REPORT

DNA partitions into triplets under tension in the presence of organic cations, with sequence evolutionary age predicting the stability of the triplet phase

Amirhossein Taghavi[1], Paul van der Schoot[2] and Joshua T. Berryman[1]*

[1] Department of Physics and Materials Science, University of Luxembourg, 162A Avenue de la Faïencerie, Luxembourg City, Luxembourg
[2] Department of Applied Physics, Theory of Polymers and Soft Matter, Technische Universiteit Eindhoven, P.O. Box 513 5600 MB, Eindhoven, The Netherlands

Abstract. Using atomistic simulations, we show the formation of stable triplet structure when particular GC-rich DNA duplexes are extended in solution over a timescale of hundreds of nanoseconds, in the presence of organic salt. We present planar-stacked triplet disproportionated DNA (Σ DNA) as a possible solution phase of the double helix under tension, subject to sequence and the presence of stabilising co-factors. Considering the partitioning of the duplexes into triplets of base pairs as the first step of operation of recombinase enzymes like RecA, we emphasise the structure–function relationship in Σ DNA. We supplement atomistic calculations with thermodynamic arguments to show that codons for 'phase 1' amino acids (those appearing early in evolution) are more likely than a lower entropy GC-rich sequence to form triplets under tension. We further observe that the four amino acids supposed (in the 'GADV world' hypothesis) to constitute the minimal set to produce functional globular proteins have the strongest triplet-forming propensity within the phase 1 set, showing a series of decreasing triplet propensity with evolutionary newness. The weak form of our observation provides a physical mechanism to minimise read frame and recombination alignment errors in the early evolution of the genetic code.

1. Introduction

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 ~1·5−1·7 (depending on sequence, pulling geometry and solution) at ~65–70 pN (Bosaeus *et al.* 2012; Liu *et al.* 2010; Vlassakis *et al.* 2008; Williams *et al.* 2002). 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) (Williams *et al.* 2001) or of ladder-like stretched and untwisted double-stranded DNA (dsDNA) have been suggested (Bosaeus *et al.* 2017; Cluzel *et al.* 1996; Konrad & Bolonick, 1996; Lebrun & Lavery, 1996; Smith *et al.* 1996). Although the structural details and degree of order are not clear or universally agreed in the literature, in general at modest extensions of long sequences not dominated by AT base pairs, we expect to see a clear two-state transition to a partly untwisted ladder-like structure, in which the base pairs remain intact but the average rise per base pair is equilibrated to a new value of ~5·8 Å, compared to the rise in unstretched B-DNA of 3·4 Å.

* Author for correspondence: Joshua T. Berryman, Campus Limpertsberg, Université du Luxembourg 162 A, avenue de la Faïencerie L-1511 Luxembourg, Luxembourg. Tel.: (+352) 46 66 44 6971; Fax: (+352) 46 66 44 36971; Email: josh.berryman@uni.lu

The stretched phase or phases are known by the umbrella label of 'S-DNA'. For GC-rich structures having strong hydrogen bonding, the base pairing is preserved in the putative S-DNA structure, and the base stacking may also be somewhat preserved by tilting and sliding of the base pairs or by opening of 'bubbles' between base pairs (Konrad & Bolonick, 1996). Reorientation of the base pairs may increase the solvent-exposed area by changing the groove width while permitting them to remain in contact such that a complete water gap does not open between them (Roe & Chaka, 2009). The detailed S-DNA structure, particularly the inclination, depends on the pulling scheme. Experimental results of Albrecht et al. show that the differences in behaviour of DNA when using different pulling schemes can depend on the loading rate, with stretching forces propagating first at high pulling rates, but shearing becoming dominant at slower pulling rates (Albrecht et al. 2008). Others have also shown that there are structural differences when using different pulling mechanisms (Bag et al. 2016; Danilowicz et al. 2009; Roe & Chaka, 2009).

The most readily available information on DNA under tension is the empirically measured force-extension curve (Rief et al. 1999; Smith et al. 1992), which provides the clear signal of some transition, but no atomistic-level information. This is supplemented by fluorescence and polarised-light studies (van Mameren et al. 2009; Nordén et al. 1992; King et al. 2013), 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 (Bag et al. 2016; Konrad & Bolonick., 1996; Lebrun & Lavery, 1996; Li & Gisler, 2009; Roe & Chaka, 2009).

DNA is often subjected to tension in its biological context, for purposes including transport, transcription and tertiary structure manipulation (Nicklas, 1988). A striking example of this is the crystal structure (*pdb:* 3cmt) of DNA bound to the RecA protein (Chen et al. 2008), a snapshot of homologous DNA recombination in action. Homologous recombination using RecA is employed for DNA repair (Kuzminov, 1999), and via other recombinases is also the fundamental process of sexual reproduction, in which copies of similar genes are exchanged between DNA strands from either parent. In the RecA structure, the extended protein-bound DNA duplex does not adopt a recognised S-like configuration, but rather disproportionates into groups of three bases, with orderly planar base stacking retained within each triplet. This triplet disproportionation has been observed in solution when bound to RecA (Nordén et al. 1992), in crystallographic structures of the RecA–DNA complex (Chen et al. 2008) and also has been suggested as a stable phase even without co-factors (Bosaeus et al. 2017). Simple energy minimisation (without thermal fluctuations or any explicit solvent) of d[GCG]$_4$ DNA under constraints of a 1·5 relative extension and also of imposed helical symmetry with unit cell 3 bp yields a structure

partitioned into triplets of three stacked base pairs separated by a stack-breaking gap of approximately one base-pair thickness (Bertucat et al. 1998). The early minimisation calculation shows that a period 3 or 'triplet' structure is physically plausible when considered purely in static chemical terms, although it has not since been experimentally or computationally observed in solution. We note that the term 'triplet' here refers to a duplex of length three bases (perhaps more properly a 'triad'), and not to a triplex of three strands.

Orderly triplet formation when complexed is in contrast to current general understanding of the structural behaviour when extended in solution, which leads us to examine whether the triplet phase can be stabilised in solution and if it could in this case 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 stable triplet structure in an aqueous solution of monatomic counterions, but do find that it is stable without specific complex to a structured enzyme, forming triplets in a solution either of terminus capped monomeric arginine peptides (Ac-Arg$^+$-NHMe Cl$^-$) or (more weakly) of the well-known intercalant ethidium bromide (Et$^+$Br$^-$).

The presence of intercalators has been observed to drive significant alterations in the quasiequilibrium force *versus* extension curve, with an effect at high concentrations of smoothing over the B to S transition and possibly of modifying the structure of the S phase (Vladescu et al. 2007). By carrying out simulated stretching experiments in the presence of DNA-binding cofactors, we intend to reduce the barrier and collectivity associated with the B–S transition, thus increasing the likelihood that the sub-microsecond simulation timescale can describe the typical laboratory-scale (milliseconds) process, and also to investigate the structural role of biologically relevant moieties (Arg) in relation to DNA under tension.

We discuss planar-stacked triplet disproportionated DNA as a solution phase of the double helix under tension, and refer to it as 'Σ DNA', with the three right-facing points of the Σ character serving as a mnemonic for the three grouped base pairs. In the same way as for 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 0 or 'the GADV world' (Ikehara et al. 2002), phase 1 or 'GADVESPLIT'

2

(Koonin & Novozhilov, 2009; Wong, 1975, 2005), then the remaining amino acids, CMFYHRKNQ), into molecular biology with the ease of Σ-formation for sequences including the associated codons. 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 gene expression.

2. Methods

2.1 Triplet propensity calculation

To calculate the triplet propensity, we proceed from datasets 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 1' 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 ⅓ 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 300 K. 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 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 re-drawing of the 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. The short script used for the calculation is included as Supplementary software.

2.2 Atomistic simulation protocols

Molecular structures were prepared using the Nucleic Acid Builder (NAB) (Macke & Case, 1998). Salt was represented using the Joung–Cheatham parameters for ions (Joung & Cheatham, 2008) and TIP3P model of water was used for solvation (Jorgensen *et al.* 1983). The ethidium molecule was represented using the GAFF (Wang *et al.* 2004) with partial charges and bond parameters assigned via the ANTECHAMBER tool (Wang *et al.* 2006). Simulations were run using the GPU-accelerated implementation of pmemd (Salomon-Ferrer *et al.* 2013) in the AMBER16

package (Case *et al.* 2016) using the AMBER 99SB+bsc0 force field parameters for the biomolecules (Hornak *et al.* 2006; Pérez *et al.* 2007). 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 0·1 M) (Table 1). The SHAKE algorithm (Ryckaert *et al.* 1977) 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 (Darden *et al.* 1993). 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.

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 0·068 nm 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 (Biebricher *et al.* 2015), binding of intercalators was essentially irreversible *in silico* except via extensionally driven conformational change. The choice to make multiple 150 ns simulations 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.

DNA was pulled using force applied to the centres of geometry of the top and bottom base pairs, such that no topological restraint was applied and force was distributed equally between 3′ and 5′ strand ends. The choice to spread the applied force was made based on an assumption that the distribution of applied force in an early biological context was likely to be relatively smooth (e.g. via hydrodynamic combing or weakly specific binding of cofactors) rather than limited to two specific atoms.

Averaging of angles (for study of DNA structural parameters) was carried out by taking the mean cosine and sine, then the arctangent of the mean values. Structures were analysed using CURVES+ (Lavery *et al.* 2009).

In order to present a dimensionless relative extension, the unstretched contour length per bp was estimated over 24

3

Table 1. *Details of the atomistic simulations*

DNA sequence	Total number of atoms	Number of Arg/EtBr	Na$^+$
[GGC]$_4$[GAC]$_4$ · [GTC]$_4$[GCC]$_4$	35 584	None	70 (24 Cl$^-$)
[GGC]$_4$[GAC]$_4$ · [GTC]$_4$[GCC]$_4$ + EtBr	32 357	88	24 (66 Br$^-$)
[GGC]$_4$[GAC]$_4$ · [GTC]$_4$[GCC]$_4$ + Arg	32 358	88	24 (66 Cl$^-$)
G$_{12}$C$_{12}$·G$_{12}$C$_{12}$	35 592	None	70 (24 Cl$^-$)
G$_{12}$C$_{12}$·G$_{12}$C$_{12}$ + EtBr	32 359	88	24 (66 Br$^-$)
G$_{12}$C$_{12}$·G$_{12}$C$_{12}$ + Arg	32 357	88	24 (66 Cl$^-$)

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.

bp as 3·46 Å (Arg Cl), 3·43 Å (EtBr) and 3·43 Å (NaCl) for the G$_{12}$C$_{12}$·G$_{12}$C$_{12}$ sequence and 3·57 Å (Arg Cl), 3·41 Å (EtBr) and 3·38 Å (NaCl) for the [GGC]$_4$[GAC]$_4$ · [GTC]$_4$[GCC]$_4$ sequence.

2.3 Base-pair step classification

Base-pair steps were classified according to the following rules (see Fig. 5, e.g structures):

- $β$: ('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 >135° and the distance donor–acceptor was <3 Å (cpptraj defaults) (Roe & Cheatham, 2013). The rise cutoff for stack breaking was made at 5·6 Å based on (Chen & García, 2013). An atomic contact was defined as a distance <3·5 Å. Examples of each conformation are shown in Fig. 2. Averages were collected over steps [3…21] of the 23 steps available in the sequences studied, in order to minimise end effects.

3. Results

3.1 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 guesses about the history of the genetic code by considering the metabolic networks leading to the different amino acids: it is hypothesised that a list of so-called 'phase 1' amino acids were present earlier in evolution than the remaining 'phase 2' amino acids, based on the complexity of the cellular machinery used in current organisms to synthesise, for example, methionine (M) from threonine (T) (Koonin & Novozhilov, 2009; Wong & Bronskill, 1979). If the genetic code in the epoch of a much simplified amino-acid alphabet already had the current structure of three base pairs per codon ('triplet code') it was therefore highly redundant at this time.

In the current triplet code, the 'phase 1' amino acids supposed to have been incorporated earliest into biology (the list of GADVESPLIT) are coded by triplets which have a specific physical tendency: the energetic cost to break base stacking at the triplet boundary is low, relative to the complete modern genetic code. In this paper, we first motivate the statistical observation of preferential triplet disproportionation in the phase 1 genetic code. We then analyse atomistic simulation data to show that disproportionation into coding triplets occurs spontaneously under tension for appropriate sequences and solution conditions.

The weak form of our observation provides a physical mechanism to minimise read frame and recombination alignment errors in the early evolution of the genetic code. We further motivate this to make the stronger claim of a possible route for the origin of the triplet genetic code, with the three base-pair structure arising from simple physical conditions in the absence of the sophisticated enzymatic machinery which later evolved to maintain the triplet code in modern organisms with a full alphabet of amino acids.

3.2 Estimation of sequence-dependent triplet-forming propensity

The free energy of base stacking in duplex DNA was long ago calculated combinatorially for the various pairs of bases, by Friedman and Honig (Friedman & Honig, 1995). Stacking energies are directional (crucially, GC·GC is stronger than CG·CG), here when writing a sequence the direction 5′ → 3′ is assumed.

From the tabulated data (Friedman & Honig, 1995), if we approximate the stacking energy for complementary duplex

4

DNA as the sum of the stacking energies for the two base steps, the weakest step is CG·CG (-10.14 kcal mol^{-1}). We therefore hypothesise 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.

To calculate the triplet propensity, we take a brute-force Monte Carlo approach, selecting a long sequence and allowing stack breaks to equilibrate according to a Boltzmann distribution, before sampling the probability for each codon to partition stack-break defects such that it has a stack break at each end but not inside it.

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 'stage 1' amino acids (Table 2).

The dramatic pattern evident in the tabulated partitioning energies is that stage 1 amino acids overwhelmingly have relatively favourable free energies to partition into triplets aligned to their codons. The exception to this pattern is interesting: arginine (R) is not listed in Wong & Bronskill's (1979) tabulation of stage 1 amino acids (Wong & Bronskill, 1979), possibly due to the large energetic cost needed to synthesise it from citrulline in modern organisms (Ratner & Petrack, 1953).

It has been advanced the CGN and AGN codons which yield arginine in the modern genetic code previously coded for the chemically similar non-canonical amino acid ornithine (Jukes, 1973), 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 1 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.

The weaker statement of this work, that the stage 1 part of the genetic code is structured so as to support a

Table 2. *Phase one amino acids (*,♀) tend to have (at least one) codon associated with them that partitions favourably at its boundaries*

Amino acid	Codon–anticodon	$\Delta G_\tau / k_B T$
G♀	GGC·GCC	1·71
A♀	GCC·GGC	1·71
S*	AGC·GCT	1·86
V♀	GTC·GAC	2·01
D♀	GAC·GTC	2·02
T*	ACC·GGT	2·07
R.	AGA·TCT	2·18
F.	TTC·GAA	2·20
E*	GAA·TTC	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	2·69
Y.	TAC·GTA	3·32
C.	TGC·GCA	3·38
H.	CAC·GTG	4·26
Q.	CAA·TTG	4·52
M.	ATG·CAT	4·56

The most favourable partitioning is for the phase 0 (DAGV) amino acids (♀).

minimisation of read-frame errors by physically favouring the partition into aligned triplets under tension, 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 (Itzkovitz & Alon, 2007), 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 observe: Table 3 shows that codons for phase 1 amino acids are significantly more able to encode this partitioning than those in phase 2. We further observe that the residues advanced by Ikehara et al. (Ikehara et al. 2002) as forming the minimal set for a functional proteome (marked ♀ in Table 3) are also those which partition most naturally into triplets.

We should note that the calculation of stacking energies by Friedman and Honig is a very elegant paper but is not the most recent attempt to measure this quantity. On the experimental side, Protozanova et al. have measured free energy of stacking for base-pair steps in a nicked DNA duplex (Protozanova et al. 2004). 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 (Pieters et al. 1989; Roll et al. 1998; Snowden-Ifft & Wemmer, 1990) which leaves

5

Table 3. *All codons, coloured by G_τ with a light colour indicating higher probability to disproportionate into triplets*

		Base 3				
		T	C	A	G	
		F.	F.	L*	L*	T
		S*	S*	S*	S*	C
T		Y.	Y.	X.	X.	A
		C.	C.	X.	Y.	G
		L*	L*	L*	L*	T
		P*	P*	P*	P*	C
C		H.	H.	Q.	Q.	A
		R.	R.	R.	R.	G
		I*	I*	I*	M.	T
		T*	T*	T*	T*	C
A		N.	N.	K.	K.	A
		S*	S*	R.	R.	G
		V†	V†	V†	V†	T
		A†	A†	A†	A†	C
G		D†	D†	E*	E*	A
		G†	G†	G†	G†	G

Base 1 (left) — *Base 2* (right)

The pattern purine-x-pyrimidine gives greatest triplet disproportionation, also having a (bulky) G or C base in the middle of the codon gives more favourable Σ formation. Because energies were found in the duplex form, complementary codons (e.g. GGC, GCC) necessarily have the same G_τ. ♀, * indicates phase 1 amino acids, ♀ indicates phase 0. X indicates a stop codon.

a question mark over the applicability of the data for our purposes. Protozanova *et al.* report values of −0·91 and −2·17 kcal mol⁻¹ 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⁻¹ (Kilchherr *et al.* 2016). Calculations by Lemkul and MacKerell (Lemkul & MacKerell, 2017) 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⁻¹. Overall then, 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. We include as Supplementary data repeated calculations of triplet formation-free energies with the Protozanova *et al.* and Lemkul and MacKerell datasets (Fig. S1). These Supplementary calculations show the same privileged status for the GADV and phase 1 amino acids as those based on the Honig dataset, although many details are different, and also in the Lemkul and MacKerrell calculation, the residue glycine loses its privileged triplet status almost entirely due to a very weak reported value for the GG·CC stack. We also include as a Supplementary software the python script used to carry out the triplet propensity calculation.

3.3 Spontaneous triplet disproportionation under tension, amplified in the presence of organic cations

Beyond the pairwise hydrophobic and electrostatic interactions of base stacking (covered by the classic calculation used to generate Tables 2 and 3) the potential importance of complex entropic, structural and solvent effects makes 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 four glycines and four prolines) and a sequence chosen to show strong triplet disproportionation based on Table 2, [GGC]₄[GAC]₄·[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.* 2002)). 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 and on the RecA structure (Chen *et al.* 2008), simulations were run both in NaCl and in a solution of Ac-Arg-NHMe Cl.

We find that for the GC-rich sequence encoding phase 1 amino acids, the triplet-disproportionated Σ conformation of DNA is observed, with the strongest triplet formation taking place in the presence of the terminus-capped arginine residues (Ac-Arg-NHMe). Figure 1 shows a regular pattern of vertical stripes (preferred base-pair step breaks) with period 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 3 is seen, even in the absence of arginine; however, this is relatively weak (as suggested by the order 1 $k_B T$ free energies of disproportionation in Table 2).

The triplet-disproportionated structures show the essential features of Σ-DNA (Fig. 2) as seen in the RecA-bound crystal structure: preserved Watson–Crick base pairing, approximately planar orientation of the bases (Fig. 3), and a large cavity every third base pair.

Extending beyond approximately 3 Å bp⁻¹ 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.

The question of planarity of the base stacking is important to monitor as a structure of stretched DNA in which formation of stack breaks is avoided by collective tilting of the stack has been associated with stretching in the 3′–3′ geometry (Roe & Chaka, 2009). Without intercalators, the average base-pair inclination in the presence and absence of intercalators up to the extension point of 1·5 was flat (Figs 3b, 3d

6
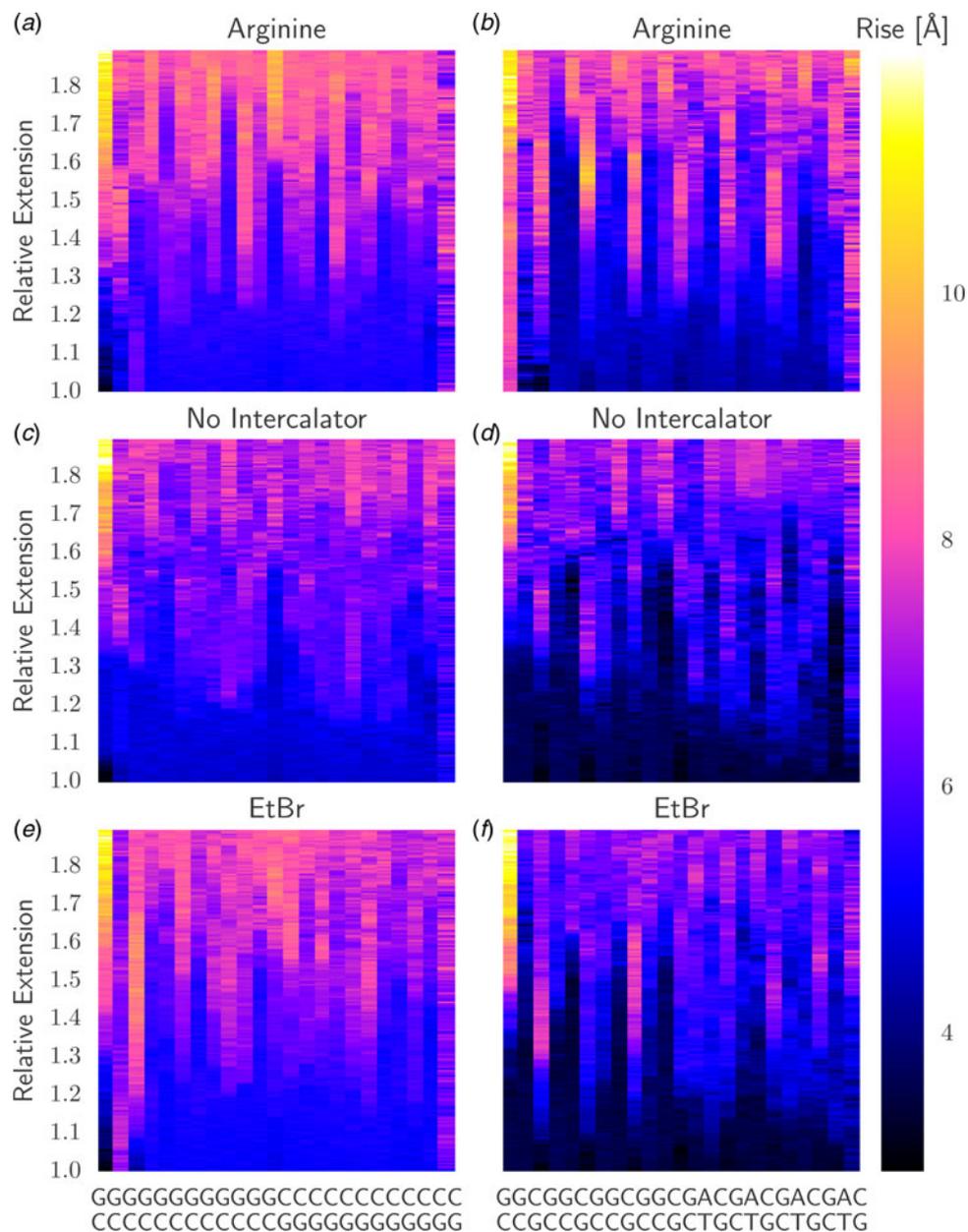
Fig. 1. Kymographs of rise per bp-step under imposed whole-DNA extension. Triplet disproportionation is strongly evident in (*b*), while the strain is spread most evenly in (*c*). Presence of arginine in a homogeneous sequence (*a*) or presence of CG steps in the absence of arginine (*d*) induce only weakly structured disproportionation.

and 3*f*), which indicates no collective tilting. For the high-entropy sequence, in the presence of intercalators, this trend continued after an extension of 1·5 but without inter-calators showed a sudden drop at around 1·7. For the low-entropy sequence ($G_{12}C_{12} \cdot G_{12}C_{12}$), 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 (Figs 3*a* and 3*c*), but in the presence of EtBr, a continuous increase is observed after extension 1·5, followed by a second flat

region after 1·6 (Fig. 3*e*). That average inclination tends to be small during DNA extension for the high-entropy sequence with or without organic cations is consistent with the results put forward from experiment (Bosaeus *et al.* 2017) as suggesting Σ formation even in free solution.

3.4 Base-pair step classification

Figure 4 shows the proportions of different local conforma-tions consistent with different phases. Examples of each conformation are shown in Fig. 5. Local conformations

7
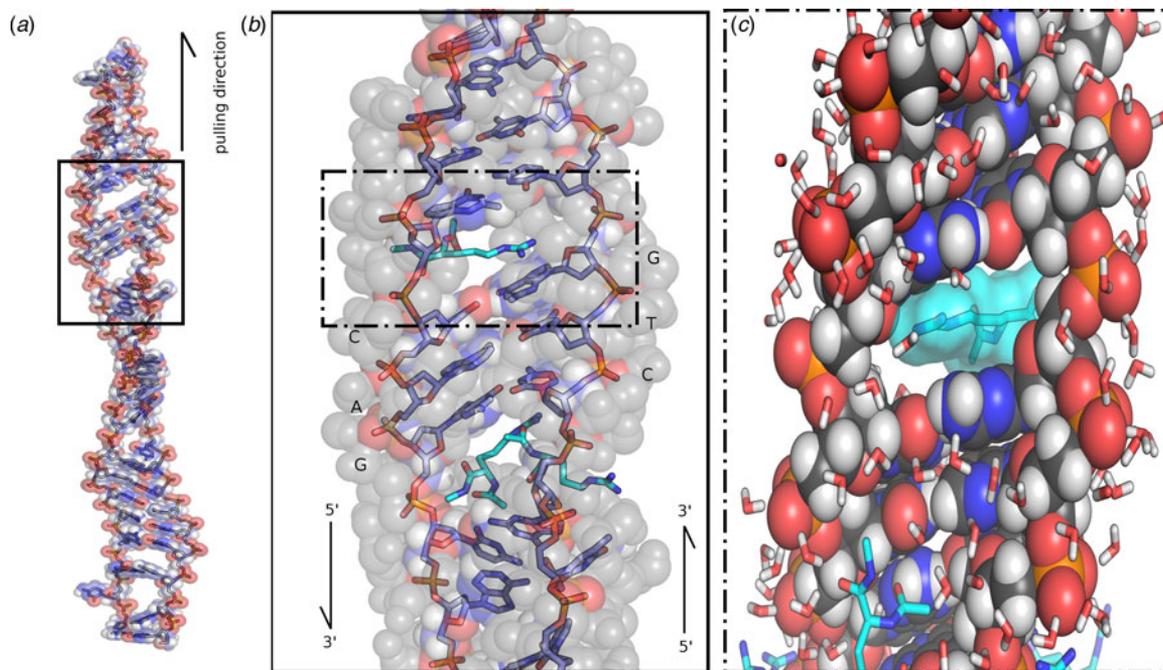
Fig. 2. The 'primordial' sequence partitions under tension predominantly at the CG steps, forming triplets (*a*), with Watson–Crick hydrogen bonding and planar base stacking preserved subject to some thermal disorder (*a, b*). Triplets are stabilised by one or two arginines (*cyan*) intercalating the 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*).

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 (Figs 4*h* and 4*l*).

4. Discussion

The mechanics of DNA under tension are affected by various factors including the counterions (Vlassakis *et al.* 2008),

sequence (Rief *et al.* 1999) and temperature (Fu *et al.* 2010). Molecular combing results show that DNA in its stretched form is not denatured, with a double-helix structure which is characterised by a diameter of 1·2 nm (Maaloum *et al.* 2011). X-ray diffraction of stretched cross-linked films of a mixed sequence of DNA show gaps of ~8 Å (André *et al.* 2008), nearly the same size as seen in the DNA–RecA complex. These experimental results suggest the existence of a stretched form of solution DNA with preserved base-pair stacking.

In order to relate the physics of DNA stretching with its function in storing and copying information, we have estimated the sequence dependence of the free energy cost in water at moderate ionic strength to unstack a given codon from its neighbours at the codon boundaries, leaving the codon itself intact as a triplet of stacked base pairs. This initial calculation reveals that although the aqueous solution environment does not strongly drive triplet partitioning, that a distinct hierarchy of triplet formation energies exists with respect to sequence features. The triplet formation energy estimates show that sequences coding for 'stage 1' amino acids hypothesised to have appeared early in evolution (plus arginine) are more likely to partition into triplets at the codon boundaries than to partition in any other pattern when under tension. In order to investigate this phenomenon, we carried out pulling simulations of DNA duplexes encoding stage 1 amino acids, in the presence of

8
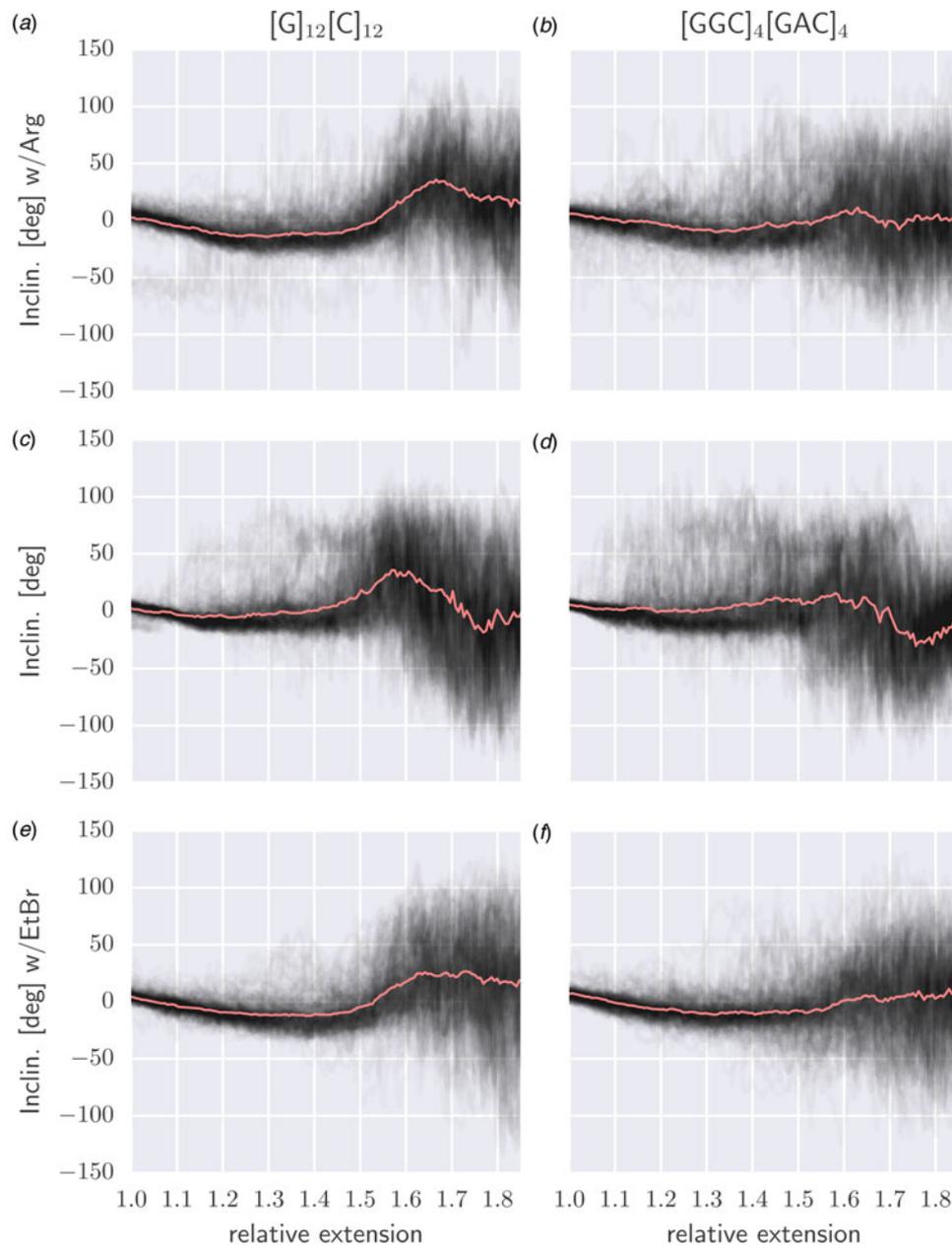
Fig. 3. Average inclination of the low- and high-entropy sequences in the presence and absence of intercalators (arginine and EtBr). Average inclination for the high-entropy sequence [GGC]$_4$[GAC]$_4$ · [GTC]$_4$[GCC]$_4$ remains relatively flat up to extension 1·5 and beyond, even without intercalant (b, d, f). For the low-entropy sequence G$_{12}$C$_{12}$·C$_{12}$G$_{12}$, average inclination remains flat up to extension of 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).

arginine and also of ethidium bromide, as well as control simulations using low-entropy sequences, and also tested aqueous conditions with monatomic salt only.

In order to observe substantial triplet disproportionation, both a bulky organic cation (i.e. arginine or ethidium) and a sequence selected from codons yielding phase 1 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, while stretching without cofactors or appropriate sequence led to disordered but not fully denatured structure consistent with experiments (Balaeff *et al.* 2011; Bosaeus *et al.* 2017) and simulations (Konrad & Bolonick, 1996).

Ikehara *et al* have shown that codons matching the pattern GNC probably constituted the original, minimal functional

9
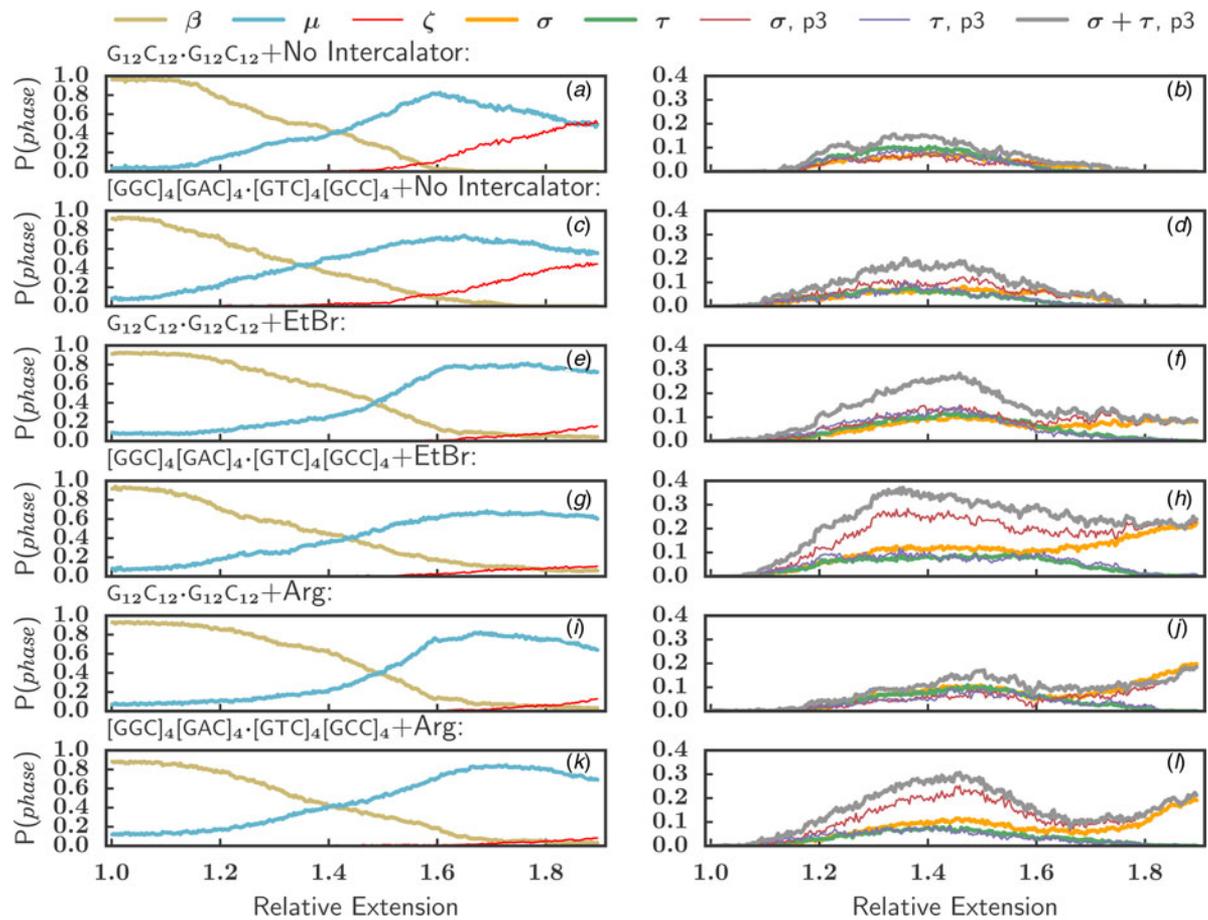
Fig. 4. Base-pair steps (four bases) were classified by local conformation as β: base-paired and stacked, μ: melted, ζ: zipper, as planar with broken stacking (σ) or as τ: tilted. The left panels (a, c–k) show the three major states of the DNA, with a melting transition over extensions 1·2–1·6, followed (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.

genetic code. These authors argue based on multiple strands of reasoning that the amino acids GADV, translated from these codons, constitute the unique minimal adequate set to generate functional globular proteins (Ikehara *et al.* 2002). We argue that it is no coincidence that the same GNC codons are those which drive maximal triplet dispro-portionation, allowing recombination and possibly protein synthesis to operate without the sophisticated enzymatic machinery which exists today. We hypothesise that a bootstrap process took place, with crude triplet dispropor-tionation facilitating recombination, driving-accelerated evolution and leading to more sophisticated protein–DNA interactions, in turn allowing expansion in further stages of the genetic code.

We attempted to reductively study the factors driving Σ for-mation of DNA, which we deem to be force (applied via fil-ament 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 timescale accessed is considerably shorter than the longest (1 s) equilibration timescale of the system.

4.1 Speculative box

We speculate that ornithine (instead of arginine) may play the role of triplet promoter in some organisms, or have done so earlier in the evolutionary process. The usurpation of ornithine codons to instead signify arginine may have led to a jump in the efficiency of recombination and an evolu-tionary explosion.

10
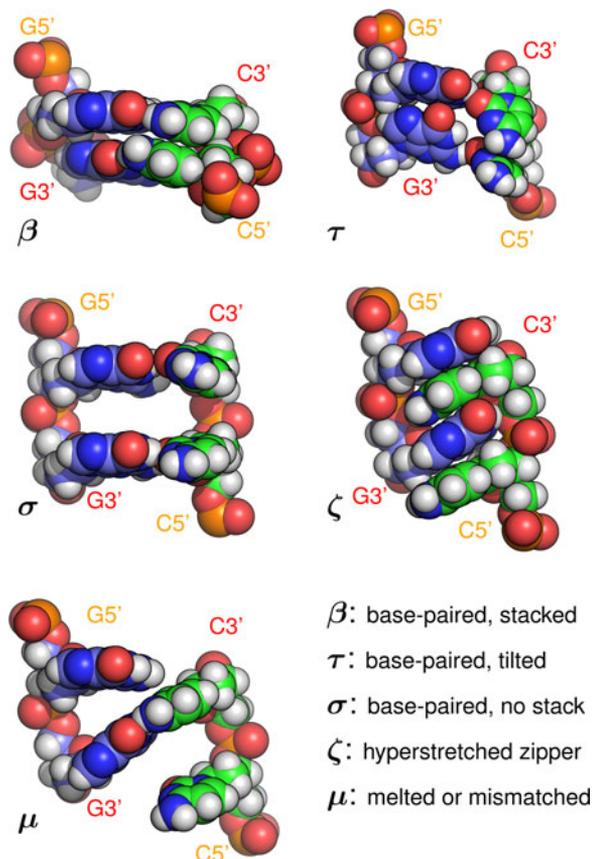
Fig. 5. Example base-pair step conformations (all of sequence GG·CC) classified by the type of stacking and hydrogen bonding present. Note that the initials β, τ, σ, ζ, μ do not refer to phases (collective structures) but local conformations. A step labelled as 'β' would, for example, be consistent with the A, B, C, Z, or Σ phases, all of which include base stacking and Watson–Crick hydrogen bonding.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S0033583517000130.

Acknowledgements

Financial Support

Conflict of Interest

None.

References

ALBRECHT, C. H., NEUERT, G., LUGMAIER, R. A. & GAUB, H. E. (2008). Molecular force balance measurements reveal that double-stranded DNA unbinds under force in rate-dependent pathways. *Biophysical Journal* **94**, 4766–4774.

ANDRÉ, A., FONTAINE-VIVE, F., MÖLLER, H. M., FISCHER, T., MARET, G., FORSYTH, V. & GISLER, T. (2008). Force-induced structural transitions in cross-linked DNA films. *European Biophysics Journal* **37**, 749–757.

BAG, S., MOGURAMPELLY, S., GODDARD III, W. A. & MAITI, P. K. (2016). Dramatic changes in DNA conductance with stretching: structural polymorphism at a critical extension. *Nanoscale* **8**, 16044–16052.

BALAEFF, A., CRAIG, S. L. & BERATAN, D. N. (2011). B-DNA to Zip-DNA: simulating a DNA transition to a novel structure with enhanced charge-transport characteristics. *Journal of Physical Chemistry A* **115**, 9377–9391.

BERTUCAT, G., LAVERY, R. & PRÉVOST, C. (1998). A model for parallel triple helix formation by RecA: single-strand association with a homologous duplex via the minor groove. *Journal of Biomolecular Structure and Dynamics* **16**, 535–546.

BIEBRICHER, A. S., HELLER, I., ROIJMANS, R. F. H., HOEKSTRA, T. P., PETERMAN, E. J. G. & WUITE, G. J. L. (2015). The impact of DNA intercalators on DNA and DNA-processing enzymes elucidated through force-dependent binding kinetics. *Nature Communications* **6**, 7304–7316.

BOSAEUS, N., EL-SAGHEER, A. H., BROWN, T., SMITH, S. B., ÅKERMAN, B., BUSTAMANTE, C. & NORDÉN, B. (2012). Tension induces a base-paired overstretched DNA conformation. *Proceedings of the National Academy of Sciences* **109**, 15179–15184.

BOSAEUS, N., REYMER, A., BEKE-SOMFAI, T., BROWN, T., TAKAHASHI, M., WITTUNG-STAFSHEDE, P., ROCHA, S. & NORDÉN, B. (2017). A stretched conformation of DNA with a biological role? *Quarterly Reviews of Biophysics* **50**, e11, 1–11.

CASE, D. A., BETZ, R. M., CERUTTI, D. S., CHEATHAM III, T. E., DARDEN, T. A., DUKE, R. E., GIESE, T. J., GOHLKE, H., GOETZ, A. W., HOMEYER, N., IZADI, S., JANOWSKI, P., KAUS, J., KOVALENKO, A., LEE, T. S., LEGRAND, S., LI, P., LIN, C., LUCHKO, T., LUO, R., MADEJ, B., MERMELSTEIN, D., MERZ, K. M., MONARD, G., NGUYEN, H., NGUYEN, H. T., OMELYAN, I., ONUFRIEV, A., ROE, D. R., ROITBERG, A., SAGUI, C., SIMMERLING, C. L., BOTELLO-SMITH, W. M., SWAILS, J., WALKER, R. C., WANG, J., WOLF, R. M., WU, X., XIAO, L. & KOLLMAN, P. A. (2016). *AMBER 2016*. San Francisco: University of California.

CHEN, A. A. & GARCÍA, A. E. (2013). High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **110**, 16820–16825.

CHEN, Z., YANG, H. & PAVLETICH, N. P. (2008). Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature* **453**, 489–494.

CLUZEL, P., LEBRUN, A., HELLER, C. & LAVERY, R. (1996). DNA: an extensible molecule. *Science* **271**, 792–794.

11

Danilowicz, C., Limouse, C., Hatch, K., Conover, A., Coljee, V. W., Kleckner, N. & Prentiss, M. (2009). 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**, 13196–13201.

Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald: an N.Log(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **98**, 10089–10092.

Friedman, R. A. & Honig, B. (1995). A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophysical Journal* **69**, 1528–1535.

Fu, H., Chen, H., Marko, J. F. & Yan, J. (2010). Two distinct overstretched DNA states. *Nucleic Acids Research* **38**, 5594–5600.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725.

Ikehara, K., Omori, Y., Arai, R. & Hirose, A. (2002). A novel theory on the origin of the genetic code: a GNC-SNS hypothesis. *Journal of Molecular Evolution* **54**, 530–538.

Itzkovitz, S. & Alon, U. (2007). The genetic code is nearly optimal for allowing additional information within protein coding sequences. *Genome Research* **17**, 405–412.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935.

Joung, I. S. & Cheatham III, T. E. (2008). Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry B* **112**, 9020–9041.

Jukes, T. H. (1973). Arginine as an evolutionary intruder into protein synthesis. *Biochemical and Biophysical Research Communications* **53**, 709–714.

Kilchherr, F., Wachauf, C., Pelz, B., Rief, M., Zacharias, M. & Dietz, H. (2016). Single-molecule dissection of stacking forces in DNA. *Science* **353**, aaf5508–aaf5508-9.

King, G. A., Gross, P., Bockelmann, U., Modesti, M., Wuite, G. J. L. & Peterman, E. J. G. (2013). 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**, 3859–3864.

Konrad, M. W. & Bolonick, J. I. (1996). 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**, 10989–10994.

Koonin, E. V. & Novozhilov, A. S. (2009). Origin and evolution of the genetic code: the universal enigma. *International Union of Biochemistry and Molecular Biology Life* **61**, 99–111.

Kuzminov, A. (1999). Recombinational repair of DNA damage in Escherichia coli and bacteriophage lambda. *Microbiology and Molecular Biology Reviews* **63**, 751–813.

Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D. & Zakrzewska, K. (2009). Curves+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Research* **37**, 5917–5929.

Lebrun, A. & Lavery, R. (1996). Modelling extreme stretching of DNA. *Nucleic Acids Research* **24**, 2260–2267.

Lemkul, J. A. & MacKerell Jr., A. D. (2017). 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**, 2053–2071.

Li, H. & Gisler, T. (2009). 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**, 325–332.

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

Maaloum, M., Beker, A. F., & Muller, P. (2011). Secondary structure of double-stranded DNA under stretching: elucidation of the stretched form. *Physical Review* **E 83**, 031903–031909.

Macke, T. & Case, D. A. (1998). Modeling unusual nucleic acid structures. In *Molecular Modeling of Nucleic Acids* (eds. N. B. Leontes and J. Santalucia Jr.), pp. 379–393. Washington, DC: American Chemical Society.

Nicklas, R. B. (1988). The forces that move chromosomes in mitosis. *Annual Review of Biophysics and Biophysical Chemistry* **17**, 431–449.

Nordén, B., Elvingson, C., Kubista, M., Sjöberg, B., Ryberg, H., Ryberg, M., Mortensen, K. & Takahashi, M. (1992). 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**, 1175–1191.

Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A. & Orozco, M. (2007). Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical Journal* **92**, 3817–3829.

Pieters, J. M., Mans, R. M., Van Den Elst, H., Van Der Marel, G. A., Van Boom, J. H. & Caltona, C. (1989). 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**, 4551–4565.

Protozanova, E., Yakovchuk, P., & Frank-kamenetskii, M. D. (2004). Stacked-unstacked equilibrium at the nick site of DNA. *Journal of Molecular Biology* **342**, 775–785.

Ratner, S. & Petrack, B. (1953). The mechanism of arginine synthesis from citrulline in kidney. *Journal of Biological Chemistry* **200**, 175–185.

Rief, M., Clausen-Schaumann, H. & Gaub, H. E. (1999). Sequence-dependent mechanics of single DNA molecules. *Nature Structural & Molecular Biology* **6**, 346–349.

Roe, D. R. & Chaka, A. M. (2009). Structural basis of pathway-dependent force profiles in stretched DNA. *Journal of Physical Chemistry B* **113**, 15364–15371.

Roe, D. R. & Cheatham III, T. E. (2013). PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation* **9**, 3084–3095.

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

Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with

constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327–341.

SALOMON-FERRER, R., GÖTZ, A. W., POOLE, D., LE GRAND, S. & WALKER, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **9**, 3878–3888.

SMITH, S., FINZI, L. & BUSTAMANTE, C. (1992). Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126.

SMITH, S. B., CUI, Y. & BUSTAMENTE, C. (1996). Over-stretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271**, 795–799.

SNOWDEN-IFFT, E. A. & WEMMER, D. E. (1990). Characterization of the structure and melting of DNAs containing backbone nicks and gaps. *Biochemistry* **29**, 6017–6025.

VAN MAMEREN, J., GROSS, P., FARGE, G., HOOIJMAN, P., MODESTI, M., FALKENBERG, M., WUITE, G. J. L. & PETERMAN, E. J. G. (2009). Unraveling the structure of DNA during over-stretching by using multicolor, single-molecule fluorescence imaging. *Proceedings of the National Academy of Sciences* **106**, 18231–18236.

VARRETTE, S., BOUVRY, P., CARTIAUX, H. & GEORGATOS, F. (2014). Management of an academic hpc cluster: The UL experience. In *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*, pp. 959–967. Bologna, Italy: IEEE.

VLADESCU, I. D., MCCAULEY, M. J., NUÑEZ, M. E., ROUZINA, I. & WILLIAMS, M. C. (2007). Quantifying force-dependent and zero-force DNA intercalation by single-molecule stretching. *Nature Methods* **4**, 517–522.

VLASSAKIS, J., WILLIAMS, J., HATCH, K., DANILOWICZ, C., COLJEE, V. W. & PRENTISS, M. (2008). Probing the mechanical stability of DNA in the presence of monovalent cations. *Journal of the American Chemical Society* **130**, 5004–5005.

WANG, J., WANG, W., KOLLMAN, P. A. & CASE, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **25**, 247–260.

WANG, J., WOLF, R. M., CALDWELL, J. W., KOLLMAN, P. A. & CASE, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry* **25**, 1157–1174.

WILLIAMS, M. C., ROUZINA, I. & BLOOMFIELD, V. A. (2002). Thermodynamics of DNA interactions from single molecule stretching experiments. *Accounts of Chemical Research* **35**, 159–166.

WILLIAMS, M. C., WENNER, J. R., ROUZINA, I. & BLOOMFIELD, V. A. (2001). Effect of pH on the overstretching transition of double-stranded DNA: evidence of force-induced DNA melting. *Biophysical Journal* **80**, 874–881.

WONG, J. T. (1975). A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences* **72**, 1909–1912.

WONG, J. T. (2005). Coevolution theory of the genetic code at age thirty. *BioEssays* **27**, 416–425.

WONG, J. T. F. & BRONSKILL, P. M. (1979). Inadequacy of prebiotic synthesis as origin of proteinous amino acids. *Journal of Molecular Evolution* **13**, 115–125.

13