. The coupling block $C_{\shortparallel}^{(\omega\nu)}$ has contributions from $C^{(\omega\omega)}$ in first order. These leverage terms persist in the reduced wlc covariance matrix $\widetilde{C}$. They are the remainder of the microscopic description of fluctuations with respect to a material frame that is offset from the average helical axis.

Consider for example a base pair step that exhibits x-displacement but no Inclination or Tip, i.e. $p_{ax} \propto d_1, \omega \propto d_3, R_{ax} = I_3$. Then (5.13) implies that any coupled Roll–Rise (C^{26}) and Roll (C^{22}) fluctuations will add to the stretching fluctuations $C_{\shortparallel}^{66}$ of the chain. In addition, the off-axis Roll–Twist fluctuation (C^{23}) contributes to twist–stretch coupling fluctuation on axis, $C_{\shortparallel}^{36}$.

When Inclination or Tip are nonzero, then due to the additional rotation R_{ax}

Table 5.1 | Comparison of wlc geometry and stiffness parameters of all six unique repetitive sequences of period two, for the MP hybrid parametrization. In the 'av' row, the values for the average step is shown. MP parameter set.

	$\frac{\pi}{\|\omega_{13}\|}$	$\frac{1}{2}h_{\shortparallel 13}$	$\beta\bar{S}^{11}$	$\beta\bar{S}^{33}$	$\beta\bar{S}^{44}$	$\beta\bar{S}^{34}$	r_{resp}
AA	10.2	0.327	144	141	976	−38.3	0.27
AC	10.4	0.333	132	142	1140	−105	0.74
AG	10.5	0.334	139	159	1120	−103	0.64
AT	10.7	0.334	111	195	975	−80.1	0.41
GG	10.9	0.338	159	186	1090	−89.9	0.48
CG	10.3	0.338	124	126	831	−78.5	0.62
$\langle\sigma\rangle$	10.5	0.334	134	153	1050	−87.9	0.57
units	bp	nm	rad^{-2}	rad^{-2}	nm^{-2}	$(\mathrm{nm\,rad})^{-1}$	$\mathrm{rad\,nm}^{-1}$

also Shift and Slide fluctuations contribute to the resulting wlc parameters. It is therefore essential to transform to an on-axis frame before averaging over the shear degrees of freedom.

5.4.3 wlc parameters of dinucleotide repeats

As a result of the coarse–graining procedure outlined above, we can extract the wlc parameters of repetitive sequences from the sequence–dependent rbc stiffness (or covariance) matrices and equilibrium offsets [Bec07].

A detailed view of wlc geometry and stiffness is given in table 5.1. The twist rate and equilibrium rise per bp vary by roughly 2 %. Their respective values for the average step, obtained by averaging the equilibrium conformation and covariance initially, closely match commonly accepted values for B-DNA.

The poly-AT repeat stands out as the most bendable sequence which is at the same time torsionally rather stiff. Another common trend in our results is that poly-G DNA is comparatively stiff with respect to bending. The values are comparable to MD studies in which elastic constants of oligonucleotides were measured, with repeats AA, AT, GC and GG [Lan00] and with AT and GC [Maz06]. There too, poly-AT is torsionally stiff but bendable. However, bending persistence lengths from [Lan00, Maz06] are up to two times bigger than either our or experimental values, possibly due to bending relaxation too slow to be seen in that simulation [Lan00]. The twisting persistence lengths in [Lan00, Maz06] are generally larger than our results by about a factor of two, and show stronger sequence–dependence,

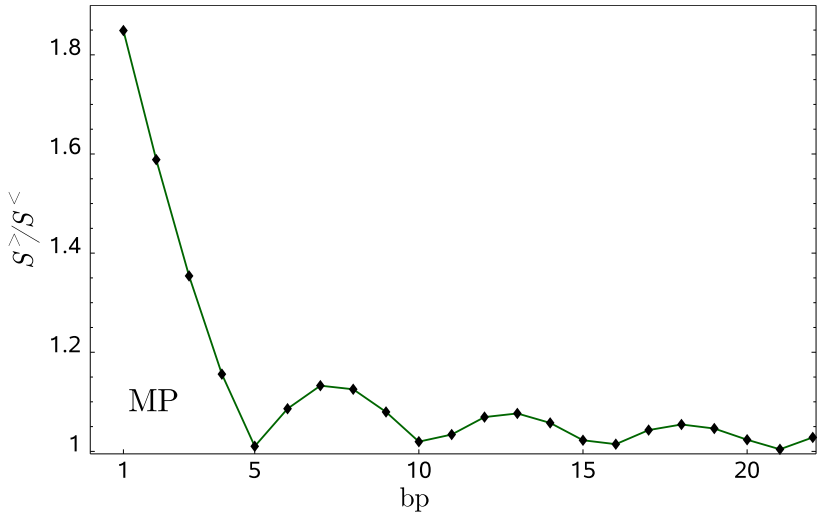

Figure 5.3 | Bending anisotropy. The ratio of larger over smaller bending stiffness decays in an oscillating fashion with compound step length. MP parameter set, average step geometry.

but with similar trends. The stretch modulus and the twist–stretch coupling depend on the sequence in a correlated way. Again comparing with [Lan00], their stretch moduli agree qualitatively but show a different sequence dependence. Also, their twist–stretch coupling constants are *positive*, unlike our and recent single–molecule experimental results [Gor06, Lio06].

The rightmost column of table 5.1 shows the ratio of overtwist over elongation in response to an external stretching force, $r_{resp} = \bar{C}^{34}/\bar{C}^{44}$. When a repetitive sequence is cut by one bp and then stretched to the original length, the "missing twist" at the last bp ranges from 29 (AA) to 20 (AC) degrees undertwist.

5.5 Anisotropic bending

A feature of short compound steps not captured by the coarse–grained wlc limit is their anisotropic bending stiffness. Using the compound covariance $\widetilde{C}_{1\,k+1}$ (see (5.3)) it is possible to quantify the decay of anisotropy for short chains. On scales much longer than a full turn, bending will be isotropic.

The ratio of the two principal bending stiffnesses as a function of chain length is shown in fig. 5.3.

The oscillatory decay results from orientational averaging over fractional turns

of the helix. Since linear response is always symmetric, the bending anisotropy has minima every *half* turn of the double helix. For exactly two full turns (21 bp), anisotropy is suppressed completely, but already a 5 bp compound step at almost a half turn is essentially isotropic. This behavior agrees nicely with that of the two principal bending stiffnesses measured in [Lan00] for oligonucleotides of increasing length. Their stiffnesses are equal at around 6 bp, in line with the fact that MD potential produces a 12 bp/turn helix structure.

5.6 Discussion

This chapter presented a way to quantitatively connect experiments on DNA elasticity on different length scales. We relate the stiffness expressed in terms of rigid base–pair deformations, obtained via an initial coarse–graining [Gon01], to the long–wavelength wlc parameters of a homogeneous chain [Bec07]. In this coarse–graining step it is essential to properly account for the helical base–pair geometry. For this purpose an on-axis version of the rigid base–pair chain was introduced, which *on average* has ideal B-DNA shape. This makes it straightforward to integrate over the shear degrees of freedom and helical phase, to finally obtain all four linear elastic constants allowed by the large–scale symmetry of the molecule [Kam97, Mar97, Mor97].

The results allow a direct comparison of the different microscopic effective potentials to single molecule and cyclization experiments. It involves no free parameter, once a microscopic rbc parameter set is specified. One finds good qualitative agreement, including the negative sign of twist–stretch coupling.

Does the rather involved computation of macroscopic parameters actually make a noticeable difference? The calculations could be simplified by disregarding the details of average helical geometry of the chain. Treating all base–pair steps as ideal B-DNA from the beginning as in [?], one would perform an average of the *off-axis* covariance matrix, over Shift, Slide and helical phase angle. Inverting this, one obtains a "naïve" stiffness matrix S_{na}. The relative error made in such a computation, $e^{ij} = (S^{ij}_{na} - \bar{S}^{ij})/\bar{S}^{ij}$ is shown in table 5.2.

While the bending and twisting stiffnesses are well approximated by the naïve guess, the error in stretch modulus and twist–stretch coupling is considerable. For these terms, leverage due to the axis offset becomes important (section 5.4). Especially the naive twist–stretch coupling is not negative enough. The effect is more pronounced with the pure MD parameter set [Lan03, Lan06b], since it has

Table 5.2 | Relative error in stiffness parameters made when using "naive" matrix elements instead of the coarse–grained parameters described above. Values are given in per cent. Average step bp parameter.

	e^{11}	e^{33}	e^{44}	e^{34}
MD	3	-13	59	50
MP	2	-7	-5	48

unusual equilibrium conformations with stronger axis offset.

The coarse–graining procedure just described involves no approximations regarding the geometry. This makes it directly applicable to alternative DNA structures, and indeed any polymer with average helical geometry, once microscopic covariance matrices are available. In fact, the more the average geometry deviates from idealized B-DNA, the greater is the need to treat the helical geometry correctly.

The main model assumption is that thermal deformation fluctuations of neighboring steps are independent. Another limitation of any rigid base–pair model is that *internal* deformation fluctuations of a base–pair such as propeller twist or buckle, are not explicit and thus effectively treated as uncorrelated between base pairs.

The framework can be extended to improve on both of these points. Nearest–neighbor correlations in base–pair parameters may be included by extending the model to a full Markov chain. Internal deformations could then be added by extending the configuration space, leading to a bi-rod [Moa05] in the continuum limit. However for either of these interesting generalizations, a microscopic parametrization is an open challenge in itself. The fact that dinucleotide step stiffness depends overall rather weakly on the flanking sequence [AB05] and the encouraging agreement with mesoscopic data, suggest that the main features of coarse–grained DNA elasticity are captured already by the presented more basic model. However, the low twist rigidity calculated here might be a result of missing negative twist correlations.

6 Coarse graining of random DNA

The correspondence of DNA elastic models on different length scales is expanded. We now turn the attention to sequence dependent *DNA elasticity as captured by the rigid base–pair model. The resulting local variability of structure and stiffness on a few-bps length scale, will translate into effective conformational and elastic properties of the mesoscopic worm–like chain model. We look at the specific case of uncorrelated, random sequence, leading to a homogeneous effective worm–like chain.*

6.1 Mapping a random sequence rbc to a homogeneous wlc

As has been known for twenty years [Tri88], the total apparent persistence length of a wlc is composed of a static part which originates from the sequence dependent equilibrium bends of the molecule, and a dynamic part induced by thermal fluctuations. Their relative contributions have been more recently measured, with incompatible results [Bed95, Vol02]. The idea of splitting the fluctuations into a static and a thermally induced part can be adapted to the case of a random sequence rbc. Extending the coarse–graining procedure to include structural variability, in this chapter, the conformational statistics of rigid base–pair chain ensembles with random, uncorrelated base sequence will be calculated. One arrives again at an effective homogeneous wlc description. On short scales, deviations from the effective wlc due to stiffness variability do occur. A quantitative estimate for these deviations will be given.

The method presented in chapter 5, consists in expressing the fluctuating conformations as deformations with respect to a helical reference structure, and then transforming to an idealized, on-axis helix. Finally, irrelevant degrees of freedom are identified and averaged over.

Two new difficulties arise in a random sequence rbc: The first is the choice of reference structure when structural disorder is present, since the chain no longer forms a regular helix in the absence of thermal fluctuations. This issue is addressed in the following sections 6.1.2, 6.1.3 and 6.1.4. The second difficulty arises

from the fact that the sequence distribution features independent *bases*, while the conformation distributions depend on the base–pair *steps*. Loosely speaking, the sequence distribution lives on the 'nodes' of the model while the conformation distribution lives on its 'links'. We explain in section 6.1.5 how this introduces effective short–range correlations.

6.1.1 Random sequence rbc

Instead of homogeneous or repetitive sequences, we now turn our attention to random sequences, as a generic approximation to the properties of natural DNA. The crucial difference is that the relaxed conformation of any realization of random DNA is no longer a regular helix, and that the relaxed conformations of consecutive steps are correlated due to sequence continuity. To get around these complications, we introduce an ensemble average over sequence randomness in addition to the thermal average at fixed sequence.

A random sequence rbc is by definition a sequence of rigid base–pair frames generated iteratively in the following way: Start with some choice of base at position $i = 1$. Then for each new base–pair $i + 1$,

1. choose a base identity b_{i+1} at random, following a fixed base distribution $p(b)$ [1].

2. Generate the bp step conformation $g_{i\,i+1}$. Due to thermal fluctuations, this conformation is also random. It follows a pdf $p(g|\sigma)$ whose center and width depend parametrically on the step sequence $\sigma_{i\,i+1} = b_i b_{i+1}$.

After $m - 1$ iterations, one ends up with a realization $\sigma_{1\,m} = b_1 \dots b_m$ of the random sequence and a corresponding realization $g_{1m} = g_{12} \dots g_{m-1\,m}$ of conformations. The random sequence rbc built up in this way has the same conformational statistics as an ensemble of thermally fluctuating rbcs, each with random but fixed sequence.

Generally, denote $\langle f(g_{1m})\rangle$ an expectation value of some function f over conformations of a thermal, random sequence rbc ensemble. It can be carried out sequentially:

$$\langle f(g_{1m})\rangle = \langle\, \langle f(g_{1m})|\sigma_{1m}\rangle \,\rangle = \sum_{b_1,\dots,b_m} p(\sigma_{1m})\, \langle f(g_{1m})|\sigma_{1m}\rangle \tag{6.1}$$

[1] In the examples below, a flat distribution is chosen, although a sequence bias can be included.

Here the *conditional expectation* $\langle\cdot|\sigma\rangle$ [2], is identical to a thermal average and $\langle\cdot\rangle$ denotes the *global* average over both thermal and sequence randomness. Averages over the sequence ensemble only, are not considered. The second equality in (6.1) follows because $\langle f(g_{1m})|\sigma_{1m}\rangle$ is already averaged over thermal fluctuations.

A random sequence rbc captures the effects of sequence dependent structure and stiffness. It is a good model for DNA under the assumptions that (a) sequences of bases are independent, that (b) thermal conformations of base–pair steps are independent, and that (c) step conformations are independent of flanking base sequence. All of these assumptions are wrong in general, but may be considered reasonable first approximations. In particular, relaxing (a) requires extra knowledge about sequence statistics. Also, no parametrizations of conformational correlations are yet available that would allow to relax (b). In MD simulation studies [Dix05, AB05], (c) was investigated, and a dependence of stiffness and equilibrium conformations on flanking base sequence was found. It is however much weaker than the dependence on the actual step sequence and can be reasonably neglected in a first approximation.

6.1.2 Irregular helix axes

The crucial step in the coarse–graining procedure is the 'on-axis transformation' described in sec. 5.3. For a homogeneous or repetitive rbc this was straightforward, since a regular reference structure is formed by the thermal equilibrium conformations.

For a typical realization of random sequence however, the thermal equilibrium conformation is already an irregular helix, which leads to the same problems of defining a centerline as discussed in sec. 5.3.1 even without thermal fluctuations. We will therefore *not* choose the approach of expressing thermal deformations of each sequence realization with respect to irregular on-axis frames. Rather, our strategy will be to describe random sequence conformations, just like those of homogeneous sequences, as deformations from some sequence–averaged, *regular* helix. The task is then to determine the geometrical parameters of this helix and the corresponding on-axis covariance.

[2] The conditional expectation of some function f is defined with respect to the conditional distribution, $\langle f|\sigma\rangle = \int f(g)p(g|\sigma)dg$

6.1.3 Thermal and sequence randomness

Consider a base pair step with sequence $\sigma_{i\,i+1}$ in a chain which fluctuates in a thermal environment. Its sequence dependent thermal mean conformation as well as the covariance matrix are moments of the conditional pdf $p(g_{i\,i+1}|\sigma_{i\,i+1})$. What changes when $\sigma_{i\,i+1}$ itself is a random variable?

To start with, it is important that the sequence dependent variability in equilibrium conformations of B-DNA bp steps is in fact *smaller* than the average thermal fluctuation size. Since only the limit of small thermal deformations has been considered throughout, it is only consistent to use the same small deformation limit for the sequence induced conformational variability.

The basic idea then is to treat sequence variability exactly on the same footing as thermally induced fluctuations; we add the sequence induced deviations from a *global* equilibrium conformation as another independent source of randomness. I.e. the basic setup is changed slightly. A random sequence step $g = g_0 \exp[\xi^i X_i]$ now fluctuates around a sequence–*independent* global center g_0. Its total fluctuations are characterized by a covariance matrix $C^{ij} = \langle \xi^i \xi^j \rangle$ resulting from *both* sequence and thermal fluctuations.

We now need to calculate the global center g_0 and the total covariance C from the thermal and sequence statistics. Recalling that $\langle\cdot\rangle$ denotes a total thermal and sequence ensemble average, we can determine g_0 by the condition that $\langle\xi\rangle = 0$, analogous to sec. 5.2.

One can split the deformation from g_0 into sequence plus thermal parts:

$$\xi = \langle\xi|\sigma\rangle + (\xi - \langle\xi|\sigma\rangle). \tag{6.2}$$

Note that the thermal equilibrium deformation $\langle\xi|\sigma\rangle$ is a random variable, depending on σ, while $(\xi - \langle\xi|\sigma\rangle)$ is the thermal deformation, another random variable.

Within a regime of linear response, the deformation energy of a step with fixed sequence σ is a quadratic function of the deviation from the thermal equilibrium value $\langle\xi|\sigma\rangle$. The associated thermal covariance matrix is sequence dependent:

$$C^{ij}(\sigma) = \left\langle\, (\xi - \langle\xi|\sigma\rangle)^i (\xi - \langle\xi|\sigma\rangle)^j \,\middle|\, \sigma \right\rangle. \tag{6.3}$$

Comparing this with the thermal fluctuations introduced in sec. 5.2.1, one sees that $g_0(\sigma) \simeq g_0(e + \langle\xi|\sigma\rangle)$. Also, (6.3) agrees with the $C(\sigma)$ used there to quadratic order in the deformations.

On the other hand, the covariance of the thermal mean values is sequence–

independent:

$$C_0^{ij} = \left\langle \langle \xi|\sigma\rangle^i \langle \xi|\sigma\rangle^j \right\rangle, \tag{6.4}$$

where the outermost expectation is effectively taken with respect to $p(\sigma)$ only, cf. (6.1).

What is the total covariance C ? The two sources of randomness are of independent physical origin, but are not independent random variables: Although the *realization* of the thermal conformation is sequence–independent, its *distribution* depends on σ. I.e, $p(\xi|\sigma)p(\sigma) = p(\xi,\sigma) \neq p(\xi)p(\sigma)$. Using the decomposition (6.2),

$$\begin{aligned} \left\langle \xi^i \xi^j \right\rangle =& \left\langle \langle \xi|\sigma\rangle^i \langle \xi|\sigma\rangle^j \right\rangle + \left\langle (\xi - \langle \xi|\sigma\rangle)^i (\xi - \langle \xi|\sigma\rangle)^j \right\rangle + \\ &+ \left\langle \langle \xi|\sigma\rangle^i (\xi - \langle \xi|\sigma\rangle)^j \right\rangle + \left\langle (\xi - \langle \xi|\sigma\rangle)^i \langle \xi|\sigma\rangle^j \right\rangle. \end{aligned} \tag{6.5}$$

Now note that $\left\langle \langle \xi|\sigma\rangle^i (\xi - \langle \xi|\sigma\rangle)^j | \sigma \right\rangle = 0$ trivially. Using this with the identity $\langle\cdot\rangle = \langle\langle\cdot|\sigma\rangle\rangle$ in (6.5), the cross–terms vanish. The simple result is that the sequential covariance and the sequence–averaged thermal covariance add up to give the total covariance C:

$$C = C_0 + \langle C(\sigma)\rangle\,. \tag{6.6}$$

This basic result [Bec07] is the generalization of the well–known relation that the inverse persistence lengths of thermal and structural disorder are additive. In fact, this relation can be recovered from (6.6), see below.

In summary, given the covariance (or stiffness) matrices and equilibrium values of all sixteen dinucleotide steps, and a distribution of relative step frequencies $p(\sigma)$, by computing g_0 and C we have characterized a single, thermally fluctuating random sequence step in terms of its center and second moment. The global equilibrium step g_0 defines a regular helix which is taken as the reference structure structure in the following. Deformations from this reference are governed by the total covariance C which includes a contribution from sequence–induced conformational variability.

6.1.4 Transformation onto the average helical axis

Having identified the regular reference structure to use, one can now begin to follow the coarse–graining procedure from chapter 5. As a first step, the total deformation fluctuations are transformed onto the *average* helical axis: $\xi_{\shortparallel} = \mathrm{Ad}\, g_{ax}^{-1}\xi$, where g_{ax} is defined by the global equilibrium g_0 via (5.4). The on-axis

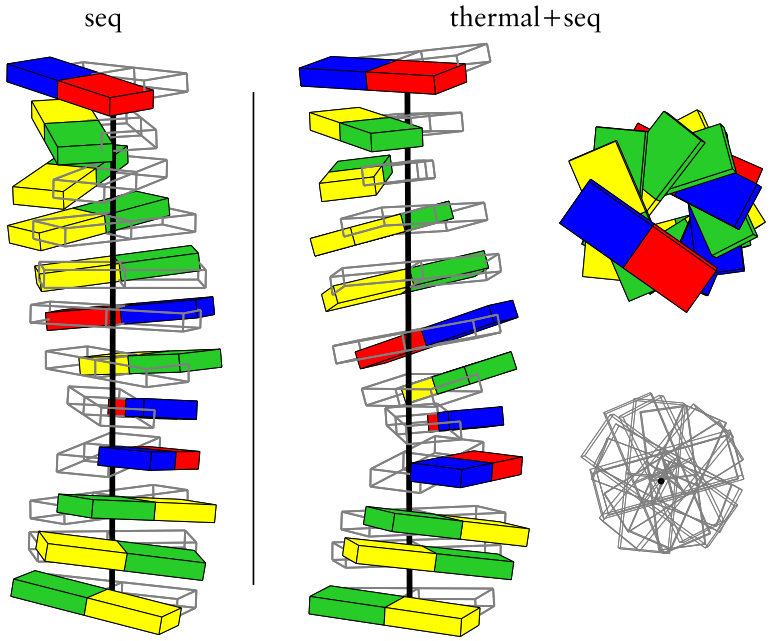

Figure 6.1 | Equivalent descriptions of a realization of a random sequence rbc. 'seq': Colored blocks represent base pairs in their thermal equilibrium conformations. Wireframe blocks represent their on-axis counterparts, which do *not* lie on a straight line without sequence averaging. 'thermal+seq': The same, but with added thermal fluctuations. The top views show the reduced helix axis offsets of the on-axis frames. (MD parameter set, base pair size scaled down by 40 % for clarity, sequence GCGTTGTGGGCT.)

deformation then still has zero mean $\langle \xi_{\shortparallel} \rangle = 0$ (but $\langle \xi_{\shortparallel} | \sigma \rangle \neq 0$) and covariance matrix

$$\mathrm{C}_{\shortparallel} = \mathrm{Ad}\, \mathrm{g}_{ax}{}^{-1}\, \mathrm{C}\, \mathrm{Ad}^{\mathsf{T}}\, \mathrm{g}_{ax}{}^{-1} = \mathrm{AD}\, \mathrm{g}_{ax}^{-1} \mathrm{C}. \tag{6.7}$$

One realization of a random sequence rbc, together with its on-axis version, is shown in fig. 6.1.

6.1.5 Correlations induced by sequence

While by assumption thermal fluctuations of neighboring steps are independent random variables, the sequences of different bases, not steps, are independent. Any realization of a random sequence of dinucleotide steps must be 'continuous', e.g. $\sigma_{12} = \mathrm{AG}$ implies that σ_{23} can only start with a G. Since the step sequences are correlated, so are the step sequence dependent static offsets $\langle \xi | \sigma \rangle$.

Consider the combined fluctuations of a short rbc consisting of m bp steps.

The joint pdf of sequence steps along the chain is the product of base pdfs, $p(\sigma_{12}, \ldots, \sigma_{m\,m+1}) = \prod_{k=1}^{m+1} p(b_k)$. This implies that correlations between thermal mean values extend up to nearest neighbor steps:

$$\left\langle \langle \xi^i_{\shortparallel k\,k+1} | \sigma_{k\,k+1} \rangle \langle \xi^j_{\shortparallel l\,l+1} | \sigma_{l\,l+1} \rangle \right\rangle = \begin{cases} C^{ij}_{0\shortparallel} & l = k \\ C^{ij}_{1\shortparallel} & l = k+1 \\ C^{ji}_{1\shortparallel} & l = k-1 \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

Here, the on-axis covariance of thermal means $C_{0\shortparallel}$ and the new on-axis nearest–neighbor term $C_{1\shortparallel}$ are defined by the left hand side (lhs). They can be computed when $p(\sigma)$ is known. No further thermal or thermal–sequential nearest–neighbor terms occur by the assumptions of the model, as can be verified by splitting the deformation in thermal and sequence parts as in (6.5).

Now combine the m base pair steps of the chain into a compound step. The compound deformation is given by eqn. (4.39),

$$\xi_{\shortparallel 1\,m+1} = \sum_{k=1}^{m} \mathrm{Ad}\, g_{0\shortparallel}^{k-m} \xi_{\shortparallel k\,k+1}. \tag{6.9}$$

What is the sequence induced covariance matrix

$$C_{0\shortparallel\,1m+1}^{ij} = \left\langle \langle \xi^i_{\shortparallel 1\,m+1} | \sigma_{1\,m+1} \rangle \langle \xi^j_{\shortparallel 1\,m+1} | \sigma_{1\,m+1} \rangle \right\rangle \tag{6.10}$$

of the compound deformation, now that nearest–neighbor correlations are present? Using (6.8), one obtains a sum of appropriately transformed single–step covariances $C_{0\shortparallel}$ and in addition a sum of nearest neighbor cross–terms involving $C_{1\shortparallel}$:

$$\begin{gathered} C_{0\shortparallel 1\,m+1} = \sum_{l=0}^{m-1} \mathrm{AD}\, g_{0\shortparallel}^{-l} C_{0\shortparallel} + \sum_{l=0}^{m-2} \mathrm{AD}\, g_{0\shortparallel}^{-l} C_{\times}, \\ \text{where } C_{\times} = C_{1\shortparallel}\, \mathrm{Ad}^{\mathsf{T}}\, g_{0\shortparallel}^{-1} + \mathrm{Ad}\, g_{0\shortparallel}^{-1}\, C_{1\shortparallel}^{\mathsf{T}} \end{gathered} \tag{6.11}$$

The cross–covariance $C_{\times}$ represents the fact that nearest neighbor equilibrium steps are correlated and their frames of reference have a relative offset equal to $g_{0\shortparallel}$.

Note that two neighboring compound steps are still correlated by sequence continuity at their interface. From (6.11) one extracts the recursion relation

$$C_{0\shortparallel 1\,l+1} = \mathrm{AD}\, g_{0\shortparallel}^{-1} C_{0\shortparallel 1\,l} + C_{0\shortparallel} + C_{\times}. \tag{6.12}$$

The same relation is obeyed by a sequence of *independent* steps with covariance matrix $\widehat{C}_0 = C_{0\shortparallel} + C_\times$. This means that except for a boundary term $C_\times$ from the beginning of the chain, a rbc with independent steps and covariance $\widehat{C}_0$ exhibits the same effective sequence induced conformational covariance as the original chain which is short range correlated by $C_\times$. The relative error in effective compound covariance is of order $1/m$. We neglect this error in the following, writing $C_{0\shortparallel 1\,m+1} = \sum_{l=0}^{m-1} \mathrm{AD}\, g_{0\shortparallel}^{-l} \widehat{C}_0$.

Finally, we can combine the independent version $\widehat{C}_0$ of the sequence induced covariance with the thermal covariance according to (6.6). The total conformation covariance of the thermally fluctuating, random sequence chain is then given by $\widehat{C} = \widehat{C}_0 + \langle C(\sigma)\rangle$.

In summary, the conformational statistics of a compound step including sequence randomness are now represented by an effective, stepwise independent on-axis rbc governed by $\widehat{C}$ that incorporates the additional requirement of sequence continuity [Bec07].

6.1.6 Averaging over shear and helical phase

The final step in the coarse–graining procedure is to average over unwanted degrees of freedom. The first average to be taken is that over the shear degrees of freedom $(\nu^1_{\shortparallel}, \nu^2_{\shortparallel})$. As explained in sec. 5.3.3, the result is that the remaining four variables $\eta = (\omega_{\shortparallel}, \nu^3_{\shortparallel})$ have a 4×4 covariance matrix $\widetilde{C}$ which equals $\hat{C}_{\shortparallel}$ with its $(\nu^1_{\shortparallel}, \nu^2_{\shortparallel})$ rows and columns deleted. It turns out that due to the particular block structure of the $\mathrm{Ad}\, g_{0\shortparallel}$ matrices, the row and column deletion may be carried out before the summation[3] in the following equation, so that for a compound step

$$\widetilde{\widehat{C}}_{1\,m+1} = \sum_{l=0}^{m-1} \widetilde{\mathrm{Ad}}\, g_{0\shortparallel}^{-l} \widetilde{\widehat{C}}\, \widetilde{\mathrm{Ad}}^{\mathsf{T}} g_{0\shortparallel}^{-l}. \tag{6.13}$$

Alternatively, one can also directly perform an average over the helical phase for an individual $\widehat{C}$ step, producing a version of the covariance that has isotropic bending as well as twist, stretch and twist–stretch coupling covariances:

$$\bar{C} = \frac{1}{2\pi} \int_0^{2\pi} \widetilde{\mathrm{Ad}}\, g_\varphi \widetilde{\widehat{C}}\, \widetilde{\mathrm{Ad}}^{\mathsf{T}} g_\varphi \, \mathrm{d}\varphi, \tag{6.14}$$

as in sec. 5.3.4. The covariance matrix $\bar{C}$ determines the long–scale conformational

[3] This is because the (ν^1, ν^2) columns of $\mathrm{Ad}\, g_{0\shortparallel}^{-l}$ contain no coupling to the (ω, ν^3) rows

statistics of the chain, i.e. the parameters of the effective wlc.

6.2 Random sequence chain conformations and numerical test

The global offset g_0 and the combined covariance matrix $\widehat{C}$ are set up such that they capture the conformational statistics of an ensemble of thermally fluctuating, random sequence rigid base–pair chains, cf. eqn. (6.1). From $\bar{C}$ one can read off the bend persistence length as $l_b = h_{\shortparallel}/\bar{C}^{11}$. The torsional modulus[4] normalized to units of length, one can call the twist persistence length $l_t = h_{\shortparallel}/\bar{C}^{33}$ (see e.g. [Mar94]). Here, the on-axis helical rise $h_{\shortparallel} = \|p_{0\shortparallel}\|$. Since they reflect sequence variability, these are *apparent* persistence lengths [Bed95, Vol02]. E.g, the square end–to–end distance, averaged over a random sequence sequence ensemble $\langle R^2 \rangle = 2l_b l$ for long contour lengths $l \gg l_b$.

Entirely analogous quantities can be defined 'at zero temperature' when thermal fluctuations are switched off, by just setting the thermal part of the covariance to 0. So if $\bar{C}_0 = \frac{1}{2\pi}\int_0^{2\pi} \widetilde{Ad}_{g_\varphi} \widehat{\widehat{C}}_0 \widetilde{Ad}^T_{g_\varphi}\, d\varphi$ is the pure sequential, helical–averaged covariance, then $l_{0,b} = h_{\shortparallel}/\bar{C}_0^{11}$ and $l_{0,t} = h_{\shortparallel}/\bar{C}_0^{33}$ are the static bend and twist persistence lengths, respectively.

The coarse–graining from rbc to wlc was tested with a simple–sampling Monte Carlo (MC) simulation according to the algorithm in sec. 6.1. The measured mean squared base–pair center end–to–end distances are shown in fig. 6.2. The theoretical curves $\langle R^2 \rangle = 2ll_b - 2l_b^2(1 - e^{-l/l_b})$ for an inextensible wlc using the computed contour and bending persistence lengths, l and l_b, fit the simulation data to within numerical error. The only deviations occur below 3 nm, where the inextensible wlc model fails to reproduce the displacement due to compression and shear modes present in the rbc. In chapter 7, an alternative way of computing the mean squared end–to–end distance is presented, see section 7.4.3. While it is less intuitive than the on–axis transformation, that method yields accurate results over the whole range of contour lengths.

In addition to the full covariance C, simulations were also carried out for structural disorder only, setting all of the $C(\sigma) = 0$. The corresponding wlc using C_0 and the next–neighbor term C_1 (see (6.8)) again fits the data. However, disregarding C_1 is clearly wrong.

Experiments that include a sequence ensemble average over conformations and thus measure apparent persistence lengths, include cryo-electron microscopy

[4] for unconstrained stretching

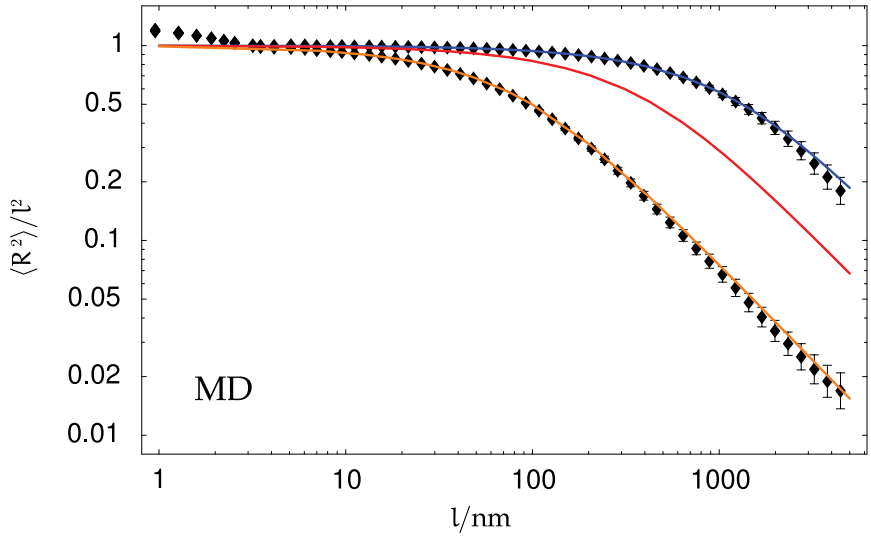

Figure 6.2 | Comparison of an MC simulation of a random–sequence rbc to the coarse–grained effective wlc. Symbols designate the measured mean squared end–to–end distances for static disorder only (upper row) and for static plus thermal fluctuations (lower row). The theoretical curves assuming a wlc model are shown from top to bottom for static disorder ($\widehat{C}_0$, blue), uncorrelated static disorder (C_0 only, red), and static plus thermal fluctuations ($\widehat{C}$, orange), respectively. MD parameter set.

of frozen conformations of oligonucleotides [Bed95], AFM tracing of adsorbed random–sequence DNA [Wig06], and cyclization of random fragments [Vol02]. Whenever such experiments are interpreted in terms of a intrinsically straight, *homogeneous* DNA, then the apparent stiffness matrix extracted from experiment corresponds to the inverse of the total covariance, $(\beta\bar{C})^{-1}$.

6.3 Response to external forces

A slightly different situation arises in force–extension experiments carried out on single molecules (e.g, [Gor06, Lio06]). Here, an external stretching force tilts the elastic energy landscape of each step along the chain, introducing a bias towards those thermal fluctuations that increase the molecule's extension. No such bias can be introduced on the sequence. Therefore the sequence randomness part of the total conformational covariance does *not* directly result in additional compliance to an external force.

What is the remaining effect of irregular sequence in micromanipulation experi-

ments? This question is discussed below in the weak static disorder limit, which is a good approximation for DNA. We adopt the basic idea of [Nel98] which is to expand the elastic Boltzmann factor $B \sim e^{-\frac{\beta}{2}(\xi-\langle\xi|\sigma\rangle)^T(\langle S\rangle+\delta S(\sigma))(\xi-\langle\xi|\sigma\rangle)}$ for weak static disorder, and to interpret the result in terms of a homogeneous chain with *renormalized* stiffness. The somewhat surprising result is that the renormalized stiffness is the inverse *total* covariance, $(\beta\widehat{C})^{-1}$[Bec07].

How does this come about? The expectation value of an observable $f(g_{1m})$, e.g, the z-extension p^3_{1m}, for a fixed sequence σ_{1m}, is given by the multiple integral

$$\langle f|\sigma_{1m}\rangle_\epsilon = \frac{1}{Z}\int\Big(\prod_{k=1}^{m-1} dV_{\xi_{k\,k+1}}\Big) f(g_{1m}) B_\epsilon e^{-\beta U(g_{1m})};$$
$$B_\epsilon = e^{-\frac{\beta}{2}\sum_{k=1}^{n-1}(\xi_{k\,k+1}-\epsilon\langle\xi|\sigma_{k\,k+1}\rangle)^T S(\xi_{k\,k+1}-\epsilon\langle\xi|\sigma_{k\,k+1}\rangle)}. \tag{6.15}$$

In this expression, Z is the partition sum and $U(g_{1m})$ is an external potential, e.g. $U = \|f\|p^3_{1m}$ for linear stretching with a force $f = \|f\|d_3$. For a start, the elastic Boltzmann weight B_ϵ, has sequence dependent offsets but a constant stiffness matrix S. The auxiliary parameter ϵ was introduced to keep track of orders in the following weak static disorder expansion:

$$\frac{B_\epsilon}{B_0} = 1 + \epsilon\sum_{k=1}^{m-1}\xi^T_{k\,k+1}\beta S\,\langle\xi|\sigma_{k\,k+1}\rangle + \frac{\epsilon^2}{2}\sum_{k=1}^{m-1} -\langle\xi|\sigma_{k\,k+1}\rangle^T\beta S\,\langle\xi|\sigma_{k\,k+1}\rangle$$
$$+\frac{\epsilon^2}{2}\Big(\sum_{k=1}^{m-1}\xi^T_{k\,k+1}\beta S\,\langle\xi|\sigma_{k\,k+1}\rangle\Big)^2 + O(\epsilon^3). \tag{6.16}$$

We proceed to calculate the global expectation value

$$\langle f\rangle_\epsilon = \big\langle\langle f|\sigma_{1m}\rangle_\epsilon\big\rangle = \sum_{b_1\ldots b_m} p(\sigma_{1m})\,\langle f|\sigma_{1m}\rangle_\epsilon\,. \tag{6.17}$$

Using (6.16) and (6.15), after interchanging sequence average and integration, the result is

$$\langle f\rangle_\epsilon = \Big\langle f\Big[1+\epsilon^2\beta^2\big(\frac{1}{2}\sum_{k=1}^{m-1}\xi^T_{k\,k+1}SC_0S\xi_{k\,k+1} + \sum_{k=2}^{m-1}\xi^T_{k-1\,k}SC_1S\xi_{k\,k+1}\big)\Big]\Big\rangle_0$$
$$+O(\epsilon^3). \tag{6.18}$$

As can be seen, in sequence average the first order term drops out. The first of

the quadratic terms from (6.16) produces a constant which is relevant only for normalization. It was discarded from (6.18).[5] The surviving second quadratic term can be seen to produce the sums involving the static covariance C_0 and nearest–neighbor covariance C_1 (see (6.11)).

The square bracket in (6.18) may be interpreted as the truncated expansion of an exponential. Written that way, eqn. (6.18) is to second order, identical to an expectation value taken without static disorder but with renormalized elastic energy [Nel98]:

$$\langle f\rangle_\epsilon = \frac{1}{\tilde{Z}}\int(\prod_{k=1}^{m-1} dV_{\xi_{k\,k+1}})\, f e^{-\beta U}\times \\ \times\, e^{-\frac{\beta}{2}\left(\sum_{k=1}^{m-1}\xi^T_{k\,k+1}(S-\epsilon^2\beta S C_0 S)\xi_{k\,k+1} - 2\sum_{k=2}^{m-1}\xi^T_{k-1\,k}\epsilon^2\beta S C_1 S\xi_{k\,k+1}\right)} + O(\epsilon^3) \quad (6.19)$$

It is an exercise in multidimensional Gaussian integrals to verify that the renormalized elastic energy in (6.19) produces the covariances

$$\langle \xi^i_{k\,k+1}\xi^j_{k\,k+1}\rangle = (\beta S)^{-1\,ij} + \epsilon^2 C_0^{ij} \text{ and } \langle \xi^i_{k-1\,k}\xi^j_{k\,k+1}\rangle = \epsilon^2 C_1^{ij} \quad (6.20)$$

to second order in ϵ, in the free case $U = 0$.

As a next step, sequence dependent stiffness can be incorporated. What changes? Splitting up the thermal covariance matrix (6.3) into its average and sequence dependent parts, $C_{th}(\sigma) = \langle C_{th}\rangle + \delta C_{th}(\sigma)$. Since C scales as $O(\xi)^2$, it is a natural choice to assign an order $O(\epsilon)^2$ to the term $\delta C_{th}(\sigma)$. In this way, the changes in width of the distribution are $O(\epsilon)$. We then replace

$$\beta S \to (\langle C_{th}\rangle + \epsilon^2\delta C_{th}(\sigma_{k\,k+1}))^{-1} = \beta S_{av} - \epsilon^2\beta^2 S_{av}\delta C_{th} S_{av} + O(\epsilon^3) \quad (6.21)$$

in eqn. (6.15), where $\beta S_{av} = \langle C_{th}\rangle^{-1}$. Repeating the expansion of B_ϵ as before, all occurrences of S in (6.16) are replaced by S_{av}. The only extra term in second order, $-\epsilon^2\beta\sum \xi^T_{k\,k+1} S_{av}\delta C_{th}(\sigma_{k\,k+1}) S_{av}\xi_{k\,k+1}$ drops out in the sequence average (6.18). Thus, sequence dependent stiffness is averaged out in this order.[6]

In summary, to second order in ϵ, the random rbc with sequence disorder in offsets and stiffness, produces the same response to external forces or torques as a

[5] Including this constant can be seen to fix the correct normalization of the elastic Boltzmann factor to second order.

[6] Note that if one treats $\delta C_{th}(\sigma) = O(\epsilon)$, the effect of stiffness variability is not as trivial. It involves correlations between stiffness and offsets which are outside the scope of this work.

homogeneous chain with a renormalized elastic energy [Bec07]. This renormalized energy corresponds to step deformation covariances which are the sum of thermal and nearest–neighbor correlated static parts. As explained in sec. 6.1.5, any chain of this kind can be mapped to an rbc with *independent* step deformations which have the *total* covariance $\widehat{\mathrm{C}} = \langle \mathrm{C}_{th} \rangle + \mathrm{C}_0 + \mathrm{C}_\times$. I.e, although the sequence disorder is quenched, its effect on the entropic elasticity of the random chain is the same *as if* the sequence randomness were an additional elastic compliance. When fitting force–extension measurements with homogeneous elastic parameters of a wlc model, the measured result corresponds to the total, or apparent stiffness $\bar{\mathrm{S}}$ and not to the bare, local stiffness $\bar{\mathrm{S}}_{av}$.

6.4 Effective worm–like chain parameters

This section gives an overiew of the results of the coarse–graining procedure for random sequence DNA [Bec07].

6.4.1 Conformational covariance of random DNA

In table 6.1 the coarse–grained wlc geometry and covariance parameters corresponding to a random sequence rbc are shown. The values are comparable to all experiments in which an ensemble average over DNA sequence is implicitly performed, see sec. 6.2.

For the crystal parameter sets, the equilibrium rise and twist are close to the commonly accepted values of 0.34 nm/step and 10.5 bp/turn. The MD rise and twist are both low, a known effect for the force field used in that study [Bev04]. The MD bending persistence length is smaller than the commonly accepted values at physiological conditions, which are around 48 nm [Vol02]. The low equilibrium Rise of the MD conformations accounts for half of this deviation. The elastic constants of the B and P parameter sets differ from the MD ones since the choice of effective temperature only fixes overall fluctuation strength, not relative stiffness of different modes, see sec. 1.6.

For all parameters sets, the twist persistence length is similar to the bend persistence length, and is smaller than the result of 58 nm extracted from cyclization data [Vol02].

No rescaling by a different effective temperature can bring all crystal stiffness parameters into reasonable agreement with MD since the various deviations occur

Table 6.1 | Random sequence wlc geometry, persistence lengths and conformational covariances for the considered rbc potentials.

	$\frac{2\pi}{\theta_{\shortparallel}}$	$h_{\shortparallel}$	l_b	l_t	$\bar{C}^{11}$	$\bar{C}^{33}$	$\bar{C}^{44}$	$\bar{C}^{34}$
B	10.1	0.334	27.1	15.2	12.	22.	0.79	0.67
P	10.5	0.334	43.4	35.7	7.7	9.4	0.86	0.85
MD	11.9	0.318	38.9	45.1	8.2	7.	1.9	1.2
MP	10.5	0.334	42.8	47.8	7.8	7.	1.	0.55
units	1	nm	nm	nm	$\frac{\text{rad}^2}{10^3}$	$\frac{\text{rad}^2}{10^3}$	$\frac{\text{nm}^2}{10^3}$	$\frac{\text{nm rad}}{10^3}$

Table 6.2 | Thermal and static contributions to the apparent persistence length for different potentials. For comparison, the l' column shows the static persistence lengths when sequence continuity is disregarded.

	l_b	$l_{b,th}$	$l_{b,0}$	$l'_{b,0}$	l_t	$l_{t,th}$	$l_{t,0}$	$l'_{t,0}$
B	27.1	29.5	327	211	15.2	15.4	1260	88.3
P	43.4	45.3	1040	575	35.7	36.3	2430	172
MD	38.9	42.	519	175	45.1	47.7	818	256
MP	42.8	44.6	1040	575	47.8	48.8	2340	172
units	nm							

in opposite directions.

6.4.2 Thermal vs. sequence randomness

Instead of combining fluctuations in a random DNA ensemble, one can consider thermal and sequence fluctuations separately. Table 6.2 shows the corresponding static and thermal persistence lengths [Tri88], whose inverse additivity follows from eqn. (6.6). In disagreement with the cryo–EM study [Bed95], the static persistence lengths are much higher than the thermal ones, leading to a correction of only a few nm. This is in accordance with the analysis based on cyclization [Vol02]. Also, the static $l_{b,0}$ for the P parameter sets correctly reproduces the value found numerically in that study, using the same parameter set. When the requirement of sequence continuity is dropped, as shown in the l' columns, static variability is strongly overestimated (for twist, more than tenfold).

Table 6.3 | Experimental stiffness parameters as given in the literature and average thermal stiffness (using the MP parameter set). The conversion factor for B, C, G, S from [Gor06] is $\beta/h_{\shortparallel}$. The conversion factors for B, C, D in [Lio06] are respectively, $\theta_{\shortparallel}^2/h_{\shortparallel}^3$, $1/h_{\shortparallel}$, $\theta_{\shortparallel}/h_{\shortparallel}^2$. Beware of a missing 1/2 factor in their first formula.

	$\beta\bar{S}^{11}$	$\beta\bar{S}^{33}$	$\beta\bar{S}^{44}$	$\beta\bar{S}^{34}$
Gore *et al.*[Gor06]	163 ± 15	327 ± 15	781 ± 150	-64 ± 15
Lionnet *et al.*[Lio06]		294	710	-47 ± 20
MP	128	149	1045	-82
units	rad^{-2}	rad^{-2}	nm^{-2}	$(\mathrm{nm\,rad})^{-1}$

6.4.3 Stiffness of random DNA

Recent single–molecule experiments at moderate applied tension have given new data on DNA stiffness [Lio06, Gor06]. All of the elastic parameters given in these articles are collected in table 6.3, together with the stiffness of a random rbc computed from the MP parameter set, see sec. 6.3. The bending modulus of 128 k_BT/rad^2 is lower than the result from [Gor06] and still on the low end of the range of $132 - 138$ k_BT/rad^2 found in previous [Wan97, Bau97, Wen02] single–molecule experiments. However, in [Sal06] a lower experimental value is reported.

The deviation in torsional rigidity is much more dramatic. Recent experimental values are about twice as high as the coarse–grained rbc results, see also [Cha04] for a review. This low twist rigidity is a feature of all parameter sets. For the crystal parameter sets one might argue this indicates that torsional deformations carry more elastic energy than bending deformations, thus 'violating' an assumed equipartition of energy. However, for the MD parameter set, this is clearly not the case; the rbc version of the simulated DNA oligomers is indeed more twistable than experimental values for DNA suggest. A speculative explanation is that there may exist negative correlations between thermal twist deformations of neighboring base pair steps which are neglected in the independent base–pair model, leading to an underestimation of twist stiffness.

Negative twist–stretch coupling has been demonstrated in [Gor06, Lio06], a feature that is reproduced with good agreement by the microscopic data, and is also visible in the local Twist-Rise coupling of the microscopic parameter sets [?].

6.5 Limits of applicability of the wlc model

As a continuous model, the wlc is defined down to arbitrarily small length scales. However the microscopic structure of DNA suggests that there must be a lower limit to its applicability. Indeed, recent experimental studies [Wig06, Lan06a] have highlighted examples of strong bending on short scales, which are in disagreement with standard wlc elasticity. At what length scale does an isotropic, homogeneous wlc fail to reproduce the behavior of a random rbc?

6.5.1 Bend angle distributions for short chains

The combined covariance matrix $\widetilde{C}_{1\,m+1}$ gives the second moment of the distribution $p(\eta_{1\,m+1})$ of deformations, observed in a random sequence, thermal ensemble of length m compound steps. Here it is not necessary that the single step deformation distributions have a Gaussian shape. Indeed such an assumption depends on the choice of coordinates, and is not justified by experiments.

Nevertheless, assume for the moment additionally that for each sequence, the single step thermal deformation distributions were in fact Gaussians in the chosen coordinates. The deformation of a specific *compound* step with sequence $\sigma_{1\,m+1}$ then again follows a Gaussian distribution $p(\eta_{1\,m+1}|\sigma_{1\,m+1})$, since in the small deformation angle approximation considered, it is the result of a convolution of the single step covariances.

Sequence randomness changes this picture. The deformation distribution of an ensemble of random compound steps $p(\eta_{1\,m+1}) = \langle p(\eta_{1\,m+1}|\sigma_{1\,m+1})\rangle$ is a sequence average of several Gaussians with different offsets and widths and thus in general *deviates* from a Gaussian shape. So a perfect Gaussian shape cannot be expected for *short* random sequence compound steps.

In a recent AFM study of DNA adsorbed to a coverslip [Wig06], bend angle distributions of DNA over short lengths have been found to favor large bend angles much more than expected from the wlc model. It is interesting to ask whether this can be explained as an effect purely of sequence randomness as outlined above. In fig. 6.3, the effective potential U_{eff} for the total bend angle $\vartheta = ((\eta^1_{1,m+1})^2 + (\eta^2_{1,m+1})^2)^{1/2}$ of random sequence compound steps of different lengths m, is shown. It was extracted from histograms of a simulation as described in section 6.1.

For compound steps shorter that 5 bp, the effective potentials stay well below

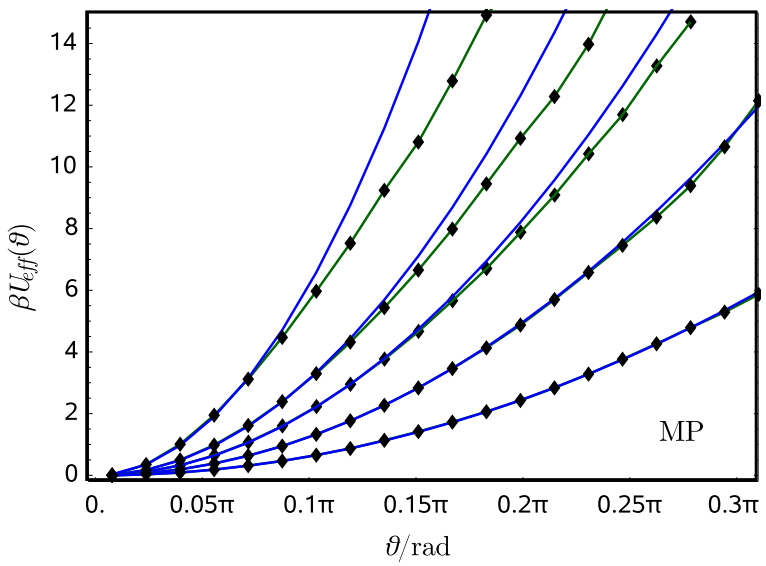

Figure 6.3 | Effective potential for the total bend angle ϑ (curve with symbols, green). The curves without symbols (blue) show the harmonic approximation to the effective potential that results of a fine–graining of an isotropic wlc with the corresponding coarse–grained persistence length. Compound step length, from left to right: 1,2,3,5,10 bp. MP parameter set.

the respective harmonic potentials that correspond to an isotropic wlc model with the coarse–grained, random DNA persistence length of $l_b = 42.8$ nm. This is the combined result of the spread is bending stiffness resulting from sequence randomness, as well as from anisotropic bending, as illustrated in fig. 5.3. However, above 5 bp the observed deviations are negligible and thus insufficient to explain the wide bend angle distributions observed in [Wig06] for DNA as long as 15 bp.

6.5.2 Short–scale stiffness variability

When the considered random chains get shorter, the effective stiffness will start to exhibit stronger fluctuations depending on sequence. The following section addresses the breakdown of the assumption of constant wlc stiffness for short chains.

The thermal covariance matrix $\widetilde{C}(\sigma_{1\,m+1})$ of a compound step with fixed sequence $\sigma_{1\,m+1}$ was calculated in sec. 6.3. While the mean thermal covariance matrix $M = \left\langle \widetilde{C}(\sigma_{1\,m+1}) \right\rangle$ is just the sequence average, the covariances of the 4×4

matrix entries are given by

$$V^{ijkl}_{1\,m+1} = \left\langle (\widetilde{C}^{\,ij}(\sigma_{1m+1}) - M^{ij})(\widetilde{C}^{\,kl}(\sigma_{1m+1}) - M^{kl}) \right\rangle . \tag{6.22}$$

This expectation can be evaluated in terms of single–step and nearest–neighbor sequential covariances of the matrix entries, analogous to the procedure for the sequence covariance itself, see eq. (6.11). The bulky result is stated here for completeness:

$$\begin{gathered}
V^{ijkl}_{1\,m+1} = \sum_{l=0}^{m-1} a^{(-l)ij}{}_{no} V_0^{nopq} a^{(-l)kl}{}_{pq} + \sum_{l=0}^{m-2} a^{(-l)ij}{}_{no} V_\times^{nopq} a^{(-l)kl}{}_{pq}, \\
\text{where } V_\times^{ijkl} = V_1^{ijno} a^{(-l)kl}{}_{no} + a^{(-l)ij}{}_{no} V_1^{nokl}, \\
V_s^{ijkl} = \left\langle \, (\widetilde{C}_{\,\sigma_{n\,n+1}} - \langle \widetilde{C}_{\,\sigma} \rangle)^{ij} (\widetilde{C}_{\,\sigma_{n+s\,n+1+s}} - \langle \widetilde{C}_{\,\sigma} \rangle)^{kl} \, \right\rangle, \; s = 0, 1, \\
\text{and } a^{(-l)ij}{}_{kl} = (\widetilde{\mathrm{Ad}}\, g_{0\,\mathrm{II}}^{-l})^{i}{}_{k} (\widetilde{\mathrm{Ad}}\, g_{0\,\mathrm{II}}^{-l})^{j}{}_{l} \text{ is closely related to } \mathrm{AD}\, g_{0\,\mathrm{II}}^{-l}.
\end{gathered} \tag{6.23}$$

Using a small fraction of this information, one can characterize stiffness variability; the relative spread of angular stiffness coefficients of compound steps over all sequences is shown in fig. 6.4. Explicitly, $\Delta S/S = (V^{iiii}_{1\,m+1})^{1/2}/M^{ii}$, where $S = S^{ii}$ and $i = 1, 3$.

Again, including the nearest neighbor cross-covariances V_1 takes sequence continuity into account. E.g, the fact that it is impossible to combine two of the comparatively soft pyrimidine–purine [Ols98] steps in a row, reduces the variability of the average stiffness across random sequence compound steps.

After one full turn, variability in stiffness is down to 5%. The effect of sequence continuity is to reduce the variability compared to a model with independent step sequences, analogous to table 6.2.

6.6 Conclusions

In addition to homogeneous or repetitive DNA as considered in chapter 5, the coarse–graining formulas have been extended to the generic case of random DNA sequence. In the rbc model, sequence randomness affects equilibrium structure as well as stiffness parameters, as described by existing microscopic parametrizations of rbc potentials.

The conformational fluctuations of random sequence DNA are directly comparable to persistence lengths measured in experiments such as cyclization and AFM

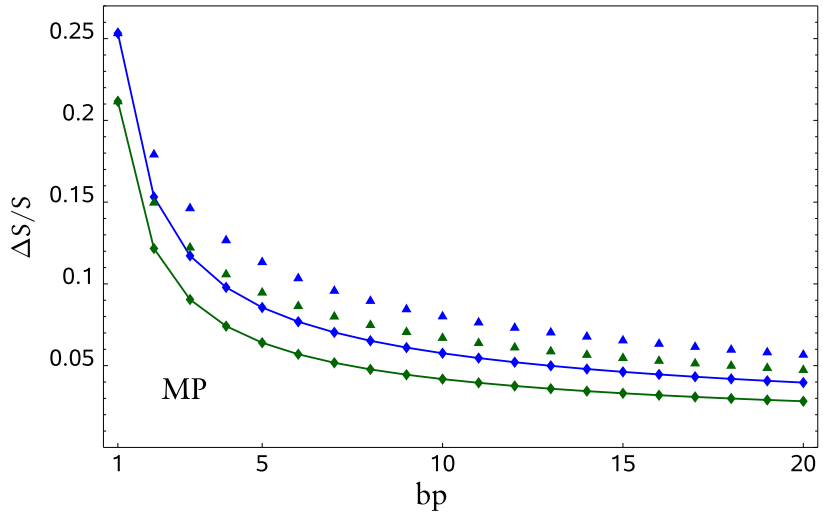

Figure 6.4 | Relative spread $\Delta S/S$ of the bend (lower curve with diamonds, green) and twist (upper curve with diamonds, blue) stiffness coefficients vs. compound step length. Ignoring sequence continuity by setting $V_\times = 0$ leads to overestimation of the stiffness variability (bend, lower green triangles; twist, upper blue triangles).

imaging of random fragments. There is good agreement in the observation that structural disorder contributes only a small correction to the total conformational statistics. The fact of sequence continuity reduces structural variability.

On short scales below a full double-helical turn, a homogeneous wlc model does not capture all features of a random rbc. Notably, the bend angle distributions of a random ensemble may have considerably bigger tails than the assumption of a Gaussian that is made in one particular incarnation of the wlc model, see chapter 7.

The variability of torsional and bending rigidities in a random ensemble of short chains reaches noticeable levels below one full turn of the double helix. Disregarding sequence continuity would lead to an overestimation of stiffness variability, similar to the structural variability.

In view of an experimental precision approaching one percent for the mesoscopic bending rigidity [Vol02], a quantitatively correct relation between mesoscopic and microscopic stiffness parameters is needed. The method [Bec07] presented in the last two chapters provides part of this link, bridging at least the gap between the base–pair scale of 1/3 nm and the scale 50nm of a persistence length of DNA.

7 Random walks on the rigid motion group

In this chapter, the continuous limit of the rigid base–pair chain is investigated. Motivated by the description of the worm–like chain in terms of a diffusion process, we construct a continuous rigid body chain as a diffusion process on the group of rigid body transformations and calculate some interesting moments of its transition function. In intrinsically superhelical DNA, these show feature that are not captured by the corresponding intrinsically straight worm–like chain.

7.1 Continuous models for DNA

DNA has a natural discrete structure in terms of its base–pairs. However when the length scale of interest is much bigger than the discretization length, a continuous model is much more appropriate: It allows a description of the molecule's shape in terms of differential equations, which is almost a prerequisite for analytical results.

In chapter 5, it was shown how to average over the helical geometry of a rbc, arriving at the elastic properties of one segment of a discrete version of a chiral, extensible wlc. Averaging also over the sequence irregularity gave a way to extend this mapping to a random sequence rbc (chap. 6). By its construction, the resulting wlc is a description valid for length scales above one helical repeat. The coarse–graining followed the order rbc → discrete wlc → wlc.

In the following, a more direct way, rbc → continuous rigid body chain (crbc) of formulating a continuous limit of the rbc will be considered, without averaging on the intermediate scale of a helical repeat. The resulting crbc does not have a chiral symmetry on short scales of a few bp. The intermediate scale where crbc and wlc may differ, is set by the helical axis offset. This regime can extend up to hundreds of bp in the case of repetitive, intrinsically superhelical DNA. On long scales, the crbc approaches the wlc, as shown below.

7.2 The worm–like chain limit

To illustrate the relation between discrete and continuous polymer models, we start by a brief consideration of the limit for the well–known worm–like chain

(Kra49, see also Yam97, Rub03). For computational simplicity, we restrict the discussion to two space dimensions.

7.2.1 Discrete versions of the worm–like chain model

A generic discrete polymer chain with inextensible contour can be defined as a sequence of beads i joined by link vectors $p_{i\,i+1}$ with constant length l_0 and summed bond angles $\theta_{k\,l} \in (-\pi, \pi)$, such that $\cos\theta_{k\,l} = l_0^{-2}\, p_{k\,k+1} \cdot p_{l\,l+1}$. The thermally fluctuating individual $\theta_{i\,i+1}$ are modeled as independent random variables, identically distributed and symmetric around $\langle\theta_{i\,i+1}\rangle = 0$.

The projection of the end–to–end vector of an n-link chain on the direction of its first link is

$$R_{\shortparallel}(n) = l_0^{-1} p_{01} \cdot \sum_{i=0}^{n-1} p_{i\,i+1} = l_0 \sum_{i=0}^{n} \cos\theta_{0i}, \tag{7.1}$$

and the bending persistence length can be defined as the expectation value of $R_{\shortparallel}$ for a long chain, $l_b = \lim_{n\to\infty}\langle R_{\shortparallel}(n)\rangle$.

Observe that $\langle\sin\theta_{i\,i+1}\rangle = 0$ and let $\langle\cos\theta_{i\,i+1}\rangle = c_\theta$. Then, rewriting the cosine of the sum $\theta_{0\,i+1} = \sum_0^i \theta_{j\,j+1}$, one has $\langle\cos\theta_{0\,i+1}\rangle = \langle\cos\theta_{0i}\cos\theta_{i\,i+1}\rangle - \langle\sin\theta_{0i}\sin\theta_{i\,i+1}\rangle = \langle\cos\theta_{0i}\rangle\, c_\theta$. By induction,

$$l_b = l_0 \sum_{i=0}^{\infty} {c_\theta}^i = \frac{l_0}{1 - c_\theta}, \tag{7.2}$$

for *arbitrary* bond angle distribution $p(\theta)$. In the limit of small variance $\nu_\theta = \langle\theta_{i\,i+1}^2\rangle \ll 1$, $c_\theta \to 1 - \nu_\theta/2$ and the persistence length $l_b \to 2l_0/\nu_\theta$. Two of the many choices of bend angle distribution are

1. the two–state chain with two possible values of each bond angle, $p(\theta) \propto \delta(\theta - \nu_\theta^{1/2}) + \delta(\theta + \nu_\theta^{1/2})$

2. the linearly elastic chain with Gaussian distribution $p(\theta) \propto e^{-\frac{\theta^2}{2\nu_\theta}}$.

Clearly, on length scales $\simeq l_0$, the two discrete models are markedly different. Only after many links $n \simeq n_= \gg 1$, the central limit theorem brings the distributions of $R_{\shortparallel}(n)$ into agreement, leading to the same persistence length.

The wlc is obtained by setting simultaneously $\nu_\theta \to \alpha\nu_\theta$, $l_0 \to \alpha l_0$ and letting $\alpha \to 0$. In this way, the ratio l_b stays constant in the limit. We define a continuous chemical distance $s = \int ds$ by setting ds so that the contour length $l = l_0 s$. Thus s

is the contour length along the chain, measured in units of the original link length l_0. In the wlc limit, the chemical distance $s_=$ at which different refined discrete models agree, tends to zero: $s_= = n_= \cdot (\alpha l_0)/l_0 = n_= \alpha \to 0$.

In other words, starting from a discrete chain, the corresponding wlc is obtained by a limiting procedure which guarantees agreement on *long* scales. But since interactions are purely local, the condition (7.2) for long–scale agreement is given in terms of the local quantities c_θ and l_0.

7.2.2 The wlc as a diffusion process

To clarify the mathematical structure of the wlc, it is worthwhile to consider the limiting process in some more detail.

When letting $\alpha v_\theta \to 0$, $\alpha l_0 \to 0$, the number of independent bond angles increases, but this increased variability is compensated by their more and more narrow distribution. The typical bond angle fluctuation decreases in size with the square root, $(\alpha v_\theta)^{1/2} \propto \alpha^{1/2}$. Therefore in the wlc limit, the tangent direction $\Theta(s) = \lim_{\mathrm{wlc}} \theta_{0\,i}$ (where $i = [s/\alpha]$), becomes a continuous function of s. In contrast, difference quotients of the tangent direction are of size $(\alpha v_\theta)^{1/2}/(\alpha l_0) \propto \alpha^{-1/2}$ and diverge in the limit: $\Theta(s)$ is nowhere differentiable.

Continuous sample paths with independent increments and linearly growing variance are a well–known characteristic property of Brownian motion. The integrated bond angle of the wlc is thus nothing but a Wiener process defined by

$$\tilde{\Theta}(s) = \int_0^s d\tilde{\Theta}(s') = \int_0^s (2l_0/l_b)^{\frac{1}{2}} dW(s'). \tag{7.3}$$

Here, dW is standard Gaussian white noise with $\langle dW(s) dW(s') \rangle = \delta(s - s')ds$. The prefactor gives the angular diffusion constant and is chosen such that (7.2) comes out right: $v_\theta = \langle \tilde{\Theta}(1)^2 \rangle = 2l_0/l_b$. The integral on the right hand side (rhs) of (7.3) is to be understood as an Itō stochastic integral.

The Langevin–like equation (7.3) also suggests an extension of the continuous model to include non-random intrinsic deformations. Adding a deterministic term $\theta_0 ds$ to the rhs results in a diffusion with a drift which corresponds to nonzero mean curvature of the chain.

The spatial conformation of the wlc can then be obtained by one further integration,

$$p(s) = \int_0^s \begin{pmatrix} \sin \tilde{\Theta}(s') \\ \cos \tilde{\Theta}(s') \end{pmatrix} ds', \tag{7.4}$$

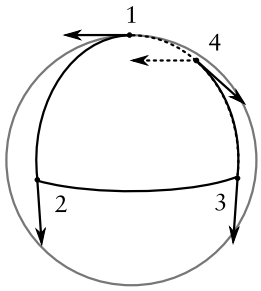

Figure 7.1 | Parallel transport on $\mathbb{S}^2$ is path dependent: The vector parallel transported along the points $1 \to 2 \to 3 \to 4$ is rotated with respect to the same vector, parallel transported along $1 \to 4$.

assuming that the initial tangent pointed in $\binom{0}{1}$-direction. I.e, the wlc contour is once continuously differentiable. Note that the integrated bond angle $\tilde{\Theta}$ can attain arbitrary real values, which can be repaired by an additional modulo operation $\Theta = \tilde{\Theta} \mod (-\pi, \pi)$.

The wlc in two dimensions, defined as a special continuum limit of a class of discrete models, is an integral over a Brownian motion on the unit circle $\mathbb{S}^1$ whose noise strength determines the large scale statistics of the chain.

7.2.3 Diffusion on the sphere

Consider now a wlc in three spatial dimensions. The unit tangent *vector* Θ lives on the unit sphere $\Theta \in \mathbb{S}^2$. Increments $\mathrm{d}\tilde{\Theta}$ of tangent vectors are elements of the respective tangent space $\mathsf{T}_\Theta S^2 \simeq \mathbb{R}^2$.

In this setting, the integral over the components of $\mathrm{d}\tilde{\Theta}$ is more difficult to interpret. The reason is that unlike the $\mathbb{S}^1$ case, there is no global way to identify tangent spaces based at different points with each other: As illustrated in fig. 7.1, the natural way to identify different tangent spaces, parallel transport of tangent vectors, depends on the chosen path! The appropriate path here is the sample path $\Theta(s)$ itself. We will not pursue this approach leading to the rather technical stochastic calculus on manifolds, see e.g. [Elw82, Eme90].

On the other hand, the Fokker–Planck equation on $\mathbb{S}^2$ corresponding to the 3-d wlc [Her52], is well known. In particular, its analogy to the Schrödinger equation of a quantum mechanical top [Sai67] has been exploited extensively. It made possible the use of tools from the quantum theory of angular momentum, see e.g. [Yam97]. Recently, this has led to exact continued fraction expansions for the

Laplace–transformed end–to–end vector distribution of the wlc [Spa04].

7.3 Continuum limit of the rigid base–pair chain

Analogous to the description of the worm–like chain as a diffusion process, we will formulate an Langevin equation for the rbc model. Although this approach appears less fruitful than the solution of the Fokker–Planck equation at first sight, the fact that the configuration space of the model is identical to the transformation group which acts on it (*SE*), actually *simplifies* the Langevin description compared to the wlc in three dimensions.[1]

This description will result in a diffusion process with values on the Lie group *SE*. The study of diffusion processes on Lie groups was originally motivated by rotational Brownian motion of particles in a thermal bath [Per28, McK60], and has been extended to matrix [Ibe76, Kar82] and to general [HD86] Lie groups. Quite generally, a continuous stochastic process on a Lie group G can be obtained as a stochastic integral over some driving process with values in the associated Lie algebra $\mathfrak{g}$. The intuitive picture for this is that random increments in the configuration $g \in G$ of the diffusing particle are 'small' group operations, parametrized by the infinitesimal generators of the group which are elements of $\mathfrak{g}$. The fact that the random increments are *multiplicative* in nature will lead to processes with multiplicative noise.

7.3.1 Choice of step coordinates

Let's take another look at the discrete rbc. If a step has a mean (or center) conformation g_0 and random deformations away from the center, it can be represented as

$$g = g_0 \exp(\tilde{\xi}^i X_i) = g_0(e + \tilde{\xi}^i X_i + O(\tilde{\xi})^2), \tag{7.5}$$

where $\langle \tilde{\xi} \rangle = 0$, as done throughout in chapter 5. For taking the continuum limit, we switch to a more symmetric formulation in terms of exponential coordinates:

$$g = \exp\big((\xi_0^i + \delta\xi^i) X_i\big). \tag{7.6}$$

Here, ξ_0 is the mean value $\xi_0 = \langle \log g \rangle$, and $\delta\xi$ represents the fluctuations, in terms of exponential coordinates. As detailed in sec. 4.2.8 and app. A.4, to first order in

[1] The same is true for inextensible and unshearable rods, which can be described by a diffusion on the rotation group $SO(3)$.

the deformations, both representations of a fluctuating step are related in a simple way: $g_0 = \exp(\xi_0^i X_i)$ and $\tilde{\xi}^i = \Omega^i{}_j \delta\xi^j$ where the matrix $\Omega = (f_1(-\operatorname{ad}\xi_0))$.

7.3.2 Diffusion on the Lie algebra

Recall that a chain of bp frames can be written as a product of homogeneous matrices. Using exponential coordinates for each step,

$$g_{0n} = \exp(\xi_{01}) \exp(\xi_{12}) \cdots \exp(\xi_{n-1\,n}), \tag{7.7}$$

where we have used the shorthand notation $\exp\xi = \exp(\xi^i X_i)$, and $\xi_{l\,l+1} = \xi_0 + \delta\xi_{l\,l+1}$. The $\delta\xi_{l\,l+1}$ denote mutually independent single step deformations. In a rbc with sequence dependent elasticity, ξ_0 and the covariance $C = \langle \delta\xi^i \delta\xi^j \rangle$ of deformations both depend on the step index l.

Observe that the discrete chain (7.7) has the property that the mean of the sum of coordinate vectors is proportional to chain length, $\langle \sum_{l=0}^{n-1} \xi_{l\,l+1} \rangle = n\xi_0$. Also, since fluctuations are independent, the variance of the sum is proportional to chain length, too: $\langle (\sum_{l=0}^{n-1} \delta\xi_{l\,l+1})^2 \rangle \propto n$. This is characteristic of a diffusion process, with drift equal to ξ_0. In fact, we can construct a corresponding continuous 'time' diffusion process $\Xi(s)$ with values in the Lie algebra *se* in a standard way, as a solution to the stochastic differential equation (sde)

$$d\Xi(s) = \xi_0 ds + B dW(s), \quad \Xi(0) = 0. \tag{7.8}$$

The continuous parameter s is the chemical distance, reaching integer values after every completed bp step, and plays the role of time.

When the drift vector ξ_0 and the fluctuation strength matrix B are constants, (7.8) describes a time–invariant diffusion, and the solution is just given by $\Xi(s) = s\xi_0 + BW(s)$, where $(W^i)_{1 \leqslant i \leqslant 6}$ are six independent, standard Wiener processes. Since $\langle dW \rangle = 0$ and $\langle dW^i(s) dW^j(s') \rangle = \delta^{ij} \delta(s - s') ds$, the covariance of Ξ is

$$\langle (BW(s))^i (BW(s))^j \rangle = B^i{}_k B^j{}_l \delta^{kl} \int_0^s \int_0^s \delta(s' - s'') ds' ds'' = s\, B^i{}_k B^j{}_l \delta^{kl}. \tag{7.9}$$

Also, since Wiener increments over disjoint intervals are independent, so are the increments $\Xi(l+1) - \Xi(l)$ for different l. Identifying $\xi_{l\,l+1} = \Xi(l+1) - \Xi(l)$, the sum of step conformation vectors $\sum_{l=0}^{n-1} \xi_{l\,l+1}$ has been *interpolated* by the

continuous process $\Xi(s)$. If the noise strength B satisfies

$$B^i{}_k B^j{}_l \delta^{kl} = C^{ij} = \langle \delta\xi^i_{l\,l+1} \delta\xi^j_{l\,l+1} \rangle, \tag{7.10}$$

then the interpolation Ξ *exactly* reproduces the discrete statistics of $\sum_0^{n-1} \xi_{l\,l+1}$ at integer values $s = n$.[2]

What changes when the step parameters ξ_0 and C are sequence–dependent? Looking at sde (7.8), the mean value over one step is $\int_l^{l+1} \xi_0(s)ds$ and can be matched to $\langle \xi_{l\,l+1} \rangle$. To match the sequence–dependent fluctuation strength, the condition is now that $\int_l^{l+1} B^i{}_k(s) B^i{}_k(s) \delta^{kl} ds = \langle \delta\xi^i_{l\,l+1} \delta\xi^j_{l\,l+1} \rangle$. A possible choice to fulfill these matching conditions is just to choose $B(s)$ and $\xi_0(s)$ to coincide with the discrete values on the interval of each original discrete step.

The reader may worry about the structure of the expressions B and ξ_0 in terms of units. After all, rotations are dimensionless while translations carry a dimension of length. A brief discussion is given in appendix A.9.

7.3.3 Diffusion on the group space

The main idea is now to lift this interpolation from *se* to the group *SE*. This is done by using the continuous process $\Xi(s)$ on the algebra to drive a diffusion on the Lie group. The result is a continuous diffusion process on the group. This interpolation on the *group* is no longer exact at the discrete 'time' intervals $s = l$.

One can think of g_{0n} (7.7) as a discrete process on the group, generated the following procedure: After each 'time' lag of $\Delta s = 1$, take a snapshot of the process $\Xi(s)$ and then multiply the exponential $\exp \Delta\Xi$ of the finite increment $\Delta\Xi(s) = \Xi(s + \Delta s) - \Xi(s)$ on the right. The result after n steps is

$$g_{0n} = g(n) = \exp(\Delta\Xi(0)) \exp(\Delta\Xi(\Delta s)) \cdots \exp(\Delta\Xi(n - \Delta s)). \tag{7.11}$$

Then, rewriting (7.7), for integer values of s,

$$g(s + \Delta s) - g(s) = g(s)(\exp(\Delta\Xi(s)) - e) = g(s) \left[\exp(\Xi(\cdot) - \Xi_c)\right]_s^{s+\Delta s} \tag{7.12}$$

where $\Xi_c = \Xi(s)$ is a constant offset: the process $\Xi - \Xi_c$ is a shifted version of Ξ which has the value 0 at s.

To approach a continuous limit, we can now choose smaller steps Δs, creating

[2] Note that $C = BB^T$ defined here is slightly different from the covariance matrix of left invariant increments used in chapter 5, there also denoted C.

finer subdivisions of the driving process ξ. In the limit $\Delta s \to 0$, the result is the sde

$$dg(s) = g(s)\,d\exp(\Xi(s) - \Xi_c)\big|_{\Xi_c=\Xi(s)}. \tag{7.13}$$

This sde has multiplicative noise, so here the question of Itō vs. Stratonovich interpretation does matter. Looking again at the discrete version (7.12), the integrand $g(s)$ is evaluated at the *beginning* of the interval, so the limit (7.13) is an Itō sde.

We use Itō's lemma to expand the differential $d\exp$. First note that the shift Ξ_c is trivial: $d(\Xi(s) - \Xi_c)\big|_{\Xi_c=\Xi(s)} = d\Xi(s)$. Then, we need to expand the exponential to second order around 0, substituting $dW^i dW^j = \delta^{ij} ds$ and $ds^2 = dW ds = 0$. The result is

$$\begin{aligned} d\exp((\Xi(s) - \Xi_c)^i X_i)\big|_{\Xi_c=\Xi(s)} &= X_i d\Xi^i(s) + \tfrac{1}{2}(B^i{}_j X_i dW^j(s))^2 \\ &= (\xi_0^i X_i + \tfrac{1}{2} C^{ij} X_i X_j) ds + B^i{}_j X_i dW^j(s), \end{aligned} \tag{7.14}$$

which can be plugged into (7.13) to get an explicit Itō sde:

$$dg(s) = g(s)\big((\xi_0^i X_i + \tfrac{1}{2} C^{ij} X_i X_j) ds + (B^i{}_j X_i dW^j(s)\big), \quad g(0) = e. \tag{7.15}$$

This is somewhat counter-intuitive, since it says that the right increment of the diffusion process on the group cannot be written purely in terms of the X_i, i.e is not an element of the Lie algebra! However, if we transform (7.15) into an equivalent Stratonovich equation (see e.g. [Ris89]), the extra drift term in (7.14) drops out again; one obtains

$$dg(s) = g(s) \circ d\xi^i X_i = g(s)(\xi_0 ds + B \circ dW(s))^i X_i, \quad g(0) = e, \tag{7.16}$$

where the standard notation $\circ d$ now indicates a Stratonovich differential. It is satisfying that when using the Stratonovich formulation which has the usual rules of variable transformation, the increment dg manifestly lies in the tangent space again.

The continuum limit constructed step by step above, is in fact a known rigorous mathematical result. The *stochastic exponential* of a continuous semimartingale on some Lie algebra (here, Ξ) is defined as the unique solution to (7.16) in the corresponding Lie group (here $g(s)$) [HD86]. It is a generalization of the usual path–ordered exponential to integrands which are stochastic processes. On the other hand, the *multiplicative integral* [McK60, Ibe76] of Ξ is defined as the continuum limit $\Delta s \to 0$ of (7.7). It has been shown [Ibe76, HD86] that both

notions agree, which is just the content of our limit construction above.

As mentioned, unlike the diffusion on the algebra, the correspondence between the original, discrete model and the continuum limit is not exact at integer values of s. The reason for that is the non-commutativity of the group. Indeed, note that if all noise terms commuted, then one could rewrite (7.7) simply as

$$g_{0n} = \exp(\xi_{01}) \cdots \exp(\xi_{n-1\,n}) = \exp(\Xi(n)), \tag{7.17}$$

so $\exp \Xi(s)$ would solve a commuting version of (7.16) and coincide with the original chain at integer s. However in the general case, $\exp(\Xi(1)) \neq g(1)$ since the lhs is an unordered exponential, while the rhs is path–ordered. Their difference originates from the non-commutativity of the random increments at different 'times'.

In summary, the process described by (7.16) is the continuum limit of the discrete rbc model, to be called a crbc. It has six continuous degrees of freedom, three linear (v) and three angular (ω) 'velocities' whose $\delta(s - s')$-correlated fluctuations around the equilibrium value $\xi_0 = (\omega_0, v_0)$ produce conformational fluctuations of the molecule.

A discrete rbc with ξ_0, C converges to the continuous rbc with the same parameters in the limit $\Delta s \to 0$. However, this limit is generally not the best–matching continuous description of the original chain because of the non-commutativity of noise terms. One can expect that the best match will have renormalized parameters. Their calculation is an interesting open problem. Similar to the motivating example of the wlc, it is unimportant whether the original, discrete distributions $p(\delta\xi)$ are Gaussians. A matching diffusion $\Xi(s)$ can always be constructed as long as the second moment of $p(\delta\xi)$ exists.

Finally, it is worth mentioning that in contrast to the wlc, the path $g(s)$ is continuous but *not continuously differentiable*, as can be seen by noting that the increment in (7.15) is δ-correlated.

7.3.4 Generator and Fokker–Planck equation

Now that the continuous crbc model is defined, we write down the corresponding Fokker–Planck equation for completeness, however not making further attempts at its solution.

The generator $\mathbf{L}$ of a diffusion process is a second–order differential operator.

Applied to a function f, it returns the initial change in expectation value g:

$$(\mathbf{L}f)(g) = \partial_s \left\langle f(g(s)) \,\middle|\, g(s) = g \right\rangle . \tag{7.18}$$

The generator can be read off from the drift term in 7.15; in terms of the left invariant basis vector fields, $\mathbf{L} = \xi_0^i L_i + \frac{1}{2} C^{ij} L_i L_j$, see also [Ibe76]. Whenever the coefficients ξ_0, C in (7.16) are constants, $\mathbf{L}$ is left invariant and the process $g(s)$ is a left invariant diffusion.

We denote by $p(g, s|g', s')$ the normalized transition probability density function to observe $g(s) = g$ when starting at $g(s') = g'$.[3] It has the usual properties $p(g, s|g', s') = \int p(g, s|g'', s'')p(g'', s''|g', s')dg''$ for intermediate 'times' s'' and $\lim_{s' \uparrow s} p(g, s|g', s') = \delta(g'^{-1}g)$.[4] Let f be an arbitrary function with compact support in *SE*. Composing conditional probabilities, the change in expectation when starting at an earlier 'time' s' is

$$\partial_s \left\langle f(g(s))|g(s') = g' \right\rangle = \left\langle (\mathbf{L}f)(g(s))|g(s') = g' \right\rangle = \int_{SE} p(g, s|g', s')\mathbf{L}f(g)dg. \tag{7.19}$$

Note that the lhs can be rewritten as $\int \partial_s p(g, s|g', s')f(g)dg$. Integrating the rhs by parts and using the fact that f is arbitrary, the transition pdf solves the partial differential equation (pde),

$$\partial_s p(g, s|g', s') = \mathbf{L}^\dagger p(g, s|g', s'), \tag{7.20}$$

which is the Fokker–Planck equation of the continuous chain. Here the Fokker–Planck operator $\mathbf{L}^\dagger = -\xi_0^i L_i + \frac{1}{2} C^{ij} L_i L_j$ acts on the 'unprimed' g-dependence. It is the adjoint of $\mathbf{L}$. The expression for $\mathbf{L}^\dagger$ is unchanged in the sequence–dependent case, since then ξ_0 and C are functions of s but not of g. One sees that the Fokker–Planck equation is left invariant, corresponding to the fact that random increments are naturally given in the local material frame $g(s)$ of the chain.

7.4 Moments as solutions to ordinary differential equations

Some interesting moments of the crbc transition probability $p(g, s|e, 0)$ can be calculated directly from the Langevin equations (7.15,7.16), without invoking any advanced machinery for solving the Fokker–Planck pde (7.20), such as harmonic

[3] $p(g, s|g', s')$ is variously known as heat kernel, propagator, or fundamental solution.
[4] $\int_{SE} \delta(g)f(g)dg = f(e)$ defines the δ-distribution on the group.

analysis on the group [Chi00, Chi01]. The basic idea is to just to take the expectation of the matrix sde governing the quantity of interest. This approach was used previously for inextensible, unshearable rods [Pan00].

We consider three quantities in detail: The mean rotation matrix, the mean end–to–end vector, and the mean squared end–to–end distance. All three can be used for defining the bending persistence length of the chain, and the three definitions give the same result for a wlc. As shown below, in the crbc, the three definitions are mutually different; they agree only in appropriate limits.

7.4.1 Mean end–to–end rotation

The end–to–end transformation of a rbc is the matrix $g(s) = \begin{bmatrix} R(s) & p(s) \\ 0 & 1 \end{bmatrix}$. The s-dependence of the expectation value of this matrix gives insight into the statistical properties of the chain. The mean rotation matrix is nothing but the matrix of direction cosine correlators

$$\left\langle R^i{}_j(s) \right\rangle = \left\langle e_j(s) \cdot e_i(0) \right\rangle, \tag{7.21}$$

which contains information on directional persistence of all rotational deformation modes along the chain.

We write down the ordinary differential equation (ode) solved by[5] $\langle g(s) \rangle = \langle g(s)|e, 0\rangle$. Taking the expectation of 7.15,

$$d\langle g(s)\rangle = \langle g(s)\rangle \, (\xi_0^i + \tfrac{1}{2} C^{ij} X_j) X_i ds, \tag{7.22}$$

$$\langle g(0)\rangle = e \tag{7.23}$$

where we used the essential fact that the Itō differential is independent,

$$\left\langle g(s) dW^i(s) \right\rangle = \langle g(s)\rangle \left\langle dW^i(s) \right\rangle = 0. \tag{7.24}$$

Note that the matrix $\langle g(s)\rangle$ is not in *SE* anymore, since the rotation part $\langle R\rangle$ is not orthogonal! A simple example for this effect is illustrated in fig. 7.2.

Correspondingly, the right increment of $\langle g(s)\rangle$ in (7.22) is not in *se*.[6] In block

[5] Here and in the following, the notation of initial condition in the angular brackets of the expectation value is suppressed.

[6] Eqn. (7.22) lives in the embedding matrix space *Aff*(3) of affine transformations of $\mathbb{R}^3$.

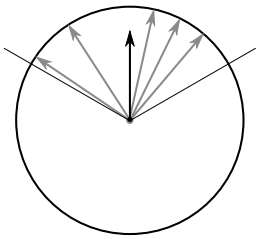

Figure 7.2 | The average of a fluctuating unit vector is shortened: $\langle\cos\theta\rangle^2 + \langle\sin\theta\rangle^2 \leqslant 1$. Therefore, it does not lie on the unit circle anymore!

form it can be written as

$$\xi_0^i X_i + \tfrac{1}{2} C^{ij} X_i X_j = M = \begin{bmatrix} M_\omega & m_\nu \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \widehat{\omega}_0 + \widetilde{\omega} & \nu_0 + \widetilde{\nu} \\ 0 & 0 \end{bmatrix}, \tag{7.25}$$

where, summing only over $1 \leqslant i, j \leqslant 3$, $\widetilde{\nu} = \frac{1}{2} C^{ij+3} \epsilon_i d_j$, and $\widetilde{\omega} = \frac{1}{2} C^{ij} \epsilon_i \epsilon_j$. The vector $\tilde{\nu}$ results from the cross–product of translational and rotational fluctuations, and the matrix $\tilde{\omega}$ is symmetric and negative definite[7], see also app. A.10.

Since (7.22) is a linear ode, its solutions can be written in terms of a matrix exponential, which has to be path–ordered in the case of s-dependent coefficients. To evaluate it explicitly, we split up (7.22) into its blocks to find

$$\partial_s \langle p(s)\rangle = \langle R(s)\rangle\, m_\nu, \tag{7.26}$$

$$\partial_s \langle R(s)\rangle = \langle R(s)\rangle\, M_\omega, \tag{7.27}$$

with the solutions $\langle R(s)\rangle = \overleftarrow{\exp} \int_0^s M_\omega(s')ds'$ and $\langle p(s)\rangle = \int_0^s \langle R(s')\rangle\, m_\nu(s')ds'$, for the initial condition (7.23).[8]

One way to define the bending persistence length in the wlc model is to take the decay length of bending correlations along the chain. How can this be done in the crbc? Let's investigate 7.27 in the case of constant coefficients in some more detail. An intuitive definition of the correlator of bending is the projection of $\langle R(s)\rangle$ on the local helical axis direction: $c_b(s) = \frac{\omega_0^T}{\|\omega_0\|} \langle R(s)\rangle \frac{\omega_0}{\|\omega_0\|}$. Looking at (7.27), the right increment M_ω of $\langle g\rangle$ has an antisymmetric part $\widehat{\omega}_0$ and a negative definite symmetric part $\widetilde{\omega}$. Together they lead to exponentially damped oscillations. The problem with the correlator $c_b(s)$ is that it is *not* an exponentially decaying

[7] In the marginal case of constrained rotations where $C^{(\omega\omega)}$ has only rank 1, $\tilde{\omega}$ is only semidefinite.

[8] The path ordering in $\overleftarrow{\exp}$ is for increasing s from left to right.

function; it still shows oscillations. This can be overcome by considering instead of c_b the correlator c_{no} of the non-oscillatory direction, $c_{no}(s) = \omega_{no}^T \langle R(s) \rangle \omega_{no}$. Here ω_{no} is defined as the unit eigenvector of M_ω with *real* eigenvalue $-1/s_{no} < 0$. For moderate noise strength, there exists exactly one such eigenvector. With this definition, from (7.27) immediately the exponential decay rule $\dot{c}_{no} = -c_{no}/s_{no}$ follows. We have found an exponentially decaying correlator. s_{no} can now be identified as the bending persistence length of the chain, given in bp units. To get an actual length, we scale with the mean helical rise $l_0 = \frac{\omega_0^T \nu_0}{\|\omega_0\|}$, so that $l_{no} = l_0 s_{no}$.

How does the decay of c_{no} compare with the on-axis bending persistence length l_b obtained in chapter 5 by a mapping of the rbc to the wlc? There, the on–axis version of the covariance was denoted $C_{\shortparallel}$, eqn. (6.7). Its $(\omega_{\shortparallel}^{1,2})$ submatrix gives the rotational fluctuations in the subspace orthogonal to ω_0. After helical phase angle averaging (sec. 5.3.4) around the ω_0 axis, $2/(C_{\shortparallel}^{11} + C_{\shortparallel}^{22}) = s_b$ gives the bending persistence length of the chain in bp units. Although this is *not* the same as s_{no}, the two definitions agree whenever ω_0 coincides with the non-oscillatory eigenvector of M_ω. One can check that in that case, $s_{no} = -\frac{\|\omega_0\|^2}{\omega_0^T \tilde{\omega} \omega_0} = s_b$. On the other hand, helical phase averaging of the covariance matrix by rotating around ω_0 automatically makes ω_0 a real eigenvector of $\tilde{\omega}$! Therefore, for isotropic bending chains, the relation $s_{no} = s_b$ (or $l_{no} = l_b$) is exact.

In conclusion, for all practical purposes in DNA, it is safe to use l_b as the bending persistence length. This is so because on one hand, the thermal fluctuations of a bps are much smaller than the equilibrium conformation of the step, so that $M_\omega = \hat{\omega}_0$+small perturbations. On the other hand on scales above a helical repeat, DNA has essentially isotropic bending.[9]

7.4.2 Mean end–to–end vector

The mean end–to–end vector $\langle p(s) \rangle$ provides an alternative way to characterize directional persistence; in section 7.2.1 the persistence length of a wlc was defined as the projection of $\langle p(s) \rangle$ on the initial direction, in the long chain limit. We will refer to this definition as the projective persistence length in this section, denoted by l_{proj}.

The solution of (7.26), for constant coefficients, reduces to

$$\langle p(s) \rangle = \int_0^s \exp(s' M_\omega) ds' m_\nu = s\, f_1(s M_\omega) m_\nu, \tag{7.28}$$

[9] with the exception of intrinsically bent sequences

where the function f_1 is defined in A.3. From its series form $f_1(z) = 1 + \frac{1}{2}z + \cdots$ one can see that the initial growth of $\langle p(s)\rangle$ is linear in s with velocity $m_\nu = \nu_0 + \tilde{\nu}$. The extra initial velocity $\tilde{\nu}$ means that coupling fluctuations influence the mean shape of the chain also for small distances.

Consider the long–chain limit $s \to \infty$ of eqn. (7.28). Clearly, for convergence, the matrix $\exp(sM_\omega)$ should show exponential decay rather than growth. This is ensured by the negative definiteness of $\tilde{\omega}$. We can directly evaluate the limit by using the formally integrated expression

$$\langle p(\infty)\rangle = \lim_{s\to\infty} \langle p(s)\rangle = \lim_{s\to\infty} \frac{\exp(sM_\omega) - e}{M_\omega} m_\nu = -M_\omega^{-1} m_\nu. \tag{7.29}$$

In between its finite limits 0 and (7.29), the mean end–to–end vector traces out a path that has the generic shape of a 'helical logarithmic spiral', resembling a regular helical shape in the beginning but then spiraling into its limiting point. This is illustrated in fig. 7.3 for arbitrarily chosen values of the mean deformation and covariance. (Cf. a similar plot in [Yam97, chapter 4] for the unshearable inextensible case.)

It turns out that there exists a critical fluctuation strength above which all remainder of a helical oscillation is extinguished. Reconsider the eigenvalues of the non-symmetric matrix $M_\omega = \hat{\omega}_0 + \tilde{\omega}$ in different limits. Without fluctuations ($\tilde{\omega} = 0$), the eigenvalues $0, \pm i\|\omega_0\|$ lead to pure oscillatory behavior in the plane normal to ω_0. In the opposite limit of strong fluctuations ($\hat{\omega} = 0$), M_ω has three real negative eigenvalues. In between, there exists a *finite* threshold fluctuation strength at which two eigenvalues just leave the negative real axis.[10] Below this fluctuation strength, all helical structure of the chain is 'forgotten'. This feature has been discussed for unshearable and inextensible rods in [Pan00].

Looking at the generic helical shape of the spiraling paths, it is clear that the projective persistence length of the wlc cannot correspond to the projection of $\langle p(s)\rangle$ on the initial tangent direction. Instead, as considered to great lengths in chapter 5, one needs to project on the initial direction of the mean helical *centerline*. Again, it is better to choose the initial non-oscillatory unit eigenvector ω_{no} instead of ω_0. We define the projective persistence length of the crbc as

$$l_{pro} = \omega_{no}{}^{\mathsf{T}} \langle p(\infty)\rangle = s_{no}\omega_{no}^{\mathsf{T}}(\nu_0 + \tilde{\nu}). \tag{7.30}$$

[10] If $\tilde{\omega}^\perp = P^\perp_{\omega_0}\tilde{\omega}P^\perp_{\omega_0}$ denotes the projection of $\tilde{\omega}$ perpendicular to ω_0, the condition for oscillatory motion in that plane is given by $\mathrm{tr}^2\,\tilde{\omega}^\perp - 4\det\tilde{\omega}^\perp < 4\|\omega_0\|^2$, at which point the eigenvalues of $P^\perp_{\omega_0}M_\omega P^\perp_{\omega_0}$ acquire an imaginary part.

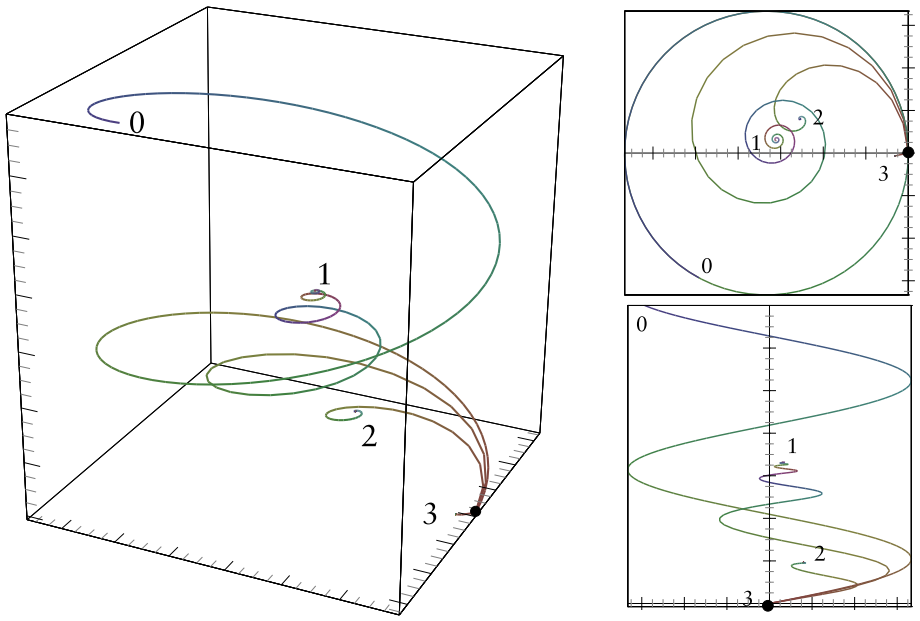

Figure 7.3 | Traces of the mean end–to–end vector $\langle p(s) \rangle$, color coded for chemical distance s. The trace numbers 0,1,2,3 correspond to fluctuation covariances scaled with a prefactor 0,0.1,1,13, respectively. The traces range form a regular helix (0) for switched–off fluctuations to eliminated oscillations in trace (3) . The latter also demonstrates that the initial tangent differs from v_0 in general.

Interestingly, this does not give the same result as the persistence length of bending correlations: $l_{pro} \neq l_{no}$. The reason is that the coupling of translational and rotational fluctuations adds an extra term $\tilde{v}$ which is absent in the pure rotational decay length l_{no}.

Does this make a difference for DNA? The conditions that led to $l_b = l_{no}$ in the previous section, have the same effect here: In the limit of small fluctuations, $\omega_{no} \to \frac{\omega_0}{\|\omega_0\|}$ and also $\tilde{v} \ll v_0$ so that $l_{pro} \to s_b \frac{\omega_0^T v_0}{\|\omega_0\|} = l_b$. Moreover, when the covariance is averaged over helical phase only twist–stretch couplings survive which means that $\tilde{v} \to 0$ for isotropic bending. So also in the isotropic bending case $l_{pro} = l_b$ is exact. For DNA on scales of a helical repeat length, both conditions are fulfilled, so here $l_{pro} = l_b$ is a good approximation.

7.4.3 Mean squared end–to–end vector

Another interesting moment of the crbc transition pdf $p(g, s|e, 0)$ is the mean squared end–to–end distance $\langle p^2(s)\rangle = \langle p^T(s)p(s)\rangle$. The chain has no long–range correlations. In the limit of long chains, it will therefore approach a Gaussian behavior, so that $\langle p^2(s)\rangle$ grows linearly in s. The prefactor is an effective diffusion constant in 3-d space, resulting from both drift and diffusion on *SE*.

In the wlc model, this diffusion constant equals $2l_0 l_b$. Therefore setting $l_{diff} = \frac{1}{2}\lim_{s\to\infty}\frac{1}{l_0 s}\langle p^2(s)\rangle$, one gets yet another definition of persistence length, which is equivalent to l_b in the wlc case.

Relating this to the crbc, the monomer length is $l_0 = \frac{\omega_0{}^T v_0}{\|\omega_0\|}$. To get a handle on $\langle p^2\rangle$, observe that the matrix

$$g^T g = \begin{bmatrix} I_3 & R^T p \\ p^T R & p^T p + 1 \end{bmatrix} \notin SE \tag{7.31}$$

contains the squared distance in its $4,4$ entry. We write down a Langevin equation for this matrix. From (7.15), using the product rule of Itō calculus,

$$\begin{aligned} d(g^T g) &= dg^T g + g^T dg + dg^T dg = \\ &= \left(M^T g^T g + g^T g M + C^{ij} X_i^T g^T g X_j\right) ds + \left(X_i^T g^T g + g^T g X_i\right) B^i{}_j dW^j(s). \end{aligned} \tag{7.32}$$

The initial condition is as usual, $(g^T g)(0) = e$.

Consider the extra drift term $C^{ij} X_i^T g^T g X_j = M'$. A straightforward calculation, using the algebraic properties of the basis matrices X_i (cf. sec. 4.2.4), and the fact that $R^T R = e$ for all s, gives the block form[11]

$$M' = \begin{bmatrix} -C^{ij}\epsilon_i\epsilon_j & -C^{ij+3}\epsilon_i d_j \\ C^{i+3j} d_i^T \epsilon_j & C^{i+3j+3}\delta_{ij} \end{bmatrix} = \begin{bmatrix} -2\tilde{\omega} & -2\tilde{v} \\ -2\tilde{v}^T & m' \end{bmatrix}; \quad 1 \leqslant i, j \leqslant 3. \tag{7.33}$$

Thus, taking the expectation value,

$$d\langle g^T g\rangle = \left(\langle g^T g\rangle M + M^T \langle g^T g\rangle\right) ds + M' ds, \tag{7.34}$$

which is an inhomogeneous linear ode.

To solve it, note first that the associated homogeneous equation is eqn. (7.34) with M' set to 0. For the initial condition $\langle g^T g\rangle(0) = A$, it has the solution

[11] As a check, one can plug this into (7.32); one then sees that the rotation part $-2\tilde{\omega}$ exactly cancels with $M_\omega^T + M_\omega = 2\tilde{\omega}$ inside the drift term.

$s \mapsto \langle g(s)\rangle^T A \langle g(s)\rangle$, where $\langle g(s)\rangle$ is a solution of the ode (7.22) with (7.23). A particular solution of the inhomogeneous equation starting at 0 is given by $s \mapsto \int_0^s \langle g(s-s')\rangle^T M' \langle g(s-s')\rangle \, ds'$. Combining, we get an explicit formula,[12]

$$\langle g^T g\rangle(s) = \langle g(s)\rangle^T \langle g(s)\rangle + \int_0^s \langle g(s')\rangle^T M' \langle g(s')\rangle \, ds'. \tag{7.35}$$

We can now plug in the explicit block form of $\langle g\rangle$ and calculate the $4,4$ matrix element to extract the mean square displacement $\langle p^T p\rangle$. The result is

$$\langle p^T p\rangle(s) = \langle p\rangle^T \langle p\rangle(s) + \int_0^s -2\langle p(s')\rangle^T \tilde{\omega} \langle p(s')\rangle - 4\langle p(s')\rangle^T \tilde{\nu} + m' \, ds'. \tag{7.36}$$

In this equation, the 'square of the mean value' $\langle p\rangle^T \langle p\rangle$ gives only a constant offset for long chains, whereas the integral term produces a linear increase in mean square displacement. The limiting behavior is

$$\frac{1}{s}\langle p^T p\rangle \overset{s\to\infty}{\longrightarrow} -2\langle p(\infty)\rangle^T \tilde{\omega} \langle p(\infty)\rangle - 4\langle p(\infty)\rangle^T \tilde{\nu} + m', \tag{7.37}$$

which can be further simplified. Plugging in (7.29), and using symmetry properties of $\tilde{\omega}$ and $\hat{\omega}_0$, the effective diffusion constant after some algebra becomes

$$2l_0 l_{diff} = -2(\nu_0 + \tilde{\nu})^T (\hat{\omega}_0 + \tilde{\omega})^{-1} (\nu_0 - \tilde{\nu}) + m'. \tag{7.38}$$

One sees that all blocks of the covariance matrix enter. In particular, in the limit of vanishing rotational diffusion $\tilde{\omega}$ but finite drift ν_0, l_{diff} diverges, since then the helical shape is persistent which leads to ballistic growth. In the opposite limit of strong rotational fluctuations, the first summand vanishes and a pure translational, isotropic diffusion with diffusion constant m' remains, a perfectly sensible result. The fact that m' occurs in the mean square displacement also means that $l_{diff} \neq l_{pro}$ in general, since the projective persistence length does not include any translational diffusion.

Consider the limit of l_{diff} in which the fluctuation terms $\tilde{\omega}, \tilde{\nu}$ are small compared to the static offsets ω_0, ν_0. Since $\hat{\omega}_0$ is singular, $(\hat{\omega}_0 + \tilde{\omega})^{-1}$ will diverge in the limit of no noise, but only on the null space of $\hat{\omega}_0$; The leading behavior is

$$(\hat{\omega}_0 + \tilde{\omega})^{-1} = \frac{\omega_0 \omega_0^T}{\omega_0^T \tilde{\omega} \omega_0} + O\left(\frac{\|\tilde{\omega}\|}{\|\omega_0\|}\right). \tag{7.39}$$

[12] As it stands, this solution is valid only for constant coefficients. It is a technical matter to extend the solution to s-dependent coefficients.

The translational term m' is small in comparison. Using this in (7.38),

$$l_{diff} \to -l_0^{-1} \frac{\nu_0^T \omega_0 \omega_0^T \nu_0}{\omega_0^T \widetilde{\omega} \omega_0} = -\frac{\omega_0^T \nu_0 \|\omega_0\|}{\omega_0^T \widetilde{\omega} \omega_0} = l_b. \tag{7.40}$$

Summarizing, in general the bending persistence length l_b , the projective persistence length l_{pro} and the diffusive persistence length l_{diff} are mutually different quantities. However in the limit where the size of fluctuations per monomer is much smaller than the drift, the pure rotational fluctuations dominate the long–scale statistics of the chain. Then the rotation–translation coupling present in l_{pro} and the translational fluctuations additionally present in l_{diff} are unimportant and $l_b \simeq l_{pro} \simeq l_{diff}$. For DNA, this is a good approximation. The relations among the different definitions of persistence length are summarized in table 7.1.

Table 7.1 | Different persistence length definitions in the crbc model.

	l_b	l_{no}	l_{pro}	l_{diff}
definition	covariance$\perp \omega_0$	non-oscillatory decay	end–to–end vector	end–to–end distance
fluctuation modes	rot $\perp \omega_0$	rot $\perp \omega_{no}$	rot, rot×trans	rot, rot×trans, trans
helical average	$\to l_b$	$\to l_b$	$\to l_b$	$\to l_b + \frac{m'}{2l_0}$
small fluctuations	$\to l_b$	$\to l_b$	$\to l_b$	$\to l_b$

The full s-dependence of the mean squared displacement can be also made more explicit than (7.35). Integrating over (7.28) and simplifying, one obtains

$$\langle p^T p \rangle(s) = 2(\nu_0 + \widetilde{\nu})^T M_\omega^{-1} \big(\exp(sM_\omega) - sM_\omega - I_3\big) M_\omega^{-1} (\nu_0 - \widetilde{\nu}) + s\, m'. \tag{7.41}$$

Comparing this with the well–known wlc result

$$\langle p^2(s) \rangle = 2l_b(\exp(-sl_0/l_b) + sl_0/l_b - 1)l_b, \tag{7.42}$$

one can draw a close analogy by identifying $-M_\omega \leftrightarrow \frac{l_0}{l_b}$ and $(\nu_0 \pm \widetilde{\nu})^T M_\omega^{-1} \leftrightarrow l_b$, and disregarding the translational diffusion term sm' which is absent in the wlc. After inserting the limiting behavior (7.39), the matrix equation (7.41) is seen to approach the scalar equation (7.42).

In fig. 7.4, plots of $\frac{1}{s}\langle p^T p \rangle(s)$ corresponding to those in fig. 7.3 are shown. The

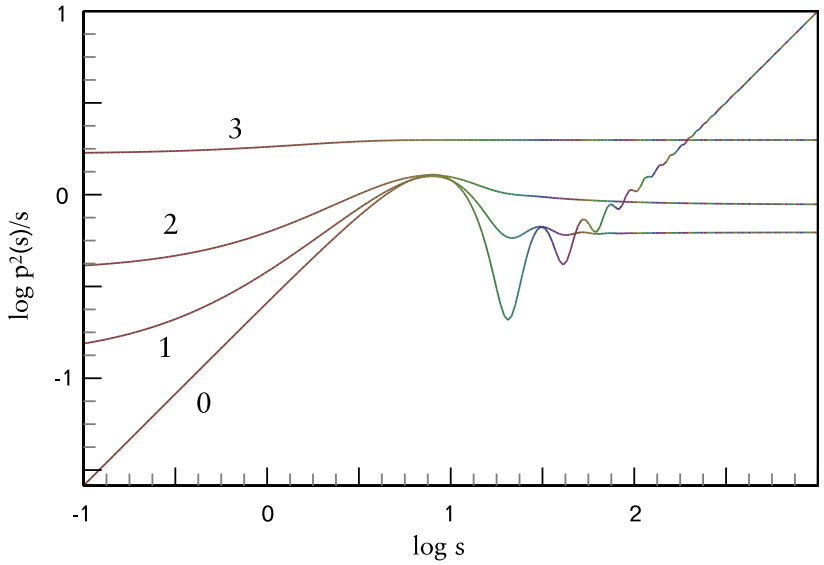

Figure 7.4 | Mean squared distance of the crbc , divided by chemical distance. The parameters used, and the color coding are the same as in fig. 7.3.

curves $1, 2, 3$ correspond to finite fluctuation strength. Their plateaus for small s values give the translational diffusion coefficient m'. The translational diffusion regime is normally not observed in DNA, as it is below the natural discretization of the molecule, but may play a role in different contexts. The plateau values at high s give the effective diffusion coefficient $2l_0 l_{\mathtt{diff}}$.

In contrast, the zero–temperature curve 0 shows ballistic growth in both limits; its shifts in y-direction correspond to the speed $\|v_0\|$ along the helix and to the monomer length l_0, respectively.

The non-monotonic behavior in s of the mean square distance is a consequence of the helical structure and can already be guessed from the traces in fig. 7.3. Curve 3 corresponds to high fluctuation strength above the threshold for oscillations; it is therefore monotonic.

Interestingly, also the effective diffusion coefficient exhibits non-monotonic behavior as a function of the fluctuation strength. From divergence at low noise strength (0) it drops to a minimum and then increases again $(1, 2, 3)$. This can be understood when considering that low bending fluctuations lead to high directional persistence, i.e. to a high diffusion constant, whereas high translation fluctuations also cause a high diffusion constant. Their competition leads to the

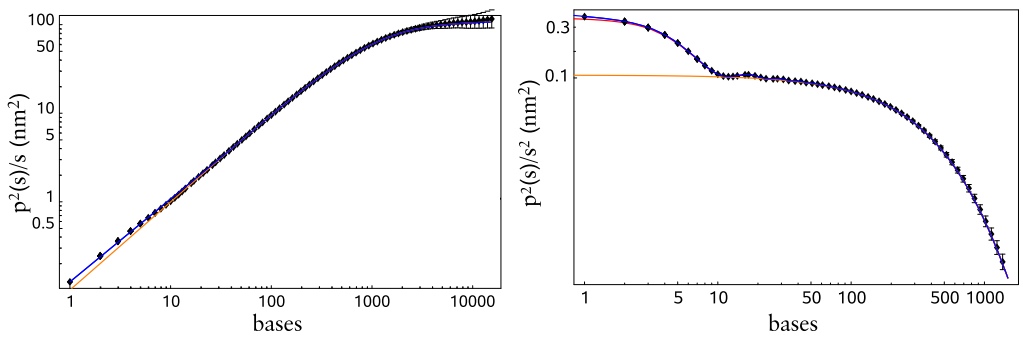

Figure 7.5 | Comparison of wlc (orange) and crbc (blue) predictions to discrete rbc simulation data (symbols). In the left panel, $\langle p^2 \rangle / s$ is shown as in 7.4. The right panel shows $\langle p^2 \rangle / s^2$ and zooms in on short lengths. The red curve corresponds to switched–off translational fluctuations. Static covariance, MD parameter set.

minimum. Its location depends on the relative strength of rotational and translational fluctuations in the covariance matrix C.

7.4.4 Numerical verification

We compare the predictions of the crbc and wlc models for the mean squared distance with a simple–sampling Monte–Carlo simulation of a discrete rbc. Essentially, we repeat the same comparison as the one made in sec. 6.2. To make the differences clear, this time a homogeneous discrete rbc which has only the *static* covariance matrix of the MD parameter set as its covariance is chosen. Thus, the data points shown in fig. 7.5 correspond to the upper row of symbols in the fig. 6.2. The curves are the wlc and crbc predictions for the mean squared displacement, eqns. (7.42) and (7.41), respectively. The parameters are not fitted but directly calculated from the covariance matrix used in the simulation.

In the left panel one sees that the wlc prediction is very good starting form a few tens of bases, while there is a small but significant discrepancy below a helical repeat. This is clearly visible in the rescaled representation in the right panel. The oscillatory behavior of $\langle p^2(s) \rangle$ cannot be captured by the wlc model, but is perfectly reproduced by the crbc prediction. The shoulder below 5 bp is not a result of translational diffusion, which can be seen from the red curve which is a version of (7.41) with $\tilde{\nu}$ and m' set to 0. Rather, it results form the mean helical geometry of the MD parameter set, which has a comparatively high axis offset.

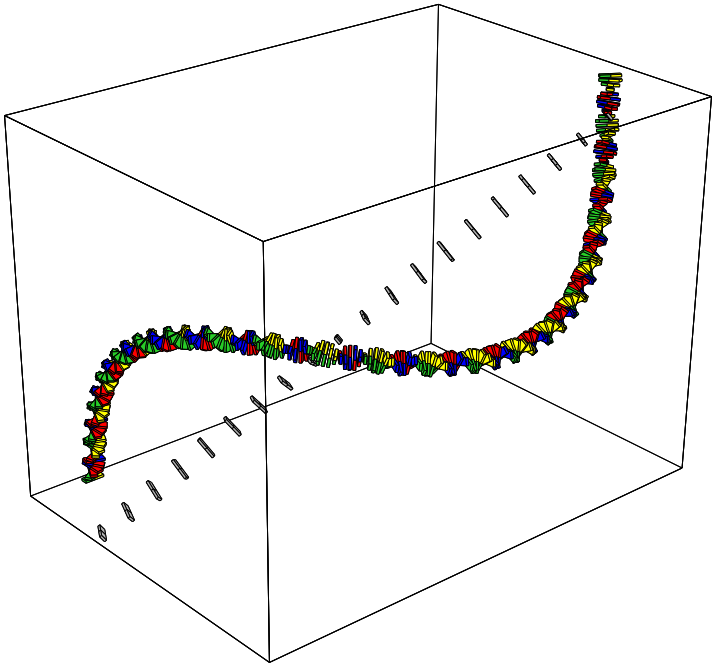

Figure 7.6 | Equilibrium conformation of 20 repeats of the sequence 'CCCCCCTTTAA'. On-axis compound steps are shown in gray. MP parameters.

7.4.5 Superhelices are described by the crbc but not by the wlc

In chapter 5, repetitive sequence rbcs were reduced to ideal B-DNA form by considering the on-axis version of the chain. The on-axis 'phantom' bases fluctuate around the helical centerline, and allow a derivation of the correct long-wavelength statistics of the chain.

However, on short to intermediate length scales, the on-axis version may have very little to do with the true conformation of the rbc. Fig. 7.6 shows an example of an 11-bp repeat, whose intrinsic conformations combine to produce a *superhelix*. The thermal conformation statistics of this repetitive DNA can be treated by combining all steps of the repeat into a compound step as explained in sec. 5.2.1. The resulting chain of repeats is homogeneous but has a rather large axis offset. The on-axis versions of the compound steps lie on the superhelical centerline, see fig. 7.6.

Clearly, on scales of the order of the superhelical repeat (220 bp in this case), the influence of the superhelical axis offset on end-to-end vector statistics is noticeable.

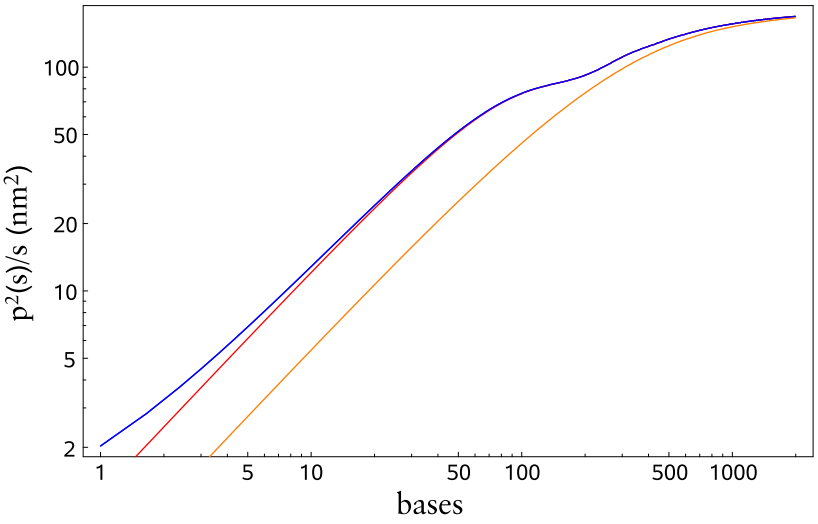

Figure 7.7 | Comparison of wlc (orange) and crbc (blue) predictions for the mean squared distance $\langle p^2 \rangle / s$. The red curve corresponds to switched–off translational fluctuations. Sequence and parameters as in fig. 7.6.

It is captured neither by the on-axis chain nor by its coarse–grained wlc counterpart. In contrast, the crbc chain does trace the superhelical oscillations. This is clearly visible in fig. 7.7 where the mean squared end-to-end distance is shown. While the wlc does reproduce the long-wavelength behavior, the true mean squared distance is increased on the scale of a superhelical repeat, as shown by the crbc result. This extra shoulder is mainly a remainder of the superhelical oscillations. Additionally, an effect of what appears as translational fluctuations of the compound step is visible below 10 bp, in the difference between the full result and the curve with suppressed translation fluctuations.

In conclusion, for repetitive sequences which produce superhelices with large axis offset, there exists an intermediate regime where the crbc model makes non-trivial predictions that the wlc cannot capture. Such superhelical repeats may be relevant in the sequence dependent positioning of nucleosomes in eukaryotic DNA, where sequential signatures of 'pre-curled' DNA have been found on a length of around 100 bp [Aud01, Aud02]. In contexts different from DNA, for example a different helical macromolecule or true rigid body diffusion in the time domain, the helical axis offset and thus the differences between wlc and crbc become important.

8 Lagrangian mechanics on the rigid motion group

In this chapter, an alternative Lagrangian formulation of the continuous rigid body chain in is considered. The equilibrium shape equations are derived, and a set of conserved quantities is found. Finally, the linear response of the chain is calculated around a known solution of the equilibrium shape. While shearable, extensible rods have long been investigated in elasticity theory, the following work is new in that emphasis is put on exploiting the underlying rigid motion group structure. This chapter should be considered an addition of mainly theoretical interest.

8.1 Lagrangian approach to random paths

In this section, we establish the correspondence between the diffusion–type description of a crbc and a formulation in terms of a local energy functional that depends on derivatives of the configuration.

We start from the wlc case where this analogy between chain conformations and particle trajectories in the Lagrangian formulation of mechanics is well known. Then, we extend the formulation to the crbc model, drawing an analogy to the Lagrangian mechanics of systems which have configurations in Lie groups as treated in, e.g. [Arn98].

8.1.1 Elastic energy of the wlc

The worm–like chain model is defined as the continuum limit of a class of discrete models with confined bending angle at each joint, see section 7.2.1. The bending confinement results from an elastic energy. E.g, using the equipartition theorem, the linear elastic model (2.) in 7.2.1, has an elastic energy per link $E(\theta) = \frac{\theta^2}{2\nu_\theta \beta}$.

To relate this discrete picture to the continuous diffusion model discussed in sec. 7.2.2, we define a quadratic energy functional $\mathcal{A}[\Theta(s)]$ as a stochastic process. Taking the limit of the sum of bond angle energies, one obtains the stochastic differential equation $d\mathcal{A}(s) = \frac{1}{2}\alpha(d\Theta(s))^2$, where α is the stiffness constant. Using

Table 8.1 | Correspondence between wlc statistics and Lagrangian mechanics.

wlc quantity	mechanics quantity
chemical distance s	time t
chain conformation $p(s)$	particle trajectory $q(t)$
elastic energy density $\mathcal{L}$	Lagrangian $\mathcal{L}$
total elastic energy $\mathcal{A}$	action $\mathcal{A}$

the Itō formula, this can be transformed into

$$d\mathcal{A}(s) = \frac{l_0 a}{l_b} dW(s) dW(s) = \frac{l_0 a}{l_b} ds. \tag{8.1}$$

One sees that the quadratic functional $\mathcal{A}$ is in fact a non-random function, linear in s.[1] To match the original mean energy density $\frac{1}{2\beta}$ along the chain, we set $a = \frac{l_b}{2\beta l_0}$. Then, (8.1) can be rewritten as

$$\mathcal{A} = \frac{1}{2}\int \frac{l_b}{2\beta l_0}(d\Theta(s))^2 = \frac{1}{2}\int \frac{l_b}{2\beta}(\partial_l \Theta)^2 dl = \frac{1}{2}\int \frac{l_b}{2\beta}(\partial_l^2 p)^2 dl. \tag{8.2}$$

Note that strictly speaking, $\partial_l \Theta$ does not exist, since the random path $\Theta(l)$ is nowhere differentiable. In integrals over a quadratic form, this derivative can be interpreted by the relation $(d\Theta(s))^2 =: (\partial_s \theta)^2 ds = (\partial_l \Theta)^2 dl$. In the case of the 3-d wlc, the bending angle corresponds to two degrees of freedom, so the mean energy density is $2 \cdot \frac{1}{2\beta}$. Therefore, the corresponding 3-d version of (8.2) has $a = \frac{l_b}{\beta}$ instead of $\frac{l_b}{2\beta}$ as stiffness constant.

The elastic energy functional (8.2) is an alternative, equivalent definition of the wlc model. It is entirely analogous to the Lagrangian function in classical mechanics or in field theory. The basic correspondences are listed in table 8.1. In the following, we adopt some of the standard notation, freely changing between the 'time' and 'chemical distance' nomenclature.

8.1.2 Elastic energy of the crbc

We now translate the construction of a continuous energy density to the crbc case. The new ingredients are the six–dimensional conformational space and the presence of drift.

In order to get a constant energy density $\frac{6}{2\beta}$ for the six degrees of freedom along

[1] This well–known fact is a result of the central limit theorem: In the continuum limit, the normalized sum of random single step variances becomes δ-distributed, i.e. non-random [Oks98].

the chain, we write

$$\mathrm{d}\mathcal{A}(s) = \frac{1}{2}(\mathrm{d}\Xi(s) - \xi_0 \mathrm{d}s)^i S_{ij} (\mathrm{d}\Xi(s) - \xi_0 \mathrm{d}s)^j, \tag{8.3}$$

where $S = (\beta C)^{-1}$ is the stiffness matrix. Using (7.8) and the Itō formula for expanding the differential $\mathrm{d}\Xi$, indeed $\mathrm{d}\mathcal{A} = \frac{1}{2} B^i{}_k B^j{}_l \delta^{kl} S_{ij} \mathrm{d}s = \frac{6}{2\beta}\mathrm{d}s$, analogous to the wlc example above.

Equation (8.3) is analogous to (8.2); we can again write it in a more traditional way using a dot $\dot{}$ to denote s derivatives of the fluctuating quantities,[2]

$$\mathrm{d}\mathcal{A}(s) = \mathcal{L}\mathrm{d}s = \frac{1}{2}(\dot{\Xi} - \xi_0)^i S_{ij} (\dot{\Xi} - \xi_0)^j \mathrm{d}s. \tag{8.4}$$

Note that the derivative $\dot{\Xi}$ is by definition nothing but the body velocity of the end frame $\dot{\Xi} = g^{-1}\dot{g} = \xi$. The energy density $\mathcal{L}$ is analogous to the Lagrangian of a classical rigid body with a rotational and translational generalized inertia tensor S_{ij}. We will use the terms Lagrangian and energy density interchangeably.

A crucial difference to the classical mechanics situation is the drift ξ_0, which reflects the material property of intrinsic shape. A way to interpret the drift is by expanding the product in (8.4); then $S\xi_0$ plays the role of an external *field* which exerts force and torque on the particle, analogous to a vector potential in electrodynamics. However, the term $S\xi_0$ is constant with respect to the body frame $g(s)$.[3] This is atypical of an external force, which should be constant in the lab frame,[4] see also section 4.3.

We can now extend the table of correspondences between Lagrangian mechanics and chain conformations in the crbc case, using some of the notions introduced in chapter 4, see tab. 8.2.

8.2 Euler–Lagrange equations

Since $\mathrm{d}\xi$ is just the component vector of $g^{-1}\mathrm{d}g(s) = X_i \mathrm{d}\xi^i(s)$, the Lagrangian depends on the frame configuration and its derivative in a very specific way:

$$\mathcal{L} = \mathcal{L}(g, \dot{g}; s) = \mathcal{L}(g^{-1}\dot{g}; s), \tag{8.5}$$

[2] Since $\frac{\Delta\xi(s) - \xi_0 \Delta s}{\Delta s}$ is of order $1/\sqrt{\Delta s}$, this expression does converge!
[3] i.e. it is a left invariant covector field
[4] i.e. right invariant

Table 8.2 | Correspondence between crbc statistics and Lagrangian mechanics.

crbc quantity	rigid body mechanics quantity
chemical distance s	time t
material frame $g(s)$	body frame $g(t)$
chain conformation $s \mapsto g(s)$	body frame trajectory $t \mapsto g(t)$
elastic energy density $\mathcal{L}$	Lagrangian $\mathcal{L}$
total elastic energy $\mathcal{A}$	action $\mathcal{A}$
material frame strain ξ	body frame velocity ξ
material frame stress μ	body linear/angular momentum μ
body frame drift ξ_0	– ? –

where the explicit s dependence represents that of $\xi_0(s)$ and $S(s)$. We will suppress its notation in the following. Since $\mathcal{L}$ depends only on left invariant vector fields it is, not surprisingly, itself left invariant: For a *constant* left offset h, $\mathcal{L}((hg)^{-1}\partial_s(hg)) = \mathcal{L}(g^{-1}\dot{g})$. Consequently, also the action is left invariant, $\mathcal{A}[g(s)] = \mathcal{A}[h\,g(s)]$. Neither $\mathcal{L}$ nor $\mathcal{A}$ are right invariant.

We can consider the action of arbitrary paths on *SE*. The path which has the highest probability according to a Boltzmann distribution with energy density $\mathcal{L}$ is that which minimizes the total energy $\mathcal{A}$. In analogy to Lagrangian mechanics, we call it the classical path. Since minimization of the total elastic energy $\mathcal{A}$ is equivalent to a stable mechanical equilibrium of the chain, the classical path is the equilibrium shape of the crbc. It is determined by the Euler–Lagrange equations of the problem together with the appropriate boundary conditions.

To find the Euler–Lagrange equations, we extremize the action. Since the Lagrangian has a simple form when given in the left invariant frame, we write a first–order variation of the path in the left invariant frame, as[5]

$$g(s) \to g(s)(e + \delta\xi(s)). \tag{8.6}$$

Note that any first order variation can be written in this basis, so we are imposing no additional restriction on the allowed variations. The body velocity changes as

$$\xi = g^{-1}\dot{g} \to (e - \delta\xi)g^{-1}(\dot{g}(e + \delta\xi) + g\delta\dot{\xi}) = \xi + [\xi, \delta\xi] + \delta\dot{\xi} + O(\delta\xi^2) \tag{8.7}$$

The variation of the action among paths with fixed initial and final points (a, b) is

[5] Here again the abbreviated notation ξ stands for the matrix $\xi^i X_i$.

then

$$\delta\mathcal{A}[g(s)] = \int \delta\mathcal{L}(\xi(s))\,ds = \int \Big(\frac{\partial}{\partial \xi^i}\mathcal{L}(\xi)\Big)(\delta\dot{\xi}^i + [\xi, \delta\xi]^i)ds. \tag{8.8}$$

Partial integration from initial (0) to final (s_f) points leads to

$$\delta\mathcal{A}[g(s)] = \int_0^{s_f} \Big(-\frac{d}{ds}\frac{\partial\mathcal{L}}{\partial\xi^i} + \frac{\partial\mathcal{L}}{\partial\xi^j}\mathrm{ad}\,\xi^j{}_i\Big)\delta\xi^i ds + \Big[\delta\xi^i \frac{\partial}{\partial\xi^i}\mathcal{L}\Big]_0^{s_f}. \tag{8.9}$$

From this expression, the Euler–Lagrange equations can be read off by considering arbitrary variations with the constraint $\delta\xi(0) = \delta\xi(s_f) = 0$. Written in explicit matrix form, they are

$$-\frac{d}{ds}S(\xi - \xi_0) + \mathrm{ad}^T(\xi)S(\xi - \xi_0) = 0. \tag{8.10}$$

In the special case of s-independent coefficients, these six equations may be simplified somewhat and written as

$$\dot{\xi} = S^{-1}\,\mathrm{ad}^T\,\xi\,S(\xi - \xi_0), \tag{8.11}$$

a system of six first–order, quadratic odes.

Recalling the discussion on the linear response of the chain in section 4.3, $\mu = S(\xi - \xi_0)$ is the stress corresponding to the strain $\xi - \xi_0$ of the molecule, expressed in the material frame. Expressed in terms of μ, the general sequence dependent Euler-Lagrange equations (8.10) become

$$\dot{\mu} = \mathrm{ad}^T\,\xi\,\mu = \mathrm{ad}^T(\xi_0 + S^{-1}\mu)\,\mu. \tag{8.12}$$

8.3 Conservation laws

The first variation may be carried out as well in a right invariant setting. This leads to a set of cyclic coordinates and conserved quantities. Proceeding in the same manner as before, the right invariant variation $g(s) \to (e + \delta\zeta(s))g(s)$ changes the body frame velocity in the following way:

$$g^{-1}\dot{g} \to g^{-1}(e - \delta\zeta)((e + \delta\zeta)\dot{g} + \delta\dot{\zeta}g) = \xi + g^{-1}\delta\dot{\zeta}g + O(\delta\zeta^2), \tag{8.13}$$

which using the Ad matrix notation, leads to

$$\delta\mathcal{A}[g(s)] = \int_0^{s_f} \Big(-\frac{d}{ds}(\mathrm{Ad}^{-1}\,g)^j{}_i\frac{\partial\mathcal{L}}{\partial\xi^j}\Big)\delta\zeta^i ds + \Big[\delta\zeta^i(\mathrm{Ad}^{-1}\,g)^j{}_i\frac{\partial\mathcal{L}}{\partial\xi^j}\Big]_0^{s_f}, \tag{8.14}$$

from which an equivalent version of the Euler–Lagrange equations is derived:

$$\frac{d}{ds}\,\mathrm{Ad}^{-\mathsf{T}}{}_{g}\, S(\xi - \xi_0) = 0. \tag{8.15}$$

Carrying out the differentiation indeed gives back 8.10. What is the interpretation of these six conserved quantities? As explained in section 4.3.4, by multiplying with $\mathrm{Ad}^{-\mathsf{T}}{}_{g}$, we can transform the material frame stress μ back to the base frame $g(0) = e$. I.e, the force and torque components, expressed in the base frame:

$$\nu = \mathrm{Ad}^{-\mathsf{T}}{}_{g}\,\mu = \mathrm{Ad}^{-\mathsf{T}}{}_{g}\, S(\xi - \xi_0), \tag{8.16}$$

are conserved. This is nothing but the statement of force balance in mechanical equilibrium, accounting correctly for the moving reference frame.

Continuing the analogy to Lagrangian mechanics, ν is a set of conserved momenta, equal to the initial momenta, and can be computed from the configuration and velocity $g, \dot{g}$ at each point. This set of conserved momenta is a direct consequence of the invariance of $\mathcal{L}$ under left translations by Noether's theorem. In the special case of pure rotational motion and no drift, the crbc is equivalent to the free motion of an asymmetric top. In this case, the conserved momenta are nothing but the conserved total angular momentum vector, given relative to the lab frame. The equations of motion $\dot{\nu} = 0$ of systems whose configuration space is a general Lie group are due to Arnol'd, see e.g. [Arn98].

We also expect to find the equivalent of conservation of energy in the chain. Defining a left invariant Hamiltonian by the usual rule,

$$\mathcal{H} = \xi^{i}\frac{\partial \mathcal{L}}{\partial \xi^{i}} - \mathcal{L} = \frac{1}{2}(\xi + \xi_0)^{\mathsf{T}} S(\xi - \xi_0); \tag{8.17}$$

note the sign change. When expressed in terms of the material stress,

$$\mathcal{H} = \frac{1}{2}\mu^{\mathsf{T}} S^{-1}\mu + \mu^{\mathsf{T}}\xi_0. \tag{8.18}$$

We can plug in (8.9) to get

$$\frac{d}{ds}\mathcal{H}(\xi) = \dot{\xi}^{i}\frac{\partial \mathcal{L}}{\partial \xi^{i}} + \underbrace{\xi^{i}\mathrm{ad}\,\xi^{j}{}_{i}}_{0}\frac{\partial \mathcal{L}}{\partial \xi^{j}} - \frac{d\mathcal{L}}{ds} = -\frac{\partial \mathcal{L}}{\partial s}, \tag{8.19}$$

so $\frac{d}{ds}\mathcal{H}(\xi(s); s) = \frac{\partial}{\partial s}\mathcal{H}(\xi(s); s)$. The Hamiltonian is conserved whenever $\mathcal{L}$ has no explicit s-dependence. Note that the Hamiltonian is *not* the same as the energy density $\mathcal{L}$; even in the constant coefficient case $\frac{\partial \mathcal{L}}{\partial s} = 0$, their difference is generally

not constant,

$$\frac{d}{ds}(\mathcal{H} - \mathcal{L}) = (\xi - \xi_0)^T S \operatorname{ad} \xi \, \xi_0. \tag{8.20}$$

We conclude that the elastic energy $\mathcal{L}$ is equidistributed along the homogeneous crbc whenever $[\xi, \xi_0] = [\xi - \xi_0, \xi_0] = 0$, i.e. when the material frame strain commutes with the equilibrium shape everywhere along the chain. As explained in section 4.2.7, this is the case exactly if the strain is an infinitesimal deformation which shares the same helical axis with the equilibrium shape.

As an example, consider a force–free equilibrium shape of a crbc which is a straight, twisted rod. E.g, such a chain is the result of an on-axis transformation as in chapter 5. We now pull on it in the direction of the axis (say, d_3) with a generalized force $\nu = (\tau, f) = (0, \|f\| d_3)$. Consider the response of the chain at $s = 0$. *If* the on-axis compliance has the property that $C_{\shortparallel}\nu = \xi(0)$ is a screw motion with helical axis d_3, then the resulting shape will be a regular helix and the energy will be equidistributed. This is automatically the case if we choose the helical phase averaged version $\bar{C}$ for the on-axis compliance. On the other hand, if the on-axis compliance does not have that property, a periodic variation of helical parameters and of the energy will result.

8.4 Linear response of the crbc

To determine the equilibrium shape of the crbc, the Euler–Lagrange equations have to be solved for given initial and final configurations. This is a hard problem due to the nonlinearity of the shape equations, and can be best solved numerically. In the following we consider not the explicit solution but the dependence of solutions on the boundary conditions.

8.4.1 Variation of the boundaries

Denote the action evaluated along the classical path (i.e. the minimal chain energy) from $g(0) = a$ to $g(s_f) = b$ by $\mathcal{A}(a, b; s_f)$. We can completely eliminate the initial point dependence: Note that due to left invariance of $\mathcal{L}$, any left translated classical path is again a classical path, from which follows $\mathcal{A}(a, b; s_f) = \mathcal{A}(e, a^{-1}b; s_f)$. This property of the action is completely analogous to that of a classical system in which the Lagrangian is translation invariant, e.g. for a particle in a uniform magnetic field, so that the action depends only on the *difference* of final and initial positions.

Thus, it is enough to consider only the initial condition $a = e$. We write the classical action starting from e as $\mathcal{A}(e, h; s_f) =: \mathcal{A}(h; s_f)$.

The question to be investigated is: How does the minimal chain energy depend on small changes in the initial and final configurations, in other words, what are the derivatives of $\mathcal{A}(a, b; s_f)$? When the final configuration but not the chemical length of the chain is varied, the chain will adopt a new shape, which is again a solution to the Euler–Lagrange equations, but corresponding to the new boundary values.

Looking back at (8.9), since we are starting from a solution of (8.10), this time only the boundary term survives in the first variation, so that

$$\delta\mathcal{A} = \left[\delta\xi^i \frac{\partial}{\partial \xi^i}\mathcal{L}\right]_0^{s_f} = \delta\xi^i(s_f)\mu_i(s_f) - \delta\xi^i(0)\mu_i(0), \tag{8.21}$$

where the body momentum $\mu(s) = S(\xi(s) - \xi_0)$. Note that $\mu(0) = \nu$ is the stress expressed in the lab frame. We call the final body stress $\mu(s_f) = \mu^f$. Then (8.21) amounts to

$$L_i\big|_{b'=b}\mathcal{A}(a, b'; s_f) = \mu_i^f \text{ and } L_i\big|_{a'=a}\mathcal{A}(a', b; s_f) = -\nu_i. \tag{8.22}$$

Because the classical action actually depends only on one argument, it is possible to express its derivatives with respect to initial and final points as derivatives of the back–translated action $\mathcal{A}(h; s_f)$. To do this, note that by definition $\mathcal{A}(a, b; s_f) = \mathcal{A}(a^{-1}b; s_f) = \mathcal{A}((b^{-1}a)^{-1}; s_f)$.

Since L_i is left invariant, we have immediately

$$L_i\big|_{b'=b}\mathcal{A}(a, b'; s_f) = L_i\big|_{h=a^{-1}b}\mathcal{A}(h; s_f). \tag{8.23}$$

For the initial point derivative, recall from (4.11) that the right invariant basis vector fields act on functions by $R_i f(g) = \frac{d}{ds}\big|_0 f((e + sX_i)g)$. One calculates

$$L_i\big|_a f(a^{-1}b) = \left.\frac{d}{ds}\right|_0 f\big((b^{-1}a(e + sX_i))^{-1}\big) = \left.\frac{d}{ds}\right|_0 f\big((e - sX_i)a^{-1}b\big) = -R_i\big|_{a^{-1}b} f. \tag{8.24}$$

We can thus rewrite (8.26) as

$$L_i\big|_{h=a^{-1}b}\mathcal{A}(h; s_f) = \mu_i^f \text{ and } R_i\big|_{h=a^{-1}b}\mathcal{A}(h; s_f) = \nu_i. \tag{8.25}$$

8.4.2 Calculation of the linear response of a crbc

After these preliminaries, we can now proceed to calculate the second derivatives of the classical action with respect to the boundary conditions. In view of (8.25), they will just give the linear response of the chain stress to deformations of the end configuration $g_f = g(s_f)$. More explicitly, we consider the non-symmetric matrix $R_i R_k \mathcal{A}(g_f; s_f)$ which can be written in a variety of different forms;

$$R_i R_k \mathcal{A} = R_i \nu_k = \mathrm{Ad}\, g_f^{-1\,j}{}_i L_j \nu_k = \mathrm{Ad}\, g_f^{-1\,j}{}_i L_k \mu_j^f, \tag{8.26}$$

where $\mu^f = \mu^f(g; s_f)$ is the final material frame stress and $\nu = \nu(g; s_f)$ is the lab frame stress which is conserved along the length s_f chain from e to g. The derivatives in (8.26) are understood to act on the end–to–end separation g_f. Recall also from sec. 4.2.5 that $[R_i, R_k]\mathcal{A} = c^l{}_{ki}\nu_l$ and $[R_i, L_j] \equiv 0$.

Here, care has to be taken when crossing *conjugate points*. Consider a solution $g(s)$ of the Euler-Lagrange equations. At every value $s_c < s_f$ where there exists not an isolated solution reaching $g(s_c)$ in the same 'time' s_c but a whole family, the matrix $R_i R_k \mathcal{A}$ becomes singular. In general, for each additional conjugate point, one additional eigenvalue matrix becomes negative. We do not consider these difficulties here, therefore the discussion is restricted to the case where the equilibrium path is a true local minimum, so that (8.26) remains positive definite.

Rather than directly varying the boundary value g_f, we look at the response of the chain when varying the stress, i.e. the inverse matrix of (8.26). We write the integrated response of the shape on the left, i.e,

$$g(s) \to (e + \delta Z(s))g(s) \tag{8.27}$$

is the accumulated first–order change in g. We can express δZ in terms of the material frame deformation. The material frame deformation $\xi = g^{-1}\dot{g}$, to first order, changes as

$$\xi(s) \to g^{-1}(e - \delta Z)((e + \delta Z)\dot{g} + \delta\dot{Z}g) = \xi(s) + \mathrm{Ad}\, g^{-1}(s)\delta\dot{Z}, \tag{8.28}$$

so that

$$\delta\dot{Z} = \mathrm{Ad}\, g\, \delta\xi = \mathrm{Ad}\, g\, C\delta\mu. \tag{8.29}$$

A crucial point is now that we are considering variations of the classical action, i.e. variations among classical paths. For this reason, the lab frame stress stays a conserved quantity along all varied paths: $\delta\nu$ is independent of s. By the group

property of the Ad matrices and the product rule,

$$\delta\mu = \delta(\mathrm{Ad}^{\mathsf{T}}\, g\, \nu) = \mathrm{Ad}^{\mathsf{T}}\, g\, \delta\nu + \mathrm{Ad}^{\mathsf{T}}\, g\, \mathrm{ad}^{\mathsf{T}}\, \delta Z\, \nu. \tag{8.30}$$

Observe that δZ itself depends on the history of $\delta\mu', s' < s$ via (8.29). In effect, eqn. (8.30) results in an ode, known as the Jacobi equation, which we will now write down. At this point it is convenient to introduce yet another variant of the ad matrices. Let $(\mathrm{ad}^{\perp}\,\tau)_{ij} = \tau_l c^l{}_{ij}$, the contraction of the structure constants with a covector. This is made so that for any $\tau \in se^*$, $V \in se$ we can interchange $\mathrm{ad}^{\mathsf{T}}\, V\, \tau = -\mathrm{ad}^{\perp}\, \tau\, V$. Using this notation and inserting (8.29),

$$\delta\dot{Z} = \mathrm{Ad}\, g\, C\, \mathrm{Ad}^{\mathsf{T}}\, g\, \delta\nu - \mathrm{Ad}\, g\, C\, \mathrm{Ad}^{\mathsf{T}}\, g\, \mathrm{ad}^{\perp}\, \nu\, \delta Z = \mathrm{AD}\, g\, C(\delta\nu - \mathrm{ad}^{\perp}\, \nu\, \delta Z). \tag{8.31}$$

(Recall that $\mathrm{AD}\, g\, C := \mathrm{Ad}\, g\, C\, \mathrm{Ad}^{\mathsf{T}}$).

The Jacobian matrix $J^{ij} = \frac{\partial \delta Z^i_f}{\partial \delta \nu_j}$ gives the first order change of the end configuration g_f, when the stress is varied. Here, both the change in end frame configuration δZ_f and the change in stress $\delta\nu$, are expressed in the lab frame. We can derive from the ode (8.31) which is a vector equation, an ode for the 6×6 matrix J, by taking partial derivatives with respect to the $\delta\nu_j$. The result is the Jacobi differential equation,

$$\dot{J} = \mathrm{AD}\, g\, C - \mathrm{AD}\, g\, C\, \mathrm{ad}^{\perp}\, \nu\, J; \qquad J(0) = 0_{6\times 6}. \tag{8.32}$$

The only external s dependence which is left in this ode is that of the known equilibrium shape $g(s)$. It is thus a linear system of odes with variable coefficients.[6] The solution can be written formally in different ways. Before doing so, note that whenever the equilibrium is stress–free, i.e. equal to the intrinsic shape, the equation can be integrated directly, and the result is

$$J(s_f) = \int_0^{s_f} \mathrm{AD}\, g(s)\, C(s) ds. \tag{8.33}$$

In the stressed case, we note that the solution of the associated homogeneous equation is the ordered matrix exponential,

$$\overrightarrow{\exp}\Big(-\int^{s} \mathrm{AD}\, g' C'\, \mathrm{ad}^{\perp}\, \nu ds'\Big), \tag{8.34}$$

where we used the abbreviated notations $g' = g(s')$, $C' = C(s')$, and the path ordering now places higher s values to the left. By the method of variation of

[6] Explicitely, $(-\mathrm{AD}\, g\, C\, \mathrm{ad}^{\perp}\, \nu)^i{}_j = \mathrm{Ad}\, g^i{}_k C^{kl} \mathrm{Ad}\, g^m{}_l \nu_p c^p{}_{im}$.

constants, we can build the solution in the inhomogeneous case. The result is

$$J(s_f) = \int_0^{s_f} \overrightarrow{\exp}\Big(-\int_s^{s_f} \mathrm{AD}\, g'C'\, \mathrm{ad}^{\perp}\, \nu ds'\Big)\, \mathrm{AD}\, g\, C ds, \tag{8.35}$$

which can be checked by differentiating. [7]

In all but the simplest examples, the matrices $\mathrm{AD}\, g'C'$ do not commute for different 'times' s'. Therefore, the solution (8.35) can only be evaluated by integrating (8.32) numerically, and is really only a formal solution.

The inverse of the Jacobian J is equal to the matrix of lab frame derivatives of the classical action,

$$(J)^{-1}_{ki} = R_i \nu_k(g_f; s_f) = R_i R_k \mathcal{A}(g_f; s_f). \tag{8.37}$$

This is what we set out to calculate.

8.4.3 Simple special cases of the linear response

Whenever the ordered exponential term in (8.35) is just the identity matrix, the Jacobian reduces to (8.33).

Transforming (8.33) to the material frame, we get

$$\frac{\partial \delta \Xi^i_f}{\partial \delta \mu^f_j} = \mathrm{AD}\, g_f^{-1} J = \int_0^{s_f} \mathrm{Ad}(g_f^{-1} g) C\, \mathrm{Ad}^{\mathsf{T}}(g_f^{-1} g) ds, \tag{8.38}$$

where $\delta\Xi_f = \mathrm{Ad}\, g_f^{-1} \delta Z_f$ is the end-frame change, expressed in the material frame. In the terminology of chapter 5, (8.38) is nothing but the covariance of the crbc interpreted as a (continuous) compound step. Indeed, (8.38) is exactly the continuous version of eqn. (5.3) derived there.

We now calculate the linear response explicitly in a very simple example. Consider a crbc that is bent intrinsically so that it closes up into a plane circle which is relaxed, fig. 8.1. For this circle, the intrinsic shape is given by $\xi_0 = (0, \kappa, 0, 0, 0, 1)$, so that κ^{-1} is the radius of the circle, and s is identical to the relaxed arc-length.

[7] Alternative formulations can be derived by noting that the integrand almost has the form $e^x \dot{x}$. One obtains the relation

$$I_6 - J(s_f)\, \mathrm{ad}^{\perp}\, \nu = \overrightarrow{\exp}\Big(-\int_0^{s_f} \mathrm{AD}\, g'C'\, \mathrm{ad}^{\perp}\, \nu ds'\Big), \tag{8.36}$$

which however cannot solved for J since $\mathrm{ad}^{\perp}\, \nu$ is a singular matrix.

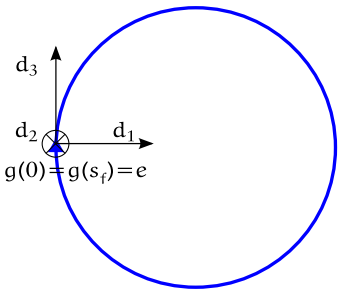

Figure 8.1 | A circle in the d_1-d_3 plane.

The equilibrium shape $g(s)$ is

$$\exp(s\xi_0) = \begin{bmatrix} \exp(s\kappa e_2) & f_1(s\kappa e_2)d_3 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \cos(s\kappa) & 0 & \sin(s\kappa) & \frac{1}{\kappa}(1-\cos(s\kappa)) \\ 0 & 1 & 0 & 0 \\ \sin(s\kappa) & 0 & \cos(s\kappa) & \frac{1}{\kappa}(\sin(s\kappa)) \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{8.39}$$

Let's for simplicity assume that the covariance matrix of the chain allows only bending around material frame d_2 axis and stretching along d_3 with compliances c_b and c_s, respectively and no coupling. Then $C = \mathrm{diag}(0, c_b, 0, 0, 0, c_s)$. Since the circle is stress–free, J is given by (8.33), which can be evaluated as

$$J = \int_0^1 \mathrm{Ad}\, g(s) C\, \mathrm{Ad}^T\, g(s) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2\pi c_b}{\kappa} & 0 & 0 & 0 & \frac{2\pi c_b}{\kappa^2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\pi}{\kappa^3}(c_b + c_s\kappa^2) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2\pi c_b}{\kappa^2} & 0 & 0 & 0 & \frac{\pi}{\kappa^3}(3c_b + c_s\kappa^2) \end{bmatrix}. \tag{8.40}$$

This combined covariance matrix can be reduced to the compliance J_{2d} of in–plane motions, by deleting the rows and columns $1, 3$ and 5. The resulting in–plane stiffness is the inverse, J_{2d}^{-1}. Note that the end frame of the circle can respond to stress also by translation along d_1, although the material cross section cannot. Also, d_3 translation and in-plane bending around d_2 are *positively* coupled. Both of these results make intuitive sense, looking at fig. 8.1.

A global measure for the in–plane compliance is given by the determinant,

$$\det J_{2d} = 2\pi^3\left(\frac{c_b^3}{\kappa^7} + \frac{2c_b^2 c_s}{\kappa^5} + \frac{c_b c_s^2}{\kappa^3}\right). \tag{8.41}$$

Interestingly, the scaling with the circle radius κ^{-1} is different for the different deformation modes. In particular, for an inextensible chain, $c_s \to 0$ and the compliance determinant scales with the seventh power of the radius.

While more general shapes will have less tractable results, the evaluation of eqn. (8.35) can always be implemented on a computer. All that is necessary is knowledge of the equilibrium shape $g(s)$ and the lab frame stress ν for that shape.

Consider now an intrinsically *straight* crbc in the d_1-d_3 plane that has constant compliance matrix $C = \mathrm{diag}(0, c_b, 0, 0, 0, c_s)$. Subjected to a pure torque $\mu^f = (0, t, 0, 0, 0, 0)$, at its end, the chain curves into a circle in the plane, and the material torque is $\mu(s) = \mu^f$ all along the chain. This can be checked by noting that the Euler–Lagrange equation (8.15) reads

$$\begin{bmatrix} R(s) & \hat{p}(s)R(s) \\ 0 & R(s) \end{bmatrix} \mu(s) = const \tag{8.42}$$

which is fulfilled if R is a rotation around the d_2 axis. Therefore, the equilibrium shape is given by (8.39) with $\kappa = c_b t$. The chain will close into a circle of radius κ^{-1} if it has length $s_f = 2\pi\kappa^{-1}$ and torque $\frac{\kappa}{c_b}$.

Now note that in eqn. (8.31), the δZ dependent term vanishes if $C\,\mathrm{ad}^{\perp}\nu = C\,\mathrm{ad}^{\perp}\mu^f \equiv 0$, which is in fact the case for our choice of parameters. In effect, the path-ordered exponential term drops out, so that like in the case of vanishing stress, J reduces to the simple result (8.33). Therefore, the results (8.40), (8.41) are valid for all homogeneous, planar circles which allow bending and stretching only.

8.5 Fluctuations

Until now, we have used the Lagrangian approach to highlight some of the features of equilibrium shapes. To consider the fluctuations of chains in a thermal ensemble, we once again invoke the Boltzmann distribution.

The statistical weight of each path $g(s)$ is given by a Boltzmann factor in the total energy, $e^{-\beta\mathcal{A}[g(s)]}$. Since the state space is now a function space, expectation values over ensembles of continuous paths have to be written as functional (or path) integrals

$$\langle \mathcal{F}[g(s)] | g(0) = e \rangle = \int \mathcal{D}g(s)\mathcal{F}[g(s)]e^{-\beta\mathcal{A}[g(s)]}, \tag{8.43}$$

where by convention the integration extends over all finite energy paths $g : (0, s_f) \to SE$ that start at e. We define the path integral by a time–sliced limit which is essentially the same as the continuum limit of (7.11). By composing the *normalized* transition pdfs of the discrete approximations to the process and taking

the limit, we make sure to obtain the correct transition pdf of the continuous chain. In this way, the division by a partition function "missing" from (8.43) is in fact included in the measure $\mathcal{D}g$.

The path integral is defined by the limit

$$\int \mathcal{D}g(s)\mathcal{F}[g(s)]e^{-\beta\mathcal{A}[g(s)]} :=$$
$$\lim_{\Delta s\to 0} \mathcal{Z}_{0n}^{-1} \int_{SE} \cdots \int_{SE} \mathcal{F}[g^{(n)}(s)]e^{-\beta\mathcal{A}_{01}} e^{-\beta\mathcal{A}_{12}} \cdots e^{-\beta\mathcal{A}_{n-1\,n}}\, dg_{01} dg_{12} \cdots dg_{n-1\,n}. \tag{8.44}$$

Here, $n = s_f/\Delta s$, and the approximate path $g^{(n)}(s)$ is given for integer values on $s/\Delta s$ by the increasing product $g^{(n)}(s) = \prod_1^{s/\Delta s} g_{l-1\,l}$.

We derive the short time action $\mathcal{A}_{i\,i+1}$ from the initial definition of the continuous chain, so that the covariance in exponential coordinates is $C\Delta s$, and the mean value is $\xi_0\Delta s$. If $\xi_{k-1\,k} = \log g_{k-1\,k}$,

$$\mathcal{A}_{k-1\,k} = \frac{1}{2}(\xi_{k-1\,k} - \xi_0\Delta s)^i \frac{S_{ij}}{\Delta s}(\xi_{k-1\,k} - \xi_0\Delta s)^j - \frac{1}{2\beta}\xi^i_{k-1\,k}\bar{A}_{ij}\xi^j_{k-1\,k}. \tag{8.45}$$

The metric factor $\bar{A}$ is a constant matrix needed to cancel the volume element in exponential coordinates, see appendix A.7. It has the effect that the single step partition sum is

$$\mathcal{Z}_{k-1\,k} = \int_{SE} e^{-\beta\mathcal{A}_{k-1\,k}} dg_{k-1\,k} = \int e^{-\frac{\beta}{2\Delta s}(\xi-\xi_0\Delta s)^T S(\xi-\xi_0\Delta s)} d^6\xi = \frac{(2\pi\Delta s)^3}{\det(\beta S)^{1/2}} \tag{8.46}$$

and the covariance $\left\langle (\xi_{k-1\,k} - \xi_0\Delta s)^i (\xi_{k-1\,k} - \xi_0\Delta s)^j \right\rangle = \Delta s\, (\beta S)^{-1\,ij}$, as required. The metric correction factor becomes unimportant in the continuum limit. Even with the metric factor included,

$$\sum \mathcal{A}_{k-1\,k} \overset{\Delta s\to 0}{\longrightarrow} \frac{1}{2}\int (\dot{\Xi} - \xi_0)^T S(\dot{\Xi} - \xi_0) ds. \tag{8.47}$$

The total partition sum is $\mathcal{Z}_{0n} = \prod_{k=1}^n \mathcal{Z}_{k-1\,k} \propto \Delta s^{3n}$. The time–sliced definition of the path integral given here relies on the facts that

1. the discrete chain $g^{(n)}(s)$ has the the energy $\sum \mathcal{A}_{k-1\,k}$

2. the discrete process $g^{(n)}(s)$ converges to the continuous $g(s)$ defined by (7.16), proved in [Ibe76, HD86].

The pdf to reach $g(s_f) = g$ when starting at $g(0) = e$ can be written as

$$p(g, s_f|e, 0) = \left\langle \delta(g^{-1}g(s_f))|g(0) = e \right\rangle = \int \mathcal{D}g(s)\delta(g^{-1}g(s_f))e^{-\beta\mathcal{A}[g(s)]}, \quad (8.48)$$

The functional integration measure in (8.48) is analogous to the standard Wiener measure for Brownian paths. The difference is that here, the paths live in a group, and that they have a drift, so that the path with highest probability is the equilibrium shape determined by ξ_0.

It is interesting to calculate transition pdf (8.48) in the approximation of Gaussian fluctuations around a known minimal shape $g(s)$ for given boundary conditions. The leading order of the transition pdf is always given by the Boltzmann factor $e^{-\beta\mathcal{A}[g(s)]} = e^{-\beta\mathcal{A}(g(s_f);s_f))}$. In Euclidean spaces, the Gaussian fluctuation correction to this factor, which is equivalent to the semiclassical approximation to the propagator in quantum mechanics, is well known. It is given by the so-called van Vleck–Morette determinant, which is the determinant of the second derivatives of the *minimal* energy with respect to the initial and final points. In our setting, this corresponds exactly to the matrix $R_i L_j \mathcal{A}$, as explained in section 8.4.1. This is the main motivation why in the previous section, the matrix $J^{-1}_{ij} = R_j R_i \mathcal{A}$ was calculated.[8]

However whether the formula

$$p(g_f, s_f|e, 0) = \det J^{-1/2} e^{-\beta\mathcal{A}(g_f;s_f)} \quad (8.49)$$

is the correct quadratic fluctuation correction also in the crbc case, is less obvious. The reason is that the crbc evolves on a Lie group, which is a curved space. Also, the Fokker–Planck operator $\mathbf{L}^\dagger$ of the crbc is *not* the same as the Laplace–Beltrami operator on the group, which precludes direct application of a variety results for the semiclassical propagator on curved spaces, see e.g. [Sch81, Gro98].

The resolution of these difficulties is an interesting open task; it would allow a (more or less) explicit calculation of the transition pdf of the crbc model to Gaussian order, from known minimal energy shapes.

[8] Note that since $\det \mathrm{Ad}\, g = 1$, in fact $\det J^{-1} = \det(R_i L_j \mathcal{A})$.

9 Outlook

In this final chapter, some interesting open questions are presented. They arose in the context of the present work, and can addressed using the methods discussed before. Overall, more questions remain open than could be answered. On the other hand, finding the right questions is arguably even more important.

9.1 Superhelical looping

The ability of DNA to form tight loops is of prime importance in various biological contexts (see [Gar07]) such as nucleosome positioning and transcription regulation via DNA looping. It depends on the free energy of cyclization, i.e, the propensity of short pieces of DNA to close up into loops, rather than to concatenate, depending on their sequence. This is a topic of active discussion [Clo04, Clo05, Du05].

As a specific example, consider the cyclization free energy of the sequence repeat 'CCCCCCTTTAA', fig. 7.6. As in section 7.4.5, we combine the repeat into a compound step, and consider a homogeneous, continuous rbc which is modeled on the compound step equilibrium geometry and compliance. By construction, the twist degree of freedom of the underlying DNA double helix is fixed in the superhelical crbc, and all original double-helical structure is then 'forgotten'. As emphasized in sec. 7.4.5, the resulting superhelical crbc has a large helix axis offset; its helix radius is 9 nm with a helical rise of 50 nm per full turn. One can now apply an external force and torque on this continuous chain to force it onto a plane circular path. The free energy of deforming the superhelix crbc into to a plane circle is an approximation to the cyclization free energy after full 11 bp repeats; the oscillations in cyclization free energy that appear by the twist degree of freedom are removed.

In comparison to 'CCCCCCTTTAA', one can consider randomly selected 11 base repeats. They have on average much less helical axis offset than that extreme example above. In fig. 9.1, the resulting elastic cyclization *energies* $\mathcal{A}(e, s_f)$ are plotted. Clearly, there is a huge sequential variation in looping energy, and the 'CCCCCCTTTAA' sequence is easier to cyclize by more than 10 $k_b T$ than the best

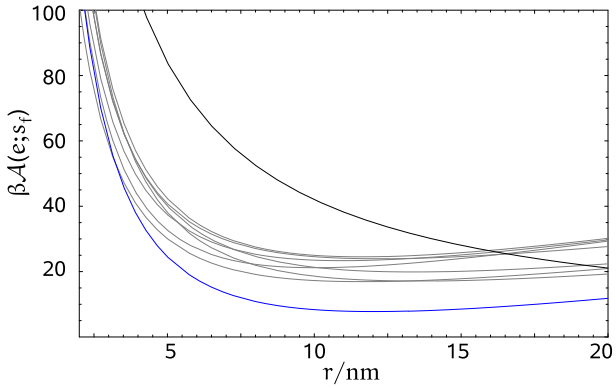

Figure 9.1 | Cyclization energies for random 11-bp repeats (gray) and for 'CCCCCCTT-TAA' (blue). The looping energy of intrinsically straight average DNA is shown in black.

random one, at the optimal radius of 12 nm. The closure energy of intrinsically straight DNA decreases with the typical wlc $1/r$ dependence.

What is the entropic correction to the cyclization free energy? Let's assume that the semiclassical expansion of the transition probability is given by a Boltzmann factor with an entropic correction as in eqn. (8.49), see sec. 8.5. The free energy of the end frame $g_f = g(s_f)$ of the chain is

$$\mathcal{F}(g_f; s_f) = \mathcal{A}(g_f; s_f) + \frac{1}{2\beta} \log\det J(g_f; s_f). \tag{9.1}$$

A zeroth approximation to the entropic contribution can be obtained by evaluating $\log\det J$ along the undisturbed superhelical shape, fig. 9.2, where the simpler form (8.33) applies. The net effect is to shift the optimal loop radius from 12 nm down to 10.2 nm in the superhelical case. In the intrinsically straight case, a free energy minimum appears at 70 nm, which is at over 6 persistence lengths in circumference and is outside the range of validity of the weakly fluctuating approximation.

The first step for a better approximation is an improvement on the plane circle approximation: The equilibrium shape of a closed loop with general equilibrium shape and stiffness, is not a plane circle in general, as can be checked by inspection of the shape equations. In a further step, the full entropic correction given by (8.35) will find a biologically useful application.

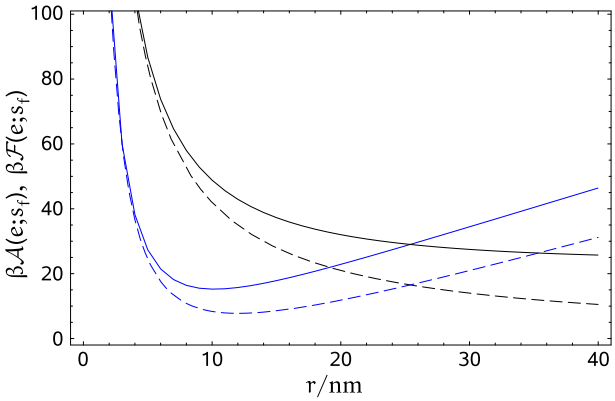

Figure 9.2 | Looping energy (dotted lines) and free energy (solid lines) for the superhelical repeat 'CCCCCCTTTAA' (blue) and for intrinsically straight, random sequence DNA (black). MP parameters.

9.2 More on indirect readout

In chapter 3, indirect readout effects in protein–DNA crystal complexes were examined at a local, base–per–base level. The model protein used there was the bacteriophage 434 repressor protein. This choice was motivated by the amount of experimental data that is available, and by the fact that a comparison between complex structures with different bound operator sequences could be made. Of course, the proposed method is really useful only when applied to other complexes of interest. We make a start here by considering two important cases where indirect readout is assumed to play a key role.

9.2.1 I-ppoI

The DNA–binding protein I-ppoI is part of the so-called His-Cys box family of homing endonucleases. It can recognize a 14-bp sequence and upon binding induces cleavage of DNA near the center of its 20-bp binding site. The protein binds as a homo-dimer (like 434 repressor) and its specific target sequence is palindromic. Although the protein can cleave target sequences that are mutated at many of the base positions, mutations in the central four–base region prevent cleavage [Jur99]. A high–resolution structure of the complex of a non-cleaving variant of the protein with DNA [Gal99] (see fig. 9.3) shows that DNA is bent by approximately 70°, localized to the central 6 bp. Like in the case of 434 repressor,

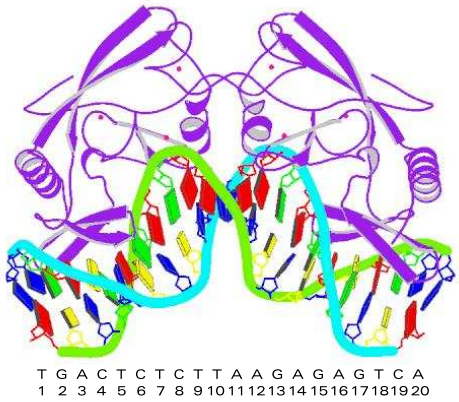

Figure 9.3 | Structure of I-ppoI.

the central region of the site has no specific contacts to the protein. In summary, the homing endonuclease I-ppoI is a candidate for indirect readout that has some similarities to 434 repressor but distorts DNA more strongly.

What can the tools developed in chapter 3 say about this complex? For a first overview, the elastic energy is shown in fig. 9.4 which is analogous to fig. 3.1. In contrast to 434 repressor, the deformation energy shows distinct features in I-ppoI. Notably, the windows around the bases 8 and 13 show peaks in elastic energy, mostly due to shearing. Indeed, a careful analysis of the structure shows these bases to be 'pushed out' of their equilibrium positions by contacts with the protein.

Between these prominent peaks and towards the free ends, the structure is more relaxed. The resulting characteristic double peak of the elastic energy is robust with respect to parametrization uncertainties.

The same basic shape persists in the sequence free energy G, evaluated for the palindromic sequence bound in the complex, shown in fig. 9.5. The peaks around bases 8 and 13 show that the native sequence has a disfavorable elastic energy at these positions. Although the sequence is symmetric and the protein as a homodimer has the same two–fold rotation axis around the central bps, the G profile is not symmetric around that point. It appears that the packing of the crystal used to solve the x-ray structure, breaks this symmetry. Of the two comparatively relaxed end regions, the one with low base numbers accommodates its sequence better.

What can local elastic optimization tell about indirect readout in this complex? Recall from chapter 3 that the information content of the elastically determined

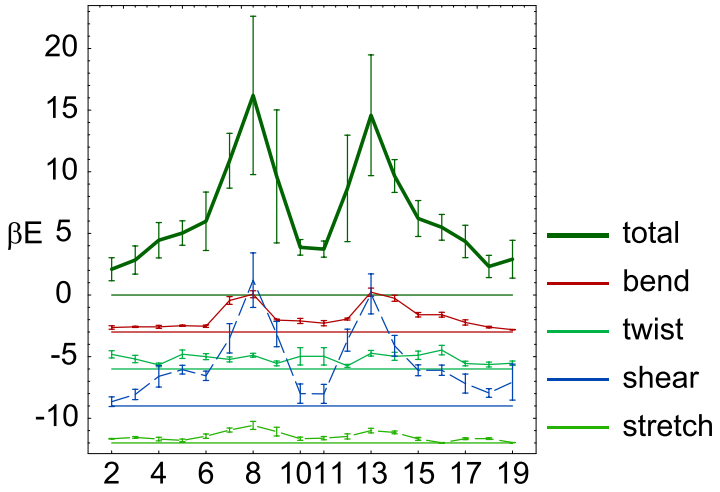

Figure 9.4 | Elastic energy E per bps in I-ppoI, split up to the partial energies for the different deformation modes. A 2 bps window was used. Lines indicate the mean and error bars indicate the spread among parametrizations.

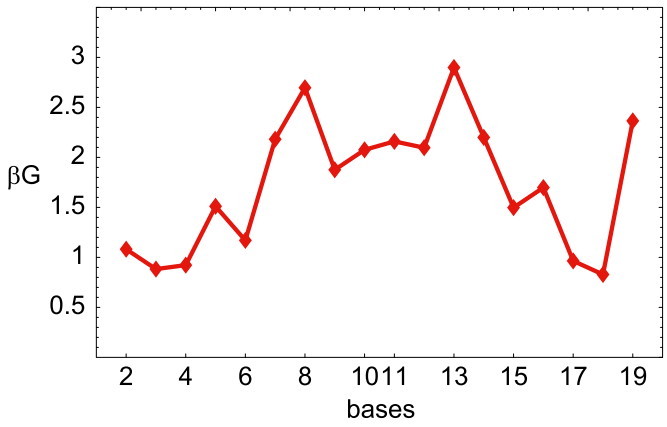

Figure 9.5 | Sequence free energy G of the native sequence, for a 2 bp moving window, given per bp. MP parametrization.

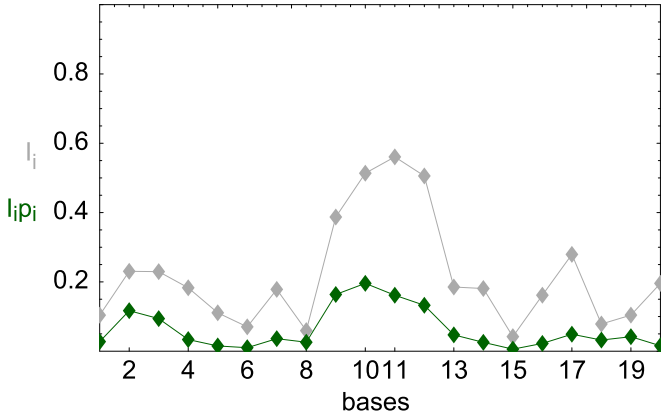

Figure 9.6 | Similarity to elastic consensus for the native bases in the I-ppoI complex. Information (gray) and scaled native probability (green) are shown for single base distributions. MP parameter set.

sequence distribution I_i, scaled with the weight of the native sequence p_i, gives a picture of how much the native sequence is optimized for the given complex structure. Plots of these quantities are shown in fig. 9.6. Although the complex is more relaxed at the ends, the elastic sequence preference for the bound sequence is strongest in the central region between the bends. This suggests that elastic optimization occurs at these sites, in agreement with the observation that there are no specific contacts of the protein in the center. Note that the bases 8 and 13 directly at the bends are elastically non-optimized. It should be mentioned that the shape of these markers depends quite strongly on the chosen parametrization in I-ppoI. E.g, the single peak in the center does not occur when using the P parametrization instead of the preferred parameter set MP (cf. chap. 2). It is replaced by two peaks around positions 7 and 14 (not shown). One feature that is however robust regarding the different parameter sets, is the total absence of elastic optimization at the equivalent positions 6 and 15.

9.2.2 The nucleosome core particle

The nucleosome is the basic building block of chromatin organization. Eucaryotic DNA is almost densely covered with histone octamers, around which DNA is wrapped roughly 1.7 times in a left–handed superhelix. The x-ray crystal structure of a nucleosome core particle containing 147 bp has been solved at near–atomic

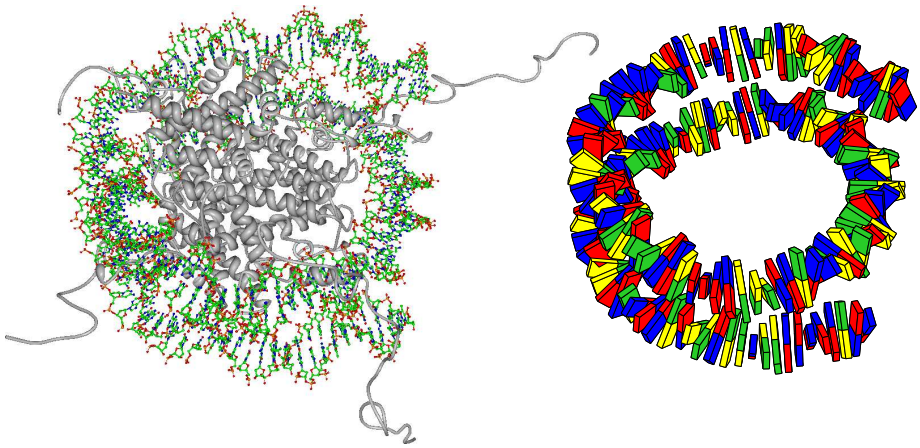

Figure 9.7 | Atomic structure (left) and DNA conformation (right) of the nucleosome core particle.

resolution [Dav02]. Fig. 9.7 shows a cartoon of the full structure and a brick representation of the path of DNA around the histone spool.

The sequence is symmetric, around bp 74, and the structure has an approximate two-fold rotation axis passing through that bp. The intrinsic structural and elastic features of DNA are known to influence the positioning of histones to different sequences. In particular, there exist empirically discovered 'positioning sequences' that are able to strongly localize histones [Clo04].

Specific chemical contacts are made only at certain points around the spool. They involve the backbone phosphates, not individual bases of the bound DNA, suggesting that elasticity may play a dominant role in positioning. These contacts are spaced 10 or 11 bp apart but are *not* symmetrically arranged with respect to the rotation axis.

What information can the local deformation free energies give about the nucleosome structure? In figure 9.8, the elastic energy E is shown. There is considerable, quasi-periodic variation between relaxed and deformed regions. Peaks in elastic energy have high energies of more than 5 $k_B T$ per bp and occur often but not always at bps with backbone contact. Depending on local deformation of the helix, a local energy maximum can also be shifted to the region between the contacts, as seen around bp 55.

Does the particular sequence used in the crystal appear particularly optimized for the structure? The profile shown in fig. 9.9 gives a global picture of the sequence

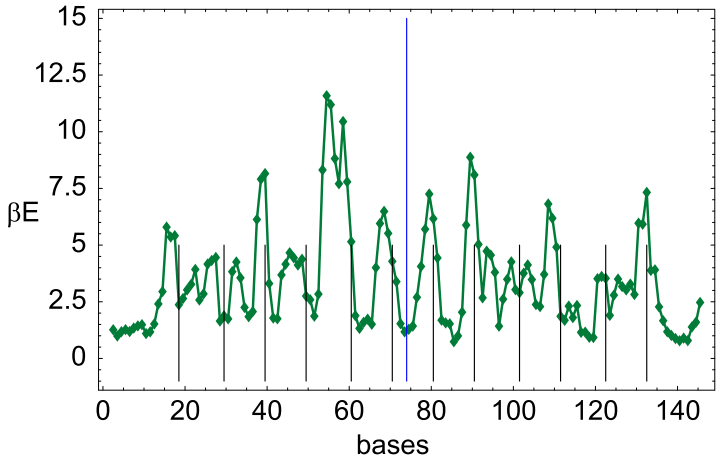

Figure 9.8 | Elastic energy E per bps in the NCP147 nucleosome core particle. A 3 bps window was used. The blue line indicates the central bp. Black lines indicate points of specific contact. MP parametrization.

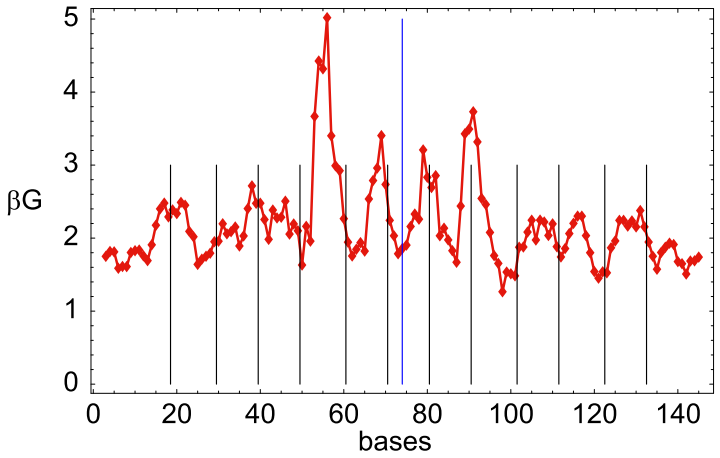

Figure 9.9 | Elastic sequence free energy energy G per bp in the NCP147 nucleosome core particle. A 4 bp window was used. Line marks as in fig. 9.8.

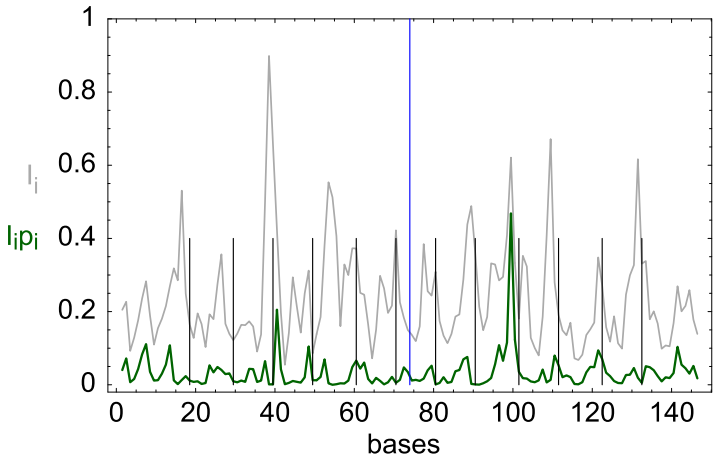

Figure 9.10 | Similarity to elastic consensus in the NCP147 complex. Information content (gray) and scaled native probability (green) are shown for dinucleotide distributions. MP parameter set.

free energy of the native sequence. Strikingly, in contrast to the E profile, points of contact are associated rather with minima than with maxima in the profile, with the exception of the most central positions.

The profile of agreement with elastic consensus $I_i p_i$, fig. 9.10 is surprisingly asymmetric. In agreement with intuition, the information content $I_{i\,i+1}$ of the elastic Boltzmann distribution for dinucleotides is mostly highest at the sites of the strongest constraints, i.e. at the histone contacts. The scaled native probability $I_{i\,i+1} p_{i\,i+1}$ profile shows that only at some of the positions, the native dinucleotide in the crystal is the optimal choice in terms of elasticity. One example is the peak around bp 101. In contrast, at bp 39 there is a strong preference for one specific dinucleotide, indicated by a peak in $I_{i\,i+1}$ but this does not coincide with the native one.

9.3 Forces and torques in crystal structures

Another intriguing possibility offered by the rigid body framework developed in this thesis, is to investigate the forces and torques that act on DNA. The basic idea here is to take the elastic energies in the rigid base–pair model seriously; when a bps is statically deformed, it reacts with a force and torque that balance the externally applied force and torque. They can be calculated for any given DNA

conformation, using the appropriate rbp stiffness matrix. Thus it is possible to take an arbitrary protein–DNA crystal structure as input and from it calculate the local distribution of forces and torques acting on the bound DNA. In this way, DNA becomes a *nanometer–scale force probe*!

Consider a rbp step inside a protein–DNA complex. Its deformation $\xi - \xi_0$ results from a combination of the *internal* tension μ_{in} in the bound rbc, and, if present, *external* forces μ_{ex} exerted through contacts with the protein,

$$\xi - \xi_0 = C(\mu_{in} + \mu_{ex}). \tag{9.2}$$

If no external forces act on base i, it will adopt its equilibrium conformation adapted to the boundary conditions at its ends. Assuming that bases $i-1$ and $i+1$ are held fixed, base i takes on a configuration such that the energy $E_{\sigma_{i-1\,i}}(\xi_{i-1\,i}) + E_{\sigma_{i\,i+1}}(\xi_{i\,i+1})$ is minimized. We have formulated the equilibrium shape equation already in the continuous case, see (8.15). In the discrete case, it takes on the form of a force balance between the two steps flanking the base i,

$$S(\sigma_{i-1\,i})(\xi_{i-1\,i} - \xi_0(\sigma_{i-1\,i})) - Ad^T\, g_{i\,i+1}^{-1} S(\sigma_{i\,i+1})(\xi_{i\,i+1} - \xi_0(\sigma_{i\,i+1})) = 0. \tag{9.3}$$

Conversely, any deviation from this local force balance means that an external force is acting additionally on the base i. More precisely,

$$S(\sigma_{i-1\,i})(\xi_{i-1\,i} - \xi_0(\sigma_{i-1\,i})) - Ad^T\, g_{i\,i+1}^{-1} S(\sigma_{i\,i+1})(\xi_{i\,i+1} - \xi_0(\sigma_{i\,i+1})) = \mu_{ex,i}. \tag{9.4}$$

Thus by calculating (9.4) from a given crystal structure, a detailed picture of forces and torques between protein and DNA can be extracted. An advantage of the particular choice of left invariant components for the deformations of each bps, is that the conjugate variable $\mu = (\tau, f)$ is a combination of true force and torque vectors, given in the material frame. In a different coordinate system, this direct interpretation of the generalized force is not possible.

We show two examples of the resulting stress profiles, using the I-ppoI and NCP147 structures introduced above. Figure 9.11 shows profiles of the total stress and the extracted, external stress μ_{ex}, split up into force and torque magnitudes. The force magnitude follows the deformation energy quite closely. One problematic point is that the magnitude reaches very high values that greatly exceed the overstretching transition observed around 65 pN in naked B-DNA. The torque is computed with respect to an axis that goes through the base pair center. It

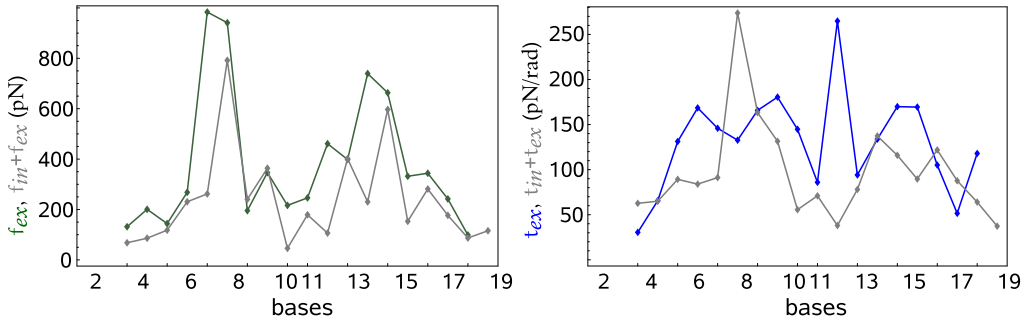

Figure 9.11 | Force ($f = \|f\|$) and torque ($t = \|\tau\|$) magnitude profiles along the I-ppoI structure. The total stress is shown in gray, the external components are colored. P parametrization.

clearly shows that the total torque magnitude does not give a good estimate of the externally applied torque; the internal torsional stress along DNA is of the same order of magnitude.

Figure 9.12 gives a three–dimensional representation of forces and torques in I-ppoI. Only the highest values are shown, compare also the profiles in 9.11. Clearly, the steps between base–pairs 6 and 7 and between 14 and 15 are pulled apart strongly by the protein. These bp steps do not coincide with the actual sites of cleavage of the functional form of the endonuclease, which occurs between bases 8 and 9 (or 11 and 12). Aside form this dominant effect, the bases at the cleavage positions 9 and 12 are twisted roughly in clockwise direction when viewed from the ends.

The same kind of analysis can provide insight also on the elastic state of DNA bound in the nucleosome core particle. In the same way as fig. 9.11, the profiles of force and torque magnitudes are shown in figure 9.13. To make clearer the trends on a scale just below a helical repeat, a moving average is used. One sees that generally, the external stress is are higher in magnitude than the total, which in the light of 9.3 means that total forces on adjacent steps tend to be oriented in opposite directions. Again the calculated forces are very high, especially when compared to the overstretching threshold. The force and torque profiles show large variations. In the left half of the complex, below the symmetry center bp 74, the positions of the peaks in external forces are not in a clear correlation with the known contact points. On the other hand, in the right half of the complex, there is quite some overlap, starting from the peak at bp 81 upwards. The torque

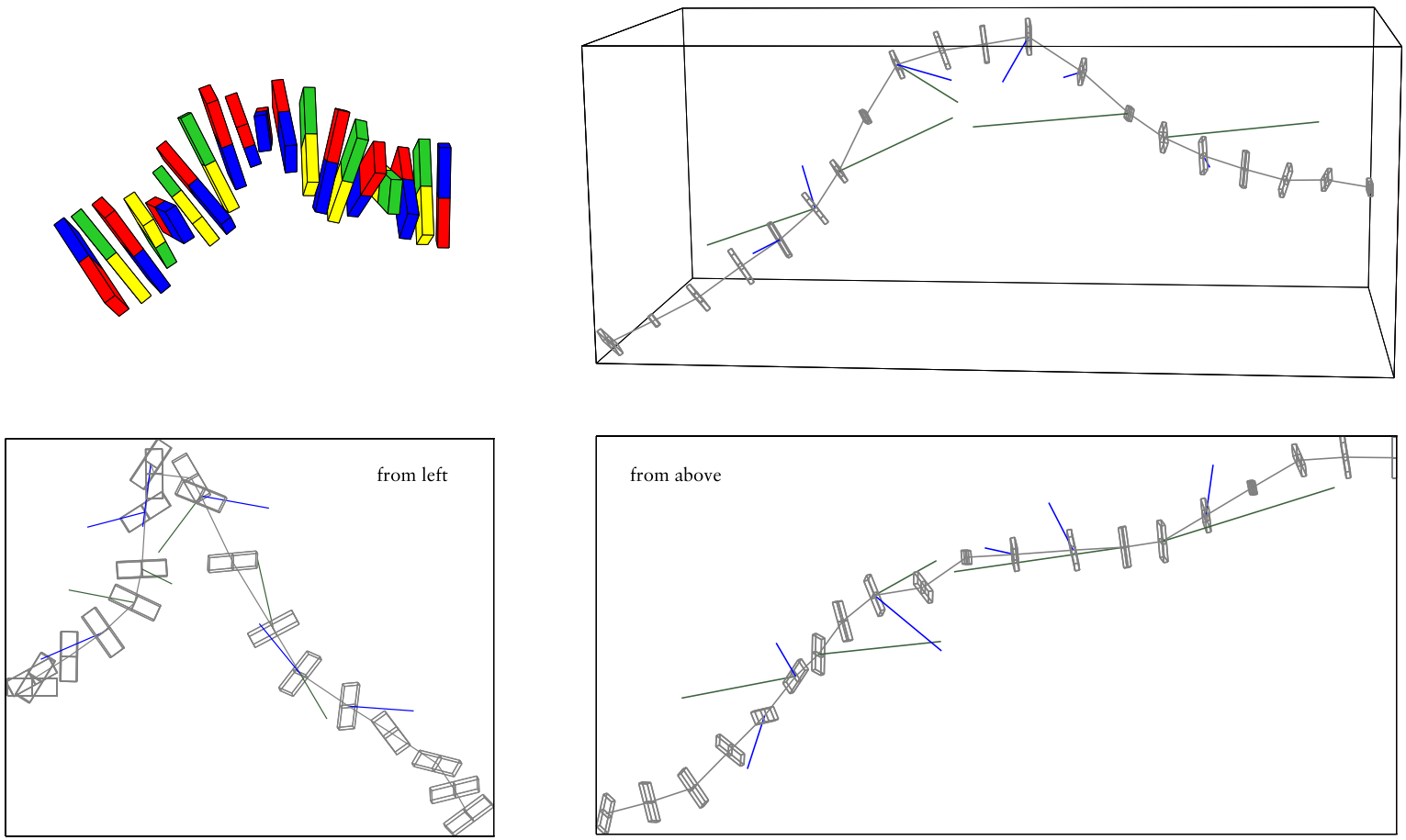

Figure 9.12 | External forces f_{ex} (green) and torques τ_{ex} (blue) acting on bound DNA. Top left: I-ppoI rbp structure. Other panels: The vectors originate at the centers of the attacked bases, which are scaled down and made transparent for clarity. A lower cutoff was used for both f_{ex} and τ_{ex}. P parametrization.

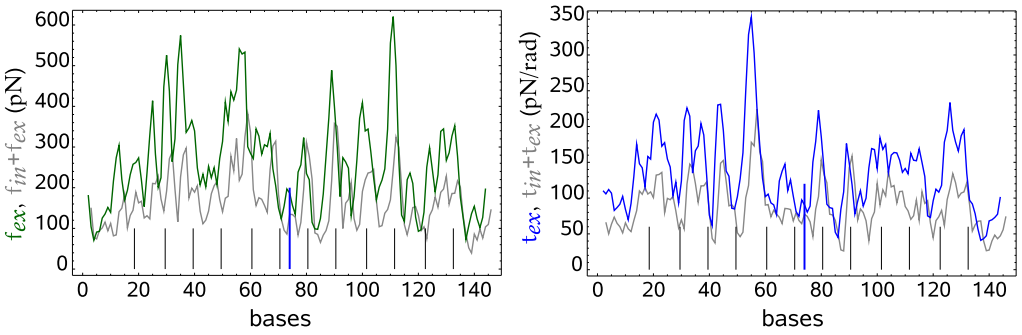

Figure 9.13 | Force and torque magnitude profiles along the NCP147 structure, plotted in the same way as in fig. 9.11. A moving average of length 3 bp is used. P parameter set.

magnitudes show a similar trend.

A three–dimensional representation of external forces in the NCP147 complex is shown in fig. 9.14. Only the most extreme forces and torques are shown. In the upper gyre, the correspondence between contact points and points of strong external forces is visible which is expected from fig. 9.13. Interestingly, a kinked path of the line connecting bp centers is not generally associated with strong external forces. It follows that a distorted bound DNA conformation often results from sequence–dependent equilibrium structure in conjunction with internal stress of DNA.

The force profiles shown here do depend quite strongly on the chosen parametrization, and the observed high forces may in part be due to the fact that experimental error in itself destroys the force balance in eqn 9.3. However, it is encouraging that the most important observed features are reasonable in light of known details of the structures. For example, in fig. 9.14 the external forces occur mostly as antagonistic pairs attaching neighboring bases, in a direction tangential to the histone surface. When the error of force determination can be controlled, and with improved elastic potentials, this method could provide a powerful new tool to measure the *elastic state* of protein–DNA complexes from x-ray crystallography, rather than just the conformation.

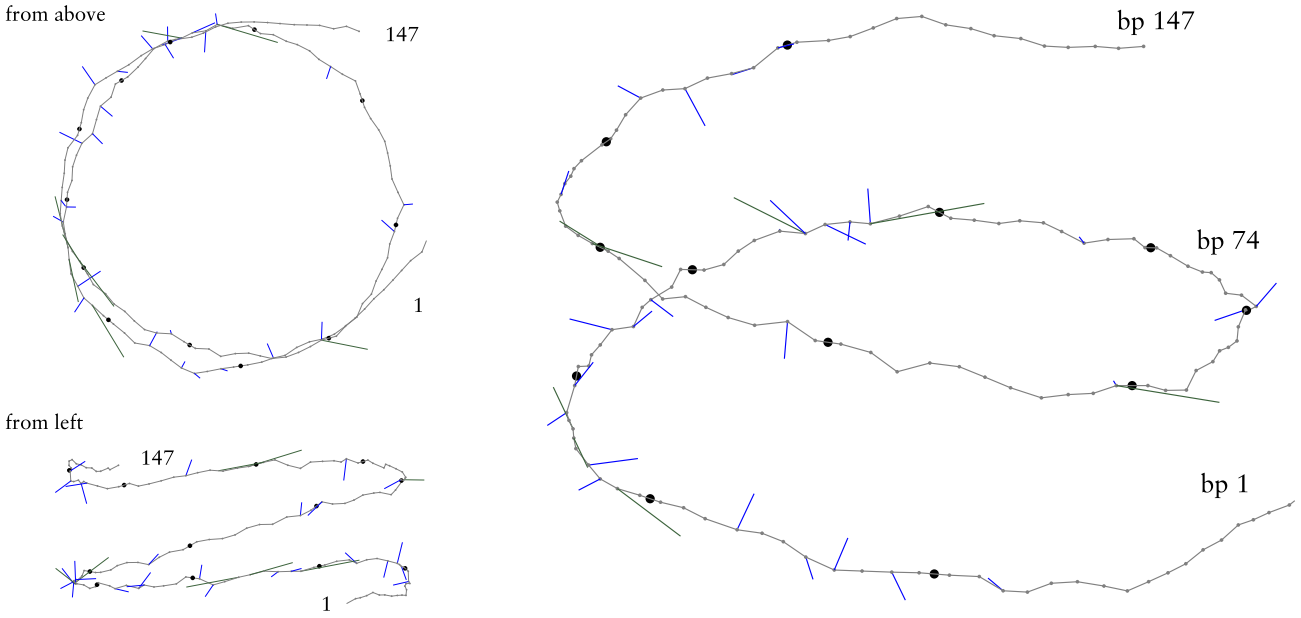

Figure 9.14 | External forces acting in the NCP147 structure, as in fig. 9.12. The base centers are represented as dots, and joined by a line. Black dots represent points of backbone contact. A lower cutoff was used for both f_{ex} and τ_{ex}. P parametrization

A Appendix

A.1 Robustness to parametrization errors

The essential feature that identifies local elastic optimization in protein-DNA complexes is a minimum in the sequence free energy G. We address here whether this feature is robust to the choice of microscopic parametrization among the available sets. In the same way as shown in fig. 3.3, the spread of calculated G profiles among parametrizations is shown in fig. A.1. Although especially for O_R1,2 the parametrization variation in E (fig. 3.3) is quite large whenever the elastic energy is high, this does not destroy the minimum in the corresponding G profiles. We conclude that the local detection of elastic sequence optimization is quite robust to the choice of parameter set.

A.2 The kernel of the adjoint map

We compute the kernel of the map $\operatorname{ad} V$, which is exactly the set of all infinitesimal motions which commute with V.

Let $V' = (\omega', v') \in \ker \operatorname{ad} V$, so that $\operatorname{ad} V\, V' = (\hat{\omega}\omega', \hat{v}\omega' + \hat{\omega}v') = 0$. Consider first the case of non-zero rotation. Since $\ker \hat{\omega} = \operatorname{span} \omega$, necessarily $\omega' = \alpha\omega$ for some real α. Using this in the second entry, we obtain $\hat{\omega}(v' - \alpha v) = 0$. In the case $\omega = 0$, one sees immediately that $\omega' = \alpha v$ but v' is arbitrary. Combining, the

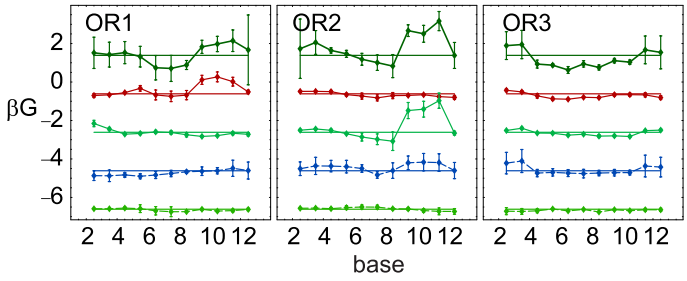

Figure A.1 | Sequence free energy as in fig. 3.7 but including parametrization error bars. Successive 2 k_BT offset, MP, 3 bps average.

two–dimensional kernel of ad V comes out to be

$$\ker \operatorname{ad}(\omega, v) = \begin{cases} \operatorname{span}((\omega, v), (0, \omega)) & \omega \neq 0 \\ \operatorname{span}((v, 0), (0, d_1), (0, d_2), (0, d_3)) & \omega = 0 \end{cases}. \tag{A.1}$$

From this result, one can easily derive the kernel of the map $\operatorname{ad}^{\mathsf{T}}$ which is a subspace of se^*. Taking the transpose of the ad matrix explicitly, one verifies that

$$\ker \operatorname{ad}^{\mathsf{T}}(\omega, v) = \begin{cases} \operatorname{span}((v, \omega), (\omega, 0)) & \omega \neq 0 \\ \operatorname{span}((0, v), (d_1, 0), (d_2, 0), (d_3, 0)) & \omega = 0 \end{cases}, \tag{A.2}$$

where the vectors v and ω have to be multiplied with appropriate scalars to get the dimensions right. E.g, $\alpha\omega$is a torque if α has dimensions of energy.

A.3 Finite matrix power series

The Cayley–Hamilton theorem states that when a d-dimensional square matrix M is plugged into its own characteristic polynomial in place of the variable, the result is $0_{d\times d}$. As a consequence, the powers $\{M^0 = I_d, M, \ldots, M^{d-1}\}$ form a matrix basis for the set of *all* powers of M. Any powers series of M can therefore be resummed such that only the powers up to M^{d-1}appear, with coefficients that are determined by the original series and by the coefficients of expansion of the higher powers in terms of the basis. We compute explicitly some of these computationally convenient finite series for cases of interest with regards to *SE*.

The exponential is defined by its power series. In the case of the antisymmetric matrix $\hat{\omega} \in so$, one directly computes $\hat{\omega}^2 = \omega\omega^{\mathsf{T}} - \|\omega\|^2 e$ and $\hat{\omega}^3 = -\|\omega\|^2\hat{\omega}$.

We get the so-called Rodrigues formula

$$\begin{aligned} \exp\hat{\omega} = \sum_{k=0}^{\infty} \frac{\hat{\omega}^k}{k!} &= e + \hat{\omega}\sum_{k=0}^{\infty} \frac{(-\|\omega\|)^{2k}}{(2k+1)!} - \hat{\omega}^2 \sum_{k=1}^{\infty} \frac{(-\|\omega\|)^{2k-2}}{(2k)!} \\ &= e + \hat{\omega}\frac{\sin\|\omega\|}{\|\omega\|} + \hat{\omega}^2\frac{1-\cos\|\omega\|}{\|\omega\|^2}. \end{aligned} \tag{A.3}$$

The matrix $-\hat{\omega}^2/\|\omega\|^2$ is a projector onto the orthogonal complement of span ω. Noting that $\hat{\omega}^2$ is symmetric we can read off a direct way to compute the logarithm of a rotation matrix $\omega = \log R$:

$$2\cos\|\omega\| + 1 = \operatorname{tr} R;\ \hat{\omega} = (2\sin\|\omega\|)^{-1}(R - R^{\mathsf{T}}). \tag{A.4}$$

Not surprisingly, the logarithm has multiple branches. Restricting to $\|\omega\| < \pi$, one covers already all of the rotation matrices except for exact half turns.

Let's now look at $V = (\omega, v) \in se$. Letting $\hat{V} = V^i X_i$, the power series

$$\exp \hat{V} = \begin{bmatrix} R & p \\ 0 & 1 \end{bmatrix} = e + \sum_{k=1}^{\infty} \frac{1}{k!} \begin{bmatrix} \hat{\omega}^k & \hat{\omega}^{k-1} v \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \exp \hat{\omega} & f_1(\hat{\omega}) v \\ 0 & 0 \end{bmatrix}, \tag{A.5}$$

where

$$f_1(z) = \int_0^1 \exp(sz)\, ds = \frac{\exp z - 1}{z} = \sum_{k=0}^{\infty} \frac{1}{(k+1)!} z^k = 1 + \frac{1}{2} z + \frac{1}{6} z^2 + \cdots. \tag{A.6}$$

Combining this with (A.3), we integrate $\exp[s\hat{\omega}]$ to get

$$f_1(\hat{\omega}) = e + \frac{\hat{\omega}}{\|\omega\|} \frac{1 - \cos \|\omega\|}{\|\omega\|} + \frac{\hat{\omega}^2}{\|\omega\|^2} \frac{\|\omega\| - \sin \|\omega\|}{\|\omega\|}. \tag{A.7}$$

We can use this result together with the finite series of $f_1(\hat{\omega})$ to give a Rodrigues formula for *SE* in terms of the first three powers of $\hat{V}$. One obtains

$$\exp \hat{V} = e + \hat{V} + \hat{V}^2 \frac{1 - \cos \|\omega\|}{\|\omega\|^2} + \hat{V}^3 \frac{\|\omega\| - \sin \|\omega\|}{\|\omega\|^3}. \tag{A.8}$$

The logarithm $V = \log g$ can be recovered as $\omega = \log R$, $v = (f_1(\hat{\omega}))^{-1} p$. The function $1/f_1$ has the power series

$$f_1^{-1}(z) = 1/f_1(z) = \sum_{k=0}^{\infty} \frac{B_k}{k!} z^k = 1 - \frac{1}{2} z + \frac{1}{6} z^2 + \cdots. \tag{A.9}$$

Here, B_k are the Bernoulli numbers, where $B_{2n+1} = 0, n \geqslant 1$. Either by resumming this series for $\hat{\omega}$, using known properties of the B_k or by inverting the projectors onto orthogonal subspaces in separately, one then gets

$$f_1^{-1}(\hat{\omega}) = (f_1(\hat{\omega}))^{-1} = e - \frac{1}{2} \hat{\omega} + \frac{\hat{\omega}^2}{\|\omega\|^2} \left(1 - \frac{\|\omega\|(1 + \cos \|\omega\|)}{2 \sin \|\omega\|} \right), \tag{A.10}$$

completing the formula for log on *SE*.

A.4 The differential of the exponential map

Besides the left and right invariant frames, exponential coordinates $q^i(g) = \log^i g$ provide yet another way of representing a vector in components. Expanding in terms of the coordinate frame, $V\big|_g = V^i \frac{\partial}{\partial q^i}\big|_g$. Finding the conversion between

the coordinate and the left invariant frame amounts to calculating the tangent map $\exp_*\big|_q : T_q T_e SE \to T_{\exp q} SE$, in left invariant components, since

$$(\exp_*\big|_q X_i) f = \partial_t\big|_0 f(\exp(q + tX_i)) = \partial_{q^i} f. \tag{A.11}$$

Following the presentation in [Sat86], we consider

$$A(s,t) = \exp(q) \exp_*\big|_{-q}(-\partial_t q)\ \exp(-q) = \exp(sq(t)) \frac{d}{dt} \exp(-sq(t)), \tag{A.12}$$

where $q \in se$. We compute $\partial_s A = [q, A] - \partial_t q \in se$, so also $A \in se$. To solve this differential equation in s, note that the homogeneous equation $\partial_s A = \operatorname{ad} q\, A$, has the solution $A(s_2) = \exp((s_2 - s_1) \operatorname{ad} q) A(s_1)$. Now by using variation of constants to solve the full inhomogeneous equation, one obtains

$$A(s) = \exp(s \operatorname{ad} q) A(0) + \int_0^s \exp((s - s') \operatorname{ad} q) ds'\, \partial_t q, \tag{A.13}$$

so since $A(0,t) = 0$, finally $A(1,t) = -f_1(\operatorname{ad} q)\partial_t q$, where f_1 is defined as in (A.6). Replacing q by $-q$, one arrives at the general relation

$$\exp_*\big|_q = l_{\exp q *}\big|_e \circ f_1(-\operatorname{ad} q). \tag{A.14}$$

Comparing this with (A.11) and recalling that $L_i\big|_g = l_{g*} X_i$, we get the expression of the coordinate vectors in the left invariant frame:

$$\partial_{q^i} = (f_1(-\operatorname{ad} q))^j{}_i L_j, \tag{A.15}$$

at the point $g = \exp q$. I.e, the matrix $f_1(-\operatorname{ad} q)$ consists of the left invariant components of the coordinate vector fields. In mathematical terms, it is the component matrix in exponential coordinates of the Maurer–Cartan form Ω on the group, defined by $\Omega : TG \to T_e G$, $V\big|_g \mapsto l_{g^{-1}*} V$. From the trigonal block structure of ad, $\det f_1(-\operatorname{ad}(\omega, v)) = \det f_1(\hat{\omega})^2$. Using the finite form of f_1 (A.7), we get for the determinant the product of its eigenvalues $(1, 1 + \frac{\|\omega\| - \sin\|\omega\|}{\|\omega\|} \pm i\frac{1-\cos\|\omega\|}{\|\omega\|})$:

$$\det(\Omega\big|_g) = \Big(\frac{2 - 2\cos\|\omega\|}{\|\omega\|^2}\Big)^2 \tag{A.16}$$

Using the group law one can see that, in exponential coordinates $\exp q = g$, $\exp q' = g'$, the tangent map of left translation from g to g' is given by the

matrix

$$(l_{g'g^{-1}*}\big|_g) = (\Omega\big|_{g'})^{-1}(\Omega\big|_g) = (f_1(-\operatorname{ad} q'))^{-1}(f_1(-\operatorname{ad} q)). \tag{A.17}$$

A.5 Lie algebra automorphisms of *se*

What is the most general change of basis of *se* that respects the commutation relations? Denote the transformation $A \in GL(6)$ and its inverse by the 6×6 block matrices

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \; A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}. \tag{A.18}$$

We require that $[AV, AV'] = A[V, V']$ for all choices of V, V'. This is equivalent to the matrix equation

$$A \operatorname{ad} V A^{-1} = \operatorname{ad}(AV), \; \forall V = (\omega, v) \in se. \tag{A.19}$$

Consider pure translations, $\omega = 0$. We retain the requirement

$$\begin{bmatrix} \widehat{A_{12}v} & 0 \\ \widehat{A_{22}v} & \widehat{A_{12}v} \end{bmatrix} = \operatorname{ad}(AV) = A \operatorname{ad} V A^{-1} = \begin{bmatrix} A_{12}\hat{v}A^{11} & A_{12}\hat{v}A^{12} \\ A_{22}\hat{v}A^{11} & A_{22}\hat{v}A^{12} \end{bmatrix}, \; v \in \mathbb{R}^3. \tag{A.20}$$

Consider the 12 block. We compute $A^{12} = -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$, so necessarily $0 = -A_{12}\hat{v}A_{11}^{-1}A_{12}$. Considering this as a composition of linear maps, we have for all v, $\ker A_{12} \supset \operatorname{im} \hat{v}A_{11}^{-1}A_{12}$. Since $\operatorname{rank} \hat{v} = 2$ for all $v \neq 0$, the rank of A_{12} can be at most 1. This implies that the rank of the 11 block of the rhs is also at most 1. Looking at the lhs, this is a 3×3 hat matrix, but the only such matrix with rank $\leqslant 1$ is the zero matrix! We conclude that $A_{12} = A^{12} = 0$.

Let now $v = 0$. We have

$$\begin{bmatrix} \widehat{A_{11}\omega} & 0 \\ 0 & \widehat{A_{11}\omega} \end{bmatrix} = \begin{bmatrix} A_{11}\hat{\omega}A^{11} & 0 \\ A_{21}\hat{\omega}A^{11} + A_{22}\hat{\omega}A^{21} & A_{22}\hat{\omega}A^{22} \end{bmatrix}, \; \omega \in \mathbb{R}^3. \tag{A.21}$$

Recall that $\operatorname{tr} \hat{\omega}^2 = -2\|\omega\|^2$, valid for hat matrices. Applying this relation, $-2\|A_{11}\omega\|^2 = \operatorname{tr} A_{11}\hat{\omega}A^{11}A_{11}\hat{\omega}A^{11} = -2\|\omega\|^2$. Therefore A_{11} is orthogonal. Also, $A_{22} = A_{11} =: R$.

This matrix has determinant $+1$: If A fulfills (A.19), then $-A$ does not. Since I_3 satisfies (A.19) and the determinant is a continuous function, all admissible choices of R have determinant $+1$.

Computing now the remaining blocks for general V, using the relation for

$R \in SO$, $R\hat{\omega}R^T = \widehat{R\omega}$, we are left with

$$\begin{bmatrix} \widehat{R\omega} & 0 \\ \widehat{A_{21}\omega} + \widehat{Rv} & \widehat{R\omega} \end{bmatrix} = \begin{bmatrix} R\hat{\omega}R^T & 0 \\ A_{21}\hat{\omega}R^T + R\hat{\omega}A^{21} + R\hat{v}R^T & R\hat{\omega}R^T \end{bmatrix}, \ \omega, v \in \mathbb{R}^3. \quad \text{(A.22)}$$

Computing $A^{21} = -R^T A_{21} R^T$, the remaining terms are:

$$\widehat{A_{21}\omega} = A_{21}R^T R\hat{\omega}R^T - R\hat{\omega}R^T A_{21}R^T = [A_{21}R^T, \widehat{R\omega}]. \quad \text{(A.23)}$$

Now since the lhs is antisymmetric, the symmetric part of $A_{21}R^T$ must vanish, we can therefore write $A_{21}R^T = \hat{p}$ for some $p \in \mathbb{R}^3$.

So putting it all together,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} R & 0 \\ \hat{p}R & R \end{bmatrix}; \quad \text{(A.24)}$$

the most general change of coordinates that respects the algebra, or Lie algebra automorphism, has the form of the Ad matrix for some rigid body transformation.

A.6 Partial diagonal forms of the *se* stiffness matrix

Let S denote a positive-definite, symmetric matrix wrt. to the standard basis $\{L_i\}$ of *se*. What is the simplest form of this matrix in some other basis that still obeys the standard commutation relations (4.6)? Under an allowed change of basis $V \mapsto V' = \mathrm{Ad}\, g^{-1} V$, see A.5, we have, written in 3×3 blocks,

$$S' = \mathrm{Ad}^T\, g\, S\, \mathrm{Ad}\, g = \begin{bmatrix} R^T(S_{11} + 2(S_{12}\hat{p})_s - \hat{p}S_{22}\hat{p})R & R^T(S_{12} - \hat{p}S_{22})R \\ R^T(S_{12} - \hat{p}S_{22})R & R^T(S_{22})R \end{bmatrix}, \quad \text{(A.25)}$$

where $\cdot_s$ denotes symmetrization. Note that since S is symmetric and positive definite, so are S_{11}, S_{22}, but S_{12} need have neither property. Clearly, we can choose R to diagonalize the 22 block, but then R is fixed. What else is possible with the remaining freedom of choosing p?

Counting the dimensions, the 3 degrees of freedom of p suffice to eliminate the 3 off–diagonal elements of S'_{11}. So generically, we can simultaneously diagonalize S_{11} and S_{22}.

Also, looking at the off-diagonal blocks, for general S_{12}, S_{22}, the freedom of p is not enough to fulfill the 9 independent equations $S'_{12} = 0$. The coupling terms between rotation and translation can therefore never be eliminated.

If we instead try to diagonalize only the off-diagonal blocks rather than S_{22}, we have p to make $(S_{12} - \hat{p}S_{22})$ symmetric (3 equations), and then can use the freedom in choosing R to diagonalize the remaining symmetric matrix, which is in general indefinite. Both blocks on the diagonal will remain non-diagonal.

Still another possibility is to make S_{12} symmetric using p, but diagonalize either S_{22} or S_{11}. They will not commute with S'_{12}, however, so that S'_{12} will not be diagonal in either case, generically.

All of the aforementioned partial diagonal forms were reproduced on the computer for randomly generated initial pos. def, symmetric S by a numerical gradient–search optimization procedure. However, no further efforts were undertaken to prove them rigorously.

Summing up, generically, the simplest forms of the metric are

1. S'_{11}, S'_{22} both diagonal,
2. $S'_{12} = S'^{\mathsf{T}}_{21}$ diagonal,
3. either S'_{11} or S'_{22} diagonal, and S'_{12} symmetric.

A.7 Volume element

The invariant volume element is given by the Jacobian determinant for the transformation between the chosen coordinate chart and the left or right invariant frames, which equals the determinant A of the Maurer–Cartan form in these coordinates. For exponential coordinates q, the result is (A.16), so that $\ln A(q) = -\frac{1}{6}\|\omega\|^2 + O(\|\omega\|^4)$. The Boltzmann distribution gets the form

$$p(q)dg(q) \propto e^{-\frac{1}{2}q^i(\beta S_{\sigma ij} + \bar{A}_{ij})q^j}\, d^6q,\ \bar{A} = \begin{bmatrix} \frac{1}{3}I_3 & 0_3 \\ 0_3 & 0_3 \end{bmatrix}. \tag{A.26}$$

In DNA, the distributions $p(\xi)$ of single steps are very narrow. Therefore when computing moments, in particular the covariance matrix $C^{ij} = \langle q^i q^j \rangle$, we can extend the integration boundaries to infinity with negligible error. Performing the integral we then get the relation $\beta S + \bar{A} = C^{-1}$. Since $\beta S \gg \bar{A}$ for typical B-DNA steps, in making the approximation $\beta S = C^{-1}$, we introduce small error of less than 1%. I.e. the stiffness matrix βS is to a very good approximation given by the inverse of the covariance.

A.8 Conversion from 3DNA coordinates

The DNA structural analysis program 3DNA [Lu03] uses a coordinate chart $\zeta = (\Omega, \tau, \rho, r_1, r_2, r_3)$, defined in [Lu97]. Here, $\theta = (\Omega, \tau, \rho)$ are Twist, Tilt and Roll angles but differ from our choice of angles. The component vector $r = (r_1, r_2, r_3)$ gives the translation with respect to the mid-frame R_m. These coordinates parametrize $g = (R, p)$ via[1]

$$\begin{aligned} R(\zeta) =& \exp((\Omega/2 - \arctan(\tau/\rho))\epsilon_3)\exp(\sqrt{\rho^2+\tau^2}\,\epsilon_2) \\ & \exp((\Omega/2 + \arctan(\tau/\rho))\epsilon_3), \\ R_m(\zeta) =& \exp((\Omega/2 - \arctan(\tau/\rho))\epsilon_3)\exp(\sqrt{\rho^2+\tau^2}/2\,\epsilon_2) \\ & \exp((\arctan(\tau/\rho))\epsilon_3), \text{ and} \\ p(\zeta) =& R_m(\zeta)r. \end{aligned} \tag{A.27}$$

Choosing exponential coordinates $\tilde{q} = (\tilde{\omega}, \tilde{v})$ based at $g_0 = g(\zeta_0)$, we can transform the coordinate frame $\{\partial_{\zeta^i}\}$ into the left invariant frame $\{L_i\}$ at g_0 by computing the Jacobian J_0 of the coordinate transition map $\zeta \mapsto \tilde{q}(g(\zeta))$ at the point g_0. After some algebra, the 3×3 blocks of $J = \frac{\partial(\tilde{q})}{\partial(\zeta)}$ are

$$\begin{aligned} \frac{\partial\tilde{\omega}^i}{\partial\theta^j} =& 1/2\,\mathrm{tr}(\epsilon_i R^T \partial_{\theta^j} R), & \frac{\partial\tilde{\omega}^i}{\partial r^j} =& 0, \\ \frac{\partial\tilde{v}^i}{\partial\theta^j} =& (R^T \partial_{\theta^j} R_{mid} q r)^i, & \frac{\partial(\tilde{v})}{\partial(r)} =& R^T R_{mid}. \end{aligned} \tag{A.28}$$

The Jacobian determinant comes out to be $\det J = \det \frac{\partial(\tilde{\omega})}{\partial(\theta)} = \frac{\sin\sqrt{\rho^2+\tau^2}}{\sqrt{\rho^2+\tau^2}}$.

A.9 Dimensional structure of the rigid base–pair chain

A basic problem in dealing with rigid body transformations is that rotation matrices are dimensionless while the translation vectors have dimensions of length [l]. One way to deal with it is to choose a fundamental lever arm length scale right from the start, to make all lengths unit-less. However, it turns out to be helpful to retain the explicit dimensional structure of the rigid body transformations. In this way one does not lose track of what quantities depend on the choice of fundamental length scale. Also, the distinct algebraic properties of translations and rotations

[1] Beware of a sign mistake in [Lu97] !

remain explicitly visible in the matrices, preventing errors. We now explain how one can make sense of a matrix group with explicit associated units.

In order to make sense of expressions like $g = \exp \xi^i X_i$ one has to take care of how to assign units to the group and to the algebra elements. Giving the matrix g an outer product unit structure ,

$$[g] = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1/[l] \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ [l] \end{bmatrix}^T = \begin{bmatrix} & & & [l] \\ & 1 \times I_3 & & [l] \\ & & & [l] \\ \frac{1}{[l]} & \frac{1}{[l]} & \frac{1}{[l]} & 1 \end{bmatrix}, \tag{A.29}$$

makes products and inverses well–defined unit-wise, see [Har94]. Lie algebra elements inherit the same structure, being infinitesimal group transformations, $[V] = [g]$. Here, the units of angle are $[\omega] = \text{rad} = 1$.

When writing $V = \xi^i X_i$as a linear combination, it is safest to assign the units to the basis matrices. We instead decide for the more intuitive choice to assign them to the vector components, so that $[\omega] = \text{rad} = 1$ and $[v] = [l]$. It may seem that this will lead to units like $[l]^2$when computing commutators $\xi^i \xi'^j [X_i, X_j]$, inconsistent with the unit structure $[V]$ of the Lie algebra. However, all inconsistent commutators are 0. In fact, also the anticommutators respect this dimensional structure, see 4.2.4. As a result, bilinear matrix products have the same units as Lie algebra elements:

$$[VW] = [\xi^i \xi'^j X_i X_j] = [V] = [g], \tag{A.30}$$

a rather stunning result if one is used to thinking in terms of scalar quantities.

Note that the adjoint matrix Ad is a map $se \to se$, so when written in terms of the standard basis, its dimensional structure is a block outer product

$$[\text{Ad } g] = \begin{bmatrix} [\omega] \\ [v] \end{bmatrix} \begin{bmatrix} 1/[\omega] \\ 1/[v] \end{bmatrix}^T = \begin{bmatrix} 1 & 1/[l] \\ [l] & 1 \end{bmatrix}. \tag{A.31}$$

What are the units associated with the crbc coefficients, considered in 7.3? To start out with, we measure the chemical distance in units of base–pair steps, $[s] = \text{bp} = 1$. The deformation ξ has mixed units of angle per bp, $[\omega] = \frac{\text{rad}}{\text{bp}} = 1$ and length per bp, $[v] = \frac{[l]}{\text{bp}} = [l]$. Correspondingly, the covariance matrix has the following outer product unit structure: $[C] = \begin{bmatrix} [\omega] \\ [v] \end{bmatrix} \begin{bmatrix} [\omega] \\ [v] \end{bmatrix}^T = [BB^T]$. For the

product BW with the dimensionless noise vector W to make sense also, the units of B have to be chosen as $[B] = \left[\begin{smallmatrix}[\omega]\\ [\nu]\end{smallmatrix}\right] \otimes \left[\begin{smallmatrix}1\\ 1\end{smallmatrix}\right]$. With these choices, the unit structure of (7.15) is meaningful,

$$[dg] = [g]\left[(\xi_0^i X_i + \tfrac{1}{2}C^{ij}X_iX_j)ds + B^i{}_j X_i dW^j(s)\right] = [g]([V][ds] + [V][dW]) = [g], \tag{A.32}$$

where it is (A.30) that saves the day.

A.10 Explicit expression for the generator

In explicit matrix notation, the generator of the diffusion process defining the continuous rbc,

$$\mathbf{L}\big|_g f = \left(\frac{\partial}{\partial s} + \frac{\partial^2}{\partial s' \partial s''}\right) f(g\,[s\xi_0^i X_i + \frac{1}{2}s's''C^{ij}X_iX_j]), \tag{A.33}$$

which can be seen by choosing f equal to the matrix entries of g, see e.g. [HD86].

Note that the X_i are effectively symmetrized here, $C^{ij}X_iX_j = \frac{1}{2}C^{ij}\{X_i, X_j\}$. Recalling the anticommutation relations (4.8), one sees that the pure translational part $C^{(\nu\nu)}$ does not contribute at all. The rotation and coupling parts do contribute, but through certain superpositions of anticommutators. Using the notation from 4.2.4, the result can be written as a block matrix

$$\frac{1}{2}C^{ij}X_iX_j = \frac{1}{2}\begin{bmatrix} C^{kl}\epsilon_k\epsilon_l & C^{kl+3}\epsilon_k d_l \\ 0 & 0 \end{bmatrix}, \quad 1 \leqslant k, l \leqslant 3. \tag{A.34}$$

In terms of the original definition (7.10) of C in the discrete model, this can also be written as $C^{kl}\epsilon_k\epsilon_l = \langle \widehat{\delta\omega}^2 \rangle$ and $C^{kl+3}\epsilon_k d_l = \langle \widehat{\delta\omega}\delta\nu \rangle$.

Bibliography

[AB05] M. Arauzo-Bravo, S. Fujii, H. Kono, S. Ahmad and A. Sarai. Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition. *J Am Chem Soc* **127**(46):16 074–89, 2005. 41, 42, 51, 84, 87

[Agg88] A. Aggarwal, D. Rodgers, M. Drottar, M. Ptashne and S. Harrison. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**(4880):899–907, 1988. 29, 30, 31

[Alb02] B. Alberts, Alexander Johnson, Julian Lewis and Martin Raff. *Molecular Biology of the Cell*. Garland, 2002. 2

[Arn98] V. I. Arnol'd and B. A. Khesin. *Topological Methods in Hydrodynamics*. Number 125 in Applied Mathematical Sciences. Springer Verlag, 1998. 126, 131

[Ash06] J. Ashworth, J. J. Havranek, C. M. Duarte, D. Sussman, R. J. Monnat, B. L. Stoddard and D. Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**(7093):656 – 659, 2006. 26

[Aud01] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton Carafa, J. F. Muzy and A. Arneodo. Long-range correlations in genomic dna: a signature of the nucleosomal structure. *Phys Rev Lett* **86**(11):2471–4, 2001. 125

[Aud02] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton Carafa and C. Thermes. Long-range correlations between dna bending sites: relation to the structure and dynamics of nucleosomes. *J Mol Biol* **316**(4):903–18, 2002. 125

[Bau97] C. Baumann, S. Smith, V. Bloomfield and C. Bustamante. Ionic effects on the elasticity of single DNA molecules. *Proc Natl Acad Sci U S A* **94**(12):6185–90, 1997. 99

[Bec06] N. Becker, L. Wolff and R. Everaers. Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Res* **34**(19):5638–5649, 2006. 31, 34, 42, 45, 47, 48, 49, 51, 52

[Bec07] N. Becker and R. Everaers. From rigid base-pairs to semiflexible polymers: Coarse-graining dna. *Physical Review E* 2007. 23, 74, 75, 76, 78, 81, 83, 89, 92, 95, 97, 103

[Bed95] J. Bednar, P. Furrer, V. Katritch, A. Stasiak, J. Dubochet and A. Stasiak. Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J Mol Biol* **254**(4):579–94, 1995. 3, 85, 93, 94, 98

[Bev04] D. Beveridge, G. Barreiro, K. Byun, D. Case, T. Cheatham, 3rd, S. Dixit, E. Giudice, F. Lankas, R. Lavery, J. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. Thayer, P. Varnai and M. 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**(6):3799–813, 2004. 24, 34, 97

[Bru02] R. F. Bruinsma. Physics of protein-dna interaction. *Physica A* **313**:211–237, 2002. 26

[Bus94] C. Bustamante, J. F. Marko, E. D. Siggia and S. Smith. Entropic Elasticity of Lambda-Phage DNA. *Science* **265**(5178):1599–1600, 1994. 72

[Cal84] C. Calladine and H. Drew. A base-centred explanation of the B-to-A transition in DNA. *J Mol Biol* **178**(3):773–82, 1984. 10

[Cal04] C. Calladine and H. Drew. *Understanding DNA*. Elsevier, 3rd edition, 2004. 10

[Cha04] G. Charvin, J. Allemand, T. Strick, D. Bensimon and V. Croquette. Twisting DNA: single molecule studies. *Contemporary Physics* **45**(5):383 – 403, 2004. 72, 99

[Chi00] G. S. Chirikjian and Y. F. Wang. Conformational statistics of stiff macromolecules as solutions to partial differential equations on the rotation and motion groups. *Physical Review E* **62**(1):880–892, 2000. 114

[Chi01] G. S. Chirikjian. Conformational statistics of macromolecules using generalized convolution. *Computational and Theoretical Polymer Science* **11**(2):143–153, 2001. 114

[Cho97] Y. Choo and A. Klug. Physical basis of a protein-DNA recognition code. *Curr Opin Struct Biol* 7(1):117–25, 1997. 25, 26

[Clo04] T. E. Cloutier and J. Widom. Spontaneous sharp bending of double-stranded DNA. *Mol Cell* **14**(3):355–62, 2004. 3, 141, 147

[Clo05] T. Cloutier and J. Widom. DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *Proc Natl Acad Sci U S A* **102**(10):3645–50, 2005. 3, 141

[Clu96] P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J. Viovy, D. Chatenay and F. Caron. DNA: an extensible molecule. *Science* **271**(5250):792–4, 1996. 72

[Col03] B. D. Coleman, W. K. Olson and D. Swigon. Theory of sequence-dependent DNA elasticity. *Journal of Chemical Physics* **118**(15):7127–7140, 2003. 7, 13

[Cor95] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J Am Chem Soc* **117**:5179–5197, 1995. 24

[Dau99] M. Daune. *Molecular Biophysics*. Oxford University Press, 1999. 26

[Dav02] C. Davey, D. Sargent, K. Luger, A. Maeder and T. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol* **319**(5):1097–113, 2002. 147

[Dic89] R. Dickerson. Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Res* **17**(5):1797–803, 1989. 11, 13, 22, 79

[Dix05] S. Dixit, D. Beveridge, D. Case, T. Cheatham, 3rd, E. Giudice, F. Lankas, R. Lavery, J. Maddocks, R. Osman, H. Sklenar, K. 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**(6):3721–40, 2005. 87

[Dre81] H. Drew, R. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura and R. Dickerson. Structure of a b-DNA dodecamer: conformation and dynamics. *Proc Natl Acad Sci U S A* **78**(4):2179–83, 1981. 8

[Dre85] H. R. Drew and A. A. Travers. Dna bending and its relation to nucleosome positioning. *J Mol Biol* **186**(4):773–90, 1985. 3

[Du05] Q. Du, C. Smith, N. Shiffeldrim, M. Vologodskaia and A. Vologodskii. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc Natl Acad Sci U S A* **102**(15):5397 – 5402, 2005. 3, 141

[Elw82] K. D. Elworthy. *Stochastic Differential Equations on Manifolds*. Cambridge University Press, 1982. 107

[Eme90] M. Emery. *Stochastic Calculus on Manifolds*. Universitext. Springer Verlag, 1990. 107

[End06] R. Endres and N. Wingreen. Weight matrices for protein-DNA binding sites from a single co-crystal structure. *Phys Rev E* **73**(6 Pt 1):061 921, 2006. 26

[Eve03] R. Everaers and M. R. Ejtehadi. Interaction potentials for soft and hard ellipsoids. *Physical Review E* **67**(4):–, 2003. 12

[Gal99] E. A. Galburt, B. Chevalier, W. Tang, M. S. Jurica, K. E. Flick, R. J. Monnat, Jr and B. L. Stoddard. A novel endonuclease mechanism directly visualized for i-ppoi. *Nat Struct Biol* **6**(12):1096–9, 1999. 143

[Gar07] H. G. Garcia, P. Grayson, L. Han, M. Inamdar, J. Kondev, P. C. Nelson, R. Phillips, J. Widom and P. A. Wiggins. Biological consequences of tightly bent dna: the other life of a macromolecular celebrity. *Biopolymers* **85**(2):115–30, 2007. 3, 141

[Gol00] R. Golestanian and T. B. Liverpool. Statistical mechanics of semiflexible ribbon polymers. *Physical Review E* **62**(4):5488–5499, 2000. 7

[Gon01] O. Gonzalez and J. Maddocks. Extracting parameters for base-pair level models of DNA from molecular dynamics simulations. *Theoretical Chemistry Accounts* **106**(1-2):76 – 82, 2001. 20, 83

[Gor06] J. Gore, Z. Bryant, M. Nollmann, M. U. Le, N. R. Cozzarelli and C. Bustamante. DNA overwinds when stretched. *Nature* **442**(7104):836 – 839, 2006. 4, 72, 82, 94, 99

[Gro97] M. Gromiha, M. Munteanu, I. Simon and S. Pongor. The role of DNA bending in Cro protein-DNA interactions. *Biophys Chem* **69**(2-3):153–60, 1997. 26, 28

[Gro98] C. Grosche and F. Steiner. *Handbook of Feynman Path Integrals*. Springer, 1998. 140

[Gro04] N. Gromiha, J. Siebers, S. Selvaraj, H. Kono and A. Sarai. Intermolecular and Intramolecular Readout Mechanism in Protein-DNA Recognition. *J Mol Biol* **224**:295, 2004. 26, 41

[Gro05] M. Gromiha. Influence of DNA stiffness in protein-DNA recognition. *J Biotechnol* **117**(2):137–45, 2005. 26

[Har94] G. Hart. The theory of dimensioned matrices. In *Proceedings of 5th SIAM Conference on Applied Linear Algebra*, pp. 186–190. 1994. 163

[HD86] M. Hakim-Dowek and D. Lépingle. *L'exponentielle stochastique des groupes de Lie*, pp. 352–374. Number 20 in Séminaire de Probabilités. Springer, lecture notes in mathematics edition, 1986. 108, 111, 139, 164

[Heg02] R. Hegde. The papillomavirus E2 proteins: structure, function, and biology. *Annu Rev Biophys Biomol Struct* **31**:343–60, 2002. 2, 72

[Her52] J. J. Hermans and R. Ullman. The statistics of stiff chains, with applications to light scattering. *Physica* **18**(11):951 – 971, 1952. 107

[Hin98] C. Hines, C. Meghoo, S. Shetty, M. Biburger, M. Brenowitz and R. Hegde. DNA structure and flexibility in the sequence-specific binding of papillomavirus E2 proteins. *J Mol Biol* **276**(4):809–18, 1998. 26, 72

[Ibe76] M. Ibero. Intégrales stochastiques multiplicatives et construction de diffusions sur un group de lie. *Bull Soc Math France* **100**:175–191, 1976. 108, 111, 113, 139

[Int03a] International Human Genome Sequencing Consortium. Building on the DNA revolution. *Science* 2003. 1

[Int03b] International Human Genome Sequencing Consortium. Human genome project information. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml, 2003. 1

[JJ00] L. Jen-Jacobson, L. Engler and L. Jacobson. Structural and thermodynamic strategies for site-specific DNA binding proteins. *Structure* **8**(10):1015–23, 2000. 26

[Jur99] M. Jurica and B. Stoddard. Homing endonucleases: structure, function and evolution. *Cell Mol Life Sci* **55**(10):1304–26, 1999. 143

[Kam97] R. D. Kamien, T. C. Lubensky, P. Nelson and C. S. OHern. Direct determination of DNA twist-stretch coupling. *Europhysics Letters* **38**(3):237–242, 1997. 7, 72, 83

[Kar82] R. Karandikar. *Approximation results for multiplicative stochastic integrals*, pp. 384–391. Number 16 in Sï¿½minaire de Probabilitï¿½s. Springer, 1982. 108

[Ken90] W. Kendall. Probability, convexity, and harmonic maps with small image .1. Uniqueness and fine existence. *Proceedings of the London Mathematical Society* **61**:371 – 406, 1990. 73

[Kne86] J. A. Knezetic and D. S. Luse. The presence of nucleosomes on a DNA template prevents initiation by rna polymerase ii in vitro. *Cell* **45**(1):95–104, 1986. 3

[Kou87] G. Koudelka, S. Harrison and M. Ptashne. Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature* **326**(6116):886–8, 1987. 2, 26, 29, 30, 31, 34, 72

[Kou88] G. Koudelka, P. Harbury, S. Harrison and M. Ptashne. DNA twisting and the affinity of bacteriophage 434 operator for bacteriophage 434 repressor. *Proc Natl Acad Sci U S A* **85**(13):4633–7, 1988. 30

[Kou91] G. Koudelka. Bending of synthetic bacteriophage 434 operators by bacteriophage 434 proteins. *Nucleic Acids Res* **19**(15):4115–9, 1991. 30

[Kou92] G. Koudelka and P. Carlson. DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 repressor. *Nature* **355**(6355):89–91, 1992. 30, 37, 45

[Kou98] G. Koudelka. Recognition of DNA structure by 434 repressor. *Nucleic Acids Res* **26**(2):669–75, 1998. 30

[Kou06] G. Koudelka, S. Mauro and M. Ciubotaru. Indirect readout of DNA sequence by proteins: the roles of DNA sequence-dependent intrinsic and extrinsic forces. *Prog Nucleic Acid Res Mol Biol* **81**:143–77, 2006. 2, 26

[Kra49] O. Kratky and G. Porod. Röntgenuntersuchung gelöster fadenmoleküle. *Rec Trav Chim Pays-Bas* **68**:1106–1123, 1949. 7, 105

[Lan00] F. Lankaš, J. Šponer, P. Hobza and J. Langowski. Sequence-dependent elastic properties of DNA. *Journal of Molecular Biology* **299**(3):695–709, 2000. 72, 75, 81, 82, 83

[Lan03] F. Lankaš, J. Šponer, J. Langowski and T. Cheatham, 3rd. DNA base-pair step deformability inferred from molecular dynamics simulations. *Biophys J* **85**(5):2872–83, 2003. 4, 21, 22, 23, 66, 83

[Lan06a] F. Lankas, R. Lavery and J. Maddocks. Kinking Occurs during Molecular Dynamics Simulations of Small DNA Minicircles. *Structure* **14**(10):1527–34, 2006. 100

[Lan06b] F. Lankaš. Sequence-dependent harmonic deformability of nucleic acids inferred from atomistic molecular dynamics. In J. Šponer and F. Lankaš, eds., *Computational studies of DNA and RNA*, volume 2 of *Challenges and Advances in Computational Chemistry and Physics*. Springer, 2006. 22, 83

[Lav89] R. Lavery and H. Sklenar. Defining the Structure of Irregular Nucleic-Acids - Conventions and Principles. *Journal of Biomolecular Structure & Dynamics* **6**(4):655–667, 1989. 75

[Leb96] A. Lebrun and R. Lavery. Modelling extreme stretching of DNA. *Nucleic Acids Research* **24**(12):2260 – 2267, 1996. 7

[Lee02] J. Lee. *Introduction to Smooth Manifolds*. Springer, 2002. 55

[Lio06] T. Lionnet, S. Joubaud, R. Lavery, D. Bensimon and V. Croquette. Wringing out DNA. *Phys Rev Lett* **96**(17):178 102, 2006. 4, 72, 82, 94, 99

[Lu97] X. Lu, M. El Hassan and C. Hunter. Structure and conformation of helical nucleic acids: rebuilding program (SCHNArP). *J Mol Biol* **273**(3):681–91, 1997. 162

[Lu99a] X. Lu, M. Babcock and W. Olson. Overview of nucleic acid analysis programs. *J Biomol Struct Dyn* **16**(4):833–43, 1999. 22

[Lu99b] X. Lu and W. Olson. Resolving the discrepancies among nucleic acid conformational analyses. *J Mol Biol* **285**(4):1563–75, 1999. 22

[Lu03] 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):5108–21, 2003. 13, 22, 64, 162

[Mad] J. H. Maddocks. Personal communication. 51

[Mar94] J. F. Marko and E. D. Siggia. Bending and Twisting Elasticity of DNA. *Macromolecules* **27**(4):981–988, 1994. 7, 78, 93

[Mar97] J. F. Marko and E. D. Siggia. Driving proteins off DNA using applied tension. *Biophysical Journal* **73**(4):2173–2178, 1997. 72, 83

[Mat88] B. Matthews. Protein-DNA interaction. No code for recognition. *Nature* **335**(6188):294–5, 1988. 25

[Mat99] A. Matsumoto and N. Go. Dynamic properties of double-stranded DNA by normal mode analysis. *Journal of Chemical Physics* **110**(22):11 070 – 11 075, 1999. 72

[Mat02] A. Matsumoto and W. Olson. Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. *Biophys J* **83**(1):22–41, 2002. 23, 72

[Maz06] A. Mazur. Evaluation of elastic properties of atomistic DNA models. *Biophys J* **91**(12):4507–18, 2006. 72, 75, 76, 81

[McK60] H. McKean. Brownian motions on the 3-dimensional rotation group. *Mem Col Sci Kyoto* **33**:25–38, 1960. 108, 111

[Mer03] B. Mergell, M. R. Ejtehadi and R. Everaers. Modeling DNA structure, elasticity, and deformations at the base-pair level. *Physical Review E* **68**(2):–, 2003. 12

[Moa05] M. Moakher and J. H. Maddocks. A double-strand elastic rod theory. *Archive for Rational Mechanics and Analysis* **177**(1):53 – 91, 2005. 7, 84

[Mor97] J. Moroz and P. Nelson. Torsional directed walks, entropic elasticity, and DNA twist stiffness. *Proc Natl Acad Sci U S A* **94**(26):14 418–22, 1997. 72, 83

[Mor05] A. Morozov, J. Havranek, D. Baker and E. Siggia. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res* **33**(18):5781–98, 2005. 18, 26, 28, 33, 41, 48

[Mur93] R. Murray, Z. Li and S. Sastry. *A mathematical introduction to robotic manipulation*. CRC Press, Boca Raton, 1993. 55

[Nel98] P. Nelson. Sequence-disorder effects on DNA entropic elasticity. *Phys Rev Lett* **80**(26):5810–5812, 1998. 95, 96

[O'H98] C. S. O'Hern, R. D. Kamien, T. C. Lubensky and P. Nelson. Elasticity theory of a twisted stack of plates. *European Physical Journal B* **1**(1):95–102, 1998. 7

[Oks98] B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer-Verlag, Berlin Heidelberg New York, 1998. 127

[Ols98] W. Olson, A. Gorin, X. Lu, L. Hock and V. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* **95**(19):11 163–8, 1998. 4, 21, 22, 23, 66, 102

[Ols01] 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**(1):229–37, 2001. 22

[Pab00] C. Pabo and L. Nekludova. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol* **301**(3):597–624, 2000. 25

[Pai04a] G. Paillard, C. Deremble and R. Lavery. Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res* **32**(22):6673–82, 2004. 26, 41

[Pai04b] G. Paillard and R. Lavery. Analyzing protein-DNA recognition mechanisms. *Structure* **12**(1):113–22, 2004. 26, 28

[Pan00] S. Panyukov and Y. Rabin. Fluctuating filaments: Statistical mechanics of helices. *Physical Review E* **62**(5):7135 – 7146, 2000. 114, 117

[Per28] F. Perrin. Etude mathématique du mouvement brownien de rotation. *Ann ENS* **45**:1–51, 1928. 108

[Pre93] C. Prevost, S. Louisemay, G. Ravishanker, R. Lavery and D. L. Beveridge. Persistence analysis of the static and dynamic helix deformations of DNA oligonucleotides - application to the crystal-structure and molecular-dynamics simulation of d(cgcgaattcgcg)2. *Biopolymers* **33**(3):335 – 350, 1993. 72

[Ric03] T. J. Richmond and C. A. Davey. The structure of DNA in the nucleosome core. *Nature* **423**(6936):145–150, 2003. 2

[Rip01] K. Rippe. Making contacts on a nucleic acid polymer. *Trends Biochem Sci* **26**(12):733–40, 2001. 72

[Ris89] H. Risken. *The Fokker-Planck Equation*. Springer-Verlag, Berlin Heidelberg New York, 1989. 111

[Rod93] D. Rodgers and S. Harrison. The complex between phage 434 repressor DNA-binding domain and operator site OR3: structural differences

between consensus and non-consensus half-sites. *Structure* **1**(4):227–40, 1993. 29

[Rub03] M. Rubinstein and R. H. Colby. *Polymer Physics*. Oxford University Press, 2003. 105

[Sai67] N. Saito, K. Takahashi and Y. Yunoki. Statistical mechanical theory of stiff chains. *J Phys Soc Jap* 1967. 107

[Sai06] L. Saiz and J. M. Vilar. Dna looping: the consequences and its control. *Curr Opin Struct Biol* **16**(3):344–50, 2006. 3

[Sal06] M. Salomo, K. Kegler, C. Gutsche, M. Struhalla, J. Reinmuth, W. Skokow, U. Hahn and F. Kremer. The elastic properties of single double-stranded DNA chains of different lengths as measured with optical tweezers. *Colloid and Polymer Science* **284**(11):1325 – 1331, 2006. 99

[Sas90a] H. Sasmor and J. Betz. Specific binding of lac repressor to linear versus circular polyoperator molecules. *Biochemistry* **29**(38):9023–8, 1990. 26

[Sas90b] H. Sasmor and J. Betz. Symmetric lac operator derivatives: effects of half-operator sequence and spacing on repressor affinity. *Gene* **89**(1):1–6, 1990. 26

[Sat86] D. Sattinger and O. Weaver. *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*. Number 61 in Applied Mathematical Sciences. Springer, 1986. 55, 158

[Sch72] R. Schleif. Fine-structure deletion map of escherichia-coli l-arabinose operon. *Proceedings of the National Academy of Sciences of the United States of America* **69**(11):3479 –, 1972. 72

[Sch75] R. Schleif and J. T. Lis. Regulatory Region Of L-Arabinose Operon - Physical, Genetic And Physiological Study. *Journal Of Molecular Biology* **95**(3):417 – 431, 1975. 3

[Sch81] L. Schulman. *Techniques and Application of Path Integration*. Wiley, 1981. 140

[Sch90] T. Schneider and R. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**(20):6097–100, 1990. 47, 48, 49

[Sch92] R. Schleif. DNA looping. *Annu Rev Biochem* **61**:199–223, 1992. 3, 72

[Sch03] H. Schiessel. The physics of chromatin. *Journal of Physics-Condensed Matter* **15**(19):R699–R774, 2003. 2

[Sch04] J. B. Schvartzman and A. Stasiak. A topological view of the replicon. *EMBO Rep* **5**(3):256–61, 2004. 2

[Seg06] 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**(7104):772–8, 2006. 72

[Shi93] L. Shimon and S. Harrison. The phage 434 OR2/R1-69 complex at 2.5 A resolution. *J Mol Biol* **232**(3):826–38, 1993. 29

[Shr90] T. E. Shrader and D. M. Crothers. Effects of DNA sequence and histone-histone interactions on nucleosome placement. *J Mol Biol* **216**(1):69–84, 1990. 3

[Sin94] R. R. Sinden. *DNA Structure and Function*. Academic Press, 1994. 2, 8, 9

[Spa04] A. Spakowitz and Z. Wang. Exact results for a semiflexible polymer chain in an aligning field. *Macromolecules* **37**(15):5814 – 5823, 2004. 108

[Sta96] A. Stasiak, V. Katritch, J. Bednar, D. Michoud and J. Dubochet. Electrophoretic mobility of DNA knots. *Nature* **384**(6605):122, 1996. 3

[Ste02] N. Steffen, S. Murphy, L. Tolleri, G. Hatfield and R. Lathrop. DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics* **18 Suppl 1**:S22–30, 2002. 26

[Str96] T. R. Strick, J.-F. Allemand, D. Bensimon, A. Bensimon and V. Croquette. The Elasticity of a Single Supercoiled DNA Molecule. *Science* **271**(5257):1835–1837, 1996. 72

[Str00] T. Strick, J. Allemand, V. Croquette and D. Bensimon. Twisting and stretching single DNA molecules. *Prog Biophys Mol Biol* **74**(1-2):115–40, 2000. 4

[Suz95] M. Suzuki, S. Brenner, M. Gerstein and N. Yagi. DNA recognition code of transcription factors. *Protein Eng* **8**(4):319–28, 1995. 25

[Tha99] A. Thastrom, P. Lowary, H. Widlund, H. Cao, M. Kubista and J. Widom. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol* **288**(2):213–29, 1999. 26

[Tri88] E. N. Trifonov, R. K. Z. Tan and S. C. Harvey. Static persistence length of DNA. In W. K. Olson, M. H. Sarma, R. H. Sarma and M. S. Sundaralinjam, eds., *Structure & Expression*, pp. 243–254. Adenine Press, 1988. 85, 98

[Vil03] J. M. Vilar, C. C. Guet and S. Leibler. Modeling network dynamics: the lac operon, a case study. *J Cell Biol* **161**(3):471–6, 2003. 3

[Vol02] M. Vologodskaia and A. Vologodskii. Contribution of the intrinsic curvature to measured DNA persistence length. *J Mol Biol* **317**(2):205 – 213, 2002. 3, 85, 93, 94, 97, 98, 103

[Wan97] M. D. Wang, H. Yin, R. Landick, J. Gelles and S. M. Block. Stretching DNA with optical tweezers. *Biophysical Journal* **72**(3):1335 – 1346, 1997. 99

[Wat53a] J. D. Watson and F. H. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**(4361):964–7, 1953. 7

[Wat53b] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**(4356):737–8, 1953. 1, 7

[Wen02] J. Wenner, M. Williams, I. Rouzina and V. Bloomfield. Salt dependence of the elasticity and overstretching transition of single DNA molecules. *Biophys J* **82**(6):3160–9, 2002. 99

[Wid97] H. Widlund, H. Cao, S. Simonsson, E. Magnusson, T. Simonsson, P. Nielsen, J. Kahn, D. Crothers and M. Kubista. Identification and characterization of genomic nucleosome-positioning sequences. *J Mol Biol* **267**(4):807–17, 1997. 26

[Wid01] J. Widom. Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys* **34**(3):269–324, 2001. 72

[Wig05] P. A. Wiggins, R. Phillips and P. C. Nelson. Exact theory of kinkable elastic polymers. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* **71**(2):021 909, 2005. 7

[Wig06] P. A. Wiggins, T. van der Heijden, F. Moreno-Herrero, A. Spakowitz, R. Phillips, J. Widom, C. Dekker and P. C. Nelson. High flexibility of DNA on short length scales probed by atomic force microscopy. *Nat Nano* **1**:137–141, 2006. 3, 94, 100, 101

[Win03] R. G. Winkler. Deformation of semiflexible chains. *Journal Of Chemical Physics* **118**(6):2919 – 2928, 2003. 7

[Yam97] H. Yamakawa. *Helical wormlike chains in polymer solutions*. Springer, 1997. 7, 105, 107, 117

[Zef02] M. Zefran and V. Kumar. A geometrical approach to the study of the cartesian stiffness matrix. *Journal of Mechanical Design* **124**(1):30 – 38, 2002. 55

Glossary

AD	adjoint matrix representation of a group acting on two–tensors
Ad	adjoint matrix representation of a group
aD	adjoint matrix representation of a Lie algebra acting on two–tensors
ad	adjoint matrix representation of a Lie algebra
β	inverse temperature
bp	base–pair
bps	base–pair step
$[\cdot]$	block matrix
C	covariance matrix
crbc	continuous rigid body chain
d_i	basis three–vector, $(d_i)^j = \delta_i^j$
$\cdot$	empty slot in a function or expression (e.g, $f(a, \cdot)$)
dsDNA	double–stranded DNA
e	group identity element, identity matrix
ϵ_i	antisymmetric 3×3 basis matrix, $(\epsilon_i)^j{}_k = \epsilon^j{}_{ik}$
$\langle A \rangle$	expectation value of A, $\int A p(A)\, dA$
$\langle A \vert B \rangle$	expectation value of A conditioned on B, $\int A p(A\vert B)\, dA$
F	conformation free energy
G	sequence free energy
g	rigid motion group element $g = (R, p)$ or its homogeneous matrix representation

$\hat{}$	antisymmetrix matrix of a 3-vector, $\hat{v} = v \times \cdot$
K	sequence and conformation joint free energy
k_B	Boltzmann's constant
lhs	left hand side
L_i	left invariant basis vector fields (l. i. frame)
MD	molecular dynamics
μ	left invariant force/torque covector components μ_i
ν	right invariant force/torque covector components ν_i
$O(x)$	order notation: $y = O(x)$ if $\lim y/x < \infty$
$o(x)$	order notation: $y = o(x)$ if $\lim y/x = 0$
ode	ordinary differential equation
p	translation group element, translation vector
$\langle \cdot, \cdot \rangle$	natural pairing of covectors and vectors, $\langle \mu, V \rangle = \mu(V)$
pde	partial differential equation
pdf	probability density function
q	exponential coordinates q^i on *SE*
R	rotation group element, rotation matrix
rbc	rigid base–pair chain
rbp	rigid base–pair
rhs	right hand side
R_i	left invariant basis vector fields (r. i. frame)
rms	root mean square
S	stiffness matrix
sde	stochastic differential equation
σ	base sequence $\sigma = b_1 \dots b_k$

$A^i B_i$	$= \sum_i A^i B_i$: implicit summation over all upper/lower index pairs
wlc	worm–like chain
ξ	left invariant vector components ξ^i
ζ	right invariant vector components ζ^i