PhD-FSTC-2018-03
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 22/02/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN PHYSIQUE

by

Amirhossein TAGHAVI
Born on 05 April 1979 in (IRAN)

MODELLING THE EXTENSIONALLY DRIVEN TRANSITIONS OF DNA

Dissertation defence committee

Dr Josh T. BERRYMAN
Université du Luxembourg

Dr Tanja SCHILLING
Professor, Universität Freiburg

Dr Jan LAGERWALL
Professor, Université du Luxembourg

Dr Albrecht OTT
Professor, Uni Saarland-Lehrstuhl für biologische Experimentalphysik

Dr Agnes NOY
University of York

Modelling the Extensionally Driven Transitions of DNA

Amirhossein Taghavi

February 2018

Abstract

Empirical measurements on DNA under tension show a jump by a factor of $\approx 1.5 - 1.7$ in the relative extension at applied force of $\approx 65 - 70$ pN, indicating a structural transition. The still ambiguously characterised stretched 'phase' is known as S-DNA. Using atomistic and coarse-grained Monte Carlo simulations we study DNA over-stretching in the presence of organic salts Ethidium Bromide (EtBr) and Arginine (an amino acid present in the RecA binding cleft). We present planar-stacked triplet disproportionated DNA as a solution phase of the double helix under tension, and dub it 'Σ DNA', with the three right-facing points of the Σ character serving as a mnemonic for the three grouped bases. Like unstretched Watson-Crick base paired DNA structures, the structure of the Σ phase is linked to function: the partitioning of bases into codons of three base-pairs each is the first stage of operation of recombinase enzymes such as RecA, facilitating alignment of homologous or near-homologous sequences for genetic exchange or repair. By showing that this process does not require any very sophisticated manipulation of the DNA, we position it as potentially appearing as an early step in the development of life, and correlate the postulated sequence of incorporation of amino acids (GADV then GADVESPLIT and then the full 20 residue set of canonical amino acids) into molecular biology with the ease of Σ-formation for sequences including the associated codons. To further investigate the dependence of stretching behaviour on the concentration of intercalating salt molecules, we present a physically motivated coarse-grained force-field for DNA under tension and use it to qualitatively reproduce regimes of force-extension behaviour which are not atomistically accessible.

Acknowledgements

I would like to thank my supervisors Dr. Joshua T. Berryman and Prof. Tanja Schilling for their support during my PhD and Prof. Bengt Nordén[1] for inspiring discussions and Prof. Paul van der Schoot[2] for valuable information and discussions.

[1]Chemistry and Chemical Engineering, Chalmers University of Technology, Gothenburg, Sweden.

[2]Department of Applied Physics, Theory of Polymers and Soft matter, Technische Universiteit Eindhoven P.O. Box 513 5600 MB Eindhoven.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Part I : DNA Structure and function

The discovery of the double helical structure of DNA [12] was iconic because the structure immediately illuminates the mechanism by which DNA can store and duplicate information. Further, the twisting and untwisting of the DNA double helix provides it with a range of peculiar mechanical properties: DNA can, for example supercoil to be packed in confinements such as viral capsids (compaction), or stretch twice its original length upon applying force with a negative twist-stretch coupling constant [13].

1.1.1 DNA molecule

The DNA double helix consists of two polynucleotide chains each composed of four different nucleotides: adenine (A), thymine (T), guanine (G) and cytosine (C). Each nucleotide has three parts: a nitrogenous base, a five-carbon sugar (which in DNA is called deoxyribose) and a phosphate group. Each nucleotide in one strand has a complementary one on the opposing strand. Nucleotides are covalently linked together in a chain and in this way form the backbone of the DNA. The chains are directional, and in a DNA duplex the directionality of the two non-covalently linked chains will be opposite. The direction of DNA replication is referred to as $5' \rightarrow 3'$, so when a duplex is pictured the first or 'sense' strand is by convention $5' \rightarrow 3'$ (following the DNA structure from the bottom to top of the page) and the complementary or 'antisense' strand is $3' \rightarrow 5'$.

Chains are connected to each other by hydrogen bonds. Each nucleotide in one strand has a complementary one on the opposing strand. A is paired with T *via* two hydrogen bonds and C is paired with G *via* three hydrogen bonds. A and G are called purines and T and C are known as pyrimidines.

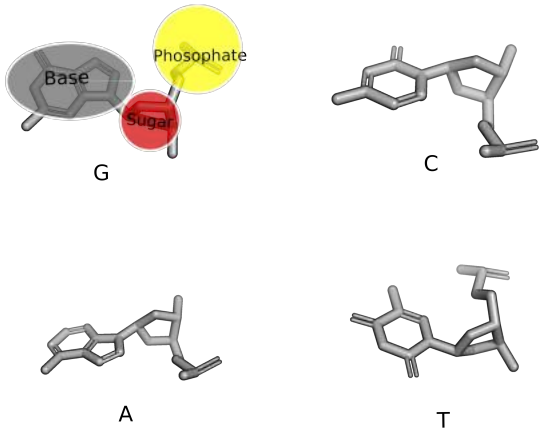


Figure 1.1: Figure shows the structure of four nucleotides, G, C, A and T. Each base (gray circle) is connected to the sugar (red circle) which for DNA is called deoxyribose and then connected to a phosphate group (yellow) to make a complete nucleotide. A and G are called purine nucleotides (bases with a double aromatic ring), C and T are called pyrimidine nucleotides (with a single aromatic ring).

Due to the phosphate groups, DNA is highly charged. In the cellular environment, this highly charged polyanion is partly neutralized by counter-ions like Na^{+} and Mg^{2+} which help to stabilize DNA structure by screening intramolecular repulsion [14,15]. This counter-ion environment is not necessarily created by monatomic salt ions: proteins, small bioorganic molecules [16] and monovalent and divalent inorganic cations can do the same job [17]. The mode and extent of neutralization of the negative charge has great impact on the structure and function of DNA.

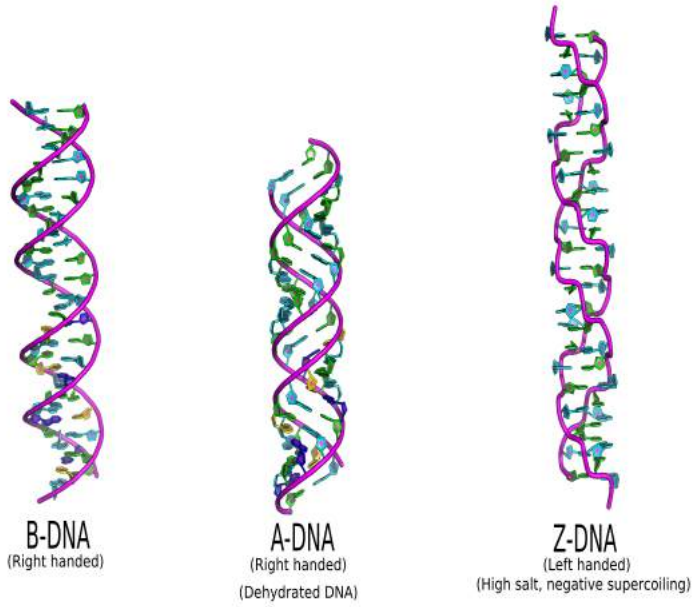


Figure 1.2: The canonical structure of DNA double helix B-DNA along with two alternate forms (A-DNA and Z-DNA) which are formed under certain conditions (low humidity and high salt concentrations, respectively). Phosphate backbone is shown in pink and bases are coloured separately(A(blue), C(green), T(yellow) and G(cyan)). DNA models were made with 3DNA [1] with the sequence d(GGCGGCGGCGGCGACGACGACGAC).

In different environmental conditions DNA can adopt different structures like A-DNA, which is the result of dehydration of DNA (Fig.1.2) [18], or Z-DNA which is a left handed structure formed in high salt conditions or negative supercoiling[1] [20,21]. Z-DNA is often associated with cellular stress or DNA-damage [22].

The sugar moiety of nucleotides can have different conformations based on the arrangement of carbon atoms of the sugar ring. Each ring has five carbons which cannot lie in the same plane. The $C2'$ atom of the sugar ring stands out of the plane on the same side as the base. This form of the sugar is called $C2'endo$. In A-DNA, sugar puckering is $C3'endo$. Structural parameters of the B-DNA are shown in Table 1.1.

[1] Z-DNA is largely found in DNA near to an active transcription locus. When RNA polymerase moves along the DNA it introduces negative torsional strain which leads to formation and stabilization of Z-DNA near the transcription start site making it a locally metastable conformation [19].

Table 1.1: Structural parameters of B-DNA.

Structural Parameter	B-DNA
direction of helix rotation	right handed
base pairs per turn	10.5
axial rise	3.32Å
pitch	34.8Å
base pair tilt	-6°
rotation per base pair	34.3°
diameter of helix	20 Å
sugar pucker	C2′ *endo*

1.1.2 DNA stabilizing forces

1.1.2.1 Stacking interactions in DNA

Besides hydrogen bonds, stacking interactions are the most important forces keeping the DNA strands together. It is believed that London dispersion forces and Coulombic interactions are the main sources of the direct stacking interactions between DNA bases, while the indirect hydrophobic interaction also plays a strong role. Stacking has a major impact on DNA functions related to bending, opening of base-pairs, and intercalation. A very important aspect of stacking interactions is their role in conformational transitions during over-stretching (especially the B to S transition) which currently appear to be mainly governed by cooperative breakage of stacking interactions. Different combinations of base-pairs have different stacking energies [10, 23]. Measuring stacking interactions experimentally or theoretically is very complicated, especially with experimental techniques, as it is hard to exclude the influence of factors like hydration, counter-ion interactions or hydrogen bonding [24–28]. Table 1.2 shows the most recent experimentally defined single-stack free energies between 10 base-pair steps [8].

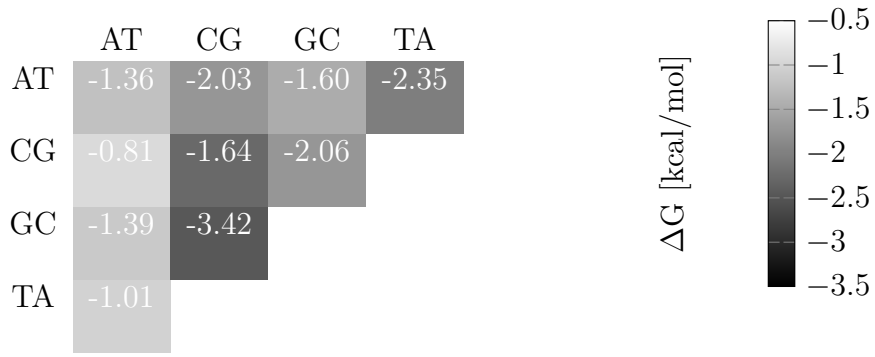


Table 1.2: Free energies of ten base-pair steps measured between two separate DNA duplexes interacting end-to-end. Here a table entry, eg. $GC \times AT = 1.39$, indicates the stacking free energy of a hypothetical duplex eg. $GA \cdot TC$ measured, however, in the absence of the backbone linkages GA and TC. (Table reproduced from data in [8].)

Further experimental studies on 30 different 300 bp long DNA duplexes with a single nick at the same place but with different neighboring bases showed that GC ($GC \cdot GC/2$) is the most stable pair (-2.4 kcal/mol) while TA ($TA \cdot TA/2$) is the least stable one (-0.05 kcal/mol) Table 1.3 [9].

Table 1.3: Stacking free energies (in kcal/mol) for different combinations of nucleotides (table reproduced from data in [9].)

	A	T	G	C
A	-1.11	-1.34	-1.06	-1.81
T	-0.19	-1.11	-0.55	-1.43
G	-1.43	-1.81	-1.44	-2.17
C	-0.55	-1.06	-0.91	-1.44

1.1.2.2 Hydrogen bonds

Hydrogen bonds in DNA have both electrostatic and covalent character and are reinforced by π polarization [29]. There is a translational and rotational entropy penalty for bases making hydrogen bonds but there is a total entropy

gain upon releasing water molecules into the bulk water surrounding the DNA molecule. In summary DNA stability is attributed to:

- Stacking interactions (dispersion forces and also electrostatic between stacked bases)
- Electrostatic repulsive interactions between phosphate groups
- Screening effect of counterions on phosphates
- Hydrophilic interactions between sugar-phosphate and water
- Hydrophobic interactions between bases and water
- Hydrogen bonding between complementary base pairs

It is a very difficult task to identify the contribution of each of these stablizing forces to the total stability of DNA, but many theoretical and experimental investigations confirm the major importance of stacking interactions among the others.

1.1.3 DNA helical parameters

One helical turn of B-DNA contains about 10.5 base pairs that are almost perpendicular to the helical axis. This whole arrangement makes a cylinder of 20 Å diameter with two grooves, a major and a minor one. In B-DNA the distance between bases is 3.4 Å , which is called the "Rise". There are a set of orientational and translational parameters which are used to describe base-pair configuration. Tilt, Roll, Twist angles and Shift, Slide, Rise translations define neighboring base-pair steps and the Buckle, Propeller Twist, opening angles and Shear, Stagger, Stretch displacements position complementary bases (Fig. 1.3). It should be mentioned that this set of parameters can be defined based on a local helix axis between two base-pairs or can also be defined globally relative to an overall helix axis. In this respect quantities can differ considerably. For example rise along the global helix axis is 2.9 Å , whereas the rise calculated from local axes will be the difference between stacked base pairs, 3.4 Å. Translational and rotational parameters are shown in Fig. 1.3.

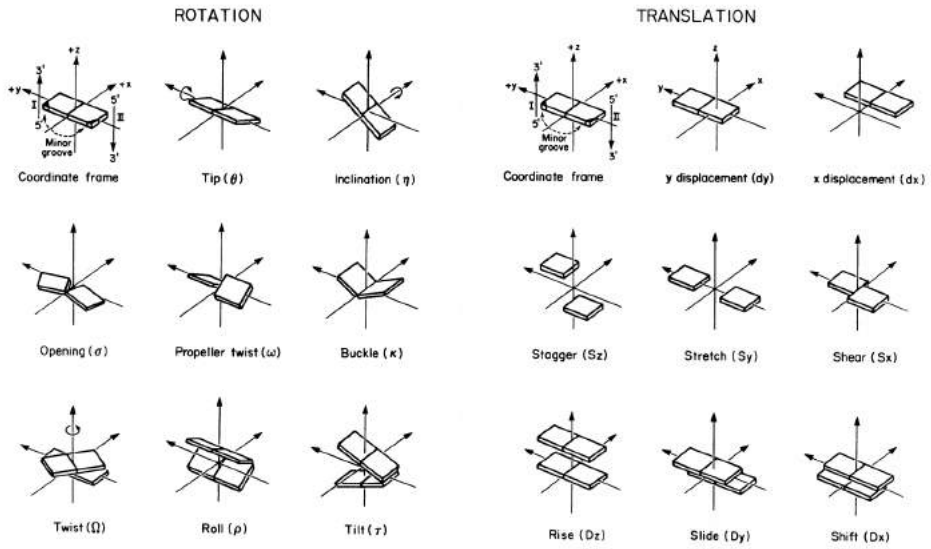


Figure 1.3: Standard definitions of various rotations and translations involving two bases of a pair or two successive base pairs (Image taken from [2]).

Moderate compression or extension of B-DNA thickens or narrows the double helix, with accompanying widening or narrowing of the minor groove and positive or negative inclination of base pairs. Detailed analysis of base-pair parameters can help to understand the exact mechanism of DNA over-stretching.

In the usual double-helical configuration, rotational and translational degrees of freedom are severely restricted. This means that parameters such as propeller twist, roll and tilt can change only within narrow ranges. The stretching of DNA can remove severe spatial constraints on rise-dependent orientation variables. As many of the steric clashes are removed upon DNA stretching, propeller twist, roll and tilt have more room to change and as a consequence the balance between van der Waals interactions and electrostatic interactions changes [30]. An analysis of over-stretched-DNA helical parameters is provided in the atomistic simulation section.

1.1.4 DNA torsion angles

Each nucleotide conformation in DNA is defined by seven torsion angles. Because covalent bonds are relatively stiff with respect to stretch, these an-

gles are commonly revealed as the important degrees of freedom particularly. Rotations around the ζ and ϵ torsions give rise to two different conformers called BI and BII where BI stands for ϵ/ζ in a trans/gauche conformation and BII for gauche/trans [31]. It is shown that BI/BII transitions are associated with base destacking [32].

Table 1.4: Nucleic acid backbone parameters, six main torsion angles (α, β, γ, δ, ϵ and ζ) around the covalent bonds defined in the figure along with the glycosidic bond χ..

Torsion angle	α	β	γ	δ	ϵ	ζ	χ
B-DNA	-30	136	31	143	-141	-161	-98

1.2 Hypotheses in relation to the Origin of life

1.2.1 RNA World

All forms of life with some minor exceptions share the same universal genetic code (UGC). It is widely believed that a "freeze" or arrest of changes in the genetic code at some point happened because any further change would impact so many coded proteins as to have lethal consequences. What we know today as the "central dogma of molecular biology" presents the flow of genetic information as in the order DNA $\rightarrow$ RNA $\rightarrow$ protein (Fig. 1.4).

Proteins are not the only molecules capable of having enzymatic action, as RNA can also form enzymes [33, 33]. It is widely believed that RNA was the first molecule familiar in modern organisms to be able to carry out self-replication and mutation and in this way evolve itself, what is called the RNA World hypothesis [34]. There is no strong evidence that RNA existed before DNA and there are serious concerns about the theory of RNA world (RNA is unstable in water above freezing due to a self-catalyzed cleavage reaction [35]) and it seems that the problems associated with this theory are far from being solved [36]. On the other hand the fact that the active site catalysing peptide bond formation lies within the core of a folded RNA enzyme ("ribozyme") does suggest the existence of an RNA world [37] or an RNA-peptide world.

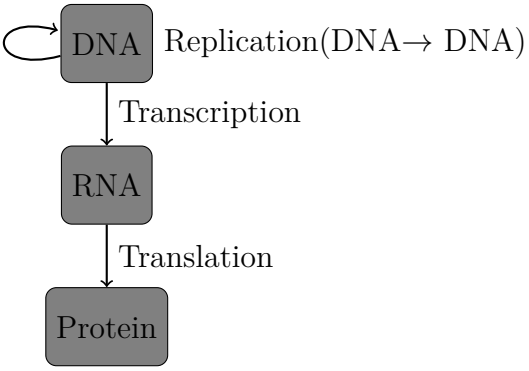


Figure 1.4: The graph shows the accepted 'central dogma of molecular biology' in which information is stored primarily in DNA, in the genetic code, which is then transferred to the RNA through the transcription process and finally manifested in the functional form, the protein molecule, through the translation mechanism.

Here we ask what was molecular biology at the dawn of life before it evolved to the sophisticated form which we know today? The RNA-protein world hypothesis can provide answers for some of the questions about the molecular biology of life at the beginning. For several reasons it is believed that RNA is evolutionary prior to DNA. First, building blocks of DNA (deoxyribonucletides) are made from RNA building blocks (ribonucleotides), second DNA replication is dependent on an RNA primer while RNA synthesis can start from a single ribonucleotide, and finally RNA synthesis machinery is less efficient than DNA (50 units per-second for RNA comparing to 500-1000 units for DNA, which provides the DNA with the advantage of fast replication).

RNA can cleave and join other RNA molecules, bind to co-factors and have enzymatic activity just like proteins, in fact it has been mentioned that many co-factors that are used by proteins are residual pieces of RNA [38]. RNA enzymatic activity creates a primitive self-replicating molecule which is able to increase the complexity of its own population and also of its chemical environment. By having autocatalytic intron[2] removal and reshuffling the exons create more complex molecules.

The main problem with the RNA world hypothesis is that prebiotic synthesis of nucleotides is difficult to reproduce under conditions which are believed to resemble those of the primitive earth, mainly due to the instability of nucleotides in water above 0°C. Another main concern which challenges the RNA world theory is the tenuously modelled chemistry for self-replication

[2]After RNA is transcribed the joined sequence is composed of introns and exons. In the process of maturation introns are removed by RNA splicing and exons are attached together to make the final copy of RNA.

of RNA. It was theorised for a long time that in early stages of life RNA was both the gene and the enzyme and was thus able to replicate itself. Recently it has been shown that a catalytic cycle of moderately complex RNA molecules *can* replicate itself and expand exponentially, even starting from micromolar concentrations in warm slightly salted water, which helps to give some credibility to the RNA world hypothesis [39]. Here we summarise the main objections to this hypothesis:

1. The complexity of the RNA makes it seem unlikely to arise prebiotically
2. RNA is inherently an unstable molecule
3. The catalytic abilities of RNA are too limited

A comprehensive review which addresses the challenges and counterarguments to the RNA world hypothesis is provided by Bernhardt [40].

1.2.2 RNA-Protein world

It has been suggested that the first proteins were short homo-oligo peptides possibly made of Lysine (because of its chemical simplicity) or Arginine/Ornithine in an abiotic process[3]. Short oligo-peptides of these residues can bind RNA by balancing its charge [41]. In the process of synthesizing the first oligo peptides those which can increase the fidelity of the process were enhanced. The fact that peptidyl transferase activity still resides in ribosomal RNA strongly implies that extension of protein chains is an RNA catalyzed process and not a protein-catalysed one, so RNA enzymology precedes protein enzymology. Based on the recent high resolution structures of ribosomes [42, 43] the core of protein synthesis machinery is composed of RNAs, and proteins decorate the outside of this core so clearly the ribosome is a ribozyme [44][4].

1.2.3 From RNA to DNA

In modern organisms DNA replication requires a plethora of proteins, strongly challenging the idea of the origin of life as "DNA". So the question here is how cell machinery has moved from RNA to DNA and why?

[3]It should be mentioned that it is believed Arginine was an evolutionary invader to usurp the Ornithine.

[4]For a recent comprehensive review reader is referred to [45].

It is believed that the first step in emergence of DNA was formation of U-DNA [46][5] and that later U was replaced by T (T-DNA). The next step should have been the appearance of enzymes which first could replicate DNA from RNA (called reverse transcriptases) and later DNA from DNA (DNA polymerases). In a general conception DNA replaced RNA as the genetic material because it is more stable than RNA and can be repaired more faithfully. The main contributor to DNA stability relative to RNA is the removal of the 2′ oxygen from the sugar moiety [47][6].

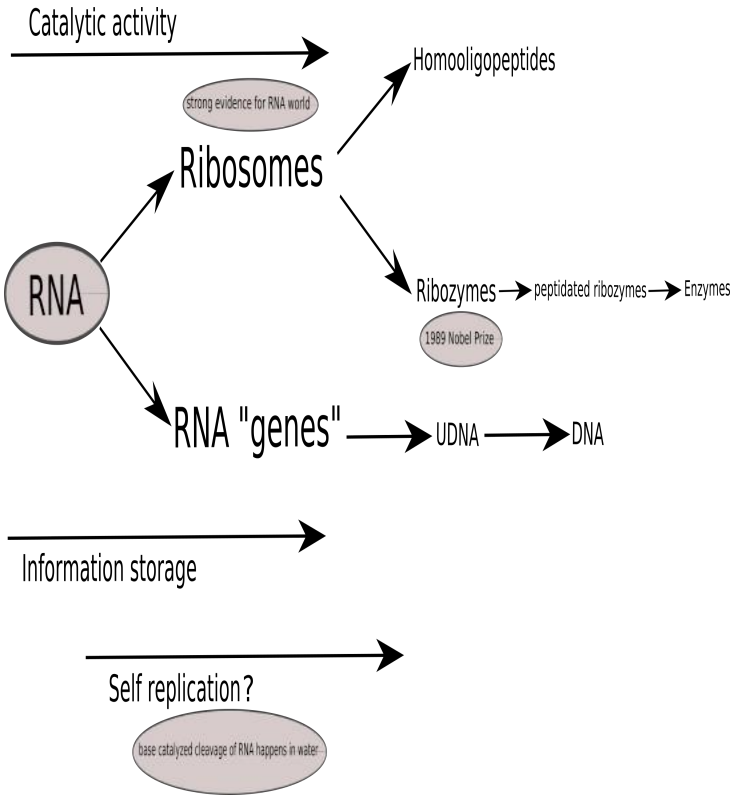


Figure 1.5: The evolution of genetic material. Supposedly the starting genetic material was RNA with the ability of catalytic activity which in modern biology is observed in ribosomes. It is also believed that the first peptides were synthesized from RNA in the form of homo-oligo peptides. RNA has also served as genetic material with the directional evolution to U-DNA and the modern DNA that we know today.

[5]This is the genome of some modern viruses.

[6]This reactive oxygen can attack the phosphodiester bond and this is the reason why RNA is so prone to strand breakage.

1.3 Evolution of the genetic code: functional orientation

We ask the question here: how was the genetic code structured in the beginning of life? Why is the genetic code in the triplet form (three base-pairs per codon) and not doublets or quadruplets? There are three groups of theories about the origin and evolution of the genetic code:

- Stereochemical theories
- Physicochemical and ambiguity reduction
- Coevolutionary theories

1.3.1 Stereochemical theories

Stereochemical theories state that the interaction between anticodons or codons and amino acids defines the origin of the genetic code [48, 49]. Experimental results have shown weak and relatively non-specific interactions between amino acids and their cognate triplets [50]. In modern biology this interaction is mediated by tRNA, the presence of which in early forms of life is highly arguable. A statistical analysis of the affinity of RNA aptamers for 8 amino acids (phenylalanine, isoleucine, histidine, leucine, glutamine, arginine, tryptophan, and tyrosine) [51] which provided evidence for this theory has been debunked by applying more rigorous statistical approaches [52].

1.3.2 Physicochemical theories

Codons for many amino acids differ by only one base from chemically similar amino acids. Physicochemical reasoning examines the selective pressure of reducing deleterious coding distance based effects between amino acids which are coded by codons differing only in one base [53]. The ambiguity reduction theory follows the same logic. It claims that those codons differing in only one base correspond to those amino acids which are physicochemicaly similar and evolution has pushed the genetic code towards a structure giving reduced possibility for dramatic mutations from a single base-pair change [54]. It is postulated for example that the minimisation of translation errors is the main theme dictating the code structure, a postulate which is argued against in

the coevolution view. The modern genetic code does seem to possess physicochemical optimality or near-optimality, however the level of optimality shown is not sufficient to uniquely determine it.

1.3.3 Co-evolution theories

The idea of co-evolution (of the genetic code together with the associated protein machinery) begins by observing that there is a biosynthetic relationship between amino acids such that there are some primary amino acids (perhaps five to ten) which emerge from simple chemistry, while other amino acids are produced in the modern cell from these precursor amino acids [3,55], requiring an expenditure of metabolic energy. Based on this theory the genetic code structure reflects the order of evolutionary appearance of amino acids, which is determined by their biosynthetic relationships: newer amino acids must 'steal' space in the genetic table, taking (less sterochemically favourable, therefore less-often used) codons which previously belonged to their chemical precursor so as to cause minimum disruption. The precursor-product relationships of amino acids are summarised in Fig. 1.6.

The co-evolution idea has received considerable corroboration within the literature [56–60]. This theory does not ignore the fact that physicochemical properties of amino acids are reflected in the genetic code but asserts that this is achieved without very strongly restricting the structure of the code [61,62]. The observation that a few position swaps of amino acids in the genetic code could have increased the code's optimisation level considerably leads us to this conclusion that physicochemical properties have been important but not fundamental [63].

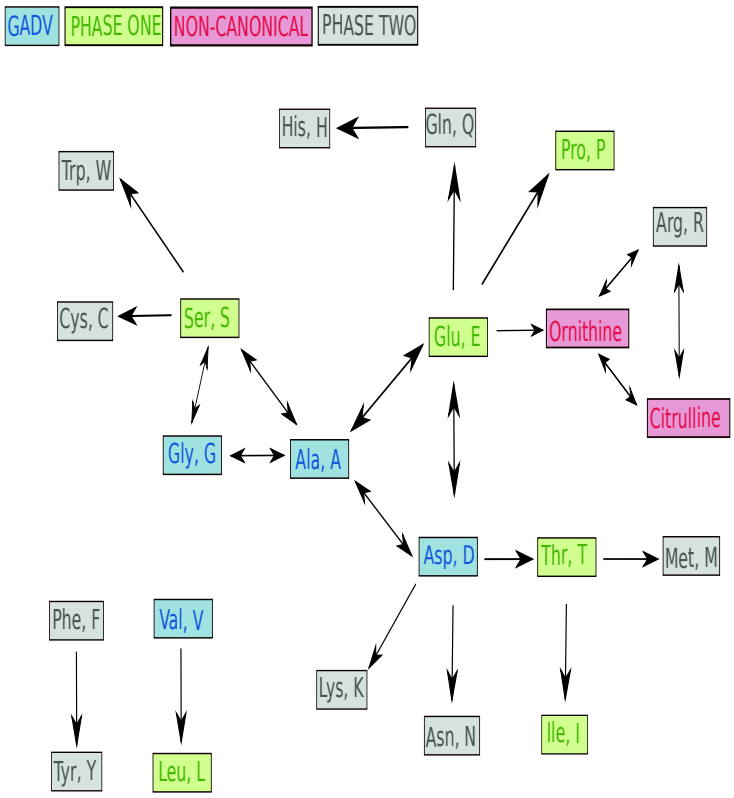


Figure 1.6: The coevolution theory states that the genetic code should reflect the biosynthetic relationship between amino acids. Here these relationships are summarized, skipping non-amino acid components of the network for the sake of simplicity, and also omitting non-canonical amino acids other than Ornithine and Citrulline. The GADV set is at the centre of the network, while the most complex amino acids are at the periphery and the ESPLIT set of intermediate-complexity amino acids are between the centre and the edge (*Graph based on Wong 1975*) [3].

1.3.4 Minimum functional genetic code: The GADV world

Following functional reasoning focused on the simplest useful system of amino acids, "SNS" and "GNC" theories arise [64]. In the SNS theory it is hypothesized that a functional early genetic code might have had the structure 'SNS' (where S (strong) signifies 'G or C' and and N signifies 'any base'), allowing 16 codons, comfortably specifying 10 amino acids with enough redundancy to meet physico-chemical restrictions which would be more important in a world with less efficient protein enzymes. By placing stronger chemical restrictions, and seeking the minimal code capable of coding for globular proteins a GNC model arises, having only four codons. In this theory the four amino acids

(Gly[G], Ala[A], Asp[D] and Val[V]) encoded by the GNC code are put forward as the minimal set for functional proteins. This GNC code draws the line of a minimal genetic code as the simplest one generating water soluble, foldable and functional proteins. Fig. 1.7 shows the supposed evolution of the genetic code[7].

[7]This criterion excludes theories which state that for example Ala coded by GCU (where U stands for Uracil) can be the minimal genetic code, because polyalanine is insoluble in water and does not form any known functional globular protein [65].

GNC primitive code

	U	C	A	G	
G	Val	Ala	Asp	Gly	C

GN[C/G] code

	U	C	A	G	
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	G

SNS primeval code

	U	C	A	G	
C	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	G
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	G

Modern genetic code

1st Base		2nd Base						3rd Base
	U		C		A		G	
U	UUU Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	UCC		UAC		UGC		C
	UUA	UCA		UAA	Stop	UGA	Stop	A
	UUG	UCG		UAG	Stop	UGG	Trp	G
C	CUU Leu	CCU		CAU	His	CGU		U
	CUC	CCC	Pro	CAC		CGC	Arg	C
	CUA	CCA		CAA	Gln	CGA		A
	CUG	CCG		CAG		CGG		G
A	AUU	ACU		AAU	Asn	AGU	Ser	U
	AUC Ile	ACC	Thr	AAC		AGC		C
	AUA	ACA		AAA	Lys	AGA	Arg	A
	AUG Met	ACG		AAG		AGG		G
G	GUU	GCU		GAU	Asp	GGU		U
	GUC Val	GCC	Ala	GAC		GGC	Gly	C
	GUA	GCA		GAA	Glu	GGA		A
	GUG	GCG		GAG		GGG		G

Figure 1.7: The supposed evolution of the genetic code. Here N stands for "anything". It is presumed that the primary genetic code was composed of four codons with the preference of starting with a "G" and ending with a "C". In the SNS theory "S" stands for a strong nucleotide (G or C). This "10 amino acid" stage of the evolution of amino acids is quite similar to Wong's "phase I" period. Highlighted amino acids show the ones in the GADV world hypothesis. At the end we have the modern genetic code.

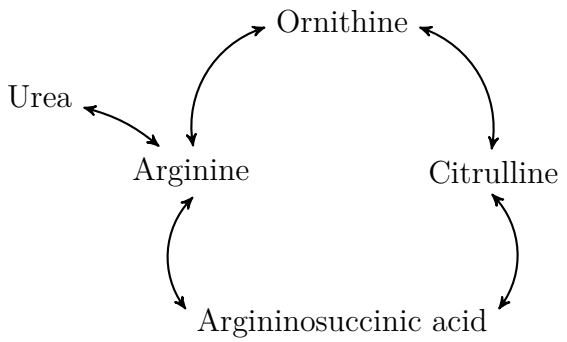


Figure 1.8: The urea cycle through which Arginine is synthesized (or broken down) in modern organisms. In normal conditions, metabolic energy is expended to generate Arginine from Ornithine and urea.

1.4 Arginine: the mysterious amino acid

An apparent anomaly in the UGC is that there are 6 codons (9.8% of the total genetic code) for Arginine while its frequency in modern proteins is usually less than 5% [66]. Therefore it is hypothesised that this structure is an evolutionary relic of some other, less metabolically expensive, amino acid which is no longer coded for directly. It has been suggested that Arginine became more common in modern organisms following the evolution of the urea cycle, and due to having a higher affinity for Ornithine tRNA, has replaced it in the codon table. Arginine has particularly important structural features: its positive charge and flexible chain allow electrostatic interactions with polyanions like DNA or RNA. The importance of this will be discussed in more detail in chapter 3. In the Arginine userpation hypothesis it is believed that in the early stages of life there was no need for positive amino acids with complex structure like Arginine or Lysine, and that Ornithine functioned as the only coded positive amino acid. As a stronger base, the replacement of Ornithine by Arginine allowed proteins to have more robust structure and a wider range of catalytic activity [67]. The helix propensity of Alanine/X-rich sequences where X is some basic amino acid increases in the order: propionic acid $<$ butyric acid $<$ ornithine $<$ lysine [68].

Although theories of the origin of life assumed that basic amino acids like Arginine and Lysine were among the least abundant amino acids, recently it has been shown that Arginine random synthesis is significantly enhanced by enviroments rich in cyanide and hydrogen sulfide [69, 70], which may be relevant to conditions on the early earth. Arginine functionality is also very

interesting as Arginine-rich oligopeptides can strongly bind to both DNA and RNA [71, 72], and proteins such as protamine[8] (which causes DNA condensation) are Arginine rich [73].

1.5 Part III: Mechanical properties of DNA

The molecular structure of DNA has been known for more than 60 years, but our knowledge about the balance of forces that determine its mechanical properties and the way it reacts to different imposed stresses, like overstretching and over-twisting, remains poor. The DNA double helix is among the stiffest of all biopolymers and responds to applied stresses in ways which are poorly represented by classical polymer theories.

DNA is one of the longest molecules in nature. A human chromosome for example is a few centimetres long. To squeeze such a lengthy molecule into a micron-size nucleus, DNA is bent and wrapped around histones, forming the bead-on-a-string structure of chromatin. DNA is typically found in a highly compact "supercoiled" configuration. The bending and torsional properties of DNA are essential to an understanding of its compactification in the nucleus.

DNA is composed of repeating structural units which consist of a ribose-phosphate to which four different groups can be linked: adenine(A), guanine(G), cytosine(C) or thymine (T). DNA differs from most polymers in that it is formed by the winding around each other of two ribose-phosphate polymer chains, interlocked by hydrogen bonds. This double helical structure prevents the relaxation of torsional stress by rotation about a single covalent bond as common with usual polymers. Moreover, the stacking of the bases on top of each other provides DNA with an unusually large flexional rigidity. It takes 50 times more energy to bend a double-stranded DNA (dsDNA) molecule into a circle than to perform the same operation on single stranded DNA (ssDNA). Moreover, the phosphates in DNA's backbone make it one of the most densely charged natural polymers known [74]. DNA-binding proteins can use the polymer's electrostatic potential to cling to DNA while they diffuse along the molecule.

Considering the central dogma of molecular biology which frames the flow

[8] Protamines are small, arginine-rich, nuclear proteins that replace histones in the haploid phase of spermatogenesis and are essential for denser packing of DNA. These proteins bind to the phosphate backbone of the DNA through the Arginine-rich domain and then DNA is folded into a toroid which is an O-shaped structure.

of information as from DNA to RNA and then to protein, the structure of DNA poses some formidable mechanical problems to the cellular machinery which has to read the genetic code buried inside the double helix. DNA polymerase enzymes need to have access to the bases and the only way is to unwind the DNA, separating two strands. Thus, as a polymerase enzyme proceeds along the molecule the DNA upstream of the transcription complex is over-wound, whereas downstream it is under-wound. The protein machinery has adapted to exploit the unique physical properties of DNA, functioning as motors capable of moving along torsionally constrained DNA molecules. Understanding of all these highly complex processes demands that first we understand the mechanical response of DNA under stress.

1.6 Elastic properties of DNA

Almost all tasks performed by DNA in the cell, including replication, transcription and various interaction with proteins are influenced by its elastic properties in relation to stretching, bending and twisting.

Following is the list of DNA elastic properties:

- Contour length L or L_c,
- End-to-end distance L_R, R_c,
- Cross-section dimensions: width b_i, height h_0, diameter d of the outer contour circumference, width b_i, height h_i of the inner contour,
- Cross-section area A_c, A_{csec},
- Helix parameter $r_{p0} = k_{p0} = 2\pi/S_0$, where S_0 is a helix pitch,
- Spring constants j, j_θ, j_s, pulling force F, F_q,
- Persistence length l_p, A_{bp},
- Kuhn segment $b_k = 2l_p$,
- Twist-stretch coupling g_p,
- Stretch modulus $E_{stm}, E_{str} = EA_c$, Young's modulus E, shear modulus G,
- Poisson's ratio ν_p,
- Specific gravity ρ_γ,
- Ultimate tensile strength(UTS) σ_t, yield stress σ_e, endurance limit σ_{en}.

At the molecular level, the Langevin force is due to thermal shocks from the environment. On the level of a nanometer size bead used in pulling experiments, these shocks correspond to a force of about 0.03 pN. Molecular motors in the cell can exert much larger forces ($f \approx 6\ pN$) in order to displace sub-micron-sizes objects. GC rich and AT rich sequences have different number of hydrogen bonds so they have different stickiness. GC rich regions require a force of 15 pN to separate strands, whereas AT rich sections require only 10 pN [75]. Large forces encountered at the molecular level, 1000 pN are associated with the rupture of covalent bonds. For biology and biotechnology the interesting range of forces is between 0.1 pN at which the molecule responds and 100 pN at which structural transitions happen. Mechanical force at the molecular level is involved in the action of many enzymes for tasks such as replication or transcription and generally for those which translocate with respect to DNA.

In many cases where DNA interacts with proteins, the double helix is severely deformed from the classical form by being bent, stretched and twisted. For example the bacterial protein RecA which has nonspecific binding to ds-DNA can lengthen the DNA by a factor of 1.5 [76]. Experiments have shown that RecA binds strongly to stretched DNA partly as a result of spontaneous thermal stretching fluctuations.

A continuous double helix has two helical curves inscribed on the same cylinder of radius R_a. A parametric description of these two helical curves is given by:

$$\begin{aligned} r_1(t) &= (R_a cost_1, R_a sint_1, r_s t_1) \\ r_2(t) &= (R_a cost_2, R_a sint_2, r_s(t_2 + \delta\psi) \end{aligned} \tag{1.1}$$

where r_s is the helix factor, and the angle ψ represents the angular shift between the two helical lines, in cut with a plane perpendicular to the central axis of the helices. If this angle is different from π, then the two helices make an asymmetric pattern with a well-defined major and minor grooves, meaning that in the direction of the long axis of the helix with radius R_a the separation of helical lines is uneven. This unevenness of separation has an important meaning for the structural properties of the double helical DNA.

1.6.1 How stiff is DNA

The WLC model [77] predicts that DNA has a local stiffness but a global flexibility. So, over nanometer length scales DNA is among the stiffest biopoly-

mers. It is not exactly clear why DNA is so stiff and this question remains controversial. It is assumed that favorable base pair stacking interactions or backbone phosphate charge repulsions are the major contributors.

Local and long-range electrostatic effects have different roles in DNA structure and stiffness. Based on the theory of Manning [14], the high negative charge density of DNA creates a layer of mobile and hydrated counterions along the DNA surface. The valence of the counterions determines the extent of neutralization of DNA charge, more than the bulk counterion concentration. The enthalpic attraction of the counterion to the DNA balances the entropic cost of localizing the ion even at very low bulk concentration. ξ, called the linear charge density parameter governs counterion binding to DNA [78]:

$$\xi \equiv \frac{e^2}{\epsilon k_B T b} \tag{1.2}$$

where e is the charge on an electron, ϵ is the bulk dielectric constant of the solvent and b is the average axial charge spacing of the DNA. The quantity $e^2/(\epsilon k_B T) = \ell_B$ is the Bjerrum length for the pure solvent, the separation distance at which the electrostatic interaction between two elementary charges (e) is comparable in magnitude to the thermal energy scale ($k_B T$).

Manning theory is expanded to include the contribution of phosphate charge to DNA stiffness [79]. In this model, bare DNA charge produces an electrostatic stretching force on DNA due to phosphate-phosphate repulsion. In this context the term "null" DNA was coined to refer to a structure of DNA with no electrostatic charge. This model for null DNA consists of fully charged DNA supplemented by an applied compression force that is equal but oppositely directed to the internal electrostatic tension present in charged DNA. This model relates the DNA persistence length P and the persistence length of the null isomer P^*, requiring also the Debye screening length κ^{-1} for the solution, the polymer radius R, the charge spacing b (inverse of the linear charge density) and the valnce Z of the counterions. [14]:

$$P = \left(\frac{\pi}{2}\right)^{2/3} R^{4/3} (P^*)^{2/3} Z^{-2} \ell_B^{-1} \left[(2Z\xi - 1) \frac{\kappa b e^{-kb}}{1 - e^{-\kappa b}} - 1 - \ln(1 - e^{-\kappa b}) \right] \tag{1.3}$$

The important result of this theory is that the relationship between the bending persistence length of DNA and the persistence length of its null isomer is multiplicative and not additive. The model also shows how the electrostatic properties of the solution govern the DNA persistence length.

While Manning theory gives a useful overview, there is significant controversy in explaining and discussing the stiffness of DNA. There is experimental evidence that DNA stiffness is dominated by base stacking rather than charge repulsion [80, 81]. Thus, balance of the forces responsible for the bending stiffness of DNA is not clearly understood. Understanding the source of these forces would allow us to engineer DNA-like polymers with altered stiffness. Besides such an understanding would shed light on how DNA compactification happens in eukaryotic DNA.

1.6.2 DNA contour length and persistence length

The persistence length provides useful information about structural rigidity of a polymer and the energetic cost of deforming it [82]. Lots of information is currently available about the persistence length of DNA, because of the importance of understanding DNA bendability [83] and because information on this lengthscale is available *via* microscopy and other relatively basic experimental techniques. It is believed that ∼50 nm is the smallest value to which the persistence length can be pushed to by neutralizing phosphate repulsions in DNA. Various methods provide different values for the persistence length of DNA, 45-53 nm (132-156 bp) and it has been shown that is largely independent of monovalent salt concentration above 20 mM [84]. It should be mentioned that divalent ions can reduce the persistence length of DNA below 50 nm [83, 85].

Under thermal excitations DNA can bend, twist or stretch. Based on the WLC model directional changes in the chain contour length cost energy which is quadratically dependent on bending angle. Above the persistence length the chain directional correlation becomes negligible [86]. It is indicated that the apparent stiffness of DNA is lower in living cells than would be predicted from *in vitro* measurements. Softening of DNA with respect to both bending and twisting is observed *in vivo* [87]. Some early studies suggested enhanced DNA softness to bending (7-fold) and twisting (2-fold) relative to naked DNA studies in solution [88]. Recent statistical mechanical studies of DNA bending and DNA looping induced by proteins revise these estimates to 1.6 fold and 3-fold [89]. In eukaryotic DNA *in vitro* and *in vivo* studies suggest a 2-fold reduction in the DNA bending persistence length from 50 to 27 nm [90].

It is hypothesized that cellular DNA in general displays unexpected flexibility because of negative supercoiling and DNA bending proteins. It has been proposed that DNA bending proteins can decrease the apparent persistence length of DNA by introducing random sites of bending and kinking [91].

These groups of proteins called architectural proteins have a definite role in enhancing the flexibility of DNA. One striking finding about these groups of proteins is that expression of structurally unrelated eukaryotic HMGB[9] proteins can revive DNA looping in *E. coli* cells lacking HU proteins[10] which are just found in prokaryotes [92]. Results of this kind suggest a fundamental mechanism for architectural proteins altering DNA *in vivo*.

1.6.3 DNA elasticity theory

The elastic stiffness of DNA can be parametrized by its contour length, L_0, persistence length L_P and elastic modulus K_0. When the length of a polymer is much longer than L_P entropic considerations dominate because of the numerous configurations that polymer may adopt. As previously mentioned these two parameters are interrelated in the WLC model. A long enough linear DNA is a flexible polymer with end-to-end mean squared distance $R_0 = (bL_0)^{1/2}$, here b is the Kuhn monomer size. The bending costs an energy per length of $k_BTA\kappa^2/2$ where $\kappa = |\partial_s^2\mathbf{r}|$ here is the curvature (the reciprocal of bending radius) and where A is the characteristic length over which a bend can be made with energy cost k_BT.

Separation of the ends of a DNA by an amount $z \ll L_0$ costs free energy $F = 3k_BTz^2/(2R_0^2)$ and requires a force $f = \partial F/\partial z = 3k_BT_z/(2AL)$. Below the characteristic force of k_BT/A, the extension z is small compared to L_0 and the linear force law is valid. Considering $1k_BT/nm = 4.1(pN)$, for DNA persistence length of 50 nm, $k_BT/A = 0.08\,pN$, so the force needed to extend DNA within the entropic regime is small.

1.7 DNA over-stretching

The development of single-molecule techniques has provided the ability to study the mechanical properties of individual macro-molecules. Optical and magnetic tweezers are the two most important techniques used in single molecule studies of DNA.

[9]These proteins are members of the high mobility group (HMG) superfamily with a unique DNA binding domain which can bind non-B-type DNA structures (like bent or kinked). Their actionas DNA chaperones influences processes in chromatin such as transcription and replication.

[10]This is a histone-like protein in bacteria. Its main function is to inhibit DNA supercoiling and to regulate DNA replication process.

Optical tweezers operate by harnessing the momentum carried by photons. In this technique ones can apply force and measure tension. Optical tweezers cover the intermediate force range from 0.1 pN to 1 nN [93] and provide angstrom position resolution and millisecond time resolution. The very first force-extension (F-X) relationships were measured for single molecules of DNA with this technique.

Magnetic tweezers, in comparison to atomic force microscopy (AFM) and optical tweezers, have many advantages which include no heating, throughput and force stability [94]. Force range of magnetic tweezers for short tethers ($< 1\ \mu m$) is ∼1-100 pN [94]. The advent of novel kinds of magnetic tweezers also allows direct measurements of torque and twist in DNA.

1.7.1 Biological implications of DNA stretching

Homologous recombination is a process in the cell which includes exchanging of DNA strands and is important for DNA repair and recombination. In prokaryotes this process involves the formation of long helical filaments of the RecA protein on DNA. These proteins cause large deformations in the double helix, extension by 50% and unwinding by 40% with respect to B-DNA relaxed structure. In this process DNA undergoes a series of structural changes that results in locally destabilizing the DNA [95].

Stretching of DNA will affect its structural features like helical pitch and backbone conformation. It also changes the conformation of base pairs. In normal DNA (B-DNA) rotational and translational degrees of freedom are restricted but as the double helix is stretched some of these stringent spatial constraints are removed. Stretching also reduces the clashes between base pairs and causes changes in orientation variables at base steps like propeller twist, roll or tilt[11]. All these conformational changes have a deep effect on DNA interaction with proteins, ligands and other macromolecules.

It has been shown that twist is one of the base-pair parameters which is most affected by stretch and it is a variable that most influences backbone conformation [96]. Most variables of DNA conformation initially show more flexibility as the rise increases, due to unstacking of bases, but fluctuation of

[11] In total there are six intra base-pair degrees of freedom of which Shear, Stretch and Stagger are translations around the X-, Y- and Z-axis respectively and Buckle, Propeller and Opening are rotations around the X-, Y- and Z-axis. There are also six inter base-pair degrees of freedom of which Shift, Slide and Rise are translations around the X-, Y- and Z-axis and Tilt, Roll and Twist are rotations around the X-, Y- and Z-axis.

these variables decreases again as the rise is increased further.

All these structural changes could have important biological relevance. For example Lebrun *et al.* have suggested that amino acid side chains of TATA-box binding protein are intercalated between base-pairs [97]. It has also been discussed that sign reversal of the propeller twist would change the position and orientation of H-bonding groups on the edges of bases, which consequently would affect the interactions with amino acids [96]. DNA stretching in chromatin facilitates its compaction and would influence site recognition by nuclear factors. Crystal structure analysis showed that 'normal' stretching occurs at the equivalent of one to two base-pairs per nucleosome which consists of 146 bp.

1.7.2 Different scenarios of DNA over-stretching

One of the most controversial topics in DNA force spectroscopy is the drastic increase in length which happens at a force of ∼65-70 pN. This phenomenon can equivalently be discussed as the formation of a plateau in the force-extension curve. In search of theories to explain this plateau, two distinct pictures arise. The first is a thermodynamic theory which argues base separation in the regions of DNA with low GC content, such that DNA over-stretching amounts to a kind of DNA melting [98, 99]. The second theory proposes that during the DNA over-stretching it is the base-stack only which is disrupted and base-pairing remains intact or almost-intact as B-DNA transforms to a new ordered or mostly-ordered form of DNA called "S" [100–103]. In a cycle of stretching and shortening hysteresis happens. Possible explanations for hysteresis in such a system provides a discriminating test of the two scenarios presented for overstretched DNA [104]. It is hypothesized that the experimentally observed asymmetry in the hysteresis could be explained only when S-DNA formation is allowed [104]. There are both experimental and theoretical sets of evidence support each proposition. In the following sections each model will be discussed in more detail.

Table 1.5: Summary of different theories concerning the behaviour of over-stretched DNA.

Theory	Explanation	Base stacking vs. pairing	Final conformation	Ref.
Thermodynamic theory	Force induced melting (FID) of DNA	Disruption or change of base-pairing	ssDNA	Rouzina *et al.* [105]
Structural transition	Some 1st-order phase transition	Disruption of base-stacking	S-DNA	Lebrun *et al.* [97]
Mixed conformations	FID and phase transition	Disruption of base stacking and pairing	ssDNA,S-DNA,B-DNA	Zhang *et al.* [106]
Alternative structured theory	Phase transition	Interdigitation of base-pairs	Zip-DNA	Balaeff *et al.* [6]
Σ-DNA theory	**Planar 2/3 stacked base-pairs**	**Triplet base-stacks with a big gap in between triplets**	**Σ-DNA**	**Current work**

An issue of concern here is the methods used to study DNA over-stretching. Different kinds of ligands which have preferential binding to ssDNA or ds-DNA are used in the experiments and it has been shown that base pairing is broken in the presence of many or all ssDNA binding ligands. Those experiments which try to represent a clear picture of DNA melting during over-stretching utilize fluorescent binding ligands [107, 108] which may be an intrinsic flaw of all studies of this kind. It is discussed that ssDNA binding ligands and proteins may bias the over-stretching transition from S-DNA towards melting by stabilizing the ssDNA [109]. There is an opposing argument that states most of these ligands or proteins bind ssDNA slowly and in consequence are unable to affect the DNA melting equilibrium [110, 111]. Table 1.5 summarizes different views on DNA overstretching conformational transitions.

1.7.3 Stretched or S-DNA

Different experimental groups have shown that when DNA is placed under tension of more than 65 pN, it transforms from its B-form to a new form or forms with extension approximately 1.7 times the B-form length [75,101,112]. This extended DNA is called "S-DNA". The Helix radius of S-DNA is shorter than B-DNA and bases are sometimes tilted [101] or remain almost untilted [113]. Tilting of the bases gives rise to a ladder-like structure in the S state of the DNA [114]. An estimation of the DNA twist stiffness measurements suggest that a not completely ladder-like form of the DNA occurs in the overstretching transition from B- to S-form. A remnant helicity of the DNA persists in its over-stretched state which is about one turn every 37.0 base pairs [115].

Theoretical analysis of DNA stretching data from some of the experiments has led to the conclusion that under physiological conditions (pH 7.5 with 150 mM Na^+) the force-extension curves of overstretched DNA could not be explained based on the formation of ssDNA [104, 116]. Analysis of the competition between force-driven formation of the S-DNA and ss-DNA ("unpeeling") showed that S-DNA and unpeeling require similar free energies and are likely to compete near 65 pN. So, it was concluded that factors that affect base pair stability could potentially determine whether the transition leads to formation of S-DNA or strand unpeeling. Particularly it is predicted that high salt concentrations and low temperatures favour S-DNA [103]. A thorough analysis of the effects of environment (salt, temperature and sequence composition) on DNA over-stretching in the absence of DNA-binding ligands or proteins has shown that for a DNA with free ends or nicks, there are two competing modes of over-stretching: (1) formation of a stretched double-stranded structure (S-DNA), or alternately unpeeling of one of the strands [117]. It should be stressed that formation of the unpeeled state is crucially dependent on the presence of a nick along the DNA, without a nick no unpeeling can occur due to topological constraints.

Aside from melting, numerical simulations of DNA over-stretching typically lead to one of two characteristic conformations, either a narrow helix with highly inclined base pairs or a flat unwound ribbon (eg. [118]). As mentioned the essential feature of S-DNA is that bases remain paired. An elegant experiment which excludes the discussed problematic effects of DNA-binding dyes [119] (*by Maaloum et al.*) has shown that the transition observed in DNA stretching is a pseudo first-order transition from B-DNA to some S-DNA where the helix diameter changes from 2.4 nm to 1.2 nm with

a ∼65% reduction in the number of turns. Maaloum *et al.* also show that stretched DNA molecule in their system is not melted into two separated strands. This work also relates the observation of ssDNA to the utilization of ssDNA-binding ligands or dyes and stresses that this kind of transition depends strongly on the ligands present.

It is interesting that even before the structure of DNA was discovered by X-ray crystallogarphy, Wilkins *et al.* suggested from an experiment that stretched DNA undergoes a transition to a structure with tilted bases which is two times longer than the relaxed molecule [120][12]. In this picture DNA still remains in a helix form but is extremely extended.

1.7.3.1 DNA topology and torsional constraints

There are three mathematical quantities which are basic to DNA topology: twist (Tw), writhe (Wr) and linking number (Lk). Twist represents the total number of double helical turns in a segment of DNA. Writhe is a property of the spatial course of the DNA and is defined as the number of times the double helix crosses itself if the molecule is projected in two dimensions. The helix-helix crossovers are assigned a positive or negative value based on the orientation of the DNA axes. The numerical term that describes the sum of the twist and writhe is the linking number:

$$Lk = Tw + Wr \tag{1.4}$$

Under torsional constraint, the sum of the twist and the writhe of the DNA molecule is a constant. The torsionally unconstrained DNA has the ability to change its twist so the linking number can change. In Fig. 1.9 it is shown how change in twist can change the linking number of DNA. In a torsionally relaxed DNA with closed ends, the structural transition due to the increasing applied force results in the formation of an underwound DNA which has ~ 37.5 bp per turn.

[12]In this study stretching of sodium thymonucleate fibres was studied by streaming solutions. At 50% humidity conditions these fibers are partly crystalline and birefringence and ultra-violet dichroism are all negative. In contrast the stretched DNA is positively birefringent and is non-crystalline and non-dichroic. It is interesting that this fiber is stable at 50% humidity but returns to the negative form when placed in a more humid place and the length shrinks from 1.5 to 1. They conclude from optical studies that purine and pyrimidine rings rotate during the extension and lie on average at about 45° to the axis if the fibre, this means a 45° tilting.

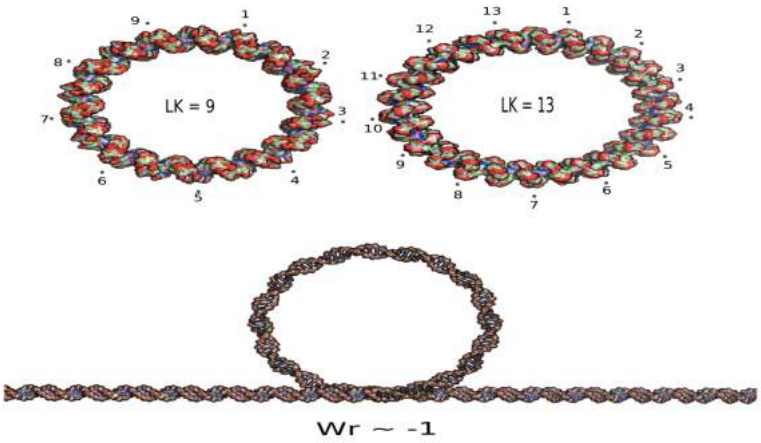


Figure 1.9: DNA can be untwisted and opened up by reducing the linking number and in this way relieve torsional tension. A circular DNA is shown with two different linking numbers imposed. Each dot represents a full turn of DNA which based on standard definitions is 10.5 base pairs. Writhe is a more difficult quantity that describes the amount of coiling of a closed curve (in this case DNA) in three dimensional space. An informal definition of writhe is as a positive or negative integer: if two strands cross and the strand underneath goes from right to left, the the writhe is positive but if the lower strand goes from left to right the writhe is negative. This definition unfortunately depends on the chosen projection, so should in theory be averaged over all projections. In practice writhe is most easily available via eqn. 1.4. *DNA models were made with NAB* [4].

1.7.4 Force-induced DNA melting

As mentioned earlier at a force of about 65 pN, dsDNA elongates to about 1.7 times the normal B-DNA contour length. In an alternative scenario to the formation of S-DNA it is speculated that dsDNA is converted to ssDNA (or melted) in the course of the over-stretching transition [121]. The peeling of one strand from its complementary strand during over-stretching is visualized by single molecule fluorescence imaging [105,108]. Pioneers of this theory have stated that force induced DNA melting can essentially explain all of the phenomena associated with the over-stretching transition [98]. However, a force-induced melting mechanism cannot explain some observations in different experiments. For example, the force response of over-stretched DNA is inconsistent with that of one ssDNA or two non-interacting ssDNA, which implies that there is no complete dissociation of two strands, a fact which is observed [119] experimentally in AFM studies of DNA over-stretching. In addition the observed fact that extended DNA immediately returns to the B-DNA with relatively modest hysteresis upon retraction implies a non-melting mechanism [103, 122–124]. Although it should be mentioned that the occurrence of hysteresis depends on solution conditions such as ion strength [112].

Summarizing the results of more than 20 years of debate we can say that if the DNA has low base-pair stability (high AT content, low salt or high temperature) peeled ssDNA is selected for torsion-unconstrained topologies, while bubble is selected for torsion-unconstrained end-closed topologies [123]. If the sequence has high base-pair stability, S-DNA is selected for both end-opened and end-closed DNA constructs. Recently it has been shown that DNA duplexes as short as 60 base-pairs can be over-stretched by 51% without strand separation [124].

1.7.5 Beyond S: Zip-DNA

S-DNA or melted DNA are not the only two documented phases of extended DNA. At very high forces, a novel zipper-like structure can manifest if melting is bypassed [125]. In zipper DNA, base-pairs are broken and the nucleobases from the opposite strands interdigitate forming a continuous aromatic stack. In a further simulation study it was shown that when large forces are applied quickly, zip-DNA self-assembles from force melted DNA, whereas when forces are applied more gradually zip-DNA is formed by passing through S-DNA as an intermediate [6].

1.7.6 Different pulling schemes

DNA can be stretched either from its $3^{'}$ or $5^{'}$ ends or from both ends and its behaviour is different in each case. Experimental results on $5^{'}$ vs $3^{'}$ stretching modes are different and inconsistent. Some experiments show no difference in rupture profile and emphasize a symmetrical stretch [126] while the most recent one shows DNA behaviour upon stretching from $3^{'}$ or $5^{'}$ to have significant differences [127][13]. Seemingly the main variable is the pattern of transmission of the stretching force around the double helix as the tension is increased gradually during the experiment or simulation.

[13]Theoretical results of Lebrun and Lavery [100] show an energetic preference for the $5^{'}$ fiber structure over the $3^{'}$ ribbon structure which might be due to the simplicity of the distance-dependent dielectric solvent model which they have used [118].

1.7.7 Concluding remarks: FIM or S-DNA?

Authors in favor of FIM (force induced melting) model present this model as the only valid model which can explain over-stretched DNA structure and dismiss the S-DNA model. A recent review on FIM gives an overview of the problem and the proposed model but does not include the critiques of FIM existing in the literature [128].

There are at least two problems with FIM theory. First is that the force-extension measurements done on ssDNA do not match those of over-stretched DNA. Second is that the extension of two separated strands to the full length of overstretched DNA requires ∼130 pN force, much larger than those observed at transition. So there are two scenarios for for the over-stretched DNA: one of S-DNA and another one of FIM. S-DNA is in most cases predicted to be thermodynamically slightly more stable than FIM. Therefore FIM should happen primarily in AT-rich regions or at DNA ends or nicks, but might also coexist as a minority phase alongside S-DNA, given that the thermodynamic penalty for melting out of S is anyway not large. As previously mentioned, usage of ssDNA binding proteins would stabilize the melted regions and lead to the conclusion that FIM is the preferred thermodynamic pathway in the presence of these proteins. The same concerns are raised in the usage of glyoxal which react with open GC base-pairs and stabilizes them in the open state [110]. Overall, stabilizing melted regions of DNA could bias the overstretching pathway from S-DNA towards FIM.

There are some experimental results of over-stretching DNA under twist which provide a simple explanation within a combination of S-, P-, B- and Z-DNA forms and it is not clear how these data can fit into the FIM model [115, 129].

In conclusion, there is experimental evidence supporting S-DNA model which contradict FIM, and vice versa. Here we try to resolve this issue by presenting new simulations and a new model for over-stretched DNA (chapters 3 and 4), discuss the role of intercalators in this process, and provide some observations relating to the role of DNA stretching in early biology.

1.8 DNA Intercalation

DNA intercalators are a group of molecules mostly containing aromatic heterocycles. These ligands intercalate between base pairs of DNA. Intercala-

tors are typically flat, but it has been shown that some biologically active compounds with greater molecular thickness like steroid hormones could fit stereospecifically between base pairs [130]. Computer modeling shows that goodness of fit of certain small molecules into DNA intercalation sites correlates with the degree of biological activity but not with strength of receptor binding. The importance of these findings is that specific sequences in DNA into which ligands best intercalate are found in the consensus sequences of genes activated by nuclear receptors, implying that intercalation is central to the mode of action of these ligands [131].

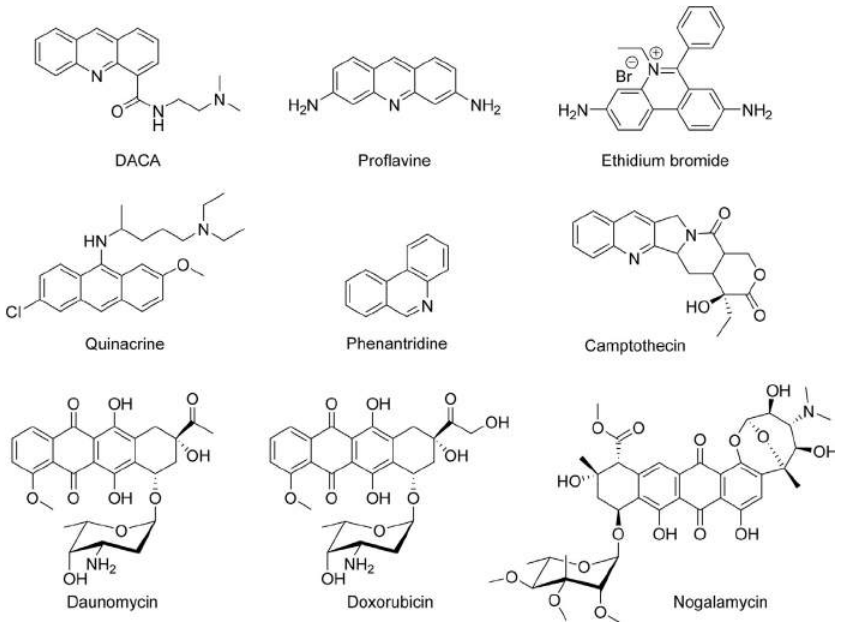


Figure 1.10: Examples of classical intercalators. (Image taken from [5]).

There are certain carcinogens like benzopyrenes which are aromatic, flat molecules with the same thickness as base-pairs, which intercalate. Intercalation is also the mode of action of anticancer drugs like actinomycin D [132] and topotecan [133].

One of the conformational changes of DNA is unwinding in which base-pairs separate and form a cavity. The shape and size of the cavity is dependent upon the degree of unwinding. When the cavity reaches the approximate width of 4 Å it becomes possible to insert flat, aromatic moieties in between the bases.

1.8.1 DNA-RecA and DNA-RAD51 interaction

In the cell, homologous recombination is a process which consists of exchanging DNA strands, creating a pair of duplexes which mostly match (are homologous) but may contain some mismatches. This process is important for DNA repair and for creating genetic diversity in sexual reproduction [134]. RecA, the main protein involved in this process (in bacteria) operates by creating a filamentous structure around the DNA. Within the active filament, the DNA is stretched by ∼50% with respect to B-DNA [135] and unwound by 15° per base-pair step, ∼40% less than its standard helical twist [136]. This is an unusual conformation for DNA which can not be attained without the help of an external force. In this case these forces are provided by interactions with the RecA protein. So what is the mechanistic reason for the DNA to be stretched and unwound to this degree?

Linear dichroism (LD) has shown that bases of DNA, in this filament, are on average roughly perpendicular to the filament axis [137] which is seen in DNA-RecA complex (Fig. 1.11). Structural studies of a tetranucleotide DNA strand bound to a RecA dimer with NMR show an average axial base separation of 5.1 Å which is stabilized by stacking interactions between sugar $C2'$ methylene groups and the following $3'$ base [138].

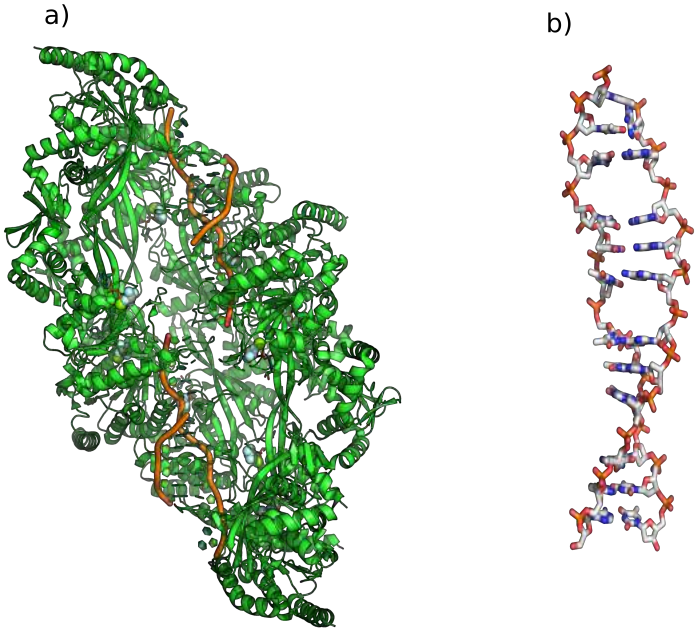


Figure 1.11: Figure shows the RecA-DNA complex (a) and separated DNA (b) (pdb:3cmt). Complexed DNA with RecA is stretched in a unique way producing triplets (b) in which base-pairs remain perpendicular relative to the helix axis.

As previously mentioned when DNA is stretched mechanically it elongates 70% with respect to B-DNA in contrast to the RecA-stretched form of DNA which is elongated by 50%. Both forms are unwound, although to a different extent. Experimental studies of DNA over-stretching have measured a helicity of 38 bp/turn for S-DNA which corresponds to a pitch of 220 Å. These numbers for the RecA-DNA complex are 18.3 bp/turn and ∼100 Å. Therefore, RecA-stretched DNA is likely to be a sort of virtual or real intermediate between B-DNA and S-DNA. It has been shown that RecA polymerizes at least 20 times faster on a double-stranded DNA in the S-form than on B-form [76, 139, 140]. It should be mentioned that initiation of RecA polymerization did not occur within 10 min of RecA addition in the absence of external force, but took place in a few seconds when a 65 pN stretching force was applied [141].

Regarding the structural changes induced in DNA by RecA filaments, it is important to mention that protein filaments constitute a dielectric medium lower than that of bulk water, so the DNA side which is in contact with protein would experience a different medium than the opposite side which is contact with water. By pulling the DNA at its $3'$ by a factor of 1.5 the same

unwinding value as that observed in the RecA filament can be achieved [142]. It should be mentioned that only at this exact degree of stretching DNA bases are perpendicular to the helix axis.

It has been shown recently that human RAD51 protein, a eukaryotic homologue of Escherichia coli RecA, as in the case of RecA, engages ssDNA and dsDNA in nucleotide triplet clusters [143].

We remark that intercalation-like gaps of a similar width have also been inferred following a result of force-induced transition in cross-linked DNA films [144].

1.9 Stretching forces in nature

The most obvious example of a stretching force being directly applied to DNA in the cell occurs during cell division. The forces exerted on chromosomes during prometaphase have been measured in grasshopper spermatocytes [145] and vary from 0.3 to 0.8 nN. Since these forces are sufficient to cause a helix to ladder transition in free DNA, it is possible that such a structural transition occurs, particularly near the centromere region, and that the altered form of the DNA plays a role in control of cell division. Hydrodynamic forces are also supposed to act upon DNA and cause DNA stretching when in a geometry of 'extensional flow' [146] which could be supposed as the primary source of DNA stretching in the dawn of life when no complicated enzymes like those in modern biology existed. DNA extension in micro-channels in experimental set ups within the range of relative extension of $\sim$1.5 is observed in various studies [146, 147]. These experiments show DNA molecules which are in high velocity gradients are stretched from the coiled conformation and aligned with the flow direction [148]. The degree of stretching and the experienced drag force depends on the initial position of the DNA molecule within the fluid. It is interesting that the mean normalized value of extension L/L_0 reaches a maximum at 1.52, in agreement with the value of extended DNA within the RecA protein filaments [147].

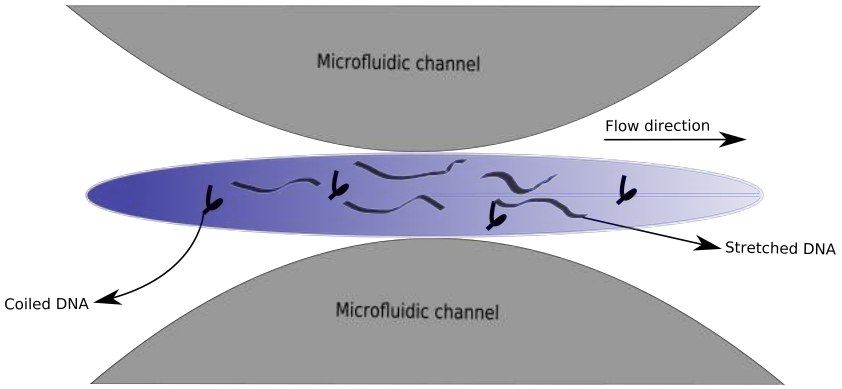


Figure 1.12: Figure shows a graphic representation of an experimental system to study flow extension in DNA dilute solution. Some DNA is aligned in the direction of the flow and stretches while some molecules might remain in the coiled conformation.

1.10 Axial distribution of DNA over-stretch deformations, formation of triplets

The distribution of the elongation in a stretched DNA is a complicated process. If we assume that B-DNA is stretched uniformly by 50% and the bases remain perpendicular to the axis this would produce a 5.1 Å base separation and in this way stacking interactions would be completely disrupted [149]. The loss of stacking energy is not favourable for DNA. If we assume that the stretch is distributed uniformly along the DNA, a 1.7 Å space would be created in between each base-pair which is not enough to even accommodate a water molecule. If we assume that the total stretch is distributed non-uniformly along the DNA axis, in a way that stacking is partially conserved and in-between is disrupted completely, we would be able to propose a mechanism for DNA elongation and its biological activity of elongation as well. In the case of RecA, base separation is non-uniform with consecutive groups of stacked base-pairs in triplet clusters followed by a big gap which reaches 8 Å.

Chapter 2

Results I

2.1 Triplet Propensity Calculation

As mentioned in the introduction, stacking interactions play an important role in stabilizing DNA. When DNA is overstretched by applying an external force, equilibrium rise value (2.8-3.4 Å) starts to increase until after a certain value (5.6 Å [150]) the stacking interaction is said to be broken. The thermodynamic stability of DNA depends on the extent to which bases stack upon each other. Stacking interactions are sequence dependent [9]. It is believed that electron correlation interactions are a significant source of stacking energies in DNA so different combinations of purines and pyrimidines (with different electronic properties) should result in significantly different stacking contributions to DNA stability. If we assume for example a a dinucleotide of A·T[1] whether this dinucleotide is flanked by purine·purine, purine·pyrimidine, pyrimidine·pyrimidine or pyrimidine·purine stacking interactions would be different and so would be the stabilization.

In this section we investigate, based on literature datasets for the sequence-dependent dinucleotide stacking free energy, the free energy associated with partitioning a DNA duplex into triplet structure with stack breaks in a regular pattern of period three.

[1]Dot indicates base-paired dinucleotide.

2.1.1 Stacking interaction

Stacking involves hydrophobic, electrostatic and dispersion components [151, 152]. It is still a controversial field of study and there is no agreement between researchers which force is dominant in stacking. For example Luo *et al.* believe nonelectrostatic interactions are dominant forces [151] while McKay *et al.* emphasize the role of attractive interactions between partial charges of bases [153]. Guckian *et al.* discuss that hydrophobic effects dominate the stacking interaction but dispersion forces and electrostatic interactions contribute substantially to the base stacking [154]. All four bases have significant charge localization and distinctive electrostatic fingerprints.

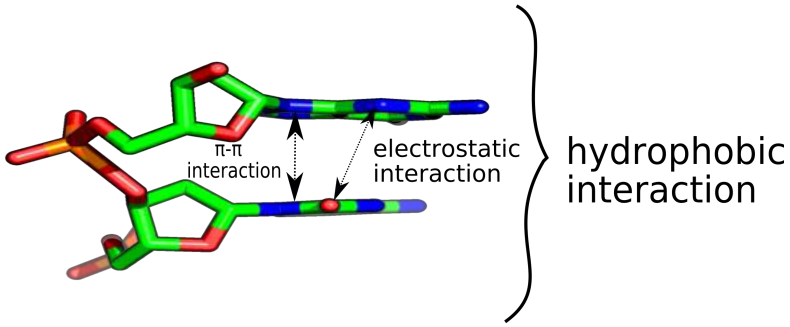


Figure 2.1: Stacking interaction between two nucleotides (CG) in an ssDNA.

2.1.2 Free energies of stacked bases

Measurement of stacking free energies in DNA is extremely difficult both theoretically and experimentally. There is as much as disagreement regarding factors contributing to stacking interactions as there is for free energy values generally over the 16 different combinations of dinucleotide stacks AA, AG, AC, etc. Typically, empirical force fields and solvent models (based on classical mechanics) are used to calculate free energy of base stacking. The current ffs (AMBER [155], CHARMM [156], GROMOS [157] and OPLS-AA [158]) used in most of the calculations employ an additive model for electrostatic

forces in which atoms have fixed partial charges. Additive representation of electrostatic interactions limit the accuracy of calculations done with empirical ffs, especially for stacking interactions in DNA [159, 160]: because of the delocalized π electrons, aromatic rings have a large (and anisotropic) polarizability which leads to a significant departure from the isotropic two-body description of interactions. It has been shown that this polarizability increases the thermal stability of DNA [161].

In Table 2.1 base stacking free energies from three different sources are summarized.

Table 2.1: Total base stacking free energies (kcal mol^{-1}) in B-DNA from three different sources. The Honig [10] and MacKerell [11] datasets are theoretical calculations and the Kamenetskii [9] data are experimental values.

Sequence	Honig(theoretical)	Kamenetskii(experimental)	MacKerell(theoretical)
AA	-6.53	-1.11	-6.02
AC	-5.0	-1.81	-5.83
AG	-7.03	-1.06	-8.38
AT	-5.32	-1.34	-7.15
CA	-4.92	-0.55	-3.48
CC	-4.36	-1.44	-2.31
CG	-5.07	-0.91	-7.59
CT	-4.64	-1.06	-5.82
GA	-7.41	-1.43	-9.69
GC	-5.76	-2.17	-11.6
GG	-7.79	-1.44	-5.81
GT	-5.93	-1.81	-5.53
TA	-5.25	-1.11	-5.07
TC	-4.78	-1.43	-4.82
TG	-5.25	-0.55	-5.21
TT	-4.92	-1.11	-3.92

2.1.3 Calculation method

To calculate the triplet propensity we proceed from the data sets giving base-pair step formation energies via a brute-force Monte Carlo approach. Initially a random sequence of 60 amino acids is drawn with equal probability

for each of the "phase one" amino acids. A DNA duplex is then defined by randomly drawing a codon for each amino acid (based on those codons available in the modern "universal" genetic code). In order to have equal sequence content in both strands, the initial sequence is then concatenated with its complement, generating a palindromic test DNA sequence of 360 bp. Defects are then inserted at random into the sequence such that 1/3 of base steps are broken, and allowed to equilibrate their positions for 150 MC steps per defect present, where an MC step is a defect repositioning attempt via Metropolis Monte Carlo at 300K. This long sequence then serves as a "reservoir" of stack breaks: in order to find the triplet propensity for a given test codon, it and its complement are inserted into the reservoir duplex and re-equilibrated for a further for 30 steps/defect before collecting statistics for the distribution of stack breaks. The triplet propensity ΔG_τ is stated simply as a negative log probability for breaks to form at each end of the test codon but not inside it. The calculation was repeated (with different reservoir sequences) until convergence, which was validated by comparing redundant codons: the ΔG_τ for all pairs such as GGA·TCC and TCC·GGA was verified to differ at worst in the third significant figure.

Stacking energies are directional (crucially, GC·GC is stronger than CG·CG), here when writing a sequence the direction $5' \rightarrow 3'$ is assumed. From the tabulated data of Friedman and Honig [10], if we approximate the stacking energy for complementary duplex DNA as the sum of the stacking energies for the two base-steps, the weakest step is CG·CG (-10.14 kcal/mol). We therefore hypothesize that a series of codons of the form GNC (where N means "anything") should have the special property of partitioning naturally at the codon boundary C to G when under tension. We make a statistical analysis based on the stack breaking free energies, to produce a measure of the propensity for a given codon, when embedded in a wider sequence, to partition extension to the codon boundaries.

2.1.4 Triplet Propensity Results

(a) Honig *et al.*

Amino Acid	codon · anticodon	$\Delta G_\tau / k_B T$
G†	GGC·GCC	1.71
A†	GCC·GGC	1.72
S*	AGC·GCT	1.86
D†	GAC·GTC	2.01
V†	GTC·GAC	2.02
T*	ACC·GGT	2.07
R.	AGA·TCT	2.16
E*	GAA·TTC	2.19
F.	TTC·GAA	2.20
N.	AAC·GTT	2.25
I*	ATC·GAT	2.30
K.	AAA·TTT	2.43
P*	CCC·GGG	2.59
L*	CTC·GAG	2.70
Y.	TAC·GTA	3.33
C.	TGC·GCA	3.38
H.	CAC·GTG	4.28
W.	TGG·CCA	4.33
Q.	CAA·TTG	4.53
M.	ATG·CAT	4.56

(b) Kamenetskii *et al.*

Amino Acid	codon · anticodon	$\Delta G_\tau / k_B T$
T*	ACC·GGT	1.38
G†	GGT·ACC	1.38
S*	AGT·ACT	1.52
A†	GCC·GGC	1.56
I*	ATC·GAT	1.57
D†	GAT·ATC	1.58
V†	GTC·GAC	1.61
N.	AAT·ATT	1.63
R.	AGA·TCT	2.39
E*	GAA·TTC	2.55
F.	TTC·GAA	2.55
K.	AAA·TTT	2.61
P*	CCC·GGG	3.19
C.	TGT·ACA	3.33
L*	CTC·GAG	3.55
M.	ATG·CAT	4.44
H.	CAT·ATG	4.45
Y.	TAT·ATA	4.57
W.	TGG·CCA	5.15
Q.	CAA·TTG	5.43

(c) MacKerell *et al.*

Amino Acid	codon · anticodon	$\Delta G_\tau / k_B T$
T*	ACT·AGT	1.30
S*	AGT·ACT	1.30
D†	GAT·ATC	1.46
I*	ATC·GAT	1.47
A†	GCT·AGC	1.47
R.	CGT·ACG	1.68
E*	GAG·CTC	1.82
L*	CTC·GAG	1.82
V†	GTC·GAC	1.91
N.	AAT·ATT	2.01
K.	AAG·CTT	2.39
Y.	TAT·ATA	3.09
F.	TTC·GAA	3.59
H.	CAT·ATG	3.92
M.	ATG·CAT	3.93
Q.	CAG·CTG	4.34
G†	GGT·ACC	4.80
P*	CCT·AGG	4.82
C.	TGT·ACA	4.97
W.	TGG·CCA	8.31

Figure 2.2: Triplet formation free energies with Honig (a), Kamanetskii (b) and McKerrel (c) datasets. The † symbol indicates a member of the GADV set, while * indicates a phase I amino acid.

2.2 Discussion

Tabulating this information for the 20 canonical amino acids (showing the triplet propensity for the most triplet-prone codon available to each amino acid) we can see a clear pattern of reduced triplet disproportionation energy for the primordial 'phase I' amino acids (Figure 2.2). This pattern is particularly evident for the Honig dataset, which made the fullest treatment of solvation forces out of the two theoretical calculations and is therefore the most relevant.

2.2.1 Roles for Triplet Disproportionation

Minimisation of read frame errors: the calculated favourable partitioning into triplets for particular codons does not necessarily relate only, or at all, to the formation of long triplet-ordered DNA structures. In transcription, recombination and also in replication of DNA, codon boundaries must be correctly identified in order to avoid read frame errors. This simple calculation using existing data shows that such errors should be reduced in those genetic codes (GADV and phase I) which are supposed to be relatively old in evolutionary history.

Origin of the triplet genetic code: the three base-pair structure of the genetic code may have arisen specifically *via* the GNC code, which was highly inefficient from a purely informational point of view, due to the physical requirement to partition codons into triplets in the absence of the sophisticated enzymatic machinery which later evolved to maintain the NNN triplet code in modern organisms.

2.2.2 Phase I amino acids: Importance of Arginine

Depending on whether a given amino acid was in the beginning supplied by the environment or produced biosynthetically (potentially a difficult distinction if early life was more 'open' than modern cellular life forms), it is categorized as either phase I or phase II [162]. It is hypothesized that life must at some point have functioned with only phase I amino acids as biosynthesis by definition presumes the presence of some existing life form. The dramatic pattern evident in the tabulated partitioning energies is that codons for phase I amino acids overwhelmingly have relatively favourable free ener-

gies to partition into triplets aligned to their boundaries. The exception to this pattern is interesting: Arginine (R) is not listed in Wong and Bronskill's 1975 tabulation of phase I amino acids [3], possibly due to the large energetic cost needed to synthesize it from citrulline in modern organisms [163][2].

It has been advanced that the CGN and AGN (where N = 'anything') codons which yield Arginine in the modern genetic code previously coded for the chemically similar non-canonical amino acid Ornithine [168], and that the function of this codon was usurped in a presumably dramatic evolutionary event when selection advantage was found in having access to the more strongly basic Arginine molecule. The original phase one list GADVESPLIT contains no basic amino acids at all making the addition of Ornithine seem valuable in order to form a good range of folded proteins, and the replacement of Ornithine with Arginine a beneficial evolutionary step in giving access to a stronger base.

Arginine stands out for a second reason: the DNA-binding recombinase RecA achieves triplet disproportionation by cradling the negatively charged DNA in a large number of positively charged R side-chains (and some K). Thus we should perhaps not be surprised if a phase of biochemical evolution in which control of triplet disproportionation is important should have some means to produce either Arginine or similar moderately bulky basic residues.

2.2.3 Minimisation of read-frame errors: too strong to be a purely steganographic effect

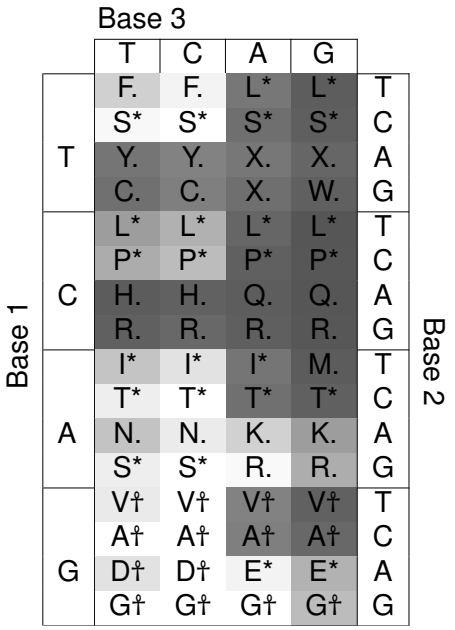
The main statement of this chapter, that the phase I part of the genetic code is structured so as to support a minimisation of read-frame errors by physically favouring the partition into codon-aligned triplets, is related to a known subtle and remarkable property of the genetic code. This property is that its redundancy is structured almost-optimally so as to support overlapping codes orthogonal to the primary code specifying amino acids [169], allowing the evolution of sequence changes altering DNA structure and interactions even within protein coding regions, without changing the coded protein. The overall flexibility of the genetic code in allowing arbitrary steganographic codes is not however sufficient to explain the strong pattern which we ob-

[2]Arginine is essential for virus replication [164,165]. The fact that viruses have emerged early in evolution of life [166] and their reproduction is essentially dependent on Arginine violates the Wong *et al.* hypothesis that Arginine was not present in early forms of life. Viral DNA synthesis continues in the absence of Arginine but formation of virions is inhibited [167].

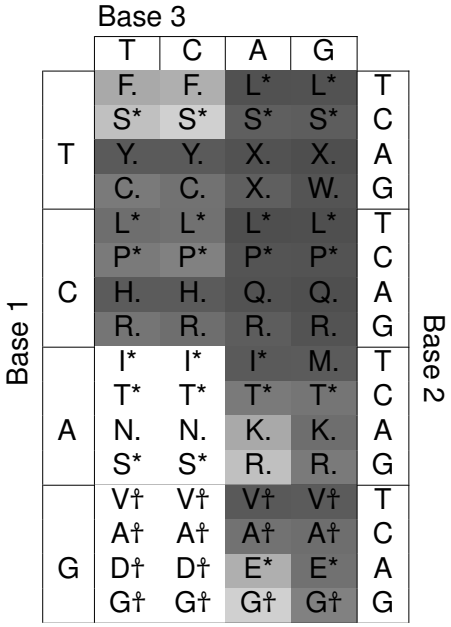
serve: Figures 2.2 and 2.3 show that codons for phase one amino acids are significantly more able to encode this partitioning than those in phase II. We further observe that the residues advanced by Ikehara *et al.* [64] as forming the minimal set for a functional proteome (marked † in Figure 2.2) are also those which partition most naturally into triplets.

We should note that the calculation of stacking energies by Friedman & Honig is a very elegant paper but is not the most recent attempt to measure this quantity. On the experimental side, Protozanova, Yavchuk & Frank-Kamenetskii have measured free energy of stacking for base-pair steps in a nicked DNA duplex [9]. It is attractive to generalise these stacking free energies for nicked DNA to the intact DNA double helix, however NMR analyses have shown that the broken phosphodiester bond pushes the conformation away from canonical B-form [170–172] which leaves a question mark over the data. Protozanova *et al.* report values of -0.91 and -2.17 kcal/mol for CG·CG and GC·GC stacks, while further experimental results for association of completely free duplex ends by Kilchherr *et al.* report -2.06 and -3.42 kcal/mol [8], (chapter 1, Table 1.2).

(a) Honig *et al.*



(b) Kamenetskii *et al.*



(c) MacKerell *et al.*

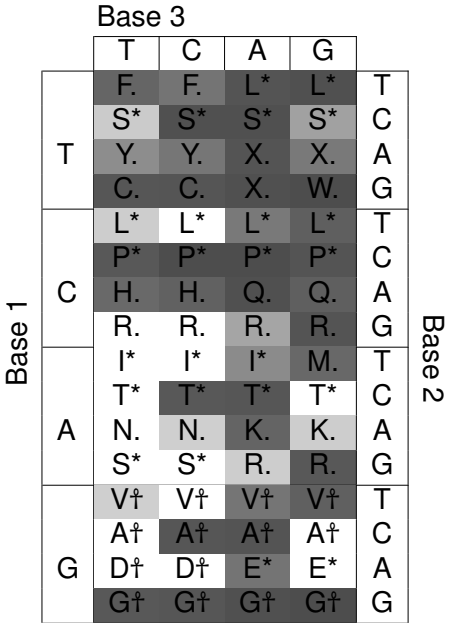


Figure 2.3: Triplet formation free energies and triplet disproportionation propensity with Honig, Kamenetskii and McKerrell data-sets. The † symbol indicates a member of the GADV set, while * indicates a phase I amino acid.

Calculations by Lemkul and MacKerell [11] make a more complex treatment of the base stacking interactions, arriving at values for the key CG·CG and GC·GC comparison of -7.59 versus -11.69 kcal/mol, but a less sophisticated treatment of the solvation. Overall, while there is significant disagreement on the magnitude (and sometimes the ranking) of the 10 stacking energies for complementary base pair steps, the key feature of a weak CG stack which points to the existence of a GNC triplet code is preserved. Repeated calculations of triplet formation free energies with the Kamenetskii *et al.* and MacKerell *et al.* data sets show the same privileged status for the GADV and phase one amino acids as those based on the Honig data-set (Fig. 2.3), although many details are different, and also in the MacKerell *et al.* calculation the residue Glycine loses its privileged triplet status almost entirely due to a very weak reported value for the GG·CC stack (Fig. 2.3).

2.2.4 Quantum corrections to stacking energy

There are many quantum mechanics (QM) calculations which show that stacking energies are overestimated in current classical mechanical force fields. For example it has been shown that AMBER-99 parameters produce artifacts which are the result of dramatically overstabilized base-base stacking [150][3]. Murata *et al.* have also found that the reversible work to unstack an aqueous purine dinucleotide is ∼5.0 kcal/mol, one order of magnitude higher than experimental results [173,174]. It is widely assumed that accurate base stacking energetics simply cannot be achieved within fixed charge classical MMs, but it has been shown that careful calibration of $1/r^6$ dispersion term performs as well as CCSD(T)[4] calculations for predicting stacking enthalpies of aromatic compounds [175].

It has been shown that for stacked configurations, the current AMBER nonbonded parameters exhibit unfavorable repulsive interactions at inter-base separation distances of 2.9-3.1 Å while CCSD(T) interaction energies are negative and favourable at these distances. This difference shows an inaccuracy in the LJ σ parameter that modulates the r^{-12} steric repulsion term of the nonbonded potential. The drawback of experiment in defining stacking energies is its inability to distinguish between base-base, base-sugar and sugar-sugar clustering. PMF calculations are used to solve this problem

[3]Base stacking in an empirical potential typically consists of a Lennard-Jones term and a Coulombic term with fixed atomic point charges.

[4]Coupled cluster is a numerical technique for describing many-body systems. This method provides an unbiased solution to the time-independent Schrödinger equation.

in a way that the energy profiles obtained are converted into stacking equilibrium constants with a distance cutoff of 5.6 Å[5] to differentiate stacked and unstacked conformations [150].

The tests used to validate nucleic acid parameters are exclusively focused on maintenance of unstretched dsDNA structural properties so they are unable to recognise if overestimated dispersion forces result in overstabilization of base stacking or not. Existing protocols for deriving MM parameters from three major families of ffs (AMBER, CHARMM, GROMOS) do not include any consideration of dispersion effects from high level QM calculations in the parametrization process.

Point charges which are included in ffs are derived from HF/6-31(G) calculations [177], but HF charge distributions are intentionally overpolarized for simulations in water. Banas *et al.* have shown that the differences between QM and MM stacking and pairing energies is the result of inaccuracy of the van der Waals ff term [174]. The two following forms are usually used for vdW interactions in ffs: (1) the standard LJ potential:

$$E_{6-12vdW} = \sqrt{\epsilon_i \epsilon_j}\Big[\Big(\frac{R_i + R_j}{r_{ij}}\Big)^{12} - 2\Big(\frac{R_i + R_j}{r_{ij}}\Big)^{6}\Big] \tag{2.1}$$

or (2) an exponential repulsion term in combination with a damped dispersion term:

$$\begin{aligned} E_{6-12vdW} = \sqrt{\epsilon_i \epsilon_j}\Big[&\frac{6}{\zeta - 6}\exp\Big(\zeta\Big(1 - \frac{r_{ij}}{R_i + R_j}\Big)\Big) \\ &- \frac{\zeta}{\zeta - 6}\frac{(R_i + R_j)^6}{r_{ij}^6 + (\frac{R_i + R_j}{\alpha})^6}\Big] \end{aligned} \tag{2.2}$$

In both forms R and ϵ parameters correspond to VDW radii and well depth respectively. In the exponential form the scaling parameter ζ is used to account for long range dispersion interaction.

Banas *et al.* have shown that ff stacking energies are overstabilized by about 25% [174]. This effect could be due to anisotropy of the dispersion (VDW) interaction [178]. While the VDW ff term is isotropic the actual dispersion of nucleic acid bases is strongly anisotropic. Consequently, dispersion interaction is relatively weakened in the stacking direction and strengthened in the H-bonding direction compared to the isotropic case. This effect is not included in the ffs, so stacking interactions are overestimated. So it is likely

[5]The distance criterion is originally proposed in [176].

that ignoring dispersion anisotropy in the ff is the source of overestimation of stacking energies. Anisotropy can also affect hydration of nucleobases and as a result cause overstabilization of stacked nucleobases [179].

It has been noted that charge penetration effects become highly attractive when rise is < 4 Å and dominate the electrostatic contribution to the interaction energy [159]. These charge penetration effects are the reason why at short ranges MM deviates significantly compared to *ab initio* methods. MM models quickly become very repulsive during close contacts which is due to insufficiently attractive MM electrostatics. So as the rise is decreased the errors in the oversimplified MM point charge model grow rapidly due to the exponential increase in the charge penetration contribution.

Distribution of the electrostatic potential in bases is dipole-like and they prefer large overlap of the rings due to dispersion attraction. The ff assumes that the electronic structure of bases is fixed but in reality it responds to polarization effects and this makes the ff electrostatic term unphysical which causes deviations from QM electrostatics at short intermolecular distances. Another problem of ffs describing stacking is that real atoms do not have radii and are not necessarily spherical which causes large anisotropic polarizability of bases in the base pairing direction. Therefore ff description of atoms as VDW spheres with a fixed radius (isotropic description) causes overestimation of base stacking which leads to overstabilization of stacked bases in MD simulations.

Extensive MD simulations with revised AMBER parameters to calculate stacking free energies confirm the recent concerns that computed stacking free energies are too favourable relative to experiment [180]. We will discuss in the atomistic simulation results section that the apparent discrepancy between simulations and experiment [113] in capturing Σ-DNA in the absence of intercalating cations (Arg or EtBr) might be due to the fact that current MM ffs are not able to describe stacking interactions properly and by overstabilizing these interactions produce kinetic or thermodynamic traps [181]. Such traps could be overcome in the presence of bulky cations and this may be a reason that the triplet disproportionation pattern is observed in simulation in the presence of these cations [182].

Chapter 3

Results: Section II

3.1 Introduction

3.1.1 Triplet Disproportionation

Under tension in aqueous solution with small or monatomic counterions, the DNA duplex stretches, unwinding if not topologically constrained, and eventually denatures. The extension against force shows a jump by a factor of $\approx 1.5 - 1.7$ (depending on sequence, pulling geometry and solution) at $\approx 65 - 70$ pN [183–185]. Several models have been proposed to explain the sudden increase in length, which is widely agreed to be the signal of a collective structural transition. The formation of regions of single stranded DNA (ssDNA) [186] or of ladder-like stretched and untwisted double stranded DNA (dsDNA) have been suggested [101, 114, 187]. At modest extensions of sequences not dominated by AT base pairs, we expect to see a partly untwisted ladder-like structure, in which the base pairs remain intact but the rise per base pair is equilibrated to a new value of $\sim$5.8 Å, compared to the rise in unstretched B-DNA of 3.4 Å. In general, this stretched phase is known as S-DNA. For GC-rich structures having strong hydrogen bonding the base pairing is preserved in the S-DNA structure, and the base stacking is also somewhat preserved by tilting and sliding of the base pairs. Tilting of the base pairs increases the solvent-exposed area while permitting them to remain in contact such that a complete water gap does not open between them.

Change in the inclination as a function of applied force strongly depends

on the pulling scheme. It is shown that in the $5'5'$ pulling mechanism, tilt angle increases gradually until the terminal H-bonds are disrupted while in the $3'3'$ pulling regime tilt angle is decreased and no early breakage of H-bonds occurs [127, 127, 188–190].

The most readily available description of DNA under tension is the empirically measured force-extension curve [75, 191], which provides a clear signal of some kind of transition but no atomistic-level information. This is supplemented by fluorescence and polarised-light studies [108, 192, 193], and by atomistic simulations which are able to provide explicit descriptions of the DNA but which are limited in the accessible timescales and system sizes [100, 114, 189]. We have in the introduction and previous chapter 2 discussed the formation of regular structures of planar, triplet-disproportionated and stacked bases, however atomistic simulations to date have shown instead the irregular formation of 'denaturation bubbles' [194, 195], different from the formation of regular triplets both in the irregularity of the spacing and in the large disruption of base planarity and base-pairing near to the solvent filled cavities formed.

3.1.2 Mechanism of action of RecA

DNA is often subjected to tension in its biological context, for purposes including transport, transcription and tertiary structure manipulation. A striking example of this is the crystal structure of DNA bound to the RecA protein [196], a snapshot of the fundamental process of sexual reproduction: the recombination of homologous DNA from two parent organisms. In this structure the extended protein-bound DNA chain does not adopt an S-like configuration, but rather disproportionates into groups of three bases, with planar base stacking retained within each triplet (Figure 3.1). This triplet disproportionation has also been observed in solution when bound to RecA [192].

RecA belongs to a family of ATPases[1] which perform homologous recombination, a process which both maintains the integrity of the genome and also creates genetic diversity. The mechanism of action of RecA is such that RecA, ATP and single-stranded DNA (ssDNA) form a helical filament that binds to double stranded DNA (dsDNA), then searches for homology and finally does the exchange of the complementary strand producing a new het-

[1] ATPases are proteins that catalyze the decomposition of ATP into ADP and release energy which is used in some enzymatic reaction.

eroduplex. In this process DNA is underwound and stretched globally but at the codon level it remains in B-form. This fact restricts the homology search to WC type base pairing.

In detail, RecA binds to ssDNA in an ATP-dependent manner and forms a helical nucleoprotein filament that has ∼6.2 RecA proteins per turn and ∼3 nucleotides per RecA protein [197]. Then DNA is underwound and stretched with ∼18.5 nucleotides per turn and an average rise of ∼5.1 Å per nucleotide [198]. The RecA-ssDNA form a presynaptic complex that searches for ssDNA-dsDNA homology and when it is found strand exchange results in the formation of a postsynaptic complex in which the complementary strand of the donor duplex is paired with the original ssDNA. Hydrolysis of ATP which is stimulated by DNA binding dissociates all DNA, releasing the new heteroduplex and a displaced ssDNA from the donor duplex [196].

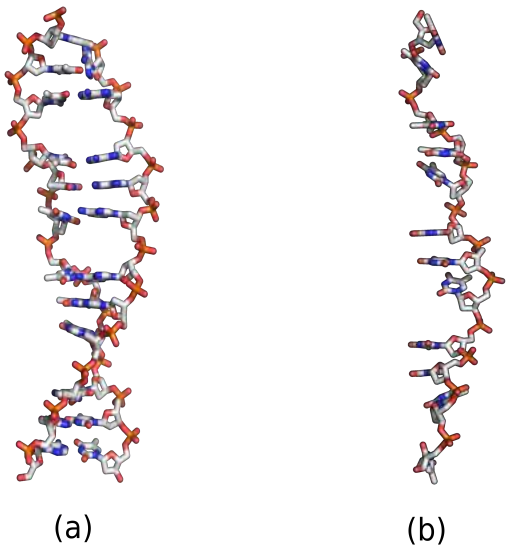


Figure 3.1: DNA encapsulated within the RecA complex is extracted from (pdb: 3cmt) for dsDNA and (pdb: 3cmw) for ssDNA. Groups of three stacked base-pairs form the peculiar feature of Σ-DNA. The triplet disproportionation is seen in both the ss- as well as ds-DNA.

3.1.3 Sigma DNA

Orderly triplet formation when complexed is in contrast to the current best estimates of the structural behaviour when extended in solution, however we are led to examine whether the triplet phase can be stabilised in solution and if so, should it be considered a canonical biologically active structure of DNA on the same footing as the A, B and Z forms. Using molecular dynamics simulations of duplex DNA with an applied force, we do not observe that the triplet structure is stable in an aqueous solution of monatomic counterions, however we do find that it is stable without specific complex to a structured enzyme, in a solution of terminus capped monomeric Arginine peptides (Ac-Arg^{+}-NHMe Cl^{-}).

We present planar-stacked triplet disproportionated DNA as a solution phase of the double helix under tension, and dub it 'Σ-DNA', with the three right-facing points of the Σ character serving as a mnemonic for the three grouped bases. As for the unstretched Watson-Crick base paired DNA structures, we remark that the structure of the Σ phase is linked to function: the partitioning of bases into codons of three base-pairs each is the first phase of operation of recombinase enzymes such as RecA, facilitating alignment of homologous or near-homologous sequences. By showing that this process does not require any very sophisticated manipulation of the DNA, we position it as potentially appearing as an early step in the development of life, and correlate the postulated sequence of incorporation of amino acids (phase one and phase two [3, 199, 200]) into molecular biology with the ease of Σ-formation for sequences including the associated codons for phase one amino acids.

We also note that the machinery of nucleotide to peptide translation occurs necessarily with reference to triplets of bases, so that further investigation into the Σ phase of single and double strands of RNA and DNA might be a valuable source of insight into the origins not only of recombination, but also of protein synthesis.

3.1.4 Sequence-Dependence of disproportionation

It is not clear what form the original genetic code had, as it is likely to have co-evolved to some extent with the associated enzymes of transcription and translation. We can make a guess about the history of the genetic code by considering the chemical complexity of the different amino acids: it is

hypothesised that a list of so-called 'phase I' amino acids were present earlier in evolution than the 'phase two' amino acids, based on the complexity of the cellular machinery used in current organisms to synthesize, for example, Methionine (M) from Threonine (T) [200,201]. If the genetic code in the time of a much simplified amino-acid alphabet already had the current structure of three base-pairs per codon it was therefore highly redundant at this time.

In the current triplet code, the 'phase I' amino acids supposed to have been incorporated earliest into biology (a list of GADVESPLIT) are coded by triplets which have a specific statistical tendency: the energetic cost to break base-stacking at the triplet boundary is low, relative to the complete modern genetic code. We motivated the statistical observation of preferential triplet disproportionation in the phase one genetic code in chapter 2. We now analyse atomistic simulation data to show that disproportionation into codon-aligned triplets occurs spontaneously under tension for appropriate sequences and solution conditions.

Beyond the pairwise hydrophobic and electrostatic interactions of base stacking (covered by the classic calculations used as input to generate Tables 2.2 and 2.3) the potential importance of complex entropic, structural and solvent effects make it necessary to carry out a full atomistic molecular dynamics investigation of DNA under tension. Given the expected importance of sequence effects, simulations were run both with a low-entropy sequence of $d[G_{12}C_{12}]$ (encoding 4 glycines and 4 prolines) and a sequence chosen to show strong triplet disproportionation based on table 2.2, $[GGC]_4[GAC]_4 \cdot [GTC]_4[GCC]_4$, encoding four repeats each of the high-scoring amino acids Gly and Asp on the first strand, then Val and Ala on the complementary strand (the GADV set of Ikehara *et al.* [64]).

3.2 Simulation protocol

3.2.1 Steered molecular dynamics(SMD)

Given a molecular process a *reaction path* can be defined along which the process proceeds in the configurational space. The progress of the simulation can be monitored and described by the *reaction coordinate*. In this context a 'potential of mean force' (PMF) can be defined: a PMF is a free energy profile along the reaction coordinate and is determined through the Boltzman-weighted average over all degrees of freedom orthogonal to the

reaction coordinate [202]. To observe relevant processes in biomolecular systems using molecular dynamics usually time-scales greater than nanoseconds are needed. SMD is a way to accelerate processes by applying external steering forces in a controlled way [203].

In a typical SMD simulation the system is steered by applying a constraint such as a time-varying harmonic potential that moves along the defined path in the configuration space. For more than three decades DNA mechanical properties have been probed experimentally *via* imposed pulling forces using AFM or laser traps [204–206], SMD is a way to complement these techniques by providing a close atomistic analogue of the experiments.

This method is like umbrella sampling in which the center of the restraint is time dependent:

$$V_{rest}(t) = \frac{1}{2}k(x - x_0(t))^2 \tag{3.1}$$

where x could be a collective or more straightforward variable such as the distance or angle between atoms, in DNA overstretching typically a simple distance restraint. Here we controlled the distance between the centre of geometry of the end two bases of the duplex and a fixed point at a large distance above the centre axis of the DNA.

3.2.2 Detailed Simulation Setup

The DNA duplexes were stretched by an additional 100 Å from their relaxed lengths, over a time period of 150 ns, giving a stretching rate of 0.029 Å ns^{-1} bp^{-1}. Because of the apparent importance of Arginine, based on table 2.2 and on the RecA structure [196], simulations were run both in NaCl and in a solution of Ac-Arg-NHMeCl, with the capped arginine molecule replacing sodium as the positive counterion.

Molecular structures were prepared using the Nucleic Acid Builder (NAB) [4]. Salt was represented using the Joung–Cheatham parameters for ions [207] and the TIP3P model of water was used for solvation [208]. The ethidium molecule was represented using the GAFF [209] with partial charges and bond parameters assigned via the ANTECHAMBER tool [210]. Simulations were run using the GPU-accelerated implementation of pmemd [211] in the AMBER16 package [212] using the AMBER 99SB+bsc0 force field parameters for the biomolecules [213, 214]. Generated DNA fragments were simulated in a rectangular periodic box with 10 Å distance from the DNA to the box boundaries. Sodium and chloride ions were added so as to give

an electrically neutral system with approximately physiological Cl^- concentration (order 0.1M). The SHAKE algorithm [215] was used to constrain all bonds involving hydrogen atoms. To calculate electrostatic interactions, the particle mesh Ewald sum was employed with a 2 fs time step [216]. The direct part of the Lennard–Jones interactions was cut off at 8 Å.

EtBr and Arginine intercalators were initialised in the bulk solvent rather than being manually inserted between base pairs: this approach is likely to underestimate the amount of intercalator bound at a given extension, but has the advantage that no bias is introduced with respect to the binding site or pose. The simulation trajectories were collected at a rate of one frame per 2 ps.

Table 3.1: Sixteen instances of 150 ns were run for each system, giving a cumulative simulation time of 14.4 μs. Effective concentration of 88 intercalants in 9 k water molecules is ∼0.5 M. 70 and 24 Na^+ ions in a box size of approximately 40 × 40 × 210 Å corresponds to a concentration of ∼0.33 and ∼0.1 M respectively.

DNA sequence	Total number of atoms	Number of Arg/EtBr	Na^+
$(GGC)_4(GAC)_4 \cdot (GTC)_4(GCC)_4$	35584	None	70(24 Cl^-)
$(GGC)_4(GAC)_4 \cdot (GTC)_4(GCC)_4$ + EtBr	32357	88	24(66 Br^-)
$(GGC)_4(GAC)_4 \cdot (GTC)_4(GCC)_4$ + Arg	35592	None	70(24 Cl^-)
$G_{12}C_{12} \cdot G_{12}C_{12}$	35592	None	70(24 Cl^-)
$G_{12}C_{12} \cdot G_{12}C_{12}$ + EtBr	32359	88	24(66 Br^-)
$G_{12}C_{12} \cdot G_{12}C_{12}$ + Arg	32357	88	24(66 Cl^-)

For each calculation, 16 independent instances were prepared and equilibrated in the B conformation for 10 ns. Pulling of the DNA then took place using steered molecular dynamics, for 16 instances of six different systems, over a time period of 150 ns (giving a pulling rate of 0.066 Å ns^{-1}). Due to the slow kinetics of intercalator dissociation in unstretched DNA, which for typical mono-intercalators is of the order of one per second ([217]), binding of intercalators was essentially irreversible in silico except via extensionally driven conformational change. The choice to make multiple 150 ns simu-

lations rather than fewer multi-microsecond runs was a decision to pursue good stochastic sampling of an explicitly non-equilibrium process, rather than attempting the computationally very challenging goal of equilibrium-like sampling, which is not achieved even on the laboratory timescale of <1 s per cycle of extension and relaxation.

3.3 Results: Spontaneous Triplet Disproportionation Under Tension, Amplified in the Presence of Organic Cations

3.3.1 Preference for GC rich sequences

We find that for CG-rich sequences encoding phase one amino acids, the triplet-disproportionated Σ-phase of DNA is observed, with the strongest triplet formation taking place in the presence of the terminus-capped Arginine residues (Ac-Arg-NHMe) (Fig. 3.2). Fig. 3.2 shows a regular pattern of vertical gaps with spacing 3 bp, over a large range of extensions. The low entropy sequence in the presence of Arginine shows some weak structure at high extensions, due to exclusion effects which disfavour binding of cations to adjacent sites. In the high entropy sequence, some structure of period three is seen, even in the absence of Arginine, however this is relatively weak (as suggested by the order-1 k_BT free energies of disproportionation in table 2.2). The triplet-disproportionated structures show the essential features of Σ-DNA (Fig. 3.3) as seen in the RecA bound crystal: preserved Watson-Crick base-pairing, approximately planar orientation of the bases and a large cavity every third base pair. Extending beyond approximately 3 Å leads to breakup of the Σ phase and also to loss of Watson-Crick hydrogen bonding, as the bases interdigitate with each other and hydrogen bond to the backbone.

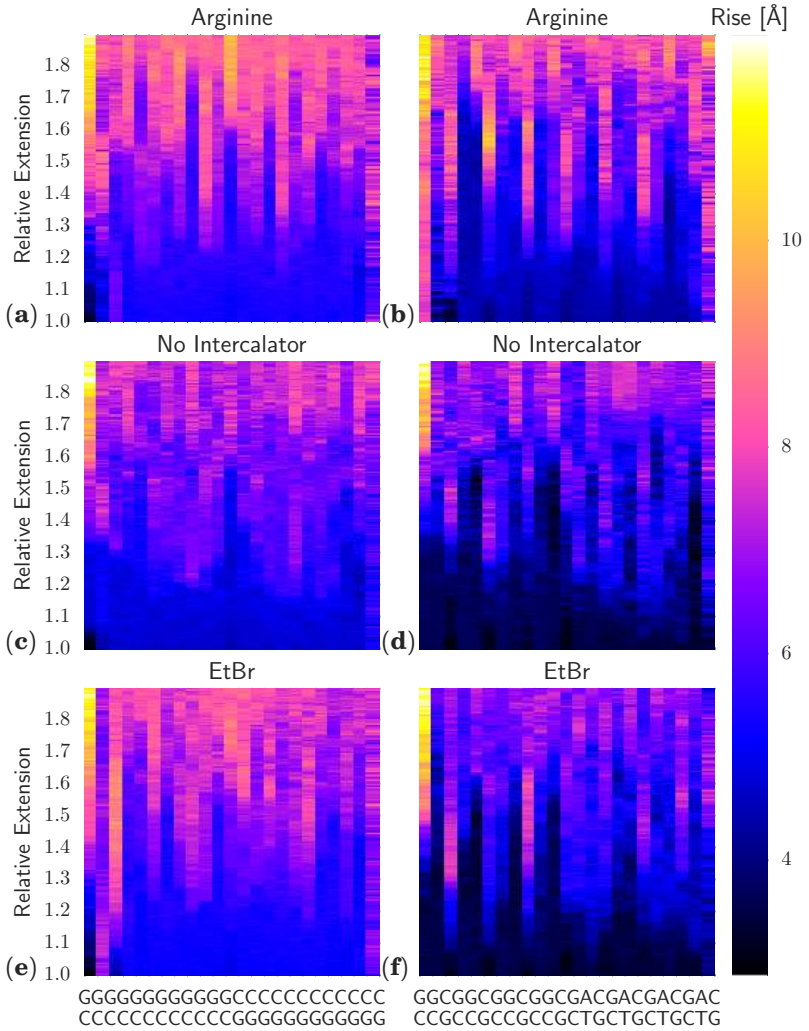


Figure 3.2: Kymographs of rise per bp-step under imposed whole-DNA extension. Triplet disproportionation is strongly evident in (**b**), while the strain is spread most evenly in (**c**). Presence of arginine in a homogeneous sequence (**a**) or presence of CG steps in the absence of arginine (**d**) induce only weakly structured disproportionation.

The average base-pair inclination in the high entropy sequence $[GGC]_4$ $[GAC]_4 \cdot [GTC]_4$ $[GCC]_4$ in the presence and absence of intercalators up to the extension point of 1.5 follows the same pattern and remains flat (Fig. 3.10 b,d,f) which indicates that base pairs were *on average* perpendicular with respect to the helix axis as oberserved experimentally by [124,137,218] although

almost all base pairs had significant positive or negative inclination. In the presence of intercalators this trend continues after an extension of 1.5 but shows a sudden drop for the duplexes in NaCl after a relative extension of 1.7. For the low entropy sequence d[$G_{12}C_{12}$], the change of average inclination up to an extension of 1.5 is the same as for the high entropy sequence. The bare sequence and the one in the presence of Arginine reach a maximum inclination at a relative extension of 1.6-1.7 and drop afterwards (Fig. 3.10 a,c) but in the presence of EtBr a continuous increase is observed after extension 1.5, followed by a second flat region after extension 1.6 (Fig. 3.10 e).

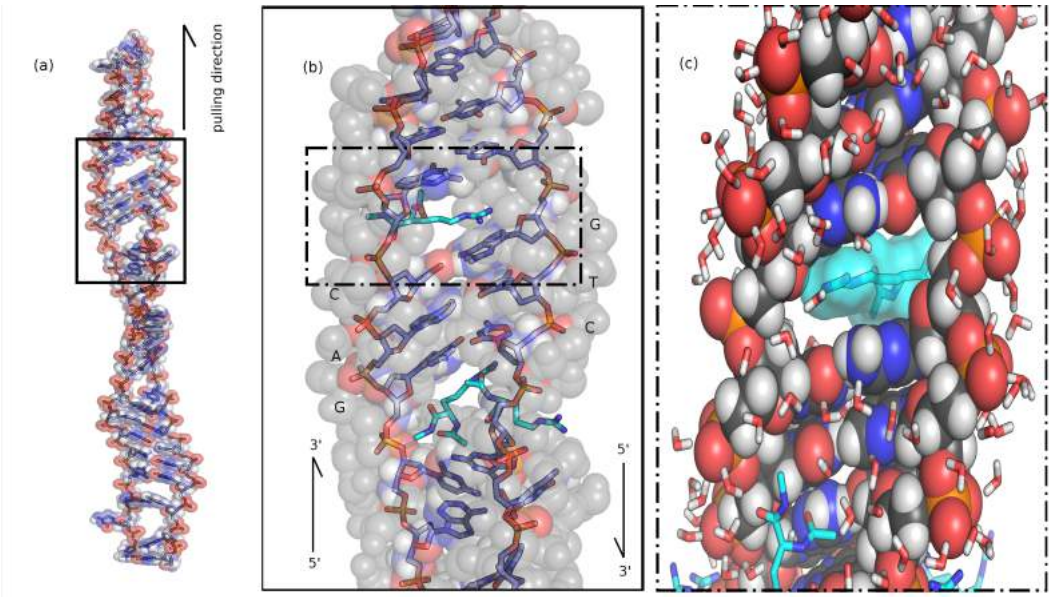


Figure 3.3: The primordial sequence partitions under tension predominantly at the CG steps, forming triplets (**a**), with Watson-Crick hydrogen bonding and planar base stacking preserved subject to some thermal disorder (**a,b**). Triplets are stabilised by one or two Arginines intercalating the stretched base steps (**b,c**) with non-specific binding that tends to place the charged end of the side-chain close to the phosphate, and partially or entirely excludes water from between the bases. (**c**) is a zoomed and rotated view of the highlighted cavity in (**b**).

3.3.2 Base-pair classification

We choose to classify base pair steps according to the following rules (Fig. 3.4):

- β:('Base-paired and stacked'): one or more Watson-Crick hydrogen bonds were preserved for each pair in the step, with rise < 5.6 Å.

- ζ:('Zipper'): one or more hydrogen bonds were present between each base and the backbone of the opposite strand.

- σ:('Space'): one or more WC hydrogen bonds were preserved for each pair in the step, rise was ≥ 5.6 Å, and at least one of the two vertical pairs of bases was completely separated, with no contact ≤ 3.5 Å.

- τ: ('Tilted'): one or more WC hydrogen bonds were preserved for each pair in the step, rise was ≥ 5.6 Å, but one or more atomic contacts remained for each vertical pair of bases.

- μ:('Melted or mismatched'): WC hydrogen bonding to the opposite base was disrupted, and not replaced with zipper hydrogen bonding.

Hydrogen bonds were defined for triangles donor-H-acceptor such that the angle at H was greater than 135° and the distance donor-acceptor was less than 3 Å (cpptraj defaults) [219]. The rise cutoff for stack breaking was made at 5.6 Å based on [150]. An atomic contact was defined as a distance < 3.5 Å. Examples of each conformation are shown in (Fig. 3.4). Averages were collected over steps [3. . . 21] of the 23 steps available in the sequences studied, in order to minimise end effects.

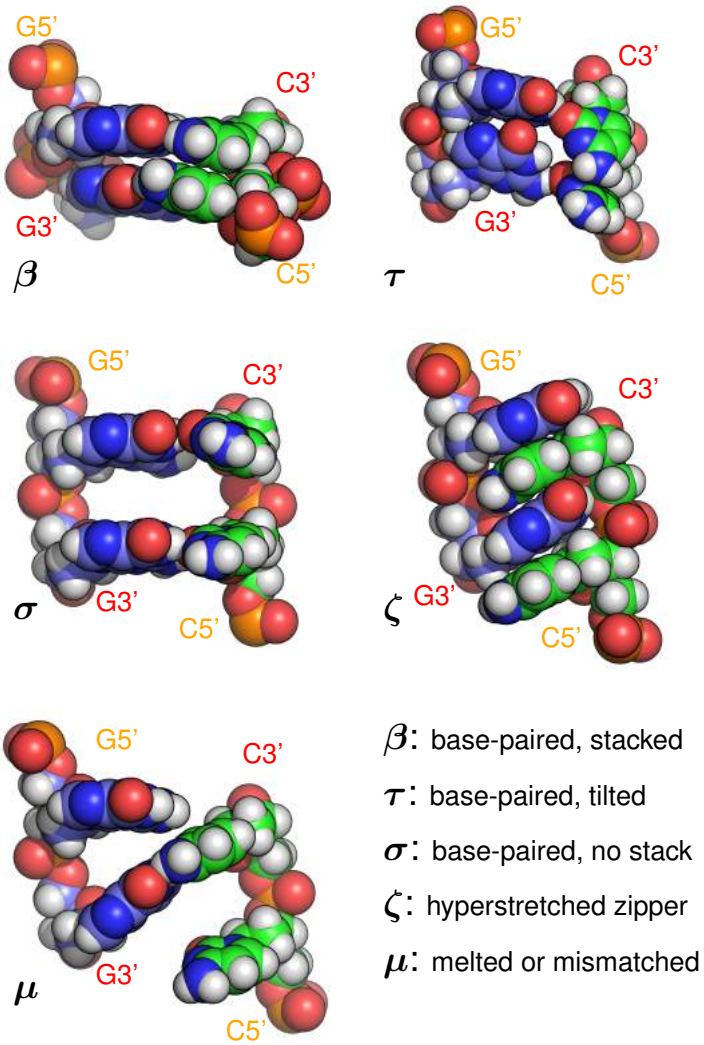


Figure 3.4: Example base-pair conformations (all of sequence GG·CC) classified by the type of stacking and hydrogen bonding present. Note that the initials β, τ, σ, ζ, μ do not refer to phases (collective structures) but local conformations. A step labelled as 'β' would for example be consistent with the A, B, C, Z or Σ phases, all of which include base stacking and Watson-Crick hydrogen bonding.

Fig. 3.9 shows the proportions of different local conformations consistent with different phases. Examples of each conformation are shown in figure 3.4. Local conformations were classified according to the physical interactions which were either broken or preserved, with Watson-Crick hydrogen bonds and planar base stacking giving way over medium extensions of 1.3-1.5 to a complex regime in which regions of melted base pairs (no or mismatched

hydrogen bonding) compete with regions in which the Σ phase dominates with WC hydrogen bonding preserved, but with stacking periodically either broken or reduced by non-collective tilting.

At extensions beyond 1.7, in the absence of intercalator molecules the zipper phase emerges. In the presence of intercalators, the zipper phase is destabilised and WC base-paired conformations are stabilised. Although the count of conformations consistent with the Σ phase increases again at high extension with intercalators, the threefold periodicity is weakened. The presence of the Σ phase (in the full sense of two β-steps alternating with one σ-step) reaches a maximum average proportion of about 30%, with the remainder at this point made up mostly of melted or indeterminate configurations, or of boundary regions between melted and structured phases (Fig. 3.9 h,l).

3.3.3 Zipper DNA

At extensions beyond 1.7 in the absence of intercalator molecules zipper DNA emerges (Fig. 3.5). In this structure bases of the DNA strands interdigitate with each other and make a single base aromatic stack. This structure was first predicted by Lohikoski *et al.* [125]. Similar motifs have been observed in experiments, although they were extended only a few base pairs [6] and in theoretical studies [220]. Zip-DNA does not require base pair complementarity.

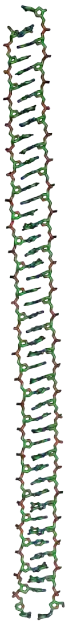


Figure 3.5: Figure shows formation of zipper DNA in the absence of intercalator molecules when extension is beyond 1.7. In this structure the bases of the DNA interdigitate, the reason that this conformation is called zip-DNA. Analysis of the electron properties of this structure shows a great magnitude of increase in π-π interactions between nucleobases compared to B-DNA [6].

(a) Zipper-DNA+ARG **(b)** Zipper-DNA+EtBr

Figure 3.6: Figure shows the formation of zipper-like DNA in extensions above 1.7 in the presence of Arginine molecules as intercalators as well as EtBr. Hydrogen bonds are more preserved in the Zipper-DNA in the presence of Arginine. Interdigitation of bases are more obvious in the presence of EtBr.

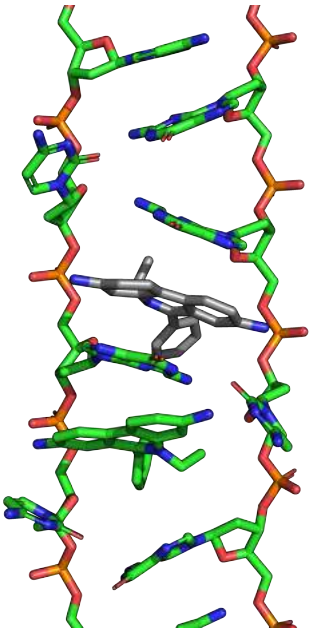


Figure 3.7: A snapshot of highly overstretched DNA in the presence of EtBr. An EtBr molecule has intercalated between base pairs (gray molecule). For more clarity EtBr molecules in the surrounding are removed.

In the presence of intercalators, including Arginine, the zipper DNA is destabilized and WC base-paired conformations are stabilized (Fig. 3.6a) while in the presence of EtBr some motifs of interdigitated base pairs are observed (Fig. 3.6b).

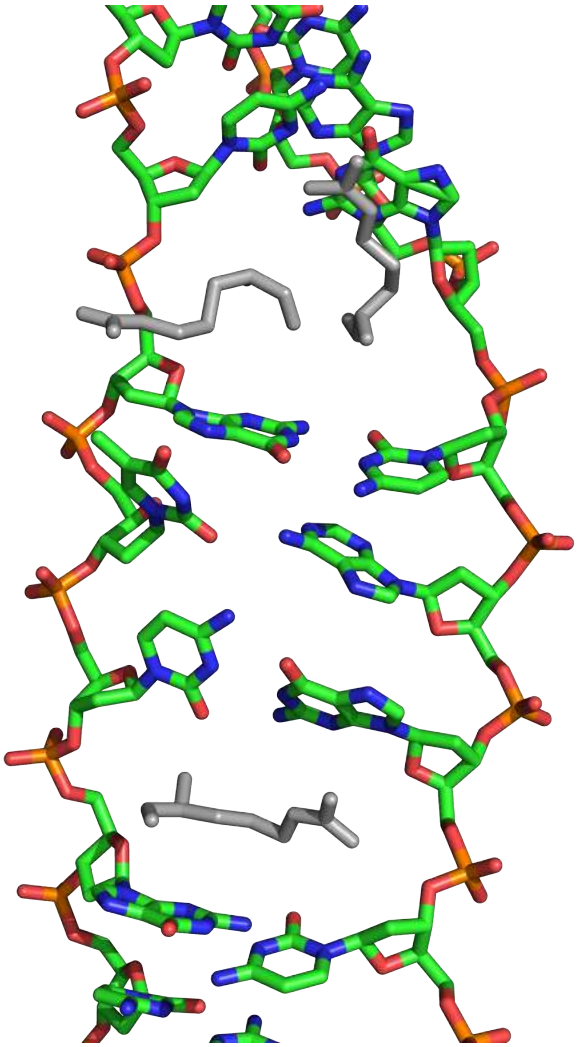


Figure 3.8: One or two Arginine molecules can intercalate in gaps created in the Σ-DNA at the relative extension of 1.5. Arginine molecules interact with DNA in a non-specific way which tends to place the charged end of the side chain close to backbone phosphate. The surrounding Arginine molecules are removed for the clarity. Arginine moieties are shown in gray.

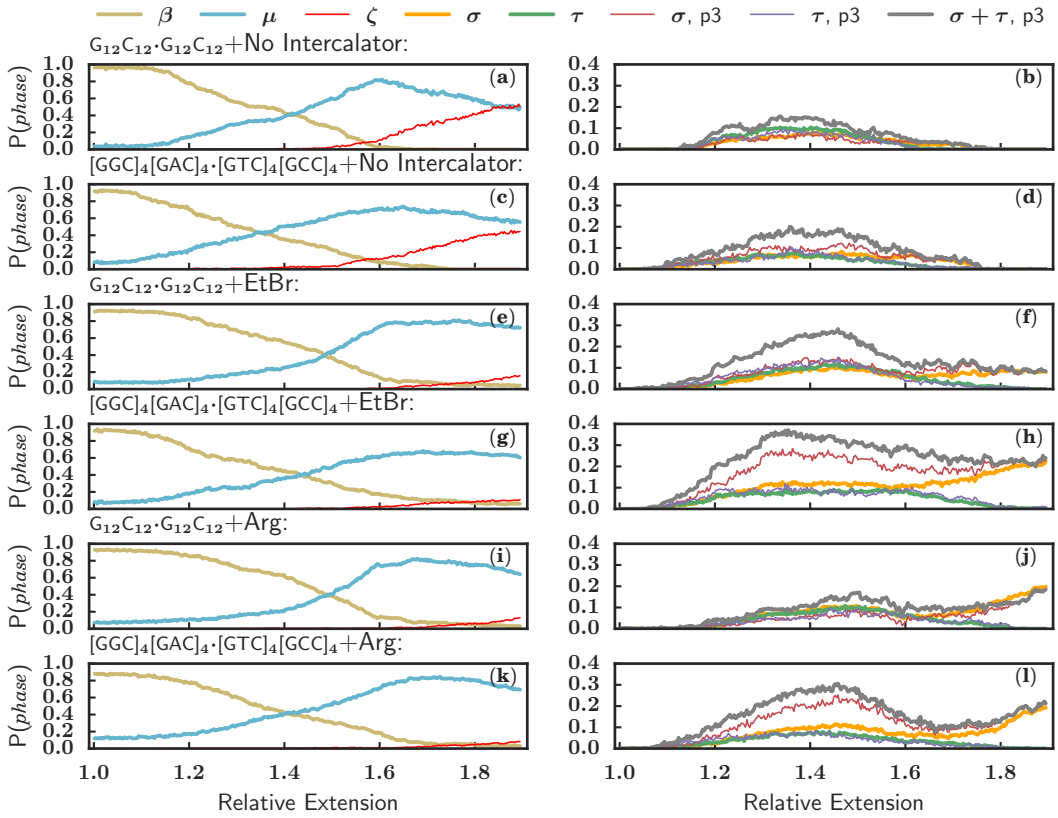


Figure 3.9: Base-pair steps (4 bases) were classified by local conformation as β: base-paired and stacked, μ: melted, ζ: zipper, as planar with broken stacking (σ) or as τ: tilted. The left panels (a„c,..,k) show the three major states of the DNA, with a melting transition over extensions 1.2-1.6, followed (in the absence of intercalator) by a hyper-stretched zipper conformation. The right panels (b,d,..,l) show the incidence of states (σ,τ) in which the rise exceeds 5.6 Å, with preserved Watson-Crick hydrogen bonding. In systems with intercalator and a triplet coding sequence (h,l); steps at the codon boundary (p3) have an enhanced proportion of σ states, peaking in the extension range 1.4-1.5.

3.3.4 Change of inclination during overstretching

DNA structure is characterized by different geometrical parameters like twist, helical rise and inclination. In B-DNA base pairs are normal to the helix axis while in A-DNA (which forms in low humidity conditions) base pairs are tilted, with contraction of the double helix. When DNA is overstretched, tilt starts to increase depending on the way that force is applied to the DNA. Pulling can be done on either the $5'5'$, $3'3'$ or the geometrical center of both ends. As the pulling force increases bases start to tilt and hydrogen bonds starts to break, this is the general view of overstretched DNA. As shown in

Fig. 3.10 this pattern changes when DNA is overstretched in the presence of intercalators. Monitoring the changes of planarity of the base stacking is important in elucidating the structural changes that DNA undergoes while stretching. Stack breaks could be avoided by collective tilting of the bases [118]. The change of average inclination for the high entropy sequence during extension with or without cations is small (Fig. 3.10) which is consistent with the recent experimental results [113]. This is the prominent feature of Σ-DNA in which stacking interactions is conserved in a triplet base-stacked pattern.

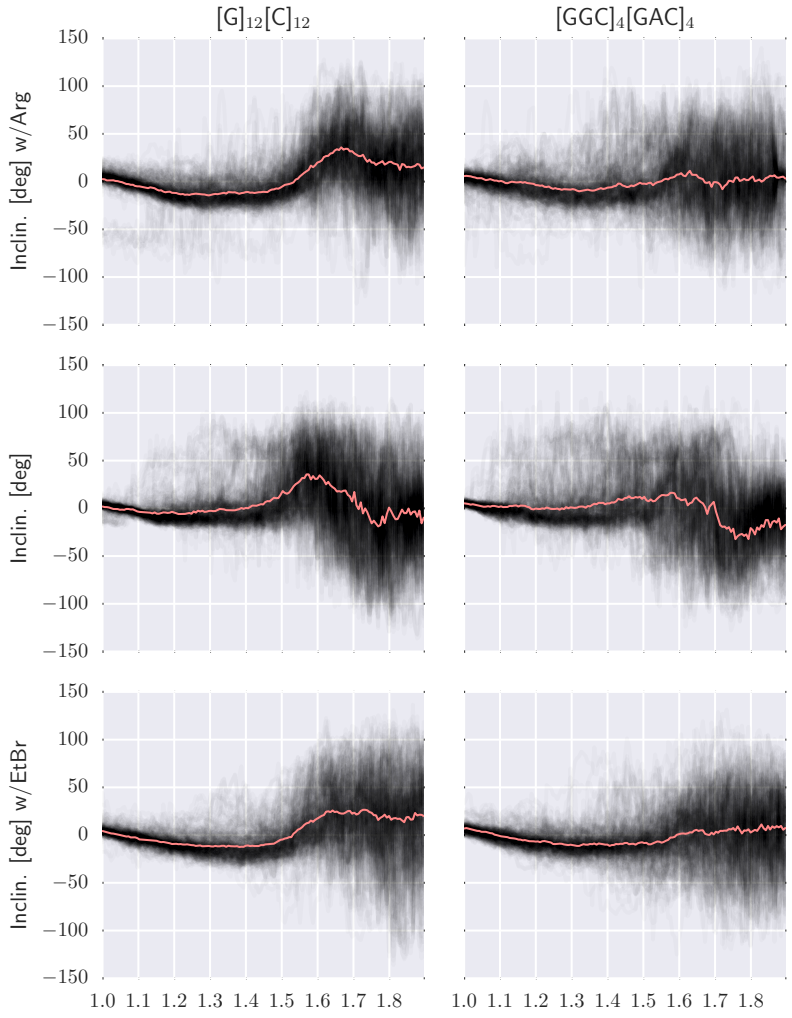


Figure 3.10: Average inclination of the low and high entropy sequences in the presence and absence of intercalators (Arginine and EtBr). Average inclination for the high-entropy sequence d[$(GGC)_4(GAC)_4$] remains relatively flat up to extension 1.5 and beyond, even without intercalant (b,d,f). For the low entropy sequence d[$(G)_{12}(C)_{12}$] average inclination remains flat up to extension 1.5 but it experiences a sudden change after the extension passes 1.5 (a,c). In the presence of EtBr inclination increases smoothly after the extension point of 1.5 and reaches the second flat region of extension beyond 1.6 (e).

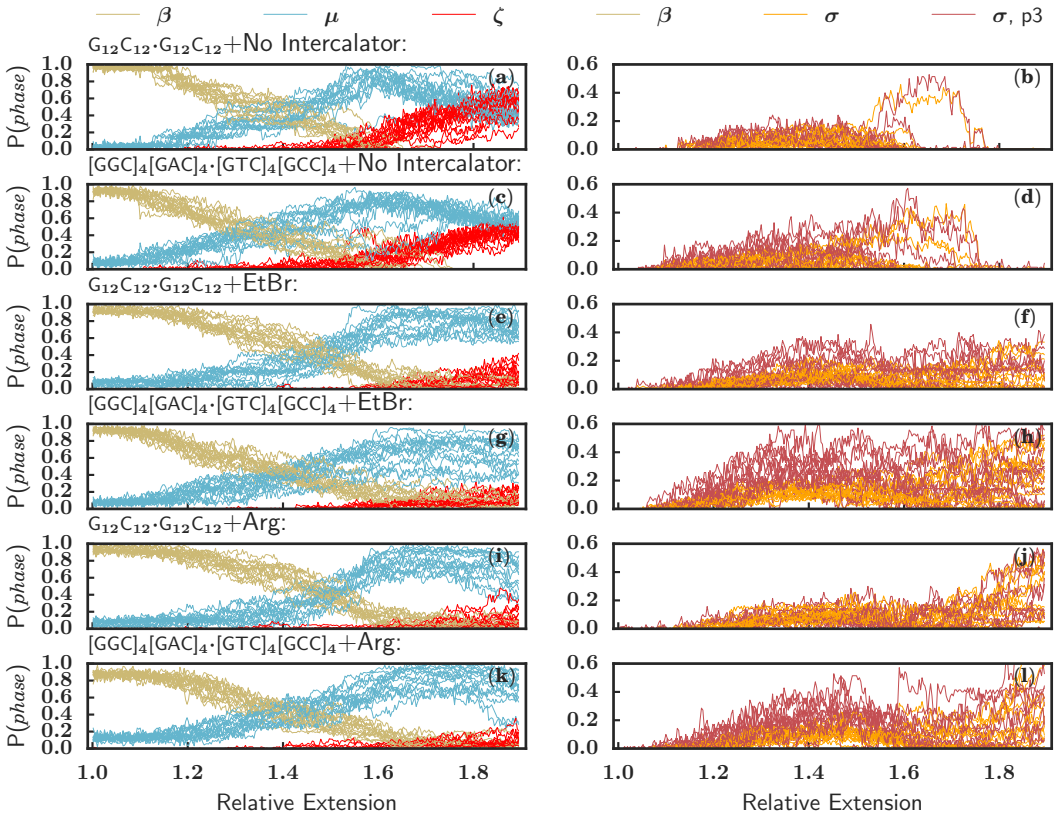


Figure 3.11: Inter-run variation of the proportion of different local conformations of different phases. Trends are consistent between instances up to extensions of > 1.5, where (especially without intercalator) strong kinetic lock-in becomes evident.

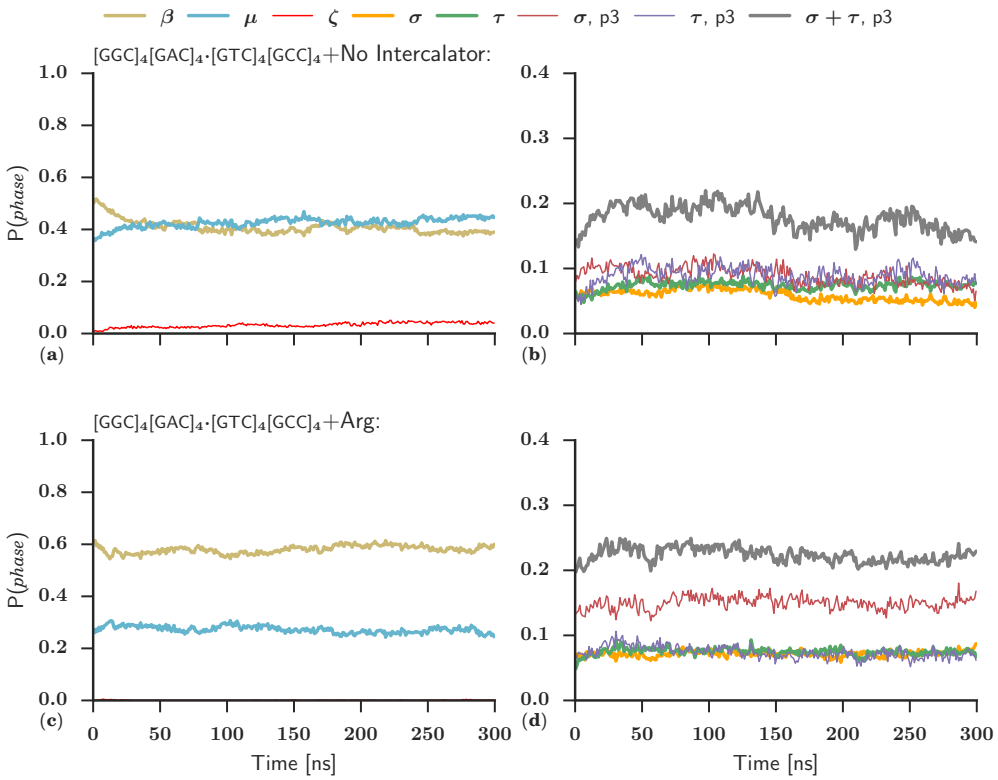


Figure 3.12: Proportion of each classified conformation versus time, at constant extension of 1.45, averaged over 16 instances and also smoothed over a 1ns window. The proportion of each conformation remains approximately constant over 300ns.

3.4 Discussion

The discussion of DNA over-extension has been consistently controversial from its beginning, as the complexity in response to a surprisingly large number of relevant variables (sequence content [75], solution conditions [184], temperature [221], intercalators, pulling geometry, pulling speed, and now also the specific codon content) which appears from an overview of the literature seems at first to be a simple case of contradictory results. Here we have tried to make a deeper examination of the codon content and solution condition variables to clarify this controversy, and especially to shed light on the elusive phase of DNA first proposed by Nordén which we have called Σ [113].

In the previous chapter (2) and in [182] we calculated that although the aqueous solution environment does not strongly drive triplet partitioning,

that a distinct hierarchy of triplet formation energies with respect to sequence features should exist. The triplet formation energy estimates showed that sequences coding for 'stage one' amino acids hypothesized to have appeared early in evolution (plus Arginine) are more likely than otherwise to partition into triplets at the codon boundaries when under tension.

In order to investigate this phenomenon we carried out pulling simulations of DNA duplexes encoding stage one amino acids, in the presence of Arginine and also of EtBr, as well as control simulations using low-entropy sequences, and in aqueous conditions with monatomic salt only. In order to observe strong triplet disproportionation both a bulky organic cation (specifically Arginine) and a sequence selected from codons yielding phase-one amino acids was required, with the combination of these two factors operating in a non-additive way to produce a solution structure of stacked base-pair triplets. Overstretching the Σ-duplex led to formation of interdigitated zipper DNA, stretching without cofactors or appropriate sequence led to disordered but not fully denatured structure consistent with other experiments [6, 113] and simulations [114].

We attempted to reductively study the factors driving Σ formation of DNA, which we deem to be force (applied via filament formation in the case of RecA, but capable of being applied in many ways), sequence content, and the presence of cationic amino acids (as part of a structured protein today, possibly not so in early biology). We have found that these factors operate collectively to drive Σ formation, in competition with tilting or melting of the DNA base pairs. We do not rule out Σ-DNA as a long-lasting stable collective structure of DNA while under tension in solution, however we are also not able to strongly confirm this as the microsecond time-scale accessed is considerably shorter than the longest (1-second) equilibration time-scale of the physical system.

We note that non-Σ conformations may still have an average inclination of zero, which casts some doubt on the claim from polarized light studies of base pairs [192] having no inclination when DNA is stretched in solution.

Chapter 4

Results: Section III

4.1 Thermodynamics of DNA Stretching in the Presence of Intercalators via a Coarse-grained Model

All-atom force fields, since their emergence in the 1980s [222], have evolved significantly and are able to reproduce structural changes from pico- to microsecond timescales [223, 224]. The longest all atom simulations on DNA reported to date have been a few microseconds long for a dodecamer sequence and hundreds of nanoseconds for a 150 bp long DNA [225], but studying some of the most important and interesting biological problems requires access to longer timescales which rules out direct use of all atom simulations. To obtain reliable statistics for some rare dynamical events like large conformational transitions or the flying in-and-out of bases (breathing motion) we need to run long timescale simulations. This is computationally demanding, and most energy is spent on microscopic fluctuations which average out during long timescales. We can use coarse-graining as a way to disregard irrelevant atomistic noise and facilitate the sampling of the more interesting long timescale behavior.

Unfortunately the field of DNA coarse graining is lagging behind those related to lipids or proteins and relatively few realistic CG models are available [226–229]. Most of the CG models available for DNA address only certain facets of DNA physics and are useful only for the designed purpose. DNA is a highly charged macromolecule, so correctly handling of electrostatic forces is difficult. Besides, many-body effects of the ionic environment

should be taken into account.

The way that microscopic details are coarsened is based either on a top-down or a bottom-up approach. In the top-down approach, the force field is chosen based on either a structural intuition or empirical data. In the bottom-up approach, the Hamiltonian is parametrized using all-atom simulations as a reference matching forces of energies for given conformations. Both approaches rest on the reliability the input data.

Studies on DNA extensibility started [120] even before the X-ray structure of DNA was defined [12, 230]. The structural response of DNA to overstretching has been controversial since the first single molecule experiments on λ-DNA [191]. To explain a sudden length increase in DNA at forces $\sim$65-70 pN, multiple theories have been presented [6, 75, 184, 221]. Two major attempts to explain the collective transition in DNA structure can be summarized in (a) the thermodynamic theory [98, 99] which proposes DNA melting and strand separation as the main mechanism for length increase and (b) the structural transition theory [100–103] which represents 'S-DNA' as the structure in which base pair stacking is broken while base pairing is maintained. The force at which overstretching happens depends on ionic strength [186], presence of nicks in the DNA backbone [129] and on the GC content of the sequence [75].

Recently we have suggested on the basis of MD simulations that for specific coding sequences (especially for codons of the pattern GNC, where N indicates 'anything'), and in the presence of appropriate intercalating cations, that a period three structure coined as Σ-DNA, with relevance to biological function [196], exists [182](see chapter 2). This conformation has been strongly suggested from experimental data even for a bare DNA sequence in a solution of monatomic counterions [113]. As the boundary CG·CG is the weakest of all base pair steps [10] (including the step GC·GC), for codons matching the GNC pattern, extension is naturally partitioned to the codon boundary (chapters 2 and 3). This observation provides a physical motivation for the early appearance of GNC codons in evolutionary history, and for the fact that they code for amino acids (GADV) which are among the most commonly appearing even in modern vertebrates [231]. For codons of other patterns this observation also holds albeit to a lesser extent, with decreasing propensity to form triplets with respect to the evolutionary newness of the amino acid earlier in the universal genetic code [182] (chapter 2).

DNA mechanical properties can be vastly altered by intercalators, which are often used as fluorescent probes or as drugs [217, 232–234]. Based on the preferred DNA binding mode, intercalators can be classified as mono-(like

ethidium bromide), bis-(like YOYO-1 [235]) or threading (like ruthenium complexes [236]); with different effects on DNA mechanical properties. The DNA force-extension profile is substantially changed in the presence of intercalators [232]. Intercalators are widely used as fluorescence probes to visualize DNA structural changes upon interaction with proteins and enzymes, therefore it is important to understand how they change DNA behaviour and consequently affect the study. Many potent anticancer drugs also bind DNA as intercalators such as (doxorubicin(adriamycin) [237], daunorubicin (daunomycin) [238], ditercalinium [239] and mitoxantrone [238, 240] (see chapter 1 Fig. 1.10)). As intercalation alters mechanical properties, stretching also effects properties of the bound intercalator, particularly the fluorescence quantum yield [241–243].

It has been shown that destacking of DNA bases is a necessary step for action of DNA recombinase enzymes such as RecA which is facilitated in the presence of EtBr [244, 245]. Several different studies have shown that in the absence of ATP, RecA filament is in the inactive (unextended) conformation but in the presence of ATP the filament adopts an extended form [246, 247]. The same extensibility is observed in the ss- and ds-DNA enveloped within the extended form of RecA filament [248]. DNA elongation which is achieved by increased spacing between bases in both ss- and ds-DNA facilitates homologous alignment. The fact that EtBr initiates the formation of DNA-RecA complex (even in the absence of ATP) implies that DNA conformation is changed so as to promote binding to RecA, in a triplet disproportionated form which we call "Σ-DNA". This disproportionation creates regular inter-bp spacing which increases the homologous pairing [249].

It is also discussed that the presence of intercalators causes significant changes in the quasiequilibrium force *versus* extension curve shifting the B to S transition to higher forces as well as decreasing the width of the transition plateau implying possible modifications in the structure of the S phase [232]. So simulated stretching of DNA in the presence of cofactors like EtBr is supposed to reduce the barrier and the collectivity associated with the structural transition of B to S. Here it is shown that overstretched DNA in the presence of (at least some) intercalator adopts a new structure in which base stacking is preserved in bunches of three base-pairs with a big gap in between each group of triplets. We have already named this new form of DNA Σ-DNA [182]. Formation of Σ-DNA in the presence of simple intercalators provides a candidate simple mechanism for DNA recombination in early forms of life where complicated enzymatic machinery like modern biology was not present.

Monte Carlo simulation is used here to provide a better understanding of DNA conformational change pathways, with emphasis on the newly discovered form of DNA (Σ-DNA). Considering the effect of intercalator concentration and the impossibility of doing MD simulations at very low intercalator concentrations (large water/EtBr ratio) we have developed a MC model to simulate the intercalation effect at very low concentrations in accordance with experimental values which are typically millimolar or less, due to the strong binding affinities (mM = 1 intercalator per 55000 water molecules).

4.2 Two-bead CG model of DNA

A two bead model based on the study of Sayar et al. [7] is introduced and then further developed. The parameters of the base model were extracted from full-atomistic DNA simulations *via* Boltzmann inversion, with no fitting for structural or mechanical properties. With only two beads the model captures the major structural features of DNA, such as: (1) the helicity and the pitch, (2) backbone directionality and (3) major and minor groove structure which results in the anisotropic bending rigidity. The accuracy of this CG model is mostly limited by the accuracy of the force field used in the full atomistic simulation which serves as a starting point.

4.2.1 The Model

The model is composed of two types of superatoms (P and B) per nucleotide, representing the collective motion of the backbone phosphate group + sugar (P) and the nucleic acid base (B). Here B superatoms are considered generic with no base specificity. The cartesian coordinates of P superatoms are chosen as the centre of mass of the atoms (O3′, P, O1P, O2P, O5′, C4′, O4′, C1′, C3′, C2′), while B superatoms are placed at the centre of mass of the atoms (N9, C8, N7, C5, C6, N3, C4) for purines and (N1, C6, C5, C4, N3, C2) for pyrimidines.

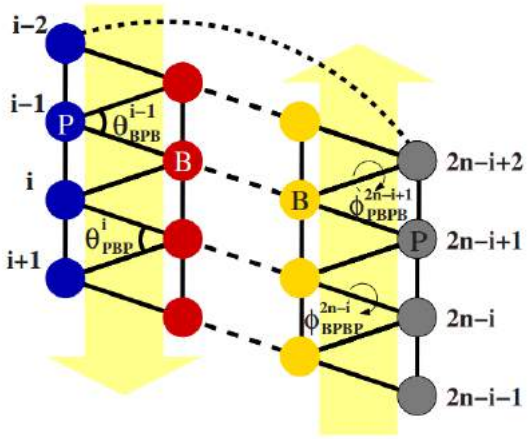


Figure 4.1: CG model of DNA molecule based on the superatoms B and P, where the former represents the phosphate backbone and the sugar group, and the latter represents the nucleic-acid bases. The superatoms P_i,B_i from the first strand and the superatoms P_{2n-i},B_{2n-i} from the second strand form the nucleic acid base pairs which are connected by hydrogen bonds in the original system. The intrastrand bonds P_iB_i, B_iP_{i+1}, P_iP_{i+1}, B_iB_{i+1} are shown by solid lines. The interstrand bonds B_iB_{2n-i} and P_iP_{2n-i} are shown by dashed lines. $\phi^{\,i}_{PBPB}$ and $\phi^{\,i}_{BPBP}$ are the dihedral angles defined by P_i, B_i, P_{i+1}, B_{i+1} and B_i, P_{i+1}, B_{i+1}, P_{i+2} respectively. Similarly, $\theta^{\,i}_{BPB}$ and $\theta^{\,i}_{PBP}$ represent the bond angles defined by B_{i-1}, P_i, B_i and P_i, B_i, P_{i+1}. The dihedral angle stiffness is explicitly included in the CG potential, whereas the bond angle stiffness arises implicitly, mostly due to the intrastrand PP and BB bonds (Image taken from [7].)

4.2.2 Interactions

The effective interactions incorporated into the model are four bonded and two dihedral potentials that maintain the local single-strand geometry: (1) harmonic bonds P_iB_i, B_iP_{i+1} and $B_{i-1}B_i$ that fix the intrastrand superatom distances as well as the angles ${\theta_{PBP}}^i$ and ${\theta_{BPB}}^i$; (2) dihedral potentials associated with the angles ${\phi_{PBPB}}^i$ and ${\phi_{BPBP}}^i$. All four harmonic bond potentials have the form:

$$V_b(r) = \frac{1}{2} K_b (r - r_0)^2 \tag{4.1}$$

where the stiffness constants, K_b, and the equilibrium bond lengths, r_0, differ as listed in Table I. In particular, the differences between P_iB_i and B_iP_{i+1} bond parameters reflects $5' \rightarrow 3'$ directionality of the molecule.

The choice of harmonic bond potentials P_iP_{i+1} and $B_{i-1}B_i$ over true angular potentials increases computational efficiency without significantly distorting the equilibrium distributions. The torsional stiffness of the dihedral angles $\phi^{\,i}_{PBPB}$ and $\phi^{\,i}_{BPBP}$ defined respectively by the superatoms

$P_iB_iP_{i+1}B_{i+1}$ and $B_{i-1}P_iB_iP_{i+1}$ is modeled by the potential,

$$V_d = K_d[1 - cos(\phi - \phi_0)] \tag{4.2}$$

Table 4.1: Force constants of the SAK model.

Interaction type	Equilibrium Positions	Force constants
P_iB_i bond	r_0=5.45Å	K_b=7.04 $k_BT/Å^2$
B_iP_{i+1} bond	r_0=6.09Å	K_b=16.14 $k_BT/Å^2$
P_iP_{i+1} bond	r_0=6.14Å	K_b=20.36 $k_BT/Å^2$
B_iB_{i+1} bond	r_0=4.07Å	K_b=15.93 $k_BT/Å^2$
PBPB dihedral	ϕ_0=3.62 rad	K_d=25.40 k_BT/rad^2
BPBP dihedral	ϕ_0=3.51 rad	K_d=27.84 k_BT/rad^2

where, again, the two stiffness coefficients, K_d for BPBP and PBPB dihedral angles and their equilibrium values, ϕ_0, are separately determined from the full-atomistic simulation data.

In addition to the intrastrand interactions, two interstrand potentials that stabilise the double-stranded structure of the model are also defined. The first interaction, which is among B_iB_{2n-i} superatoms, reflects the hydrogen bonding between the nucleic acid bases A-T or G-C. The second interstrand interaction is a pair-specific potential among P_i and P_{2n-i} and stabilises the positioning of the two strands on opposite sides of the helical axis by maintaining the particular arrangement of the four superatoms which represent complementary nucleotides. In this model all the hydrogen bonds are represented by a single inter-BB interaction, which is not sufficient to prevent such folding of the strands. Therefore, one other interstrand interaction is included in the Hamiltonian: a Lennard-Jones excluded volume potential between all superatom pairs (except P_iP_{i+2}) that do not otherwise interact maintains the self-avoidance of the DNA chains. The excluded volume of the superatoms is represented via a repulsive Lennard-Jones interaction:

$$U_{LJ}(r) = \begin{cases} 4\left[\left(\frac{r_0}{r}\right)^{12} - \left(\frac{r_0}{r}\right)^{6} + 0.25\right] & r < r_0 \\ 0 & r \geq r_{cut} \end{cases} \tag{4.3}$$

The upward shift of the U_{LJ} is to avoid a jump discontinuity at r_{cut}. In this way the truncated potential would be exactly zero at the cut off distance.

For all B-B and B-P pairs r_0 = 5.35 Å and r_{cut} = 6 Å, whereas for all P-P pairs these values are doubled. Superatom pairs that are bonded and all P_iP_{i+2} pairs are excluded from these Lennard-Jones interactions.

The force constants and the equilibrium values for bond and dihedral potentials which are used in this model are obtained from the thermal fluctuations of the associated superatoms via Boltzmann inversion. The fluctuation data is obtained from molecular dynamics (MD) trajectories of full-atomistic benchmarking study [250]. Boltzmann inversion of the probability distributions for each type of bond length, bond angle, and dihedral angle yields potentials of mean force (PMFs). These PMF curves were used by the authors of our reference model to obtain force constants for the intrastrand bonded interactions by means of harmonic fits.

The B_iB_{2n-i} interaction is incorporated into the model via a tabulated potential kindly provided by Sayar *et al.* In the reference paper, the tabulated potential is obtained via a piece-wise continuous curve fit to the PMF data. The numerical values of the potential at a number of points are stored in a table and used in our MC program. The P_iP_{2n-i} PMF curve also displays an equilibrium separation. In this model only the repulsive part of this interaction is modeled via a tabulated potential.

Based on the initials of the authors, we call this force field (ff) "SAK" [7]. In summary the effective interactions included in this force field are:

- Harmonic bonds
- Dihedral torsions
- Repulsive-only Lennard-Jones interactions
- A tabular potential specific to paired bases stabilizing the double-helix structure

Electrostatics are not modeled explicitly. As the model is parametrised it is suitable for physiological salt concentration [7].

4.2.3 Modifications to the SAK force field (SAK*ff and SAKI)

In order to model DNA under tension, the SAK ff was modified slightly (we call the new form SAK*ff). A cutoff of 5.6 Å was added to the (otherwise harmonic) base stacking attraction, allowing bases to de-stack after reaching a separation great enough that water could hypothetically enter between them, screening any dispersion forces and reducing the energetic cost to further separate the bases.

In the same spirit, a distance-dependence was added to the dihedral terms of the Hamiltonian, such that for a B-B super atom distance greater than 5.6 Å the energetic cost to untwist this dihedral angle (BPBP) is reduced quickly to zero, reflecting the physical origin of DNA twist as arising from the competition between the backbone length and the shorter length scale of the base stacking.

The forcefield was then further modified to include the effect of intercalation (SAKI), which is also introduced mathematically here. An integer valued hidden state s is added to the model per base pair, such that it can have three values $s \in \{0(B-DNA), 1(S-DNA), 2(I-DNA)\}$:

(1) **Bonds:**

$$
\begin{array}{llll}
\multicolumn{4}{c}{I=0} \\
\hline
V_b(r) = \left\{\min(\frac{1}{2}k_b(r-r_0)^2, V_{max})\right\} & & r < r_{destack} & \text{SAKff} \\
\multicolumn{4}{c}{I=1} \\
\hline
V_b(r) = \left\{\min(k_b(\frac{1}{2}(r-r_0)^2 + (r_{destack}-r_0)(r-r_{destack})) + \lambda), V_{max}+\lambda)\right\} & & r > r_{destack} & \text{SAK*ff} \\
\multicolumn{4}{c}{I=2} \\
\hline
V_b(r) = \left\{\min(\frac{1}{2}k_b(r-r_{intercal})^2 + \epsilon, V_{max})\right\} & & & \text{SAKI}
\end{array}
\tag{4.4}
$$

where K_b is the stiffness constant, r_0 is the equilibrium bond length, $r_{destack}$ is the bond length beyond which bases start to destack, $r_{intercal}$ is defined based on the maximum bond length of intercalated bases ($\sim$6.0Å).

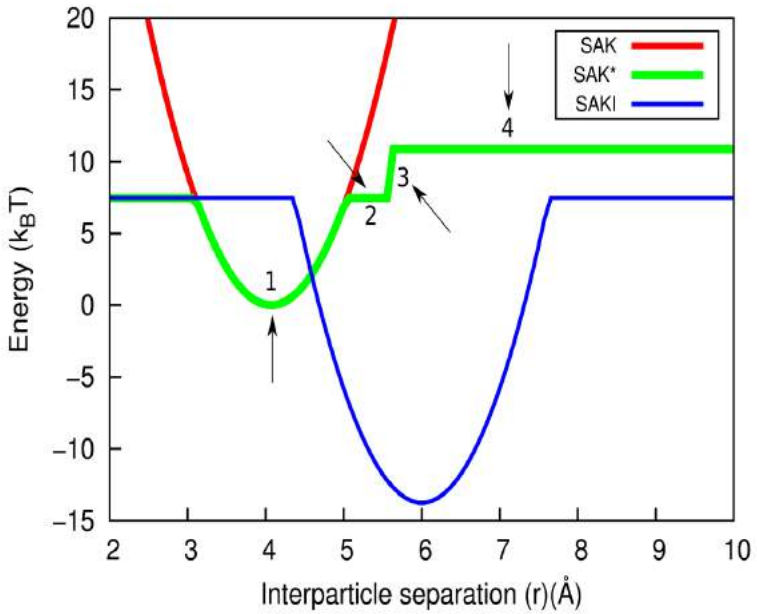


Figure 4.2: Original SAK bond potential along with modified SAK*ff for stretched DNA and SAKI for intercalated DNA are shown based on the equation 4.4.

A combined harmonic-nonharmonic potential as shown in Fig. 4.2 is composed of 7 regions:

1. Harmonic part
2. Stack melted/water entering
3. Force into zipper formation
4. Zipper

(2) **Dihedral torsions:** For a dihedral BPBP or PBPB if r is defined as the distance B to B:

$$V_d = \begin{cases} K_d[1 - cos(\phi - \phi_0)] & r < r_{destack} \\ K_d[1 - cos(\phi - \phi_0)]/[5(r - r_{destack}) + 1] & r > r_{destack} \end{cases} \quad (4.5)$$

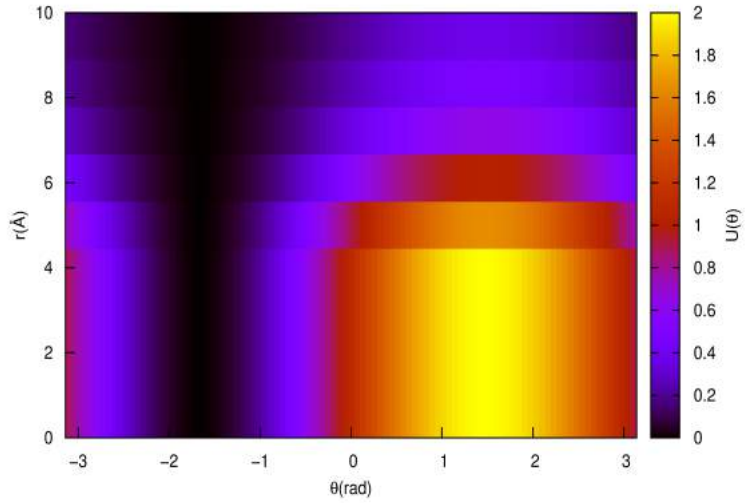


Figure 4.3: As the DNA is overstretched and rise reaches the threshold of 5.6 Å chirality of the DNA molecule starts to diminish.

In Fig.4.3 changes of the dihedral torsion is shown. As the DNA is overstretched and undertwisted the chirality of the dihedral angle starts to diminish.

4.3 Monte Carlo simulation of DNA

The Monte Carlo sampling applied here proceeds according to the Metropolis prescription, which involves optimizing the sampling of the configuration space and hence convergence via importance sampling consistent with Boltzmann statistics and microscopic reversibility. Monte Carlo moves are not bound to be local. They can be tailored to alter large portions of a chain, thereby promising efficient equilibration. Five different Monte Carlo moves are designed for this study of which is explained briefly.

4.3.1 A brief overview of the Monte Carlo method

In equilibrium statistical mechanics thermodynamic properties are calculated as ensemble averages over all points x in a high dimensional configuration space Γ. In the canonical ensemble the average of an observable $A(x)$ is given by

$$\langle A\rangle = \int dx\ A(x)P_{eq}(x) = \frac{1}{Z}\int dx\ A(x)[\exp(-\beta U(x))] \tag{4.6}$$

In general, the integral cannot be solved analytically. Monte Carlo (MC) simulations provide a numerical approach to this problem by generating a random sample of configuration space points $x_1, ..., x_m, ..., x_M$ according to some distribution $P_s(x)$. $\langle A\rangle$ is then estimated by:

$$\bar{A} = \frac{\sum_{m=1}^{M} A(x_m)e^{-\beta U(x_m)}/P_s(x_m)}{\sum_{m=1}^{M} e^{-\beta U(x_m)}/P_s(x_m)} = \frac{\sum_{m=1}^{M} A(x_m)W(x_m)}{\sum_{m=1}^{M} W(x_m)} \tag{4.7}$$

Here the "weight" $W(x) = P_{eq}(x)/P_s(x)$ in introduced. It should be noted that while $\langle A\rangle$ is a number, $\bar{A}$ is still a random variable. Whether $\bar{A}$ represents a good estimate for $\langle A\rangle$ depends on the total number M of configurations used and for a given M on the choice of $P_s(x)$.

$P_s(x)$ should approximate $P_{eq}(x)$ as closely as possible to obtain meaningful results from MC simulations. Regarding this fact, two approaches are considered:

1. *Static MC methods*: Static methods generate a sequence of statistically independent configuration space points from the distribution $P_s(x)$. In this case one has to tune the algorithm cleverly so that weights $W(x)$ do not get out of hand.

2. *Dynamic MC methods*: These methods generate a sequence of *correlated* configuration space points via some stochastic process which has $P_{eq}(x)$ as its unique equilibrium distribution. This process is usually taken to be a Markov process [251]. A Markov process is one which has no "memory". That is, the probability for the occurrence of the future configuration x depends only on the present configuration x' and not on the other configurations that the process visited in the past.

4.3.2 Translational move

In most Monte Carlo studies a strategy is defined by the translational displacement of a single particle. Here, the step size is adjusted to attain a 30-50% acceptance rate, which produces satisfactory convergence. The method is called adaptive step-sizing. This move is done through iterations of the

following steps for N particles:

1. Randomly pick one of N particles.
2. Perturb each of the x, y, z coordinates by a predefined step size.
3. Compute the change in potential energy due to the particle move.
4. Apply Metropolis criterion to accept or reject the movement.

4.3.3 Crankshaft move: rigid rotation

In this move a consecutive group of N particles are selected starting and ending at P super-atoms. An axis of rotation is selected based on the centre of mass of the start and end of the cluster and rotated by an angle τ. The angle of rotation is changed adaptively during the simulation to give an acceptance rate of $\sim$30%. A Rotation matrix in 3D based on Euler angles is used for this move.

4.3.4 Topology changing rotation

Response of DNA to over-stretching depends on whether it is torsionally relaxed or constrained. In this regard a complex MC move is designed which helps to untwist or over-twist the DNA. The accepted move causes a quarter-turn opening of the DNA and decrease or increase of the twist. Quarter-turns ($\frac{\pi}{2}$) are made possible by rotations at the periodic boundaries.

4.3.5 Extension move

In this move force is applied on DNA by altering the periodic boundary conditions along the z-axis. F_i is the external force applied on the DNA with the extension step dz so the term Fdz contributes to the change of internal energy of the DNA molecule. The term

$$F\Delta z \tag{4.8}$$

is included in the Boltzman factor of the Metropolis acceptance criterion,

$$e^{(\beta F \Delta z)} \tag{4.9}$$

where Δz is the extension step. So ΔE would be

$$min\Big[1, e^{-\beta(\Delta E - F\Delta z)}\Big] \tag{4.10}$$

4.3.6 intercalation move

Assuming the work done on DNA is in the form of mechanical work (extension) and chemical work (intercalating particles), this move takes into account the chemical potential (μ) of the intercalator (EtBr). Thus the thermodynamics of the system is affected by the imposed force [103] which is the additional source of work done on the system as well as by the interacting particles which bring the chemical potential into play [252]. In the same way as the mechanical work the chemical potential term,

$$\mu \sum_i N_i \tag{4.11}$$

is added to the standard Metropolis criterion, where μ is the chemical potential and N is the number of intercalated super-atoms, changing the energy of the system at different intercalating particle concentrations, such as the acceptance term becomes

$$min\Big[1, e^{-\beta(\Delta E + \mu \sum_i N_i)}\Big] \tag{4.12}$$

4.3.7 Pulling Scheme

Pulling experiments are done in different schemes. Force can be exerted on $3'3'$, $5'5'$ or $3'5'$ ends of a DNA molecule. Differences are found in the response of the overstretched DNA with regard to the pulling mechanisms [127]. In our model extension force is applied via the periodic boundaries of the simulation, a pulling mechanism which suppresses unpeeling. The initial setup of the configuration is such that the DNA is topologically closed, however a Monte carlo move is included such that the linking number can relax to the equilibrium value for the given extension. The effect of different pulling mechanisms on the mechanical response of DNA is out of the scope of the current study as we have limited the pulling scheme to uniform application of force via the PBC.

In Fig. 4.7 a snapshot of the DNA configuration at $F = 150$ pN is shown as it is forced through the overstretching transition. As the DNA is stretched, the base steps do not increase isotropically. As expected for an anharmonic model, as individual base steps pass the threshold for the BB stacking interaction, they then extend further, allowing neighbouring bases to relax.

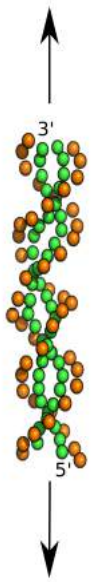


Figure 4.4: Pulling force is applied via the periodic boundaries. This pulling mechanism prevents the unpeeling of DNA strands.

The lengthscale of the stretched BB-rise is significant in that this is comparable to the PP bead equilibrium separation in the SAK and SAK* force fields: on breaking of the BB stacking interaction, the force is then transferred to the sugar-phosphate backbone after liberating about 2 Å of extension.

4.4 Mapping an atomistic model onto the CG model: analysis of DNA structural parameters

In order to be able to use standard programmes for structural analysis of DNA, which are implemented with respect to atomistic descriptions, we have used the NAB module of the Amber suite [253] to map an atomistic model onto the CG model of DNA. The P and B superatoms are aligned to onto O3′, P, O1P, O2P, O5′, C4′, O4′, C1′, C3′, C2′ and C4, C5, C6, C8, N3, N7, N9 for purines(GUA) or C2, C4, C5, C6, N1, N3 for pyrimidines (CYT). The

workflow is such that each frame of the CG model is mapped onto a corresponding atomistic model and then fed into 3DNA for further analysis (Fig. 4.5). As 3DNA does not conveniently account for periodic DNA structures, the first and last two base-pairs are ignored in the analysis. It has already been shown that rise, roll and twist are the step parameters of importance in the intercalation process [254].

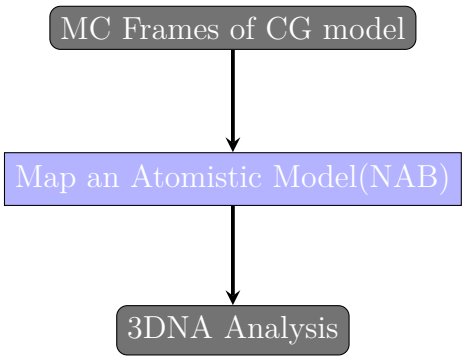


Figure 4.5: For convenient visualisation and analysis we map specific atom groups for all-atom DNA onto the superatoms of the SAK* ff model.

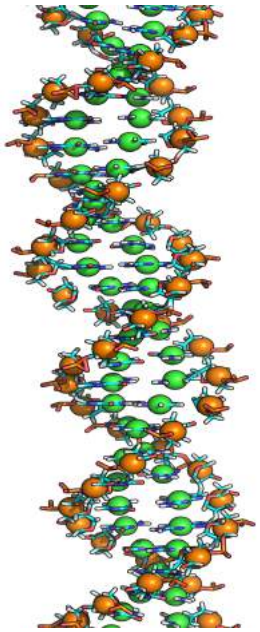


Figure 4.6: Mapping an atomistic model onto the CG model. P and B superatoms are shown in orange and green respectively and atomistic model is represented with sticks.

4.5 Intercalators and force-extension curves

The influence of intercalating molecules on the mechanical behavior of DNA can be studied by investigating force-extension curves. Several experimental studies have been performed on the shape of the force-extension curve of dsDNA as a function of intercalator concentration [217, 232].

Both studies have measured several force-extension curves for varying concentrations of particles. The curves are equilibrium curves; they are measured on time scales that allow the system to equilibrate. The curves with relatively high concentrations of intercalator show different qualitative behavior than the zero-concentration curves. The experimentally observed effect of intercalator-induced shift of the overstretching transition, towards higher forces, is reproduced (Fig. 4.11).

When particles bind to DNA via intercalation, they alter the local structure of the DNA molecule in a different way than overstretching does. A planar molecule or a planar part of a molecule is inserted between normally neighboring base pairs in a plane perpendicular to the helical axis. It was experimentally observed [255, 256] that a bound intercalator inhibits other

intercalators from binding at adjacent binding sites on the DNA molecule. This is called the "neighbor-exclusion" principle [257]. This exclusion principle is not caused by direct repulsion between intercalated molecules, but is a result of intercalator-induced structural changes in the dsDNA. An important consequence of the exclusion principle would be the fact that a saturated DNA molecule has only half its binding sites occupied. This is in accordance with the experimentally observed 1.5-fold elongation (instead of 2-fold) of the DNA molecule at saturation [258].

Yan and Marko [259] predicted that at high stretching forces the maximum binding could be increased to one intercalating molecule per base pair. Indeed, Vladescu [232] showed that at high stretching forces the overall contour length of saturated DNA (0.68 nm per base pair) was twice as long as for B-DNA (0.34 nm per base pair). This violation of the neighbor-exclusion principle is attributed to the fact that the exclusion is mediated by structural changes in the DNA backbone. Apparently the strong stretching forces then cancel these structural changes.

4.5.1 Parametrization of ff SAK-Intercalators (SAKI)

Based on the theoretical studies of van der Schoot *et al.* [260] we assign a free energy penalty ϵ to a segment in the intercalated state, which shows that B-DNA is the preferred state in the absence of any force. As the CG model of DNA does not discriminate between bases, this parameter (along with other parameters which will be explained) can be considered as averages over all nucleobases in the DNA.

From experimental data of the overstretching transition [112] it is clear that it is cooperative in nature. In other words, as soon as the helix overstretches at some location along the DNA molecule, it does so everywhere. This is indicative of the existence of a free energy cost upon the creation of an "interface" between BDNA and overstretched DNA. At this interface there is a segment in the dsDNA molecule that adopts properties of both B-DNA and overstretched DNA. This intermediate state is energetically unfavorable, resulting in a free energy cost for such an interface. In terms of our model, such an interface exists between a base-pair in state 0 (B-DNA) and a base-pair in state 1 (stretched DNA). We assign a free energy penalty to each such interface and call it the cooperativity parameter and use the symbol λ for it.

dsDNA stretches locally to twice its normal contour length if an interca-

lator is bound. If an intercalator binds the dsDNA, the free energy penalty λ might be overcome by the free energy bonus for binding. At moderate stretching forces, a bound intercalator prohibits other intercalators to bind at adjacent binding sites on the dsDNA. The neighbor-exclusion principle leads to an 1.5-fold elongation of the contour length at saturation, but as mentioned previously, it is theoretically predicted that at high stretching forces the maximum binding could be increased to one intercalator per base pair [260].

4.5.1.1 Cooperativity parameters for the three-state model

A free energy is associated with changing a base-pair from B-DNA to intercalated DNA. The free energy penalty for state $S_i = 2$ (intercalated) is more complicated than the penalty for state $S_i = 1$ (stretched), because particle binding is involved. The intercalator studied in this model is assumed to be mono-intercalator, one intercalator can bind per base-pair. The free energy penalty related to changing a base-pair from state 0 to state 2 and binding an intercalator, ΔE, is given by:

$$\Delta G = \epsilon - \mu \tag{4.13}$$

The first term, ϵ is the free energy penalty that is associated with changing the dsDNA from state 0 to state 2. This free energy contains contributions from binding of the intercalator and deformation of the DNA. It also contains a contribution related to the interaction of unbound intercalators with the surrounding solution [261]. The second term is the chemical potential, μ, of free intercalators in solution, which are available for binding to the DNA molecule which is defined as the Gibbs free energy per particle. The chemical potential term expresses that more particles bind to the dsDNA if more particles are available in the solution. Assuming the solution is ideal, the chemical potential μ is given by

$$\mu = k_B T \ln\left(\frac{C}{55.6}\right) \tag{4.14}$$

In the experimental setup to study DNA overstretching in the presence of intercalators [262], only one dsDNA molecule was available, thus, the number of binding locations is much smaller than the number of free intercalators. So we can treat μ as a constant that only depends on the experimentally controlled intercalator concentration.

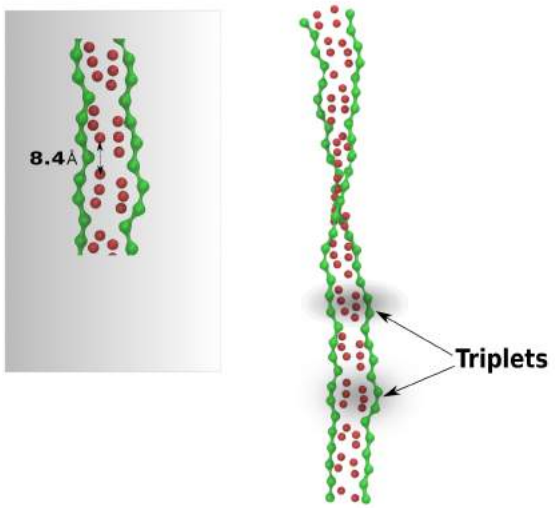


Figure 4.7: A snapshot of the base-pairs in the overstretched DNA. The DNA is underwound and stretched globally but locally it adopts a B-DNA like conformation. The Σ triplet repeating unit of stacked base-pairs is evident. Each triplet is separated from adjacent triplet base-pairs by a gap. Base-pair structure of the triplets closely resembles B-DNA with a base pair separation of ~ 3.6Å. Backbone atoms (green) are shown in a continuous manner to discriminate them from bases (red). The step going from one triplet to the next has a stretched base-base distance of ~ 8.4 *Å* (inset shows the big gap between each triplet cluster with the next one).

Now we have a three state model of DNA, any of these states can be neighbors to each other, so we need more than one cooperativity parameter to correctly describe DNA interaction with the intercalator. The Neighbor-exclusion principle inhibits two neighboring base-pairs from both being in state 2 but in the high force regime the exclusion principle can be violated. A free energy penalty for a 2/2-interface, δ is introduced in the model, which is positive, to explain the experimental observations; at low forces the penalty prevents a 2/2-interface from occuring, but at high forces the amount of work done becomes important. State 2 is the longest state, so if the stretching force is sufficiently large it might overcome the free energy penalty δ and allow neighboring particles to both be in state 2.

A lower value of δ shifts the overstretching transition to higher forces and makes it easier for two intercalators to be next to each other. Based on the experimental results a range of values is selected for δ which is $2.5 < \delta < 5.5$ k_BT.

Based on the work of Biebricher *et al.* [217] a free energy penalty, η, is introduced for a 1/2-interface, showing that intercalated dsDNA molecules do not overstretch cooperatively. A positive value of η is essential for the

shift of the overstretching force [260]. This suggests that intercalated DNA is stabilized against overstretching. So, a base-pair in the ground state (state 0, B-DNA) has more difficulty overstretching (to state 1) if other base-pairs are in the intercalated state (state 2).

Table 4.2: The cooperativity parameters of the 3-state CG-DNA model. The row number gives the state of the i th base-pair, while the column number gives the state of base-pair $i+1$. The symbols in the table give the free energy penalties that are associated with the interfaces between segment i and $i+1$. In this study λ, δ and η are all positive.

	$S_{i+1}=0$	$S_{i+1}=1$	$S_{i+1}=2$
$S_i=0$	0	λ	0
$S_i=1$	λ	0	η
$S_i=2$	0	η	δ

The free energy penalty δ, penalizes interfaces between two intercalated base-pairs and is thus responsible for the neighbor-exclusion principle. Extensive simulations for parametrization of the force-field showed that the window of chemical potentials that shift the overstretching transition is a function of δ. The second variable which needed to be parametrized is ϵ (free energy binding of the intercalators). The binding free energy of dsDNA intercalation is a free energy bonus of at least $\sim 20\ k_BT$ and the question is why this interaction free energy is so large?

Atomistic simulations showed that stacking energies between ethidium and A-T base pairs, at a mutual distance of 0.33 nm (approximately the distance between ethidium and the nearest base pair in intercalated dsDNA), are in the order of $\sim 38\ k_BT$ [263]. Šponer *et al.* showed the stacking energies between different base pairs are in the order of $-11\ k_BT$ to $-19\ k_BT$, depending on the type of nucleotide [264]. The difference between these interactions gives an indication of the binding free energy for intercalation, which is in the order of $-20\ k_BT$. Řeha and co-workers attributed this large interaction free energy to electrostatic and dispersion forces, as the intercalator moiety is charged and polarizable [263]. Hydrophobic forces are also suggested to play a role in ethidium-dsDNA interactions [265]. So ethidium molecules break their interaction with water molecules upon intercalation.

4.5.2 DNA over-stretching in the absence and presence of intercalators

The most frequently collected observable in DNA-stretching is the force-extension curve. Fig. 4.8 shows a calculated curve for the modified force field (SAK* ff), accounting for base destacking and weakening of angular restraints with unstacking. The free energy penalty of λ can reproduce the experimental results regarding the start point of collective transition which causes the DNA to over-stretch to $\sim$1.7 its original length as well as the extent of the plateau.

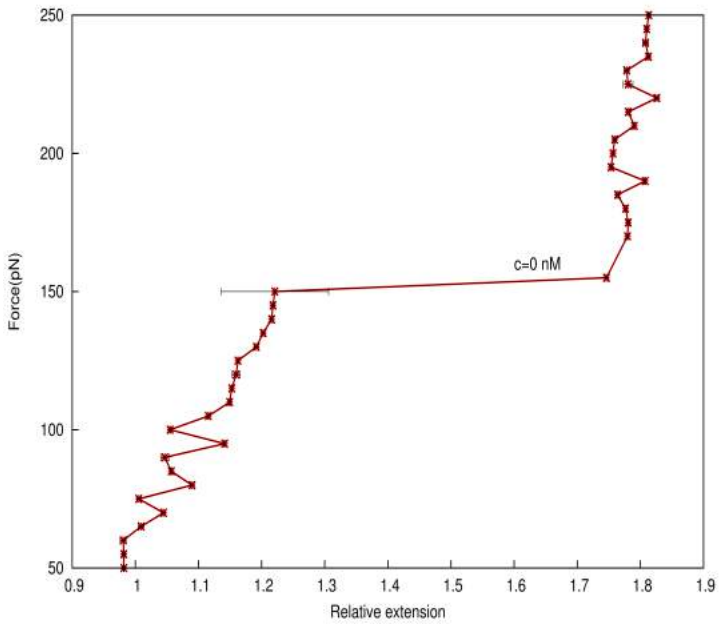


Figure 4.8: Force-extension curve for duplex DNA overstretching obtained from one Monte Carlo simulation. Each data point corresponds to a single simulation at constant force.

Due to the simplicity of the model and the environment, particularly the lack of base specificity and ignorance of thermal agitations over-stretching happens at higher forces comparing to experiment, in good agreement with atomistic simulations and qualitative but less good agreement with experiments [261, 266]. Here we have chosen a 24 base-pair long DNA to observe

the structural transition. Our reasons are that first we can compare our results with our MD simulations [182](see chapter 3) and second, as it has been shown, the cooperativity length for the B to S transition in DNA is ∼22-25 base pairs [267, 268].

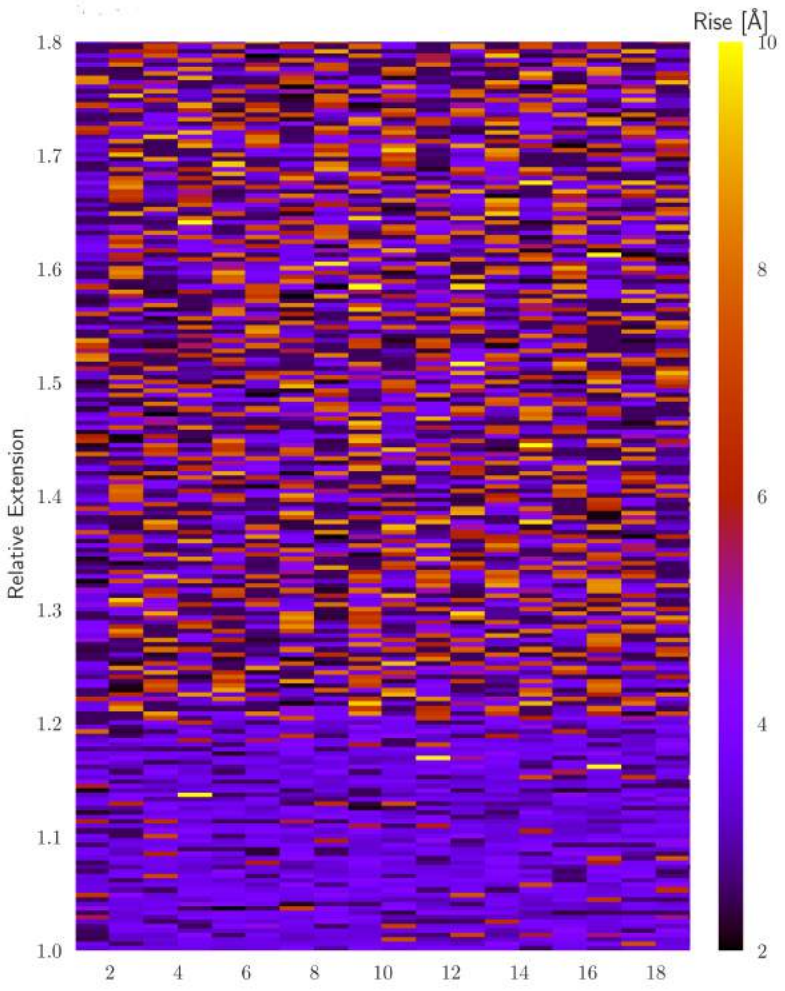


Figure 4.9: Figure shows the kymograph of rise per base-pair for the DNA extension in the absence of intercalators for a single simulation.

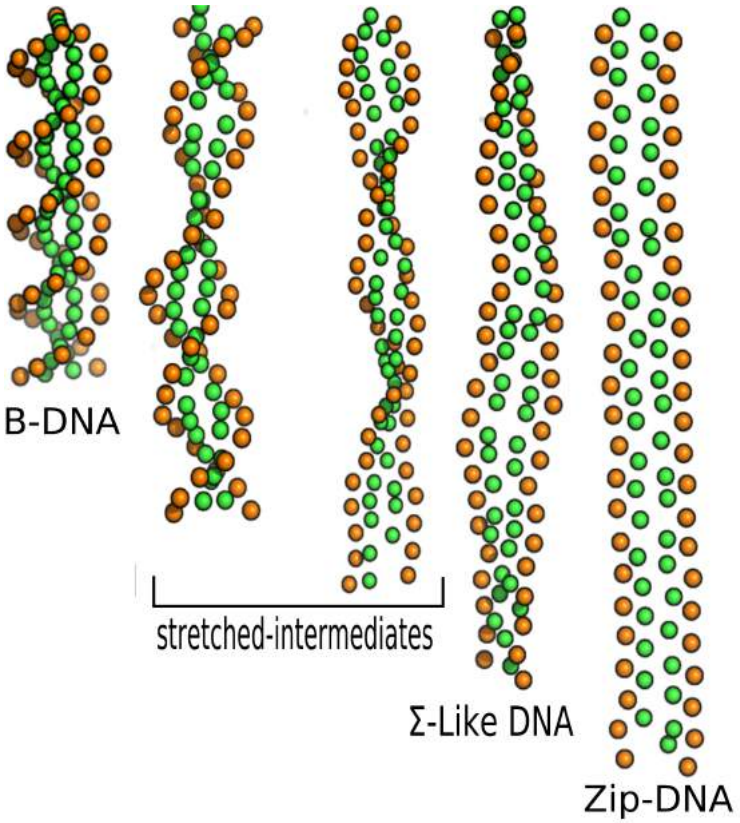


Figure 4.10: Snapshots of overstretched DNA conformation in the absence of the intercalator. DNA untwisting happens as the force is increased gradually until the DNA is completely untwisted and forms the "zipper-like DNA". Green and orange beads represent the bases and the backbone respectively.

As shown in the Fig. 4.10 DNA over-stretching in the absence of the intercalator results in the gradual unwinding of the DNA until the linking number drops to zero and the so-called "Zip-DNA" is formed. In the absence of an intercalator over-stretching is locally anti-cooperative but shows no long-range pattern of disproportionation into triplets.

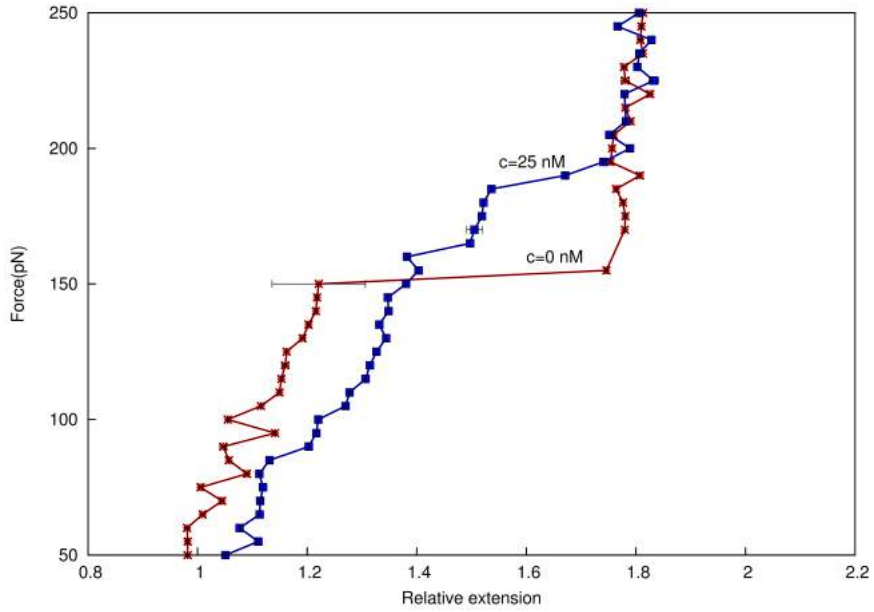


Figure 4.11: Calculated force-extension curve of DNA over-stretching in the absence of intercalator and 25nM concentration of intercalator. As expected the transition force is shifted to higher values in the presence of intercalator and the width of the transition plateau decreases in qualitative agreement with experimental results. A collective transition to over-stretched DNA happens at 150 pN in the absence of the intercalator. Early simulation results have found that when the twist is allowed to drop, the over-stretching transition force is ∼150 pN.

In the Fig. 4.12 it is shown that over-stretched intercalated DNA (25 nM of EtBr) adopts a new conformation consistent with Σ-DNA. This structure shows triplet disproportionation. These results are in accordance with our previous proposition regarding the formation of Σ-DNA [182] and with early predictions of Takahashi *et al.* [245] which proposed RecA binding to dsDNA is accelerated in the presence of intercalator EtBr. Here our results show EtBr could facilitate RecA binding to DNA through formation of triple base-pair stacks.

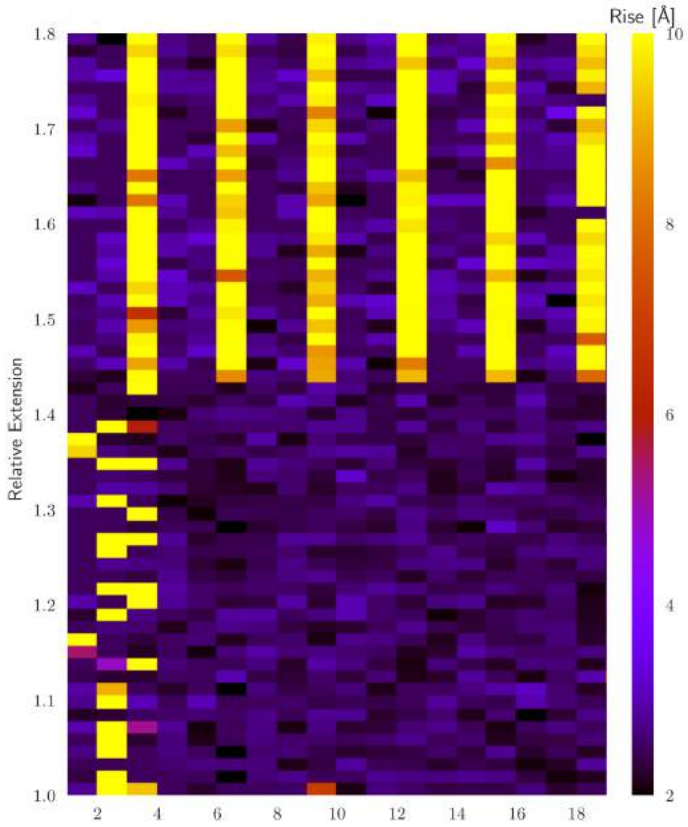


Figure 4.12: Kymograph of rise per base-pair of DNA over-stretching in the presence of intercalators for a single simulation. Triplet disproportionation is evident in the strained DNA.

Fig. 4.13 shows that DNA over-stretching in the presence of an intercalator (EtBr (25 nM)). The structure is inhomogeneous in a way that is consistent with Σ-DNA or triplet disproportionation. These results provide an explanation of the stimulation of complex formation between RecA-DNA in the presence of the EtBr [244, 245, 269] and are consistent with previously reported MD results of DNA over-stretching in the presence of a bulky intercalator (see chapter 3 [182]). It seems that the interaction of EtBr with DNA provides the RecA with the conformation which is required to form the final complex, an inhomogeneously stretched structure.

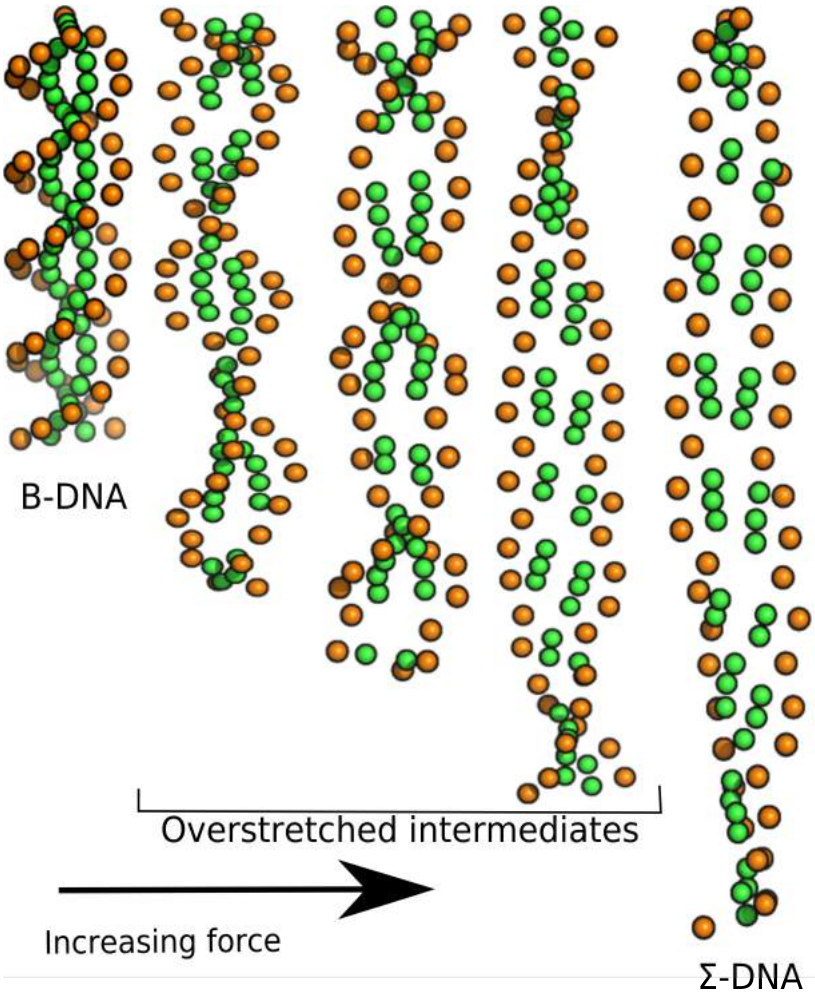


Figure 4.13: DNA over-stretching in the presence of an intercalator produces a unique conformation in which stretch is spread nonhomogeneously. Stacking interaction is preserved within the triplets with the rise parameter within the range of B-DNA. Stacking is broken between consecutive triplets providing a big gap assumed to be necessary for base flipping and homology search in recombination.

4.5.3 Rationale for Unequal Partitioning of Extension

The unequal partition of extension in the the SAKI model could be attributed to the free energy penalties associated with the interaction of intercalant with DNA. We can compare this by defining the case in which unequal partition is energetically favourable to equal partition, the term on the right indicates the energetic cost of partitioning the extension for three base-pair steps into only one step, illustrating this with the following counterexample of a harmonic potential, where the left side shows the cost of three extensions of size δ,

while the right side shows the cost of one extension with size 3δ.

$$\sum_{i=1,3} k\delta^2 < k(3\delta)^2 \tag{4.15}$$

$$3k\delta^2 < 9k\delta^2 \tag{4.16}$$

in the harmonic case the partitioning is clearly unfavourable. If we set a threshold x_{max} however:

$$\sum_{i=1,3} k\delta^2 < kx_{max}^2 \tag{4.17}$$

$$3\delta^2 < x_{max}^2 \tag{4.18}$$

we then see a clear and interesting expression for the situation in which the unequally partitioned extension becomes favourable: $x_{max} < \delta\sqrt{3}$, or the threshold for the hydrophobic interaction is less than $\sqrt{3}$ times the extension per base pair. From the backbone chemistry we have a second criterion, that $3\delta \leq \sim 6.0$ Å: this prevents us from employing a partitioning into more than three base pairs [113, 124].

In physical DNA, the partitioning into triplet repeats (rather than, for instance a situation in which a one-third section of the DNA is extended and the remainder is relaxed) could arise from kinetic factors in that extension is transferred to neighbouring base-steps when a given step passes its cutoff interaction distance. In the Monte Carlo model however the triplet partitioning is found to be an equilibrium structure, so it is necessary to consider the apparent anti-cooperativity of step formation. The short-range anti-cooperativity of step formation sits together with the apparent global cooperativity of the S-DNA transition, shown by the sudden step in the system-wide extension. Based on the current results, we coin a new name for this conformation of DNA which conveys triplet disproportionation, Σ-DNA.

4.5.4 Shifting the overstretching transition

The first intercalator-induced effect that we analyze is the striking change in the force extension curves; the shift of the overstretching force towards higher forces as a function of the intercalator (Fig. 4.11). Intercalated base pairs have a different length and a different free energy penalty than non-intercalated base pairs, thus they influence the force-extension curve. How-

ever, for a shift in the overstretching force to occur another parameter of the model is essential: the cooperativity parameter η. Thus η represents a free energy penalty associated with an intercalated base pair neighboring an overstretched base pair.

If η was equal to 0, states 1 and 2 could independently occur. This would lead to a macroscopic state where states 0, 1 and 2 are all present. At forces larger than the original overstretching force, state 1 has a lower free energy than state 0, so most base-pairs in state 0 are excited in state 1, while the intercalated base-pairs (state 2) are unaffected. Furthermore, the presence of intercalated base-pairs would break the cooperativity of the overstretching transition, because it allows the chain to have base-pairs in state 0 and base-pairs in state 1 in the chain simultaneously, without paying the free energy penalty λ.

However, this picture completely changes when η is larger than 0. This penalizes a 1/2-interface, and states 1 and 2 cannot freely mix. State 1 now pays a free energy penalty for interfaces with both state 0 and state 2.

So, why does the overstretching force increase with μ? This is also an immediate consequence of η. It is caused by the fact that the overstretching transition does, in the presence of intercalators, not start from a chain with all segments in state 0. Instead, some base-pairs are in state 2 at the original overstretching force. Not only the base-pairs in state 0, but also the base-pairs in state 2, need to change into state 1 at the overstretching transition. However, changing a base-pair from state 2 to state 1 costs more free energy than changing a base-pair from state 0 to state 1 at that force. Thus, changing the partly intercalated chain into a completely overstretched chain is more difficult than changing a completely state 0 chain to a completely overstretched chain. So the force needs to do more work before the overstretching transition is realized, and the overstretching transition shifts towards higher forces.

4.6 Conclusions

Homologous pairing through disproportionation of DNA is a process common to very different organisms, often mediated by structurally quite different proteins [249]. The common factor in homologous pairing is structural deformation of the DNA, forming triplet stacked base-pairs, which provides a framework for efficient homology search and recombination.

Here we show that physically motivated changes to existing models of DNA can lead to a model which captures the untwisting and over-stretching behaviour of DNA in a computationally inexpensive coarse-grained model. We show that triplet disproportionation of base-pairs, a common structural change of DNA in homologous recombination, can be driven using the combination of simple intercalator molecules and applied force. By adding a three-valued hidden-variable to our Hamiltonian (in order to track the intercalation and stretching status for given bases) and making trivial adjustments to the interaction potentials, effective only away from their minima, we find that the modelled DNA assumes a new conformation commensurate with the Σ DNA structure. In the absence of intercalators, over-stretching of our model shows only weak periodicity, an observation which is in accord with reported MD simulations [182]. DNA over-stretching in the presence of an intercalator results in the formation of a unique DNA conformation in which stacking is preserved with a pattern of three base-paired B-DNA like bases with conserved stacking.

The observation based on modelling that Σ-DNA structure is favoured only in the presence of intercalator is consistent with the possibility that intercalation stabilizes the base-pairs against melting under extensional force [270][1].

4.7 Disproportionated stretched DNA: a global mechanism?

The Σ DNA structure is not specific to RecA but also observed in other structurally unrelated proteins which are involved in recombination [249]. dsDNA extended structure with big gaps in between triplet stacked base pairs is a prerequisite for homologous recombination as this process involves base pair switching. The fact that SAK*-I force field, when parametrized to model DNA stretching in the presence of the intercalator EtBr, is able to manifest a triplet disproportionated conformation suggests that the strand-extension step of the operation of recombinase enzymes is driven by quite simple physics which the small-molecule intercalators are able to reproduce.

It is interesting that different proteins involved in homologous pairing (HP) which are structurally and evolutionary distinct like ATP-independent HP protein, yeast mitochondrial Mhr1, RecT from a cryptic temperature

[1] It is also interesting that the intercalated DNA is stretched ∼1.5-fold.

bactriophage, bacterial RecO; and ATP-dependent HP proteins like Rad51 and RecA use a common extended DNA structure as an intermediate for HP [249].

Chapter 5

Conclusions

5.1 The Sigma Hypothesis

The crystal structure of RecA-ssDNA and RecA-dsDNA complexes shows a novel and interesting extended structure of DNA: groups of triplet base stacks with a conformation near B-form in a near perpendicular orientation to the helical axis [196]. These structures proved the initial speculations of nucleobase perpendicularity in the recombination complex proposed by Nordén.

It was recently shown using tweezers that a GC rich sequence (60% GC) can undergo a reversible overstretching transition to form a unique conformation which is extended ∼1.5 and remains base-paired [124]. This is the same extension which is seen in DNA complexed with RecA protein [196]. It was suggested that this range of extension is attainable based on a period-three stacking of bases [95] in accordance with the crystal structure of DNA in complex with RecA which is also extended inhomogeneously.

It was further claimed that the stretched structure of a GC rich sequence at a relative extension of 1.5 should probably be a very similar to the one which is found in RecA-DNA complex; a non-homogeneous stretched DNA with triplet base-stacks [113]. In a discussion with an author of that paper (Nordén) we have coined for this structure the name Σ-DNA. It should be mentioned that such a structure is not consistent with experimental data when pulling is done either at the $5'5'$ ends or at the $5'3'$ ends, only at $3'3'$. The authors of the experimental paper have argued that this extended form of DNA is a stable or metastable structure which can be formed in free

solution[1].

5.2 Findings of this Thesis

5.2.1 Sequence Dependence of Triplet Formation

By making a rough calculation of the free energy cost to partition the DNA duplex into a Σ structure, we found a strong and interesting sequence dependence of this energy in that sequences for amino acids supposed to be older in evolutionary history require less free energy to partition into the triplet form. We also find that this energy is still positive, even at imposed extension, which appears to count against the Σ hypothesis in its strong form that the ordered Σ phase can form in free solution without cofactors.

The trend in triplet propensity appears immediately relevant to recombination, and to suggest an enhanced or important role for triplet disproportionation or for Σ formation in the early Earth, however there exist further possible explanations for the trend in triplet propensity. It is possible that the local formation of individual codon triplets, for instance as part of transcription, is or was the determining factor in selecting for a genetic code with the observed pattern of triplet propensities. The fact that a three-base stacking arrangement is also observed in mRNA-ribosome-tRNA complex [271] supports the assumption that triplet base stacks can act as a recognition element and the rather general idea that the origin of the genetic code could have a physical explanation related to triplet formation in some way.

5.2.2 Detailed Observation of Sigma formation, enhanced by Arginine

To make a more detailed analysis of the Σ formation process, atomistic simulations were carried out with a choice of cofactors and sequences. An evolutionarily old sequence with high triplet propensity indeed showed very strong triplet disproportionation, although only in the presence of free Arginine, an amino acid present in the RecA binding cleft and believed by many to have some important role in interacting with nucleic acids in an early nucleic

[1] Over-stretching experiments can be done in the presence of urea to reduce the hydration effect and put the emphasis on stacking interactions.

acid/peptide world.

Although the formation of period three structure was strongly enhanced in a non-additive way by the choice of ancient sequence and Arginine co-factor, the resulting structure did not appear as symmetrical and orderly as would be expected, for instance, from an X-ray crystal structure. A certain amount of thermal disorder is expected from molecular dynamics simulations, reflecting the real situation of biomolecules in solution, however if Σ can be characterised as a solution phase based on the simulation data is must be seen as one which has a high inherent disorder or which is very close in energy to a more disordered or partially melted S phase: statistically classifiying triplets as Σ or non-Σ, a peak population of approximately 25% Σ was observed (around the extension of 1.5) with the remainder of triplets (groups of three bp *i.e.* of four steps) including tilted, melted or mismatched bases, or breaks not aligned to the codon boundary.

5.2.3 Coarse-Grained Modelling of DNA Extension

Atomistic simulation of DNA is expensive, slow and very limited in the lengthscales which can be probed. In the atomistic simulations carried out, a duplex of 24bp (and the associated solution) consumes a significant proportion of the memory of the computing accelerators employed, while being only just larger than the cooperativity length of 22bp required to see a sharp transition [272]. While a four bead/bp model of DNA may seem to be less rigorous than one which includes all atoms, in some ways it can be more so in that it permits more fully equilibrated simulations to be run, and also should permit analysis of the scaling behaviour of extension-related phase transitions to be made.

A simple coarse-grained model for the behaviour of duplex DNA under tension was developed (using a physically motivated Hamiltonian and parameters) and demonstrated to give correct force-extension behaviour. It was observed that this model produced a triplet disproportionation transition to a Σ-like phase, which should permit its use in the future to make more ambitious and thermodynamically rigorous analyses of this transition. The order of chapters in this thesis is different to the chronological sequence of events: the observation of triplet disproportionation in the coarse-grained model in fact provided the inspiration to run the calculations documented in the first two results chapters, and viewing a coarse-grained simulation structure also inspired the term Σ-DNA.

5.3 Sigma-DNA in multiple pairing processes?

It is interesting that different proteins involved in homologous pairing (HP) which are structurally and evolutionary distinct like ATP-independent HP protein, yeast mitochondrial Mhr1, RecT from a cryptic temperature bacteriophage, bacterial RecO; and ATP-dependent HP proteins like Rad51 and RecA use a common extended DNA structure as an intermediate for HP. The meiotic recombination protein Dmc1 also uses triplet formation [273], defining this process as the basis of sexual reproduction. It is also worth mentioning that while RecA and Rad51 untwist the DNA while extending it, Mhr1 promotes HP without untwisting the DNA, although unfortunately at this time there is no detailed information on how this takes place.

HP involves base pair exchanges, which requires a break of base stacking (and pairing), for which the Σ structure may be a widely occuring intermediate state. Our findings of a buried or fossilised physical tendency for DNA to break easily into triplets probably relate to most homologous pairing processes, however those mediated *via* Mhr1 would seem likely to be an exception.

Σ conformation is seen in crystal structures of DNA complexed with Rad51. Linear dichroism studies of DNA-Rad51 complex suggest that tyrosine is intercalated between DNA bases and this facilitates the formation of triplet disproportionated DNA [274]. Although the somewhat tyrosine-like (planar and aromatic) intercalator EtBr was found in the atomistic simulations here not to strongly promote Σ formation, the Rad51 results are an interesting suggestion that Arginine may not be unique in its effects.

5.4 Towards a Minimal Recombination System

The full ATP-hydrolysing mechanism of recombination *via* such proteins as RecA and Rad51 in modern organisms seems to be implausibly sophisticated for proteins constructed only of GADVR, or even of GADVESPLITR, however the process of recombination is (while not fundamental to life) very important in preventing DNA damage and also in accelerating evolution by permitting the exchange of genetic information. Did earlier organisms take advantage of their simpler and more redundant genetic code to operate a simpler recombination mechanism? The results from atomistic simulation show

that given a simplified genetic code, simple disordered Arginine peptides are sufficient to accelerate at least part of the recombination process, reducing the free energy cost to form triplets and effectively acting as catalysts for the first stage of the strand exchange mechanism.

5.5 New Stretching Studies With an Evolutionary Perspective

This work presents for the first time a link between nucleic acid stretching physics and evolutionary history. It is to be hoped that this new perspective will lead to significant further work. In closing, several spurs to curiosity remain:

- Does the enhanced triplet disproportionation of ancient DNA generalise to U-DNA, or RNA?
- Can the newly discovered physical tendencies of GNC-code nucleic acid polymers make the design of any sort of minimal life-systems more straightforward?
- Can the solution Σ phase be confirmed with any atomistic-level experimental information?
- If classical forcefields are corrected for their known deficiencies, will the Σ phase be stabilised, or disappear?

Bibliography

[1] XJ Lu and WK Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, 2003.

[2] RE Dickerson. Definitions and nomenclature of nucleic acid structure components. *Nucleic acids research*, 17(5):1797–1803, 1989.

[3] JTF Wong. A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences*, 72(5):1909, 1975.

[4] TJ Macke and DA Case. Modeling unusual nucleic acid structures. ACS Publications, 1998.

[5] A Rescifina, C Zagni, MG Varrica, V Pistarà, and A Corsaro. Recent advances in small organic molecules as DNA intercalating agents: Synthesis, activity, and modeling. *European journal of medicinal chemistry*, 74:95–115, 2014.

[6] A Balaeff, SL Craig, and DN Beratan. B-DNA to zip-DNA: simulating a DNA transition to a novel structure with enhanced charge-transport characteristics. *The Journal of Physical Chemistry A*, 115(34):9377–9391, 2011.

[7] M Sayar, Barış A, and A Kabakçıoğlu. Twist-writhe partitioning in a coarse-grained DNA minicircle model. *Physical Review E*, 81(4):041916, 2010.

[8] F Kilchherr, C Wachauf, B Pelz, M Rief, M Zacharias, and H Dietz. Single-molecule dissection of stacking forces in DNA. *Science*, 353(6304):aaf5508, 2016.

[9] E Protozanova, P Yakovchuk, and MD Frank-Kamenetskii. Stacked–unstacked equilibrium at the nick site of DNA. *Journal of molecular biology*, 342(3):775–785, 2004.

[10] RA Friedman and B Honig. A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophysical journal*, 69(4):1528–1535, 1995.

[11] JA Lemkul and AD MacKerell Jr. Polarizable force field for DNA based on the classical drude oscillator: I. refinement using quantum mechanical base stacking and conformational energetics. *Journal of Chemical Theory and Computation*, 13(5):2053–2071, 2017.

[12] JD Watson and FHC Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[13] MY Sheinin and MD Wang. Twist–stretch coupling and phase transition during DNA supercoiling. *Physical Chemistry Chemical Physics*, 11(24):4800–4803, 2009.

[14] GS Manning. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quarterly Reviews of Biophysics*, 11(2):179–246, 005 1978.

[15] LD Williams and LJ Maher III. Electrostatic mechanisms of DNA deformation. *Annual review of biophysics and biomolecular structure*, 29(1):497–521, 2000.

[16] BG Feuerstein, N Pattabiraman, and L J Marton. Spermine-DNA interactions: a theoretical study. *Proceedings of the National Academy of Sciences*, 83(16):5948–5952, 1986.

[17] B Jayaram and DL Beveridge. Modeling DNA in aqueous solutions: theoretical and computer simulation studies on the ion atmosphere of DNA. *Annual review of biophysics and biomolecular structure*, 25(1):367–394, 1996.

[18] F DiMaio, X Yu, E Rensen, M Krupovic, D Prangishvili, and EH Egelman. A virus that infects a hyperthermophile encapsidates A-form DNA. *Science*, 348(6237):914–917, 2015.

[19] A Rich and S Zhang. Z-DNA: the long road to biological function. *Nature reviews. Genetics*, 4(7):566, 2003.

[20] AH Wang, GJ Quigley, FJ Kolpak, JL Crawford, JH Van Boom, G van der Marel, and A Rich. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, 282(5740):680, 1979.

[21] A Nordheim, EM Lafer, LJ Peck, JC Wang, BD Stollar, and A Rich. Negatively supercoiled plasmids contain left-handed Z-DNA segments as detected by specific antibody binding. *Cell*, 31(2):309–318, 1982.

[22] G Wang, LA Christensen, and KM Vasquez. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proceedings of the National Academy of Sciences*, 2006.

[23] C Alhambra, FJ Luque, F Gago, and M Orozco. Ab initio study of stacking interactions in A-and B-DNA. *The Journal of Physical Chemistry B*, 101(19):3846–3853, 1997.

[24] MM Warshaw and I Tinoco. Optical properties of sixteen dinucleoside phosphates. *Journal of molecular biology*, 20(1):29–38, 1966.

[25] I Stokkeland and P Stilbs. A multicomponent self-diffusion NMR study of aggregation of nucleotides, nucleosides, nucleic acid bases and some derivatives in aqueous solution with divalent metal ions added. *Biophysical chemistry*, 22(1-2):65–75, 1985.

[26] KM Guckian, BA Schweitzer, RX Ren, CJ Sheils, PL Paris, DC Tahmassebi, and ET Kool. Experimental measurement of aromatic stacking affinities in the context of duplex DNA. *Journal of the American Chemical Society*, 118(34):8182, 1996.

[27] DP Aalberts, JM Parman, and NL Goddard. Single-strand stacking free energy from DNA beacon kinetics. *Biophysical journal*, 84(5):3212–3217, 2003.

[28] JM Huguet, CV Bizarro, N Forns, SB Smith, C Bustamante, and F Ritort. Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proceedings of the National Academy of Sciences*, 107(35):15431–15436, 2010.

[29] C Guerra, T van der Wijst, and FM Bickelhaupt. Substituent effects on hydrogen bonding in Watson–Crick base pairs. A theoretical study. *Structural Chemistry*, 16(3):211–221, 2005.

[30] J Mazur, RL Jernigan, and A Sarai. Conformational effects of DNA stretching. *Molecular Biology*, 37(2):240–249, 2003.

[31] WK Olson, D Swigon, and BD Coleman. Implications of the dependence of the elastic properties of DNA on nucleotide sequence. *Philosophical Transactions of the Royal Society of London. Series*

A: Mathematical, Physical and Engineering Sciences, 362(1820):1403–1422, 2004.

[32] A Balaceanu, M Pasi, PD Dans, A Hospital, R Lavery, and M Orozco. The role of unconventional hydrogen bonds in determining bii propensities in b-dna. *The Journal of Physical Chemistry Letters*, 8(1):21–28, 2016.

[33] C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983.

[34] W Gilbert. Origin of life: The RNA world. *Nature*, 319(6055), 1986.

[35] SA Benner. Paradoxes in the origin of life. *Origins of Life and Evolution of Biospheres*, 44(4):339–343, 2014.

[36] MP Robertson and GF Joyce. The origins of the RNA world. *Cold Spring Harbor perspectives in biology*, 4(5):a003608, 2012.

[37] Poul Nissen, Jeffrey Hansen, Nenad Ban, Peter B Moore, and Thomas A Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289(5481):920–930, 2000.

[38] HB White. Coenzymes as fossils of an earlier metabolic state. *Journal of Molecular Evolution*, 7(2):101–104, 1976.

[39] TA Lincoln and GF Joyce. Self-sustained replication of an RNA enzyme. *Science*, 323(5918):1229–1232, 2009.

[40] HS Bernhardt. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). *Biology direct*, 7(1):23, 2012.

[41] N Maizels and AM Weiner. Peptide-specific ribosomes, genomic tags, and the origin of the genetic code. In *Cold Spring Harbor symposia on quantitative biology*, volume 52, pages 743–749. Cold Spring Harbor Laboratory Press, 1987.

[42] N Ban, P Nissen, J Hansen, P B Moore, and TA Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481):905–920, 2000.

[43] MM Yusupov, Gulnara Zh , A Baucom, K Lieberman, TN Earnest, JHD Cate, and HF Noller. Crystal structure of the ribosome at 5.5 Å resolution. *science*, 292(5518):883–896, 2001.

[44] TA Steitz and PB Moore. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends in biochemical sciences*, 28(8):411–418, 2003.

[45] MP Robertson and GF Joyce. The origins of the RNA world. *Cold Spring Harbor perspectives in biology*, 4(5):a003608, 2012.

[46] I Takahashi and J Marmur. Replacement of thymidylic acid by deoxyuridylic acid in the deoxyribonucleic acid of a transducing phage for bacillus subtilis. *Nature*, 197(4869):794–795, 1963.

[47] P Forterre, J Filée, and H Myllykallio. Origin and evolution of DNA and DNA replication machineries. In *The genetic code and the origin of Life*, pages 145–168. Springer, 2004.

[48] M Shimizu. Molecular basis for the genetic code. *Journal of molecular evolution*, 18(5):297–303, 1982.

[49] M Yarus. An RNA-amino acid complex and the origin of the genetic code. *The New Biologist*, 3(2):183–189, 1991.

[50] C Saxinger, C Ponnamperuma, and C Woese. Evidence for the interaction of nucleotides with immobilized amino-acids and its significance for the origin of the genetic code. *Nature*, 234(49):172–174, 1971.

[51] M Yarus, JG Caporaso, and R Knight. Origins of the genetic code: the escaped triplet theory. *Annu. Rev. Biochem.*, 74:179–198, 2005.

[52] A Ellington, M Kharpov, and CA Shaw. The scene of a frozen accident. *Rna*, 6(4):485–498, 2000.

[53] TM Sonneborn. Degeneracy of the genetic code: extent, nature, and genetic implications. *Evolving genes and proteins. Academic Press, New York*, pages 377–397, 1965.

[54] WM Fitch and K Upper. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. In *Cold Spring Harbor symposia on quantitative biology*, volume 52, pages 759–767. Cold Spring Harbor Laboratory Press, 1987.

[55] J Wong, Siu-Kin Ng, Wai-Kin M, T Hu, and H Xue. Coevolution theory of the genetic code at age forty: pathway to translation and synthetic life. *Life*, 6(1):12, 2016.

[56] M Di Giulio. Origin of glutaminyl-tRNA synthetase: an example of palimpsest? *Journal of molecular evolution*, 37(1):5–10, 1993.

[57] M Di Giulio. The coevolution theory of the origin of the genetic code. *Journal of molecular evolution*, 48(3):253–254, 1999.

[58] CT Zhu, XB Zeng, and WD Huang. Codon usage decreases the error minimization within the genetic code. *Journal of Molecular Evolution*, 57(5):533–537, 2003.

[59] M Archetti. Codon usage bias and mutation constraints reduce the level of errorminimization of the genetic code. *Journal of Molecular Evolution*, 59(2):258–266, 2004.

[60] L Klipcan and M Safro. Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. *Journal of theoretical biology*, 228(3):389–396, 2004.

[61] M Archetti. Codon usage bias and mutation constraints reduce the level of errorminimization of the genetic code. *Journal of Molecular Evolution*, 59(2):258–266, 2004.

[62] M Di Giulio. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *Journal of Molecular evolution*, 29(4):288–293, 1989.

[63] M Di Giulio and M Medugno. The level and landscape of optimization in the origin of the genetic code. *Journal of molecular evolution*, 52(4):372–382, 2001.

[64] K Ikehara, Y Omori, R Arai, and A Hirose. A novel theory on the origin of the genetic code: a GNC-SNS hypothesis. *Journal of molecular evolution*, 54(4):530–538, 2002.

[65] B Bechinger. Membrane insertion and orientation of polyalanine peptides: a 15 N solid-state NMR spectroscopy investigation. *Biophysical journal*, 81(4):2251–2256, 2001.

[66] FJR Taylor and D Coates. The code within the codons. *Biosystems*, 22(3):177–187, 1989.

[67] J Oró, SL Miller, C Ponnamperuma, and RS Young. *Cosmochemical Evolution and the Origins of Life: Proceedings of the Fourth International Conference on the Origin of Life and the First Meeting of the International Society for the Study of the Origin of Life, Barcelona, June 25–28, 1973, Volume I: Invited Papers and Volume II: Contributed Papers*. Springer Science & Business Media, 2013.

[68] S Padmanabhan, EJ York, JM Stewart, and RL Baldwin. Helix propensities of basic amino acids increase with the length of the side-chain. *Journal of molecular biology*, 257(3):726–734, 1996.

[69] DAM Zaia, C Thaïs BV Zaia, and H De Santana. Which amino acids should be used in prebiotic chemistry studies? *Origins of Life and Evolution of Biospheres*, 38(6):469–488, 2008.

[70] BH Patel, C Percivalle, DJ Ritson, CD Duffy, and JD Sutherland. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nature Chemistry*, 7(4):301–307, 2015.

[71] R Tan and AD Frankel. Structural variety of arginine-rich RNA-binding peptides. *Proceedings of the National Academy of Sciences*, 92(12):5282–5286, 1995.

[72] DP Mascotti and TM Lohman. Thermodynamics of oligoarginines binding to RNA and DNA. *Biochemistry*, 36(23):7272–7279, 1997.

[73] R Balhorn, L Brewer, and M Corzett. DNA condensation by protamine and arginine-rich peptides: Analysis of toroid stability using single DNA molecules. *Molecular reproduction and development*, 56(S2):230–234, 2000.

[74] C Bustamante, Z Bryant, and SB Smith. Ten years of tension: single-molecule DNA mechanics. *Nature*, 421(6921):423–427, January 2003.

[75] M Rief, H Clausen-Schaumann, and HE. Gaub. Sequence-dependent mechanics of single DNA molecules. *Nat Struct Mol Biol*, 6(4):346–349, April 1999.

[76] JF Leger, J Robert, L Bourdieu, D Chatenay, and JF Marko. RecA binding to a single double-stranded DNA molecule: a possible role of DNA conformational fluctuations. *Proceedings of the National Academy of Sciences*, 95(21):12295–12299, 1998.

[77] M Doi and SF Edwards. *The theory of polymer dynamics*, volume 73. oxford university press, 1988.

[78] D Stigter. Evaluation of the counterion condensation theory of polyelectrolytes. *Biophysical journal*, 69(2):380–388, 1995.

[79] RW Wilson and VA Bloomfield. Counterion-induced condensation of deoxyribonucleic acid. A light-scattering study. *Biochemistry*, 18(11):2192–2196, 1979.

[80] Micaela Caserta, Eleonora Agricola, Mark Churcher, Edwige Hiriart, Loredana Verdone, Ernesto Di Mauro, and Andrew Travers. A translational signature for nucleosome positioning *in vivo*. *Nucleic Acids Research*, 37(16):5309–5321, 2009.

[81] J Virstedt, T Berge, RM. Henderson, MJ. Waring, and AA. Travers. The influence of DNA stiffness upon nucleosome formation. *Journal of Structural Biology*, 148(1):66 – 85, 2004.

[82] GS Manning. Three persistence lengths for a stiff polymer with an application to DNA B-Z junctions. *Biopolymers*, 27(10):1529–1542, 1988.

[83] CG Baumann, S B Smith, VA Bloomfield, and C Bustamante. Ionic effects on the elasticity of single DNA molecules. *Proceedings of the National Academy of Sciences*, 94(12):6185–6190, 1997.

[84] JR Wenner, MC Williams, I Rouzina, and VA Bloomfield. Salt dependence of the elasticity and overstretching transition of single DNA molecules. *Biophysical journal*, 82(6):3160–3169, 2002.

[85] AK Mazur and M Maaloum. Atomic force microscopy study of DNA flexibility on short length scales: smooth bending versus kinking. *Nucleic Acids Research*, 2014.

[86] Jiro Shimada and Hiromi Yamakawa. Ring-closure probabilities for twisted wormlike chains. Application to DNA. *Macromolecules*, 17(4):689–698, 1984.

[87] NA. Becker, JD. Kahn, and LJ Maher III. Bacterial repression loops require enhanced DNA flexibility. *Journal of Molecular Biology*, 349(4):716 – 730, 2005.

[88] SM Law, GR Bellomy, PJ Schlax, and MT Record Jr. *In vivo* thermodynamic analysis of repression with and without looping in lac constructs: Estimates of free and local lac repressor concentrations and of physical properties of a region of supercoiled plasmid DNA *in vivo*. *Journal of Molecular Biology*, 230(1):161 – 173, 1993.

[89] Y Zhang, AE McEwen, DM Crothers, and SD Levene. Analysis of *in vivo* LacR-mediated gene repression based on the mechanics of DNA looping. *PLoS One*, 1(1):e136, 2006.

[90] L Ringrose, S Chabanis, PO Angrand, C Woodroofe, and AF Stewart. Quantitative comparison of DNA looping *in vitro* and *in vivo*: chromatin increases effective DNA flexibility at short distances. *The EMBO Journal*, 18(23):6630–6641, 1999.

[91] AA Travers, SS Ner, and MEA Churchill. DNA chaperones: a solution to a persistence problem? *Cell*, 77(2):167–169, 1994.

[92] NT Sebastian, EM Bystry, NA Becker, and LJ Maher III. Enhancement of DNA flexibility *in vitro* and *in vivo* by HMGB BoxA proteins carrying Box B residues. *Biochemistry*, 48(10):2125–2134, 2009.

[93] JR Moffitt, YR Chemla, SB Smith, and C Bustamante. Recent advances in optical tweezers. *Annu. Rev. Biochem.*, 77:205–228, 2008.

[94] H Chen, H Fu, X Zhu, P Cong, F Nakamura, and J Yan. Improved high-force magnetic tweezers for stretching and refolding of proteins and short DNA. *Biophysical journal*, 100(2):517–523, 2011.

[95] C Prévost and M Takahashi. Geometry of the DNA strands within the RecA nucleofilament: role in homologous recombination. *Quarterly reviews of biophysics*, 36(04):429–453, 2003.

[96] A Sarai, RL Jernigan, and J Mazur. Interdependence of conformational variables in double-helical DNA. *Biophysical journal*, 71(3):1507–1518, 1996.

[97] A Lebrun, Z Shakked, and R Lavery. Local DNA stretching mimics the distortion caused by the TATA box-binding protein. *Proceedings of the National Academy of Sciences*, 94(7):2993–2998, 1997.

[98] I Rouzina and VA Bloomfield. Force-induced melting of the DNA double helix 1. Thermodynamic analysis. *Biophysical journal*, 80(2):882–893, 2001.

[99] A Hanke. Denaturation transition of stretched DNA. *Biochemical Society Transactions*, 41(2):639–645, 2013.

[100] A Lebrun and R Lavery. Modelling extreme stretching of DNA. *Nucleic Acids Research*, 24(12):2260–2267, 1996.

[101] P Cluzel, A Lebrun, C Heller, and R Lavery. DNA: an extensible molecule. *Science*, 271(5250):792, 1996.

[102] H Clausen-Schaumann, M Rief, C Tolksdorf, and HE Gaub. Mechanical stability of single DNA molecules. *Biophysical Journal*, 78(4):1997 – 2007, 2000.

[103] S Cocco, J Yan, JF Léger, D Chatenay, and JF Marko. Overstretching and force-driven strand separation of double-helix DNA. *Physical Review E*, 70(1):011910, 2004.

[104] S Whitelam, S Pronk, and PL Geissler. There and (slowly) back again: entropy-driven hysteresis in a model of DNA overstretching. *Biophysical journal*, 94(7):2452–2469, 2008.

[105] MC Williams, L Rouzina, and MJ McCauley. Peeling back the mystery of DNA overstretching. *Proceedings of the National Academy of Sciences*, 106(43):18047–18048, 2009.

[106] X Zhang, H Chen, S Le, I Rouzina, PS Doyle, and J Yan. Revealing the competition between peeled ssDNA, melting bubbles, and S-DNA during DNA overstretching by single-molecule calorimetry. *Proceedings of the National Academy of Sciences*, 110(10):3865–3870, 2013.

[107] Chandrashekhar M, V Subramaniam, C Otto, and ML Bennink. Force spectroscopy and fluorescence microscopy of dsDNA–YOYO-1 complexes: implications for the structure of dsDNA in the overstretching region. *Nucleic acids research*, 38(10):3423–3431, 2010.

[108] J van Mameren, P Gross, G Farge, P Hooijman, M Modesti, M Falkenberg, GJL Wuite, and EJG Peterman. Unraveling the structure of DNA during overstretching by using multicolor, single-molecule fluorescence imaging. *Proceedings of the National Academy of Sciences*, 106(43):18231–18236, 2009.

[109] O Krichevsky. DNA overstretched state: S-DNA form or force-induced melting?: Comment on "biophysical characterization of DNA binding from single molecule force measurements" by Mark C. Williams *et al.* *Physics of life reviews*, 7(3):350–352, 2010.

[110] L Shokri, MJ McCauley, I Rouzina, and MC Williams. DNA overstretching in the presence of glyoxal: Structural evidence of force-induced DNA melting. *Biophysical Journal*, 95(3):1248 – 1255, 2008.

[111] K Pant, RL Karpel, L Rouzina, and MC Williams. Salt dependent binding of T4 gene 32 protein to single and double-stranded DNA: single molecule force spectroscopy measurements. *Journal of molecular biology*, 349(2):317–330, 2005.

[112] SB Smith, Y Cui, and C Bustamente. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science*, 271(5250):795, 1996.

[113] N Bosaeus, A Reymer, T Beke-Somfai, T Brown, M Takahashi, P Wittung-Stafshede, S Rocha, and B Nordén. A stretched conformation of DNA with a biological role? *Quarterly Reviews of Biophysics*, 50, 2017.

[114] Michael K and J Bolonick. Molecular dynamics simulation of DNA stretching is consistent with the tension observed for extension and strand separation and predicts a novel ladder structure. *Journal of the American Chemical Society*, 118(45):10989–10994, 1996.

[115] JF Léger, G Romano, A Sarkar, J Robert, L Bourdieu, D Chatenay, and JF Marko. Structural transitions of a twisted and stretched DNA molecule. *Phys. Rev. Lett.*, 83:1066–1069, Aug 1999.

[116] H Chen, H Fu, and CG Koh. Sequence-dependent unpeeling dynamics of stretched DNA double helix. *Journal of Computational and Theoretical Nanoscience*, 5(7):1381–1386, 2008.

[117] H Fu, H Chen, JF Marko, and J Yan. Two distinct overstretched DNA states. *Nucleic acids research*, pages 5594–5600, 2010.

[118] DR Roe and AM Chaka. Structural basis of pathway-dependent force profiles in stretched DNA. *The Journal of Physical Chemistry B*, 113(46):15364–15371, 2009.

[119] M Maaloum, AF Beker, and P Muller. Secondary structure of double-stranded DNA under stretching: Elucidation of the stretched form. *Phys. Rev. E*, 83:031903, Mar 2011.

[120] MHF Wilkins, RG Gosling, and WE Seeds. Physical studies of nucleic acid: Nucleic acid: An extensible molecule? *Nature*, 167(4254):759–760, 1951.

[121] JF Marko and ED Siggia. Stretching DNA. *Macromolecules*, 28(26):8759–8770, 1995.

[122] H Fu, H Chen, JF Marko, and J Yan. Two distinct overstretched DNA states. *Nucleic Acids Research*, 38(16):5594–5600, 2010.

[123] DH Paik and TT Perkins. Overstretching DNA at 65 pN does not require peeling from free ends or nicks. *Journal of the American Chemical Society*, 133(10):3219–3221, 2011.

[124] N Bosaeus, AH El-Sagheer, T Brown, SB Smith, B Åkerman, C Bustamante, and B Nordén. Tension induces a base-paired overstretched DNA conformation. *Proceedings of the National Academy of Sciences*, 109(38):15179–15184, 2012.

[125] R Lohikoski, J Timonen, and A Laaksonen. Molecular dynamics simulation of single DNA stretching reveals a novel structure. *Chemical physics letters*, 407(1):23–29, 2005.

[126] K Hatch, C Danilowicz, V Coljee, and M Prentiss. Demonstration that the shear force required to separate short double-stranded DNA does not increase significantly with sequence length for sequences longer than 25 base pairs. *Physical Review E*, 78(1):011920, 2008.

[127] C Danilowicz, C Limouse, K Hatch, A Conover, VW Coljee, N Kleckner, and M Prentiss. The structure of DNA overstretched from the 5' 5' ends differs from the structure of DNA overstretched from the 3' 3' ends. *Proceedings of the National Academy of Sciences*, 106(32):13196–13201, 2009.

[128] KR Chaurasiya, T Paramanathan, MJ McCauley, and MC Williams. Biophysical characterization of DNA binding from single molecule force measurements. *Physics of Life Reviews*, 7(3):299 – 341, 2010.

[129] A Sarkar, JF Léger, D Chatenay, and JF Marko. Structural transitions in DNA driven by external force and torque. *Phys. Rev. E*, 63:051903, Apr 2001.

[130] AH Hardin, SK Sarkar, Y Seol, GF Liou, N Osheroff, and KC Neuman. Direct measurement of DNA bending by type IIA topoisomerases: implications for non-equilibrium topology simplification. *Nucleic Acids Research*, 39(13):5729–5743, 2011.

[131] R Palchaudhuri and PJ Hergenrother. DNA as a target for anticancer compounds: methods to determine the mode of binding and the mechanism of action. *Current opinion in biotechnology*, 18(6):497–503, 2007.

[132] H Chen, X Liu, and DJ Patel. DNA bending and unwinding associated with actinomycin D antibiotics bound to partially overlapping sites on DNA. *Journal of molecular biology*, 258(3):457–479, 1996.

[133] BL Staker, MD Feese, M Cushman, Y Pommier, D Zembower, L Stewart, and AB Burgin. Structures of three classes of anticancer agents bound to the human topoisomerase I-DNA covalent complex. *Journal of medicinal chemistry*, 48(7):2336–2345, 2005.

[134] MJ Campbell and RW Davis. Toxic mutations in the RecA gene of E. coli prevent proper chromosome segregation. *Journal of molecular biology*, 286(2):417–435, 1999.

[135] A Stasiak. Three-stranded DNA structure; is this the secret of DNA homologous recognition? *Molecular microbiology*, 6(22):3267–3276, 1992.

[136] A Stasiak and E Di Capua. The helicity of DNA in complexes with RecA protein. *Nature*, 299(5879):185–186, September 1982.

[137] B Nordén, C Elvingson, M Kubista, B Sjöberg, H Ryberg, M Ryberg, K Mortensen, and M Takahashi. Structure of RecA-DNA complexes studied by combination of linear dichroism and small-angle neutron scattering measurements on flow-oriented samples. *Journal of molecular biology*, 226(4):1175–1191, 1992.

[138] T Nishinaka, Y Ito, S Yokoyama, and T Shibata. An extended DNA structure through deoxyribose-base stacking induced by RecA protein. *Proceedings of the National Academy of Sciences*, 94(13):6623–6628, 1997.

[139] GV Shivashankar, M Feingold, O Krichevsky, and A Libchaber. RecA polymerization on double-stranded DNA by using single-molecule manipulation: the role of ATP hydrolysis. *Proceedings of the National Academy of Sciences*, 96(14):7916–7921, 1999.

[140] R Lavery, A Lebrun, JF Allemand, D Bensimon, and V Croquette. Structure and mechanics of single biomolecules: experiment and simulation. *Journal of Physics: Condensed Matter*, 14(14):R383, 2002.

[141] M Hegner, SB Smith, and C Bustamante. Polymerization and mechanical properties of single RecA–DNA filaments. *Proceedings of the National Academy of Sciences*, 96(18):10109–10114, 1999.

[142] G Bertucat, R Lavery, and C Prèvost. A model for parallel triple helix formation by Ree A: Single-strand association with a homologous Duple via the minor groove. *Journal of Biomolecular Structure and Dynamics*, 16(3):535–546, 1998.

[143] J Xu, L Zhao, Y Xu, W Zhao, P Sung, and HW Wang. Cryo-EM structures of human RAD51 recombinase filaments during catalysis of DNA-strand exchange. *Nature Structural & Molecular Biology*, 24(1):40–46, 2017.

[144] A André, F Fontaine-Vive, HM Möller, T Fischer, G Maret, VT Forsyth, and T Gisler. Force-induced structural transitions in cross-linked DNA films. *European Biophysics Journal*, 37(6):749–757, 2008.

[145] RB Nicklas. The forces that move chromosomes in mitosis. *Annual review of biophysics and biophysical chemistry*, 17(1):431–449, 1988.

[146] KW Hsiao, C Sasmal, J Ravi P, and CM Schroeder. Direct observation of DNA dynamics in semidilute solutions in extensional flow. *Journal of Rheology*, 61(1):151–167, 2017.

[147] G Juarez and PE Arratia. Extensional rheology of DNA suspensions in microfluidic devices. *Soft Matter*, 7(19):9444–9452, 2011.

[148] PG De Gennes. Coil-stretch transition of dilute flexible polymers under ultrahigh velocity gradients. *The Journal of Chemical Physics*, 60(12):5030–5042, 1974.

[149] VB Zhurkin, G Raghunathan, NB Ulyanov, RD Camerini-Otero, and RL Jernigan. A parallel DNA triplex as model for the intermediate in homologous recombination. *Journal of molecular biology*, 239(2):181–200, 1994.

[150] AA Chen and AE García. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 110(42):16820–16825, 2013.

[151] R Luo, HSR Gilson, MJ Potter, and MK Gilson. The physical basis of nucleic acid base stacking in water. *Biophysical journal*, 80(1):140–148, 2001.

[152] J Šponer, J Leszczynski, and P Hobza. Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers*, 61(1):3–31, 2001.

[153] SL McKay, B Haptonstall, and SH Gellman. Beyond the hydrophobic effect: Attractions involving heteroaromatic rings in aqueous solution. *Journal of the American Chemical Society*, 123(6):1244–1245, 2001.

[154] KM Guckian, BA Schweitzer, RX Ren, CJ Sheils, DC Tahmassebi, and ET Kool. Factors contributing to aromatic stacking in water: evaluation in the context of DNA. *Journal of the American Chemical Society*, 122(10):2213–2222, 2000.

[155] TE Cheatham III, P Cieplak, and PA Kollman. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics*, 16(4):845–862, 1999.

[156] N Foloppe and AD MacKerell Jr. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of computational chemistry*, 21(2):86–104, 2000.

[157] TA Soares, PH Hünenberger, MA Kastenholz, V Kräutler, T Lenz, RD Lins, C Oostenbrink, and WF van Gunsteren. An improved nucleic acid parameter set for the GROMOS force field. *Journal of computational chemistry*, 26(7):725–737, 2005.

[158] GA Kaminski, RA Friesner, J Tirado-Rives, and WL Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487, 2001.

[159] TM Parker, EG Hohenstein, RM Parrish, NV Hud, and CD Sherrill. Quantum-mechanical analysis of the energetic contributions to π stacking in nucleic acids versus rise, twist, and slide. *Journal of the American Chemical Society*, 135(4):1306–1316, 2013.

[160] TM Parker and CD Sherrill. Assessment of empirical models versus high-accuracy ab initio methods for nucleobase stacking: Evaluating the importance of charge penetration. *Journal of chemical theory and computation*, 11(9):4197–4204, 2015.

[161] H Rosemeyer and F Seela. Modified purine nucleosides as dangling ends of DNA duplexes: the effect of the nucleobase polarizability on stacking interactions. *Journal of the Chemical Society, Perkin Transactions 2*, (4):746–750, 2002.

[162] JT Wong. Emergence of life: from functional RNA selection to natural selection and beyond. *Front Biosci*, 19:1117–1150, 2014.

[163] S Ratner and B Petrack. The mechanism of arginine synthesis from citrulline in kidney. *Journal of Biological Chemistry*, 200(1):175–185, 1953.

[164] LC Archard and JD Williamson. The effect of arginine deprivation on the replication of vaccinia virus. *Journal of General Virology*, 12(3):249–258, 1971.

[165] MD Sanchez, AC Ochoa, and TP Foster. Development and evaluation of a host-targeted antiviral that abrogates herpes simplex virus replication through modulation of arginine-associated metabolic pathways. *Antiviral research*, 132:13–25, 2016.

[166] MHV van Regenmortel and BWJ Mahy. *Desk encyclopedia of general virology*. Academic Press, 2010.

[167] RW Tankersley. Amino acid requirements of herpes simplex virus in human cells. *Journal of bacteriology*, 87(3):609–613, 1964.

[168] TH Jukes. Arginine as an evolutionary intruder into protein synthesis. *Biochemical and biophysical research communications*, 53(3):709–714, 1973.

[169] S Itzkovitz and U Alon. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome research*, 17(4):405–412, 2007.

[170] JML Pieters, RMW Mans, H van den Elst, GA van den Marel, JH van Boom, and C Altona. Conformational and thermodynamic consequences of the introduction of a nick in duplexed DNA fragments: an NMR study augmented by biochemical experiments. *Nucleic acids research*, 17(12):4551–4565, 1989.

[171] EA Snowden-Ifft and DE Wemmer. Characterization of the structure and melting of DNAs containing backbone nicks and gaps. *Biochemistry*, 29(25):6017–6025, 1990.

[172] C Roll, C Ketterlé, V Faibis, GV Fazakerley, and Y Boulard. Conformations of nicked and gapped DNA structures by NMR and molecular dynamic simulations in water. *Biochemistry*, 37(12):4059–4070, 1998.

[173] K Murata, Y Sugita, and Y Okamoto. Free energy calculations for DNA base stacking by replica-exchange umbrella sampling. *Chemical physics letters*, 385(1):1–7, 2004.

[174] P Banás, D Hollas, M Zgarbová, P Jurecka, M Orozco, TE Cheatham III, J Sponer, and M Otyepka. Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins. *Journal of Chemical Theory and Computation*, 6(12):3836–3849, 2010.

[175] S Grimme. Do special noncovalent π–π stacking interactions really exist? *Angewandte Chemie International Edition*, 47(18):3430–3434, 2008.

[176] J Norberg and L Nilsson. Temperature-dependence of the stacking propensity of adenylyl-3', 5'-adenosine. *Journal of Physical Chemistry*, 99(35):13056–13058, 1995.

[177] WD Cornell, P Cieplak, CI Bayly, IR Gould, KM Merz, DM Ferguson, DC Spellmeyer, T Fox, JW Caldwell, and PA Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[178] M Zgarbová, M Otyepka, J Šponer, P Hobza, and P Jurečka. Large-scale compensation of errors in pairwise-additive empirical force fields: comparison of AMBER intermolecular terms with rigorous DFT-SAPT calculations. *Physical Chemistry Chemical Physics*, 12(35):10476–10493, 2010.

[179] C Maffeo, B Luan, and A Aksimentiev. End-to-end attraction of duplex DNA. *Nucleic acids research*, 40(9):3812–3821, 2012.

[180] RF Brown, CT Andrews, and AH Elcock. Stacking free energies of all DNA and RNA nucleoside pairs and dinucleoside-monophosphates computed using recently revised AMBER parameters and compared with experiment. *Journal of chemical theory and computation*, 11(5):2315–2328, 2015.

[181] JD Tubbs, DE Condon, SD Kennedy, M Hauser, PC Bevilacqua, and DH Turner. The nuclear magnetic resonance of CCCC RNA reveals a right-handed helix, and revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics. *Biochemistry*, 52(6):996–1010, 2013.

[182] A Taghavi, P van der Schoot, and JT Berryman. DNA partitions into triplets under tension in the presence of organic cations, with sequence evolutionary age predicting the stability of the triplet phase. *Quarterly reviews of biophysics*, 50, 2017.

[183] MC Williams, L Rouzina, and VA Bloomfield. Thermodynamics of DNA interactions from single molecule stretching experiments. *Accounts of chemical research*, 35(3):159–166, 2002.

[184] J Vlassakis, J Williams, K Hatch, C Danilowicz, VW Coljee, and M Prentiss. Probing the mechanical stability of DNA in the presence of monovalent cations. *Journal of the American Chemical Society*, 130(15):5004–5005, 2008.

[185] N Liu, T Bu, Y Song, W Zhang, J Li, W Zhang, J Shen, and H Li. The Nature of the force-induced conformation transition of dsDNA studied by using single molecule force spectroscopy. *Langmuir*, 26(12):9491–9496, 2010.

[186] MC Williams, JR Wenner, L Rouzina, and VA Bloomfield. Effect of pH on the overstretching transition of double-stranded DNA: evidence of force-induced DNA melting. *Biophysical Journal*, 80(2):874–881, 2001.

[187] A Lebrun and R Lavery. Modelling extreme stretching of DNA. *Nucleic acids research*, 24(12):2260–2267, 1996.

[188] S Bag, S Mogurampelly, WA Goddard III, and PK Maiti. Dramatic changes in DNA conductance with stretching: structural polymorphism at a critical extension. *Nanoscale*, 8(35):16044–16052, 2016.

[189] H Li and T Gisler. Overstretching of a 30 bp DNA duplex studied with steered molecular dynamics simulation: Effects of structural defects on structure and force-extension relation. *The European Physical Journal E*, 30(3):325–332, 2009.

[190] R Lavery, A Lebrun, JF Allemand, D Bensimon, and V Croquette. Structure and mechanics of single biomolecules: experiment and simulation. *Journal of Physics: Condensed Matter*, 14(14):R383, 2002.

[191] SB Smith, L Finzi, and C Bustamante. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science*, 258(5085):1122–1126, 1992.

[192] B Nordén, C Elvingson, M Kubista, B Sjöberg, H Ryberg, M Ryberg, K Mortensen, and M Takahashi. Structure of RecA-DNA complexes studied by combination of linear dichroism and small-angle neutron scattering measurements on flow-oriented samples. *Journal of molecular biology*, 226(4):1175–1191, 1992.

[193] GA King, P Gross, U Bockelmann, M Modesti, GL Wuite, and EJG Peterman. Revealing the competition between peeled ssDNA, melting bubbles, and S-DNA during DNA overstretching using fluorescence microscopy. *Proceedings of the National Academy of Sciences*, 110(10):3859–3864, 2013.

[194] SA Harris, ZA Sands, and CA Laughton. Molecular dynamics simulations of duplex stretching reveal the importance of entropy in determining the biomechanical properties of DNA. *Biophysical journal*, 88(3):1684–1691, 2005.

[195] J Řezáč, P Hobza, and SA Harris. Stretched DNA investigated using molecular-dynamics and quantum-mechanical calculations. *Biophysical journal*, 98(1):101–110, 2010.

[196] Z Chen, H Yang, and NP Pavletich. Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature*, 453(7194):489–494, May 2008.

[197] MM Cox. The bacterial RecA protein: structure, function, and regulation. In *Molecular Genetics of Recombination*, pages 53–94. Springer, 2007.

[198] X Yu, SA Jacobs, SC West, T Ogawa, and EH Egelman. Domain structure and dynamics in the helical filaments formed by RecA and Rad51 on DNA. *Proceedings of the National Academy of Sciences*, 98(15):8419–8424, 2001.

[199] J Wong. Coevolution theory of the genetic code at age thirty. *BioEssays*, 27(4):416–425, 2005.

[200] EV Koonin and AS Novozhilov. Origin and evolution of the genetic code: the universal enigma. *IUBMB life*, 61(2):99–111, 2009.

[201] JTF Wong and PM Bronskill. Inadequacy of prebiotic synthesis as origin of proteinous amino acids. *Journal of molecular evolution*, 13(2):115–125, 1979.

[202] S Park and K Schulten. Calculating potentials of mean force from steered molecular dynamics simulations. *The Journal of chemical physics*, 120(13):5946–5961, 2004.

[203] B Isralewitz, M Gao, and K Schulten. Steered molecular dynamics and mechanical functions of proteins. *Current opinion in structural biology*, 11(2):224–230, 2001.

[204] C Bustamante, JF Marko, ED Siggia, and S Smith. Entropic elasticity of lambda-phage DNA. *Science*, 265(5178):1599–1600, 1994.

[205] C Bustamante, Z Bryant, and SB. Smith. Ten years of tension: single-molecule DNA mechanics. *Nature*, 421(6921):423–427, January 2003.

[206] KC Neuman and A Nagy. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature methods*, 5(6):491, 2008.

[207] IS Joung and TE Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The journal of physical chemistry B*, 112(30):9020–9041, 2008.

[208] WL Jorgensen, J Chandrasekhar, JD Madura, RW Impey, and ML Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[209] J Wang, RM Wolf, JW Caldwell, PA Kollman, and DA Case. Development and testing of a general AMBER force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[210] J Wang, W Wang, PA Kollman, and DA Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, 25(2):247–260, 2006.

[211] R Salomon-Ferrer, AW Gotz, D Poole, Scott Le G, and RC Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *Journal of chemical theory and computation*, 9(9):3878–3888, 2013.

[212] DA Case, RM Betz, DS Cerutti, TE Cheatham III, TE Darden, RE Duke, TJ Giese, H Gohlke, AW Goetz, N Homeyer, S Izadi, P Janowski, J Kaus, A Kovalenko, TS Lee, S LeGrand, P Li, C Lin, T Luchko, R Luo, B Madej, D Mermelstein, KM Merz, G Monard, G Nguyen, HT Nguyen, I Omelyan, A Onufriev, DR Roe, A Roitberg, C Sagui, CL Simmerling, WM Botello-Smith, J Swails, RC Walker, J Wang, RM Wolf, X Wu, Xiao L, and P.A. Kollman. AMBER 2016. *University of California, San Francisco*, 2016.

[213] V Hornak, R Abel, A Okur, B Strockbine, A Roitberg, and C Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[214] A Pérez, I Marchán, D Svozil, J Sponer, TE Cheatham, CA Laughton, and M Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*, 92(11):3817–3829, 2007.

[215] JP Ryckaert, G Ciccotti, and HJC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[216] T Darden, D York, and L Pedersen. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[217] AS Biebricher, I Heller, RFH Roijmans, TP Hoekstra, EJG Peterman, and GJL Wuite. The impact of DNA intercalators on DNA and DNA-processing enzymes elucidated through force-dependent binding kinetics. *Nature communications*, 6, 2015.

[218] SP Edmondson and WC Johnson. Base tilt of B-form poly [d (G)]-poly [d (C)] and the B-and Z-conformations of poly [d (GC)]-poly [d (GC)] in solution. *Biopolymers*, 25(12):2335–2348, 1986.

[219] DR Roe and TE Cheatham III. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, 9(7):3084–3095, 2013.

[220] C Sathe, A Girdhar, JP Leburton, and K Schulten. Electronic detection of dsDNA transition from helical to zipper conformation using graphene nanopores. *Nanotechnology*, 25(44):445105, 2014.

[221] H Fu, H Chen, JF Marko, and J Yan. Two distinct overstretched DNA states. *Nucleic acids research*, pages 5594–5600, 2010.

[222] B Tidor, KK Irikura, BR Brooks, and M Karplus. Dynamics of DNA oligomers. *Journal of Biomolecular Structure and Dynamics*, 1(1):231–252, 1983.

[223] A Pérez, FJ Luque, and M Orozco. Dynamics of B-DNA on the microsecond time scale. *Journal of the American Chemical Society*, 129(47):14739–14745, 2007.

[224] R Lavery, K Zakrzewska, D Beveridge, TC Bishop, DA Case, T Cheatham, S Dixit, B Jayaram, F Lankas, and C Laughton. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic acids research*, 38(1):299–313, 2010.

[225] CK Materese, A Savelyev, and GA Papoian. Counterion atmosphere and hydration patterns near a nucleosome core particle. *Journal of the American Chemical Society*, 131(41):15005–15013, 2009.

[226] EJ Sambriski, DC Schwartz, and JJ De Pablo. A mesoscale model of DNA and its renaturation. *Biophysical journal*, 96(5):1675–1690, 2009.

[227] PD Dans, A Zeida, MR Machado, and S Pantano. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *Journal of chemical theory and computation*, 6(5):1711–1725, 2010.

[228] TE Ouldridge, AA Louis, and JPK Doye. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *The Journal of chemical physics*, 134(8):02B627, 2011.

[229] RC DeMille, TE Cheatham III, and V Molinero. A coarse-grained model of DNA with explicit solvation by water and ions. *The Journal of Physical Chemistry B*, 115(1):132–142, 2010.

[230] RE Franklin and RG Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171:740–741, 1953.

[231] NS Bogatyreva, AV Finkelstein, and OV Galzitskaya. Trend of amino acid composition of proteins of different taxa. *Journal of bioinformatics and computational biology*, 4(02):597–608, 2006.

[232] ID Vladescu, MJ McCauley, I Rouzina, and MC Williams. Mapping the phase diagram of single DNA molecule force-induced melting in the presence of ethidium. *Physical review letters*, 95(15):158102, 2005.

[233] AA. Almaqwashi, T Paramanathan, P Lincoln, I Rouzina, F Westerlund, and MC. Williams. Strong DNA deformation required for extremely slow DNA threading intercalation by a binuclear ruthenium complex. *Nucleic Acids Research*, 42(18):11634–11641, 2015.

[234] AA Almaqwashi, J Andersson, P Lincoln, I Rouzina, F Westerlund, and MC Williams. DNA intercalation optimized by two-step molecular lock mechanism. *Scientific Reports*, 6:37993, 2016.

[235] B Kundukad, P Cong, JRC van der Maarel, and PS Doyle. Time-dependent bending rigidity and helical twist of DNA by rearrangement of bound HU protein. *Nucleic Acids Research*, 41(17):8280–8288, 2013.

[236] A Mihailovic, I Vladescu, M McCauley, E Ly, MC Williams, EM Spain, and ME Nuñez. Exploring the interaction of ruthenium (II) polypyridyl complexes with DNA using single-molecule techniques. *Langmuir*, 22(10):4699–4709, 2006.

[237] EF Silva, RF Bazoni, EB Ramos, and MS Rocha. DNA-doxorubicin interaction: New insights and peculiarities. *Biopolymers*, 107(3), 2017.

[238] M Trieb, C Rauch, FR Wibowo, B Wellenzohn, and KR Liedl. Cooperative effects on the formation of intercalation sites. *Nucleic acids research*, 32(15):4696–4703, 2004.

[239] Q Gao, LD Williams, M Egli, D Rabinovich, SLe Chen, GJ Quigley, and A Rich. Drug-induced DNA repair: X-ray structure of a DNA-ditercalinium complex. *Proceedings of the National Academy of Sciences*, 88(6):2422–2426, 1991.

[240] CC Wu, YC Li, YR Wang, TK Li, and NL Chan. On the structural basis and design guidelines for type II topoisomerase-targeting anticancer drugs. *Nucleic acids research*, 41(22):10630–10640, 2013.

[241] CU Murade, V Subramaniam, C Otto, and ML Bennink. Interaction of oxazole yellow dyes with DNA studied with hybrid optical tweezers and fluorescence microscopy. *Biophysical journal*, 97(3):835–843, 2009.

[242] A Sischka, K Toensing, R Eckel, SD Wilking, N Sewald, R Ros, and D Anselmetti. Molecular mechanisms and kinetics between DNA and DNA binding ligands. *Biophysical journal*, 88(1):404–411, 2005.

[243] ML Bennink, OD Schärer, R Kanaar, K Sakata-Sogawa, JM Schins, JS Kanger, B G de Grooth, and J Greve. Single-molecule manipulation of double-stranded DNA using optical tweezers: interaction studies of DNA with RecA and YOYO-1. *Cytometry*, 36(3):200–208, 1999.

[244] RJ Thresher and JD Griffith. Intercalators promote the binding of RecA protein to double-stranded DNA. *Proceedings of the National Academy of Sciences*, 87(13):5056–5060, 1990.

[245] SK Kim, B Norden, and M Takahashi. Role of DNA intercalators in the binding of RecA to double-stranded DNA. *Journal of Biological Chemistry*, 268(20):14799–14804, 1993.

[246] J Heuser and J Griffith. Visualization of RecA protein and its complexes with DNA by quick-freeze/deep-etch electron microscopy. *Journal of molecular biology*, 210(3):473–484, 1989.

[247] X Yu and EH Egelman. Structural data suggest that the active and inactive forms of the RecA filament are not simply interconvertible. *Journal of molecular biology*, 227(1):334–346, 1992.

[248] A Stasiak and E Di Capua. The helicity of DNA in complexes with RecA protein. *Nature*, 299(5879):185–186, 1982.

[249] T Masuda, Y Ito, T Terada, T Shibata, and T Mikawa. A non-canonical DNA structure enables homologous recombination in various genetic systems. *Journal of Biological Chemistry*, 284(44):30230–30239, 2009.

[250] DL Beveridge, SB Dixit, G Barreiro, and KM Thayer. Molecular dynamics simulations of DNA curvature and flexibility: Helix phasing and premelting. *Biopolymers*, 73(3):380–403, 2004.

[251] DP Landau and K Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2014.

[252] D Keller, D Swigon, and C Bustamante. Relating single-molecule measurements to thermodynamics. *Biophysical journal*, 84(2):733–738, 2003.

[253] TJ Macke and DA Case. Modeling unusual nucleic acid structures. ACS Publications, 1998.

[254] A Mukherjee, R Lavery, B Bagchi, and JT Hynes. On the molecular mechanism of drug intercalation into DNA: a simulation study of

the intercalation pathway, free energy, and DNA structural changes. *Journal of the American Chemical Society*, 130(30):9747–9755, 2008.

[255] Donald M Chothers. Calculation of binding isotherms for heterogeneous polymers. *Biopolymers*, 6(4):575–584, 1968.

[256] W Bauer and J Vinograd. Interaction of closed circular DNA with intercalative dyes: II. the free energy of superhelix formation in SV40 DNA. *Journal of molecular biology*, 47(3):419–435, 1970.

[257] HM Berman and PR Young. The interaction of intercalating drugs with nucleic acids. *Annual review of biophysics and bioengineering*, 10(1):87–114, 1981.

[258] E Nordmeier. Absorption spectroscopy and dynamic and static light-scatteing studies of ethidium bromide binding to calf thymus DNA: implications for outside binding and intercalation. *Journal of physical chemistry*, 96(14):6045–6055, 1992.

[259] J Yan and JF Marko. Effects of DNA-distorting proteins on DNA elastic response. *Physical Review E*, 68(1):011905, 2003.

[260] K Schakenraad, AS Biebricher, M Sebregts, B ten Bensel, EJG Peterman, GJL Wuite, I Heller, C Storm, and P van der Schoot. Hyperstretching DNA. *Nature Communications*, 8(1):2197–, 2017.

[261] P Cluzel, A Lebrun, C Heller, R Lavery, JL Viovy, D Chatenay, and F Caron. DNA: An extensible molecule. *Science*, 271(5250):792–794, 1996.

[262] GJL Wuite, RJ Davenport, A Rappaport, and C Bustamante. An integrated laser trap/flow control video microscope for the study of single biomolecules. *Biophysical Journal*, 79(2):1155–1167, 2000.

[263] D Řeha, M Kabelác, F Ryjácek, J Šponer, JE Šponer, M Elstner, S Suhai, and P Hobza. Intercalators. 1. nature of stacking interactions between intercalators (ethidium, daunomycin, ellipticine, and 4‘, 6-diaminide-2-phenylindole) and DNA base pairs. Ab initio quantum chemical, density functional theory, and empirical potential study. *Journal of the American Chemical Society*, 124(13):3366–3376, 2002.

[264] J Šponer, J Leszczyński, and P Hobza. Nature of nucleic acid-base stacking: nonempirical ab initio and empirical potential characterization of 10 stacked base dimers. comparison of stacked and H-bonded

base pairs. *The Journal of Physical Chemistry*, 100(13):5590–5596, 1996.

[265] F Gago. Stacking interactions and intercalative DNA binding. *Methods*, 14(3):277–292, 1998.

[266] A Garai, S Mogurampelly, S Bag, and PK Maiti. Overstretching of B-DNA with various pulling protocols: Appearance of structural polymorphism and S-DNA. *The Journal of chemical physics*, 147(22):225102, 2017.

[267] P Bianco, L Bongini, L Melli, M Dolfi, and V Lombardi. Piconewton-millisecond force steps reveal the transition kinetics and mechanism of the double-stranded DNA elongation. *Biophysical journal*, 101(4):866–874, 2011.

[268] L Bongini, V Lombardi, and P Bianco. The transition mechanism of DNA overstretching: a microscopic view using molecular dynamics. *Journal of The Royal Society Interface*, 11(97), 2014.

[269] E Tuite, SK Kim, B Norden, and M Takahashi. Effects of intercalators on complexation of RecA with duplex DNA. *Biochemistry*, 34(50):16365–16374, 1995.

[270] J Lipfert, S Klijnhout, and NH Dekker. Torsional sensing of small-molecule binding using magnetic tweezers. *Nucleic acids research*, 38(20):7122–7132, 2010.

[271] M Selmer, CM Dunham, FV Murphy, A Weixlbaumer, S Petry, AC Kelley, JR Weir, and V Ramakrishnan. Structure of the 70s ribosome complexed with mRNA and tRNA. *Science*, 313(5795):1935–1942, 2006.

[272] D Argudo and PK Purohit. Equilibrium and kinetics of dna overstretching modeled with a quartic energy landscape. *Biophysical journal*, 107(9):2151–2163, 2014.

[273] JY Lee, T Terakawa, Z Qi, JB Steinfeld, S Redding, YH Kwon, WA Gaines, W Zhao, P Sung, and EC Greene. Base triplet stepping by the RAD51/RecA family of recombinases. *Science*, 349(6251):977–981, 2015.

[274] A Reymer, K Frykholm, K Morimatsu, M Takahashi, and B Nordén. Structure of human RAD51 protein filament from molecular modeling and site-specific linear dichroism spectroscopy. *Proceedings of the National Academy of Sciences*, 106(32):13248–13253, 2009.