Quarterly Reviews of Biophysics

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

--Manuscript Draft--

Manuscript Number:	QRBP-D-17-00009
Full Title:	DNA partitions into triplets under tension in the presence of organic cations, with sequence evolutionary age predicting the stability of the triplet phase
Short Title:	Triplet Phase of Ancient DNA Sequences
Article Type:	Report
Corresponding Author:	Joshua Berryman LUXEMBOURG
Corresponding Author's Institution:	
First Author:	Amirhossein Taghavi
Order of Authors:	Amirhossein Taghavi
	Paul van der Schoot
	Joshua Berryman
Abstract:	Using atomistic simulations of GC-rich DNA duplexes we show the formation of stable triplet structure when 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 solution phase of the double helix under tension, subject to 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 emphasize the structure-function relationship in Σ DNA. We supplement atomistic calculations with thermodynamic arguments to show that codons for 'phase one' 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 one set, showing a series of decreasing triplet propensity with evolutionary newness.
Keywords:	dna stretching; S-DNA; sigma DNA; triplet disproportionation; evolution; origin of life

Title page

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

Authors' names:

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

Correspondence:

Joshua T. Berryman,

Address : Campus Limpertsberg, Université du Luxembourg 162 A, avenue de la Faïencerie L-1511 Luxembourg

Telephone (+352) 46 66 44 **6971**

Fax (+352) 46 66 44 **36971**

E-mail: josh.berryman@uni.lu

Word count: 4033

Abstract

Using atomistic simulations of GC-rich DNA duplexes we show the formation of stable triplet structure when 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 solution phase of the double helix under tension, subject to 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 emphasize the structure-function relationship in Σ DNA. We supplement atomistic calculations with thermodynamic arguments to show that codons for 'phase one' 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 one set, showing a series of decreasing triplet propensity with evolutionary newness.

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 pN (Williams *et al*., 2002; Vlassakis *et al*., 2008; Liu *et al*., 2010; Bosaeus *et al*., 2012). 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 (Cluzel *et al*., 1996; Konrad and Bolonick, 1996; Lebrun and Lavery, 1996a; Smith *et al*., 1996). 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 ~ 5.8 Å, compared to the rise in unstretched B-DNA of 3.4 Å.

This stretched phase or phases is known by the umbrella label of ‘S-DNA’. For GC-rich structures having strong hydrogen bonding the base pairing is preserved in the 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. Reorientation 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. The detailed S-DNA structure, particularly the inclination, depends on the pulling scheme. When the 5´ ends of each strand are pulled, tilt angle increases gradually until the terminal H-bonds are disrupted, while in the 3´3´ pulling regime the tilt angle is decreased and no early breakage of H-bonds is seen (Lavery *et al*., 2002; Li and Gisler, 2009; Danilowicz *et al*., 2009; Bag *et al*., 2016).

The most readily available information on DNA under tension is the empirically measured force-extension curve (Smith *et al*., 1992; Rief *et al*., 1999), which provides the clear signal of some transition, but no atomistic-level information. This is supplemented by fluorescence and polarised-light studies (Nordén *et al.*, 1992; van Mameren *et al.*, 2009; 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 (Lebrun and Lavery, 1996b; Konrad and Bolonick, 1996; Li and Gisler, 2009). Simple energy minimisation of d(GCG)$_4$ DNA under extension (without thermal fluctuations or explicit solvent) yields partition into four base-stacked triplets (Bertucat *et al.*, 1998), however subsequent fully dynamic simulations have shown instead the irregular formation of 'denaturation bubbles' (Harris *et al.*, 2005; Rezác *et al.*, 2010), 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.

DNA is often subjected to tension in its biological context, for purposes including transport, transcription and tertiary structure manipulation (Nicklas, 1998) . 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 the fundamental process of sexual reproduction: the recombination of homologous DNA from two parent organisms. In this 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 crystallogrphy structure of RecA-DNA complex (Chen *et al.*, 2008) and also has been suggested as a stable phase even without co-factors (Bosaeus *et al.*, current work).

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 vs. 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.*, 2008). 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 real (millisecond) 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 ones 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 zero (the GADV world) (Ikehara *et al.,* 2002), phase one and phase two (Wong, 1975; Wong, 2005;Koonin and Novozhilov, 2009), 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.

METHODS

Molecular structures were prepared using the Nucleic Acid Builder (NAB) (Macke and Case, 1998). Salt and water were represented using the Joung-Cheatham (Joung and Cheatham III, 2008) and TIP3P (Jorgensen *et al*., 1983) parameters, with the AMBER15 forcefield (Ivani *et al*., 2015) used for DNA and the AMBER14SB forcefield used for peptides (Maier *et al*., 2015). 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 (Götz *et al*., 2012) in the AMBER16 package (Case *et al*., 2016). For each calculation, 16 independent replicas were prepared and equilibrated in the B conformation for 10 ns. Pulling of the DNA then took place using steered molecular dynamics, over a time period of 150 ns (giving a pulling rate of 0.68 Å ns^{-1}). 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. 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 for 24 bp was estimated giving values of 3.46 Å (Arg Cl), 3.43 Å (EtBr) and 3.43 Å (NaCl) for the [$G_{12}C_{12}$] 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.

RESULTS

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 metabolic networks leading to the different amino acids: it is hypothesised that a list of so-called ‘phase one’ 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) (Wong and Bronskill, 1979; Koonin and Novozhilov, 2009). If the genetic code in the epoch of a much simplified amino-acid alphabet already had the current structure of three basepairs per codon it was therefore highly redundant at this time.

In the current triplet code, the ‘phase one’ amino acids supposed to have been incorporated earliest into biology (a list of ADEGILPSTV) 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 one 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.

The free energy of base-stacking in duplex DNA was long ago calculated combinatorically for the various pairs of bases, by Friedman and Honig (Friedman and Honig, 1995). Although free energy calculations for nucleic acids remain subtle and difficult two decades after this initial work, as the results are in accordance with chemical intuition we can be confident in the ranking of the different pairs, and the absolute values are in any case less important for the following discussion. From this tabulated data we can see that the weakest step is CG (-4.36 kcal/mol), and the strongest is GG (-7.79). 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 remains CG-CG (-8.72 kcal/mol) and the strongest is GG-CC, with an energy of -12.3 kcal/mol.

Based on the stacking energy of complementary pairs, it is possible to arrive at the energetic cost to separate a given codon from its neighbours as a triplet of stacked base-pairs. If we consider the duplex xGGCy-pGCCq (coding for GLY on strand 1, ALA on the complementary strand, where x,y,p,q are bases from the adjacent codons), then the stacks needed to find the triplet disproportionation energy G_τ are xG, Cy, pG, Cq. In order to select x, y we randomly choose two amino acids from the set of phase one residues ADEGILPSTV, and then randomly choose a codon for the given amino acid subject to the constraint that if a codon beginning in G or ending in C (or both) is available in the genetic code, this codon is preferred. The bases p, q are then selected as complementary to x, y. After sampling 10^6 amino acid pairs, it is then possible to tabulate the average energy to partition into a triplet associated with a given codon (G_τ).

If we assume that the DNA is subject to tension such that it must partition in some way, and that the partitions must be somewhat evenly spread (here we arbitrarily assume two breaks per five base-pairs) then we can present a relative free energy ΔG_τ by comparing against the alternative pairs of sites at which to break base-stacking. Combinatorics gives 1/2 ($N^2 – N$)

such site combinations for a stretch of N steps, where N is 1 less than the number of base pairs. Here 1/2 (N^2 - N) = 6, leaving 5 site pairs for step breakage not including the pair which defines a 'Σ' triplet. To get a probability, the comparison should be to a Boltzmann-weighted sum of all alternative energies, i.e.:

$$p(\tau) = \frac{e^{-G_\tau/k_B T}}{e^{-G_\tau/k_B T} + \sum_{i=1..5} e^{-G_i/k_B T}}$$

Tabulating this information for the 20 canonical amino acids, we can see a clear pattern of reduced triplet disproportionation energy for the primordial 'stage one' amino acids (Table 1).

Table 1.

The dramatic pattern evident in the tabulated partitioning energies is that stage one 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 and Bronskill's 1979 tabulation of stage one amino acids (Wong and Bronskill, 1979), possibly due to the large energetic cost needed to synthesize it from citrulline in modern organisms (Ratner and Petrack, 1953).

It has been advanced 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 (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 list ADEGILPSTV contains no basic amino acids at all making the addition of Ornithine seem valuable in order to form 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 a similar moderately bulky basic residues.

The weaker statement of this work, that the stage one part of the genetic code is structured so as to support a 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 and 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 2 shows that codons for phase one 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 2) are also those which partition most naturally into triplets.

Table 2.

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 1 and 2) 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 4 glycines and 4 prolines) and a sequence chosen to show strong triplet disproportionation based on Table 1, d[$(GGC)_4(GAC)_4$], 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). 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 1 and on the RecA structure (Chen *et al*., 2008), simulations were run both in NaCl and in a solution of Ac-Arg-NHMe Cl, with the capped Arginine molecule replacing sodium as the positive counterion.

We find that for the GC-rich sequence 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. 1 shows a regular pattern of vertical bases 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 *order1* $k_B T$ free energies of disproportionation in Table 1).

Fig. 1

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.

Fig. 2

Extending beyond approximately 3 Å/bp 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 average base-pair inclination in the high entropy sequence d[(GCC)$_4$(GAC)$_4$] in the presence and absence of intercalators up to the extension point of 1.5 follow the same pattern and remain flat (Fig. 3b,d,f) which indicates base-pair perpendicularity with respect to the helix axis (Nordén *et al.*, 1978; Edmondson and Johnson, 1986, Bosaeus *et al.*, 2012). In the presence of intercalators this trend continues after extension point 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. 3a,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. 3e). 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.* Current work) as suggesting Σ formation even in free solution.

Fig. 3

DISCUSSION

DNA behaviour under tension is affected by factors like 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 characterized by a diameter of 1.2 nm (Maaloum *et al*., 2011). X-ray diffractions of stretched cross-link 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 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 separate a given codon from its neighbours as a triplet of stacked base-pairs, and found 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 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 Ethidium Bromide, 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 (i.e. Arginine or Ethidium) 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 experiments (Balaeff *et al.*, 2011; Bosaeus *et al.* Current work) and simulations (Konrad *et al.* 1996).

Ikehara and co-workers have shown that codons matching the pattern GNC (where N signifies "anything") probably constituted the original, minimal functional 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 disproportionation, allowing recombination and possibly protein synthesis to operate without the sophisticated enzymatic machinery which exists today. We hypothesize that a bootstrap process took place, with crude triplet disproportionation facilitating recombination, driving accelerated evolution and leading to more sophisticated protein-DNA interactions, in turn allowing expansion in stages of the genetic code.

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 evolutionary explosion.

Acknowledgements

This project was financially supported by the University of Luxembourg. Computer simulations presented in this paper were carried out using the HPC facility of the University of Luxembourg (Varrette et al., 2014). We thank Bengt Nordén for considerable guidance, criticism and inspiration.

Financial support

This work was supported by the University of Luxembourg (grant number R-AGR-0136).

Conflict of interest

“None”

References

BAG, S., MOGURAMPELLY, S., GODDARD III, W. A., AND MAITI, P. K. (2016). Dramatic changes in DNA conductance with stretching: structural polymorphism at a critical extension. *Nanoscale*, 8(35):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(34), 9377–9391.

BERTUCAT, G., LAVERY, R., AND 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(3), 535–546.

BOSAEUS, N., EL-SAGHEER, A. H., BROWN, T., SMITH, S. B., ÅKERMAN, B., BUSTAMANTE, C., AND NORDÉN, B. (2012). Tension induces a base-paired overstretched DNA conformation. *Proceedings of the National Academy of Sciences*, 109(38), 15179–15184.

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., A. KOVALENKO, 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. AND KOLLMAN P. A. (2016), AMBER 2016, University of California, San Francisco.

CHEN, Z., YANG, H., AND PAVLETICH, N. P. (2008). Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature*, 453(7194):489–494.

CLUZEL, P., LEBRUN, A., HELLER, C., LAVERY, R. (1996). DNA: an extensible molecule. *Science*, 271(5250):792.

DANILOWICZ, C., LIMOUSE, C., HATCH, K., CONOVER, A., COLJEE, V. W., KLECKNER, N., AND 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(32), 13196–13201.

EDMONDSON, S. P., JOHNSON JR, W. C. (1986). 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, 2335–2348.

FRIEDMAN, R. A. AND HONIG, B. (1995). A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophysical Journal*, 69(4):1528–1535.

FU, H., CHEN, H., MARKO, J. F., YAN, J. (2010) Two distinct overstretched DNA states. *Nucleic Acids Research*, 38(16), 5594-600.

GÖTZ, A. W., WILLIAMSON, M. J., XU, D., POOLE, D., LE GRAND, S., AND WALKER, R. C. (2012). Routine microsecond molecular dynamics simulations with amber on gpus. 1. generalized born. *Journal of Chemical Theory and Computation*, 8(5), 1542–1555.

HARRIS, S. A., SANDS, Z. A., AND LAUGHTON, C. A. (2005). Molecular dynamics simulations of duplex stretching reveal the importance of entropy in determining the biomechanical properties of DNA. *Biophysical Journal*, 88(3), 1684–1691.

IKEHARA, K., OMORI, Y., ARAI, R., AKIKO 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. AND ALON, U. (2007). The genetic code is nearly optimal for allowing additional information within protein coding sequences. *Genome Research*, 17(4), 405–412.

IVANI, I.; DANS, P. D.; NOY, A.; PÉREZ, A.; FAUSTINO, I.; HOSPITAL, A.; WALTHER, J.; ANDRIO, P.; GOÑI, R.; BALACEANU, A.; PORTELLA, G.; BATTISTINI, F.; GELPÍ, J. L.; GONZÁLEZ, C.; VENDRUSCOLO, M.; LAUGHTON, C. A.; HARRIS, S. A.; CASE, D. A.; OROZCO, M. Parmbsc1: A Refined Force Field for DNA Simulations. *Nature Methods* 2015, 13, 55–58.

JORGENSEN, W. L., CHANDRASEKHAR, J., MADURA, J. D., IMPEY, R. W., AND KLEIN, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2), 926–935.

JOUNG, I. S. AND 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(30), 9020–9041.

JUKES, T. H. (1973) Arginine as an evolutionary intruder into protein synthesis. *Biochemical and Biophysical Research Communications*. 53(3):709-14.

KING, G. A., GROSS, P., BOCKELMANN, U., MODESTI, M., WUITE, G. J. L., AND 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(10), 3859–3864.

KONRAD, M. W. AND 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(45):10989–10994.

KOONIN, E. V. AND NOVOZHILOV, A. S. (2009). Origin and evolution of the genetic code: the universal enigma. *International Union of Biochemistry and Molecular Biology Life*, 61(2):99–111.

LAVERY, R., LEBRUN, A., ALLEMAND, J.F., BENSIMON, D., AND CROQUETTE, V. (2002). Structure and mechanics of single biomolecules: experiment and simulation. *Journal of Physics: Condensed Matter*, 14(14):R383.

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. AND LAVERY, R. (1996). Modelling extreme stretching of DNA. *Nucleic Acids Research*, 24(12):2260–2267.

LI, H. AND 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(3), 325–332.

LIU, N., BU, T., SONG, Y., ZHANG, W., LI, J., ZHANG, W., SHEN, J., AND LI, H. (2010). The nature of the force-induced conformation transition of dsDNA studied by using single molecule force spectroscopy. *Langmuir*, 26(12), 9491–9496.

MACKE, T. AND CASE. D. A. Modeling unusual nucleic acid structures. In Molecular Modeling of Nucleic Acids, N.B. LEONTES AND J. SANTALUCIA, JR., eds. (Washington, DC: American Chemical Society, 1998), pp. 379-393.

JAMES A. MAIER, J. A., MARTINEZ, C., KASAVAJHALA, K., WICKSTROM, L., HAUSER, K. E., AND SIMMERLING C. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation. 11* (8), pp 3696–3713.

NORDÉN, B., SETH, S., TJERNELD, F. (1978) Renaturation of DNA in ethanol-methanol solvent induced by complexation with methyl green. *Biopolymers*, 17, 523–525.

NORDÉN, B., ELVINGSON, C., KUBISTA, M., SJÖBERG, B., RYBERG, H., RYBERG, M., MORTENSEN, K., AND 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(4), 1175–1191.

NICKLAS, R. B. (1988). The forces that move chromosomes in mitosis. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):431–449.

RATNER, S. AND PETRACK, B. (1953). The mechanism of Arginine synthesis from citrulline in kidney. *Journal of Biological Chemistry*, 200(1):175–185.

ŘEZÁČ, J., HOBZA, P., AND HARRIS, S. A. (2010). Stretched DNA investigated using molecular-dynamics and quantum-mechanical calculations. *Biophysical journal*, 98(1), 101–110.

RIEF, M., CLAUSEN-SCHAUMANN, H., AND GAUB, H. E. (1999). Sequence-dependent mechanics of single DNA molecules. *Nature Structural & Molecular Biology*, 6(4), 346–349.

SMITH, S., FINZI, L., AND BUSTAMANTE, C. (1992). Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science*, 258(5085), 1122–1126.

SMITH, S. B., CUI, Y., AND BUSTAMENTE, C. (1996). Over-stretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science*, 271(5250), 795.

VAN MAMEREN, J., GROSS, P., FARGE, G., HOOIJMAN, P., MODESTI, M., FALKENBERG, M., WUITE, G. J. L., AND 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(43), 18231–18236.

VARRETTE, S., BOUVRY, P., CARTIAUX, H., AND 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), pages 959– 967, Bologna, Italy. IEEE.

VLADESCU, I. D., MCCAULEY, M. J., ROUZINA, I., WILLIAMS, M. C. (2005). Mapping the phase diagram of single DNA molecule force-induced melting in the presence of ethidium. *Physical Review Letters*. 95(15), 158102.

VLASSAKIS, J., WILLIAMS, J., HATCH, K., DANILOWICZ, C., COLJEE, V. W., AND PRENTISS, M. (2008). Probing the mechanical stability of DNA in the presence of monovalent cations. *Journal of the American Chemical Society*, 130(15), 5004–5005.

WANG, J., WANG, W., KOLLMAN, P. A., AND CASE, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2), 247–260.

WANG, J., WOLF, R. M., CALDWELL, J. W., KOLLMAN, P. A., AND CASE, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174.

WILLIAMS, M. C., ROUZINA, I., AND BLOOMFIELD, V. A. (2002). Thermodynamics of DNA interactions from single molecule stretching experiments. *Accounts of Chemical Research*, 35(3), 159–166.

WILLIAMS, M. C., WENNER, J. R., ROUZINA, I., AND BLOOMFIELD, V. A. (2001). Effect of pH on the overstretching transition of double-stranded DNA: evidence of force-induced DNA melting. *Biophysical Journal*, 80(2), 874–881.

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

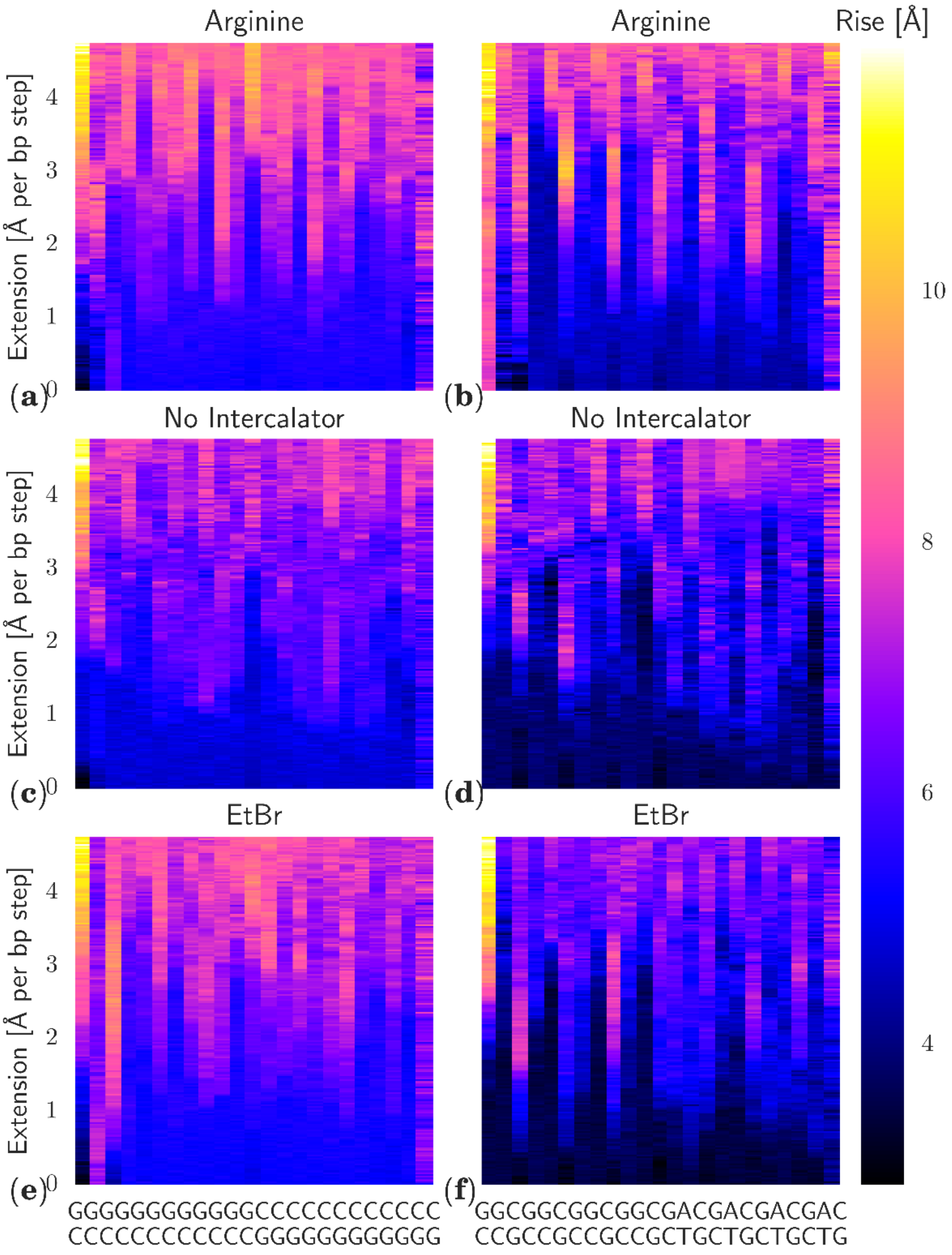
WONG, J.T.F. (1975). A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences*, 72(5), 1909-1912.

WONG, J.T.F. AND BRONSKILL, P. M. (1979). Inadequacy of prebiotic synthesis as origin of proteinous amino acids. *Journal of Molecular Evolution*, 13(2), 115–125.

Amirhossein Taghavi, Paul van der Schoot and Joshua T. Berryman

DNA partitions into triplets under tension….

Figure 1.

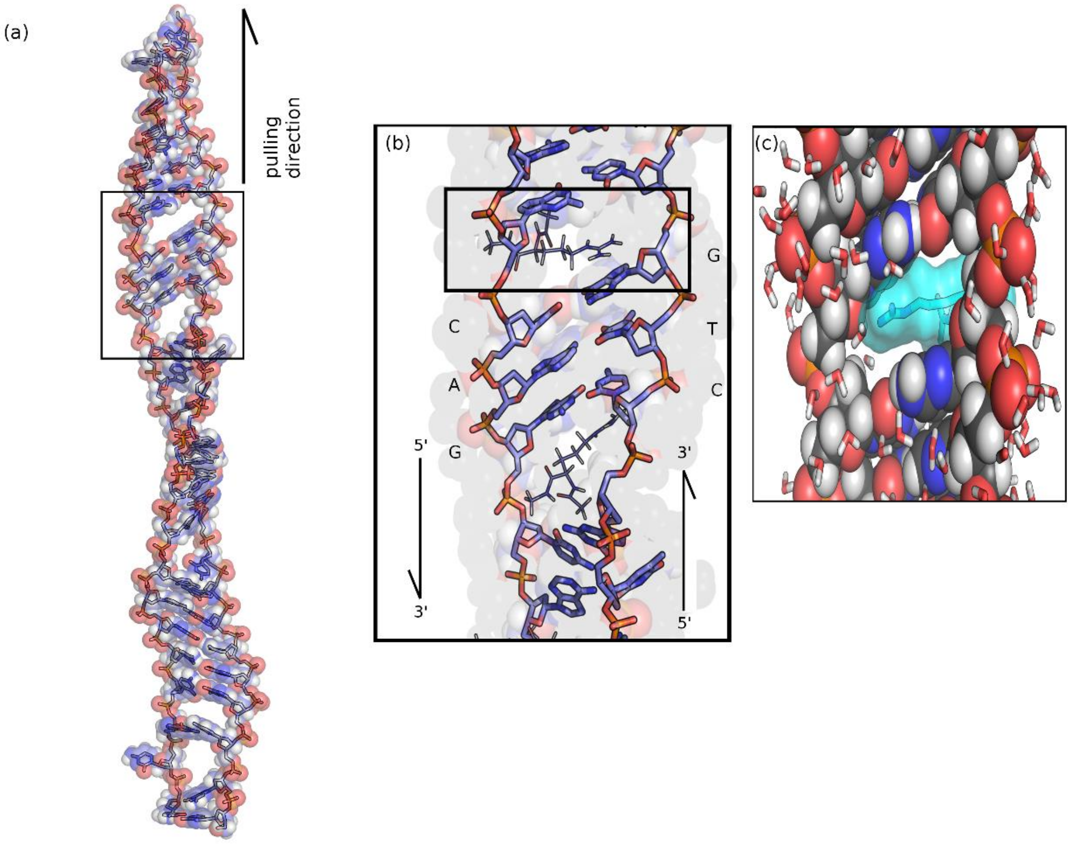


Triplet phase of ancient DNA sequence

Amirhossein Taghavi, Paul van der Schoot and Joshua T. Berryman

DNA partitions into triplets under tension….

Figure 2.



Triplet phase of ancient DNA sequence

Amirhossein Taghavi, Paul van der Schoot and Joshua T. Berryman

DNA partitions into triplets under tension….

Figure 3.

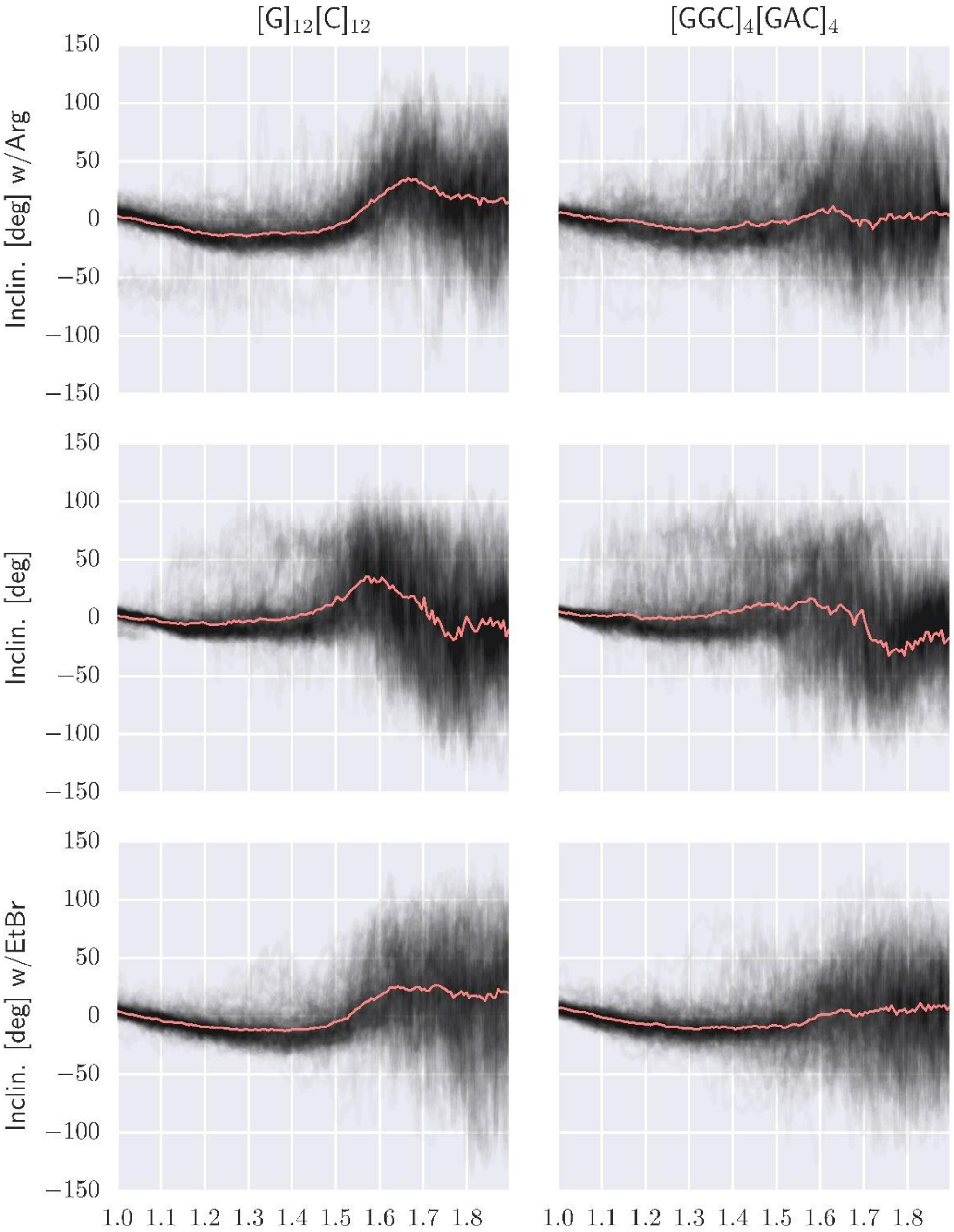


Table I.

AA	Codon	ΔG	p(τ)
G♀	GGC	-1.500	0.4211
A♀	GCC	-1.500	0.4211
S*	AGC	-0.024	0.4127
V♀	GTC	-0.909	0.2735
D♀	GAC	-0.909	0.2735
T*	ACC	-0.435	0.2657
N.	AAC	-0.433	0.2165
R.	AGA	-0.419	0.2719
P*	CCC	-0.177	0.1900
I*	ATC	-0.143	0.1881
E*	GAG	0.022	0.1425
L*	CTC	0.022	0.1425
K.	AAA	0.232	0.0750
F.	TTT	0.232	0.0750
Y.	TAT	1.252	0.0072
X.	TAG	1.716	0.0017
C.	TGT	1.832	0.0028
M.	ATG	2.302	0.0007
H.	CAT	2.302	0.0007
Q.	CAG	2.505	0.0002

Table II.

		Base 3			
Base 1	Base 2	T	C	A	G
T	T	F.	F.	L*	L*
	C	S*	S*	S*	S*
	A	Y.	Y.	X.	X.
	G	C.	C.	X.	Y.
C	T	L*	L*	L*	L*
	C	P*	P*	P*	P*
	A	H.	H.	Q.	Q.
	G	R.	R.	R.	R.
A	T	I*	I*	I*	M.
	C	T*	T*	T*	T*
	A	N.	N.	K.	K.
	G	S*	S*	R.	R.
G	T	V♀	V♀	V♀	V♀
	C	A♀	A♀	A♀	A♀
	A	D♀	D♀	E*	E*
	G	G♀	G♀	G♀	G♀

Amirhossein Taghavi, Paul van der Schoot and Joshua T. Berryman

DNA partitions into triplets under tension….

Figure 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 homogenous sequence (**a**) or presence of CG steps in the absence of Arginine (**d**) induce only weakly structured disproportionation.

Figure 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 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 an axial view of the highlighted cavity in (**b**).

Figure 3.

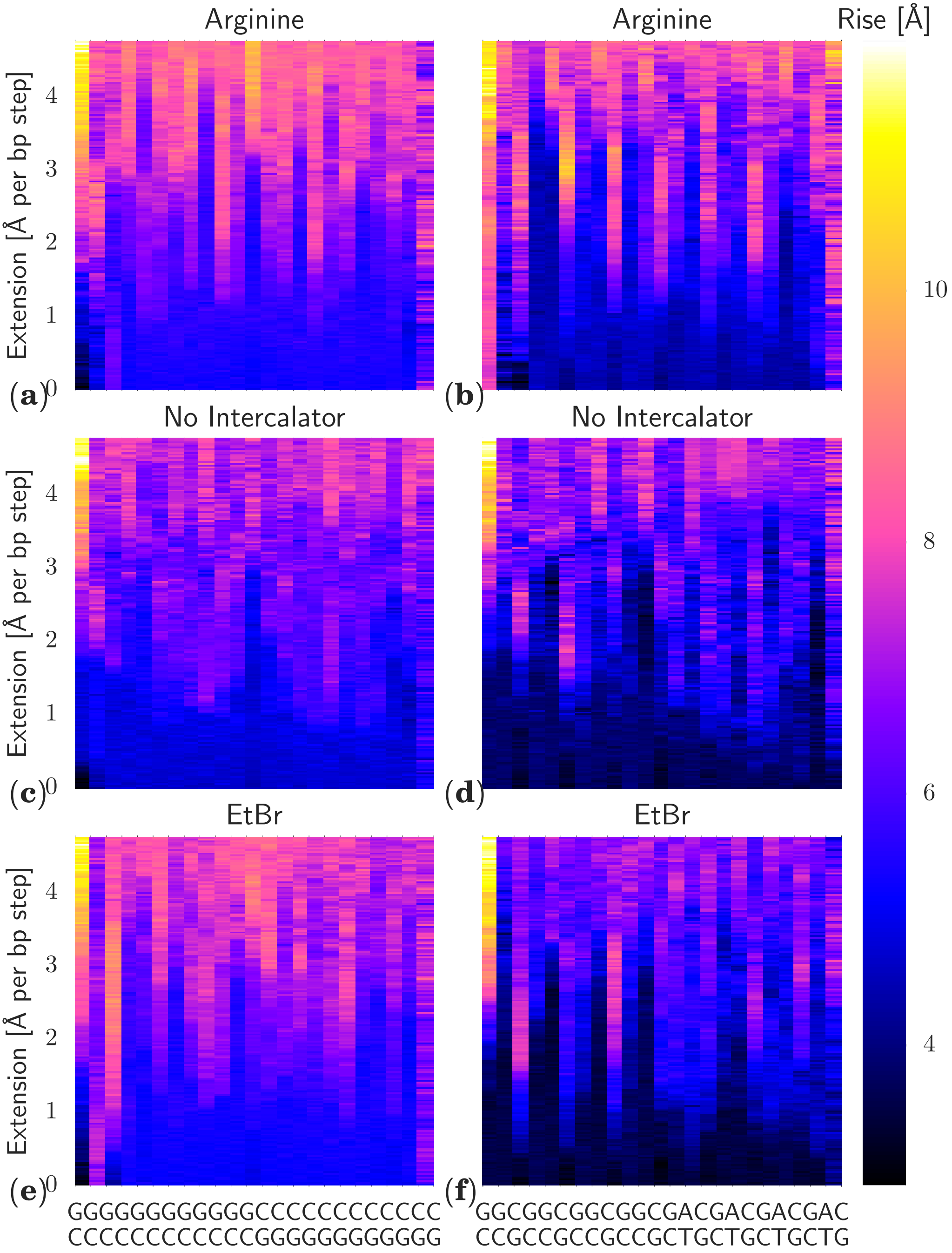
Average inclination of the low and high entropy sequences in the presence and absence of intercalators (Arginine and EtBr). Average inclination for the high-entropy sequence d[(GGC)$_4$(GAC)$_4$] remains relatively flat up to extension 1.5 and beyond, even without intercalant (**b, d, f**). For the low entropy sequence d[G$_{12}$C$_{12}$] average inclination remains flat up to extension 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**).

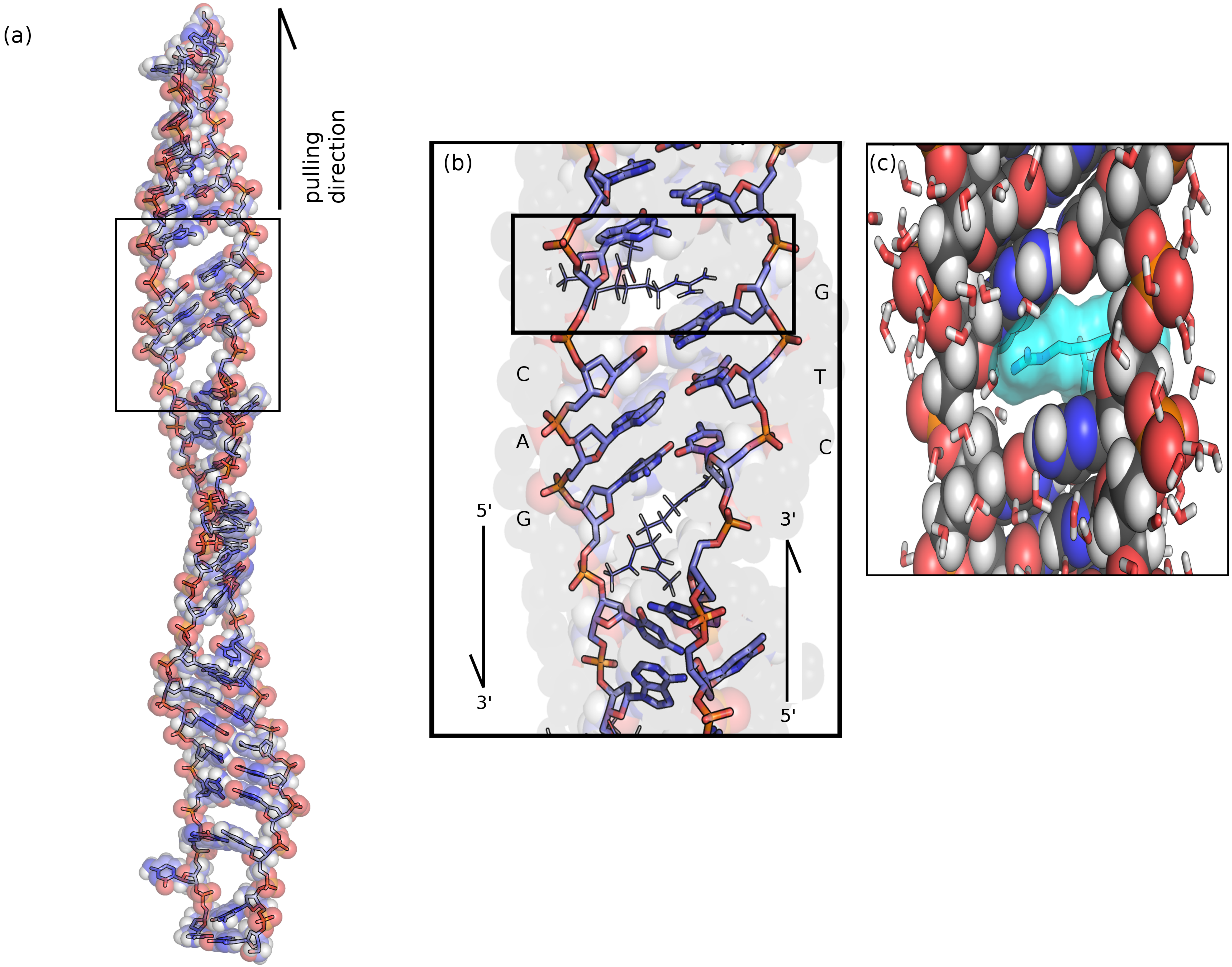
Table 1.

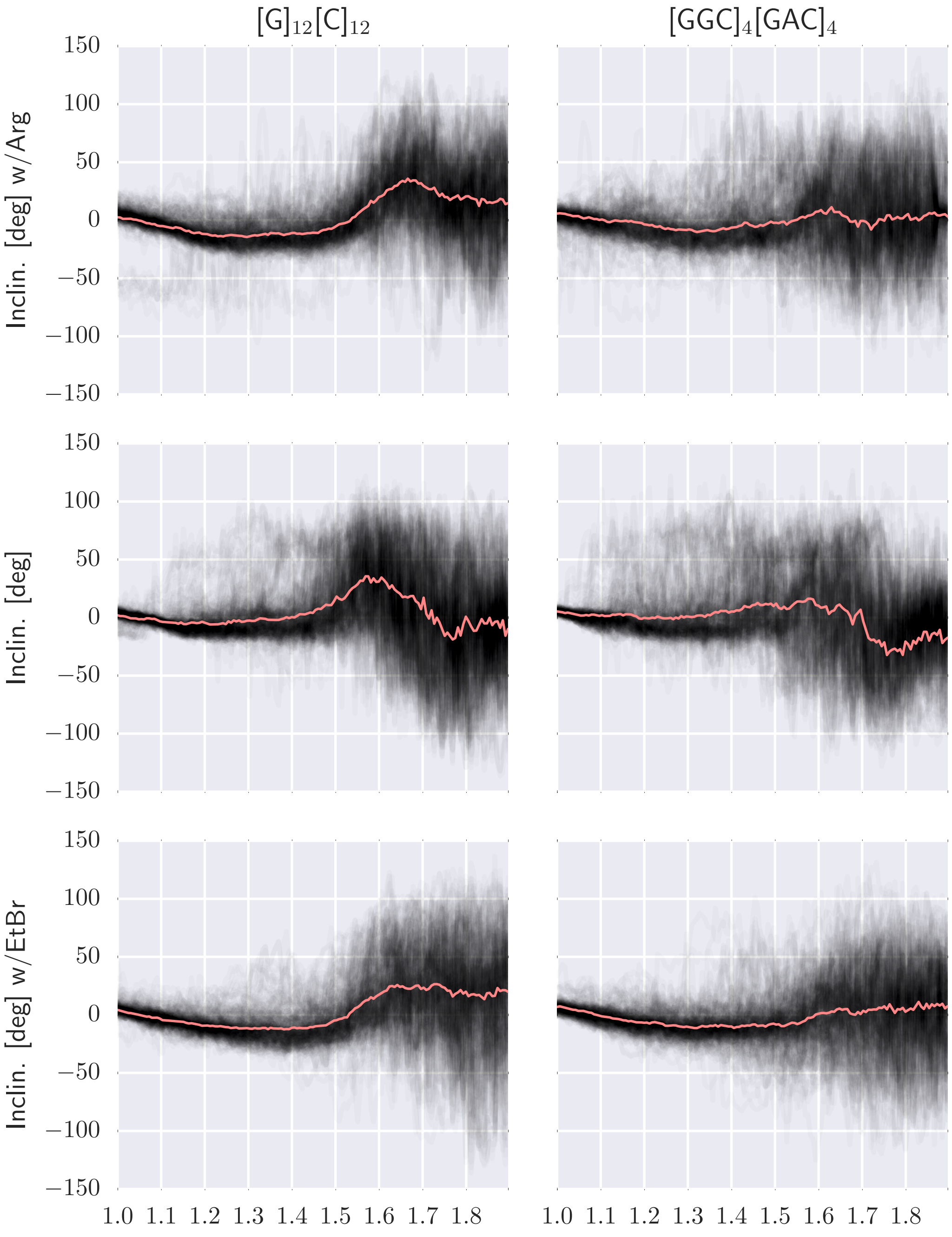
Phase one amino acids (*,♀) tend to have (at least one) codon associated with them that partitions favourably at its boundaries. The most favourable partitioning is for the phase zero (DAGV) amino acids (♀). X indicates a stop codon, other letters are standard amino acid abbreviations. Energy units are *kcal/mol*.

Table 2.

All codons, colored by $p(\tau)$ with a light color indicating higher p. 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 (eg. GGC,GCC) necessarily have the same $p(\tau)$. ♀, * indicates phase one amino acids, ♀ indicates phase zero.


Arginine
Arginine
Rise [Å]
No Intercalator
No Intercalator
EtBr
EtBr
Extension [Å per bp step]
Extension [Å per bp step]
Extension [Å per bp step]
0
1
2
3
4
(a)
(b)
(c)
(d)
(e)
(f)
10
8
6
4
GGGGGGGGGGGCCCCCCCCCCC
CCCCCCCCCCCGGGGGGGGGGG
GGCGGCGGCGGCGACGACGACGAC
CCGCCGCCGCCGCTGCTGCTGCTG

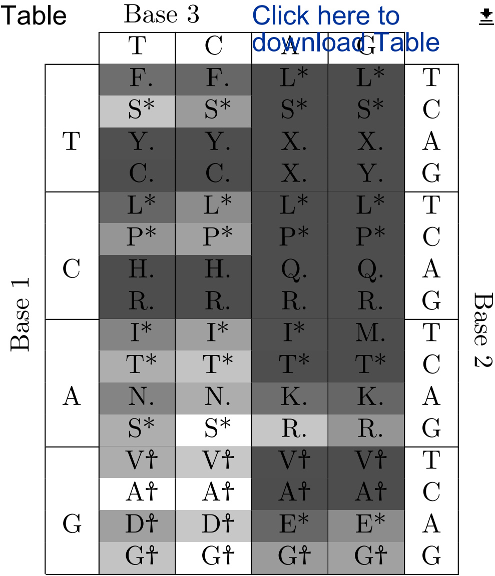

(a)
pulling direction
(b)
G
C
T
A
C
5'
G
3'
3'
5'
(c)


$[G]_{12}[C]_{12}$
$[GGC]_4[GAC]_4$
Inclin. [deg] w/Arg
Inclin. [deg]
Inclin. [deg] w/EtBr
150
100
50
0
−50
−100
−150
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8

AA	Codon	ΔG	$p(\tau_{\ddagger})$
G☥	GGC	-1.500	0.4211
A☥	GCC	-1.500	0.4211
S*	AGC	-1.024	0.4127
V☥	GTC	-0.909	0.2735
D☥	GAC	-0.909	0.2735
T*	ACC	-0.435	0.2657
N.	AAC	-0.433	0.2165
R.	AGA	-0.419	0.2719
P*	CCC	-0.177	0.1900
I*	ATC	-0.143	0.1881
E*	GAG	0.022	0.1425
L*	CTC	0.022	0.1425
K.	AAA	0.232	0.0750
F.	TTT	0.232	0.0750
Y.	TAT	1.252	0.0072
X.	TAG	1.716	0.0017
C.	TGT	1.832	0.0028
M.	ATG	2.302	0.0007
H.	CAT	2.302	0.0007
Q.	CAG	2.505	0.0002

Table

Click here to download Table



Base 1	Base 2	Base 3: T	C	A	G
T	T	F.	F.	L*	L*
	C	S*	S*	S*	S*
	A	Y.	Y.	X.	X.
	G	C.	C.	X.	Y.
C	T	L*	L*	L*	L*
	C	P*	P*	P*	P*
	A	H.	H.	Q.	Q.
	G	R.	R.	R.	R.
A	T	I*	I*	I*	M.
	C	T*	T*	T*	T*
	A	N.	N.	K.	K.
	G	S*	S*	R.	R.
G	T	V☥	V☥	V☥	V☥
	C	A☥	A☥	A☥	A☥
	A	D☥	D☥	E*	E*
	G	G☥	G☥	G☥	G☥