Chem

CellPress

Article

Structural Basis of the Mispairing of an Artificially Expanded Genetic Information System

Linus F. Reichenbach,[1,4] Ahmad Ahmad Sobri,[2,4] Nathan R. Zaccai,[2,4] Christopher Agnew,[2] Nicholas Burton,[2] Lucy P. Eperon,[3] Sara de Ornellas,[1] Ian C. Eperon,[3*] R. Leo. Brady[2*] & Glenn A. Burley[1,5*]

[1] Dr L. F. Reichenbach, Dr S. de Ornellas, Dr G. A. Burley Department of Pure and Applied Chemistry, University of Strathclyde. Thomas Graham Building, 295 Cathedral Street, Glasgow, G1 1XL, United Kingdom. E-mail:

[2] Dr C. Agnew, Mr A. A. Sobri, Mr N. Burton, Dr N. R. Zaccai, Prof. R.L. Brady, School of Biochemistry, University of Bristol, Bristol, BS8 1TD, United Kingdom. E-mail:

[3] Dr L. P. Eperon, Prof. I. C. Eperon, Leicester Institute of Structural & Chemical Biology and Department of Molecular & Cellular Biology, University of Leicester. Henry Wellcome Building, Lancaster Road, Leicester, LE1 7RH, United Kingdom.

[4] These authors contributed equally.

[5] Lead Contact

* Correspondence: I.C.E. eci@le.ac.uk R.L.B:l.brady@bristol.ac.uk G.A.B. glenn.burley@strath.ac.uk

SUMMARY

The synthetic nucleotide P pairs with Z within DNA duplexes by a unique hydrogen-bond arrangement relative to naturally occurring Watson-Crick base-pairs. The loss of this synthetic genetic information by PCR results in the conversion of P-Z into a G-C base-pair. Here we show structural and spectroscopic evidence that the loss of this synthetic genetic information occurs *via* G-Z mispairing. Remarkably, the G-Z mispair is both plastic and pH-dependent, forming a two-hydrogen bonded 'slipped' pair at pH 7.8 and a three-hydrogen bonded Z-G pair when the pH is above 7.8. This study highlights the need for robust structural and functional methods to understand the mechanisms of mutation when developing next generation synthetic genetic base-pairs.

INTRODUCTION

Nucleic acids are the fundamental repository of genetic information found in living organisms. Unique to this family of biomacromolecules is the ability of DNA to act as a template for the replication and storage of genetic information, with levels of fidelity approaching a single-base substitution every 10^6 nucleotides for proof-reading-deficient DNA polymerases.[1] Maintaining the genetic integrity (replication) and high-fidelity information transfer (transcription and translation) of the primary sequence of nucleic acid molecules is contingent on the discrimination of two hydrogen-bond regimes: pairing of A with T *via* two hydrogen bonds and G with C (Fig. 1a) *via* three hydrogen bonds.[2,3] Base-pairing that deviates from conventional Watson-Crick profiles is a source of mutation and, subsequently, adaptive evolution.

The development of an expanded genetic repertoire might have many uses, but it would require the design of synthetic base-pairs that pair orthogonally and exclusively with each other and not with naturally-occurring nucleotides.[4-11] The Hirao and Romesberg groups have developed synthetic base-pairs which rely on complementarity of shape rather than hydrogen-bonding to pair preferentially with each other. Both systems enable replication fidelities approaching that of natural Watson-Crick base-pairs.[4,5,10,12-19] A particular mutational hallmark of these hydrophobic base-pairs is the gradual conversion of the synthetic base-pair into an A-T pair observed both *in cellulo* and *via* repeated rounds of PCR although the structural basis for the loss of synthetic information is at present not known.[16,18]

An alternative design approach is an artificially expanded synthetic genetic information system (AEGIS) where synthetic base-pairs present a unique arrangement of hydrogen bonds relative to Watson-Crick base-pairs.[20-22] The most recent incarnation of this strategy is the P-Z base-pair (Fig. 1b).[20; 23-25] This design archetype has been used in applications ranging from aptamer development through to their development as primers in DNA diagnostics applications.[26-28] A recent structural study of an oligodeoxyribonucleotide (ODN) duplex containing four P-Z base-pairs reveals a structure that adopts a conventional B-type conformation when in complex with a host protein (*N*-terminal fragment of Moloney murine leukemia virus reverse transcriptase, MMLV-RT). However, an A-type duplex was formed in complex with MMLV-RT using ODN duplex containing six consecutive P-Z base-pairs,[23; 25] which suggests a level of structural plasticity, although it is unclear whether this is due to the sequence context of the ODN and/or host protein binding.

Another unique feature of the P-Z base-pair is the loss of this synthetic genetic information occurs *via* conversion into a G-C base-pair under PCR amplification conditions.[21; 22] The putative mechanism for this loss is that the conjugate base of the Z nucleoside (*i.e.*, deprotonation of N_1-H of Z, pK_a ~ 7.8)[21] might be an anionic mimic of C, resulting in the competitive incorporation of dGMP relative to dPMP opposite a Z-containing DNA template.[20; 21] Critical for future applications of synthetic genetic systems is the need for a comprehensive understanding of the molecular determinants of mispairing of synthetic nucleotides with naturally occurring nucleotides when incorporated into duplex DNA. Using a combination of PCR analysis of base pair fidelity, X-ray crystallography, UV-vis thermal melts and CD spectroscopy, we show that G-Z mispairing is remarkably plastic; pairing *via* a mixture of canonical (*i.e.,* three hydrogen bonds) and non-canonical (*i.e.,* two hydrogen bonds) regimes that is aligned with the pK_a of the Z N_1-H (Fig. 1c).[22]

RESULTS

Experimental Design

This work had two principal aims: *(i)* to determine the structural and thermodynamic differences between a 'matched' P-Z pair and 'mismatched' G-Z pair as a function of the pK_a of the Z N_1-H; and *(ii)* to understand if the fidelity and amplification of P-Z information is influenced by pH. The experimental design that was used to address our first aim focused on obtaining crystal structures of dodecamer ODN duplexes containing P-Z or G-Z base-pairs in the absence of a host protein. This was an important consideration as host protein binding could alter the pairing profile particularly with short oligonucleotides complicating their validity and interpretation. The archetypical Dickerson dodecamer $d(CGCGAATTCGCG)_2$ (ODN1) for which a crystal structure has previously been determined in the absence of host protein and presence of Ca^{2+} (PDB: 463D)[29] was used as a comparison for our structural and spectroscopic studies as this self-complementary sequence forms B-type DNA duplexes with natural as well as modified nucleotides.[30-34] The $d(CGC\mathbf{P}AATT\mathbf{Z}GCG)_2$ sequence (ODN2) was prepared by solid phase DNA synthesis using established protocols and consists of two P-Z base-pairs in a self-complementary duplex in positions 4 and 9.[22] The dodecamer sequence $d(CGC\mathbf{\underline{G}}AATT\mathbf{Z}GCG)_2$ (ODN3), prepared using the same methodology, contained a double G-Z mismatch in the same positions. Finally, the structural and thermodynamic properties of ODN1-3 were compared to a naturally-occurring duplex containing a double G-T mismatch in positions 4 and 9 (ODN4, PDB: 113D).[35] All four ODNs formed well-defined B-type duplexes. Unlike ODN4, which was solved in the presence of Mg^{2+} [22], ODN2-3 did not readily crystallise at pH < 7.8 in the presence or absence of that cation. ODN2-3 were therefore solved in the presence of Ca^{2+}, at pH 7 (for ODN2), at pH 7.8 (for ODN3a) and at pH 8.5 (for ODN3b). The previously published ODN1 structure (at pH 7, and in the presence of Ca^{2+}) [29] had similar crystal packing (indexed as space group R3 for ODN1 and H3 for ODN2 and ODN3 which are equivalent).

Crystal structural analysis of DNA duplexes containing P-Z and G-Z base-pairs

Within the crystal lattices, the duplex structures of ODN2-3 are asymmetric leading to non-identical conformations for each of the P-Z (ODN2) or G-Z (ODN3) pairs. An overlay of ODN1-3 (Fig. 2a) indicates broad similarity with no substantial distortion away from the B-type conformation as a consequence of the insertion of the artificial bases in both ODN2 and ODN3 (main chain RMSD < 0.6 Å). A three-hydrogen bonded P-Z pairing was observed in ODN2 with hydrogen bond lengths of 2.7-2.8 Å, which is comparable to G-C hydrogen bond lengths observed elsewhere in this 2.2 Å resolution structure (Fig. 2b-c). The hydrogen bonding distances were also equivalent to those reported in an earlier P-Z containing structure;[23] and suggests that the adjacent bases do not affect the P-Z pairing (Table S4). The inclusion of two P-Z pairs in ODN2 is accompanied by a slight widening of the major groove by approximately 1 Å. In both ODN2 and ODN3a-b, the Z-NO_2 group is closely planar to the remainder of the aromatic ring system. Consequently, one of the Z-NO_2 oxygens is positioned where it could establish an intramolecular hydrogen bond with the exocyclic Z-NH_2 (N=O···H-N bond length 2.7 Å), thus forming a pseudo bicyclic ring system.[23, 25] Since the Z-NO_2 group projects into the major groove, this additional steric bulk might contribute to the slight widening of the groove observed in ODN2 and other Z-P sequences.

Geometric analysis of ODN3a (pH 7.8) suggests distortion relative to ODN1 is greatest around the G-Z mismatch pairs (Table S5). For example, the G-Z pair located at position 4 in ODN3a has propeller twists of approximately 10° greater than those observed in either of the ODN1 or ODN2 structures, and similarly about 3° increases in buckle angles. This substantially reduces the co-planarity of the base pairs and alters their interconnecting hydrogen bonds. G-Z pairing in ODN3a also exhibits shearing of up to 1.5 Å which is significantly greater than the shearing observed for P-Z pairs (0.3 Å in this study) in ODN2, although the analysis is partially limited by the lower resolution of the G-Z pH 7.8 structure (2.5 Å).

There is a striking difference in the G-Z base-pairing profile between the structures at pH 7.8 (ODN3a, Fig. 2c) and pH 8.5 (ODN3b, Fig. 2d). The observed electron density indicates that the G-Z in position 4 of ODN3a is present as a slipped or 'wobble-like' pair containing two hydrogen bonds: (G)O6-N1(Z) at 3.0 Å, and (G)N1-O2(Z) at 3.1 Å (Fig. 2d). In this arrangement, both the ring N-1 atoms of G and Z must be protonated. This pairing is reminiscent of G-T mispairing pairing observed in ODN4, however the distortion of the base pair is less pronounced (Table 1). In contrast, as with the G-C and Z-P interactions in ODN1 and ODN2 respectively, the G-Z base-pair present in position 9 of ODN3a forms three hydrogen-bonds in a pseudo G-C Watson-Crick arrangement: (G)N1-N1(Z) 3.1 Å, (G)O6-N4(Z) 3.1 Å, and (G)N2-O2(Z) 3.2 Å. It is likely that the anionic form of Z is present as exemplified by the longer bond lengths relative to those of the nearest natural base pairs (2.8-3.0 Å). Taken collectively, our crystal structural analysis has shown that G-Z pairing is highly pH-dependent.

Since the ODN3a crystals were grown at a pH corresponding to the pK_a of the N_1-H of Z (*i.e.,* 7.8)[22] where a mixed population of the neutral and anionic forms of Z is possible, we conclude that Z can pair with G in two different hydrogen-bond arrangements. When crystals of ODN3b were grown at pH 8.5, only a single type of triple hydrogen-bonded arrangement was observed (Fig. 2e), which is consistent with both Z nucleotides being present in their deprotonated form. Although these observed structures are consistent with the expected protonation states of Z based on its pKa measured as a free base, we cannot rule out that there might be a shift of the pK_a when incorporated in B-type DNA.[36, 37]

Duplex stability of G-Z pairing is highly pH-dependent

UV-vis melting and Circular Dichroism (CD) experiments were then conducted to further explore how the pH dependence of G-Z mispairing impacts the thermal stability and structure of ODN duplexes in solution. While ODN1 and ODN4 exhibited virtually no pH dependence by UV-vis melt, the P-Z-containing ODN2 exhibited slight (2 °C) destabilization as the pH was raised from 5.5 to 7.5 (Fig. 3a). In contrast, the duplex melt of ODN3 was highly pH-dependent and resulted in a 12 °C stabilization as the pH was increased from 5.5 to 8.0 (Table S1). No further increase in the duplex melt was observed

above 8.0, which indicates that the thermal stability of duplex ODN3 is aligned with the pK_a of the N-1 of Z. There is also a significant difference in the duplex melt of ODN3 and ODN4, which contains two G-T mispairs. No pH dependence was observed for ODN4, which is significantly destabilized relative ODN1 and ODN3. All four ODNs exhibited a characteristic right-handed B-type duplex according to CD (*i.e.*, negative peak around 245 nm and positive peak between 260 and 280 nm)[38] at both pH 5.5 and 7.5 using a sodium cacodylate buffer(Fig. 3b). Taken collectively, UV-vis melt and CD show that G-Z mispairing is both highly pH dependent and significantly more stable than a corresponding G-T mispair. These results align with the pairing plasticity seen in the crystal structures of ODN1-3.

Analysis of maintenance of Z-P base-pairs during PCR

We set out to explore the capacity of the Z-P information to be maintained in the absence of the corresponding modified triphosphate (*i.e.*, dZTP when P is present in the DNA template or dPTP when Z is present in the DNA template). Taq polymerase was chosen as this polymerase has been shown previously to incorporate dZTP and dPTP[26]. However, extensive analyses by PCR, using a plasmid template and primers containing Z or P, closely following published conditions[39], showed that the other nucleotides readily misincorporated during PCR (Figure S35/36).

To test whether the propensity to lose P-Z information during PCR amplification might arise from the pH-dependent ionization state of Z, we analysed the pH-dependency of PCR using a range of P-Z DNA templates. To maintain consistency with our UV-melt data, PCR was then conducted using the same cacodylate buffer composition from pH 5.5 to 7.5. Again there was no evidence that amplification was reduced in the absence of the cognate dPTP and dZTP (Fig. 4). To maximize the possibility of observing a pH-dependent effect we investigated the PCR amplification of a strand containing three consecutive P-Z base-pairs. PCR amplification was unaffected by the omission of the complementary dPTP and dZTP (Fig. S37). We conclude that AEGIS nucleotides do not allow faithful propagation of a template containing P or Z nucleotides across a pH range used in PCR.

DISCUSSION

These experiments were designed to determine the structural, spectroscopic and functional impact of pH on P-Z "matched" pairing versus G-Z "mismatched" pairing. We discuss three major conclusions that emerged from our results.

(i) A G-Z mismatch leads to significant local distortion within a B-form duplex at neutral pH - The crystal structures of ODN2-3 clearly illustrate the ability of Z-containing ODNs to pair both with artificial P and natural G nucleotides whilst maintaining an overall B-form duplex (Fig. 2). An unexpected observation is the ODN3 structure obtained at pH 7.8 captures both two (Wobble-like) and three (pseudo Watson-Crick) hydrogen-bond arrangements of the G-Z mispair, highlighting the pairing plasticity of Z mispairing with G at physiologically-relevant pH. We infer that our difficulty to obtain reproducible diffraction-quality crystals of ODN3 at a pH < 7.8 is likely due to the instability of the two hydrogen-bonded G-Z slipped pair to promote well-ordered crystallization. This is supported by our UV-vis thermal melt data, which show a strong stabilization effect under slightly basic conditions (Fig. 3a). It is notable that significant accompanying geometric distortions are limited to the G-Z base pairings and their immediately adjacent bases. Further, the ODN3 structure at pH 7.8 implies that a transition between the twin and triple hydrogen-bonded pairing arrangements can be readily accessed without disruption of the adjacent duplex structure. These observations imply that G-Z mismatches would be readily tolerated in thermal cycling conditions used in PCR amplification.[20; 22] One explanation of the results might be that the effective local pH in the active site of the enzyme is closer to the optimal pH (~8.3) for the reaction.

(ii) Stabilization of the Z anion is the source of mis-pairing - UV-melt experiments confirm that the stability of the G-Z mispair is pH-dependent; *i.e.*, contingent on a negative charge forming on the Z nucleobase. The Z-NO_2 group and the O2 carbonyl provide extensive conjugation of the resultant anion, which could bias the negative charge either within

interior of the duplex (Z N-1$^-$)[40] or projected into the major (Z NO_2^-) or minor (Z O-2$^-$) groove as depicted schematically in Fig. 5. Furthermore, the Z NH_2 group in the 6-position could provide additional stabilization of this anion *via* an intramolecular hydrogen bond with the Z NO_2 in the 5-position. A similar intramolecular hydrogen bonding profile of a negatively-charged nucleobase has also been observed with 5-carboxycytosine where the carboxylate group also projects into the major groove and forms an intramolecular hydrogen-bond with the adjacent exocyclic amine.[41] Thus, the combined structural characteristics of a strongly electron-withdrawing Z NO_2 group increasing the acidity of the Z N_1-H, extensive stabilization of the anion by resonance and the possible presence of an intramolecular hydrogen bond between the Z NO_2 and the exocyclic Z NH_2 provide a unique synthetic nucleotide where base-pairing stability can be tuned within a physiologically-relevant window.[42]

(iii) G-Z mis-pairing compromises orthogonality of P-Z pairing - One of the most important applications of synthetic genetic systems is in PCR, where the orthogonal nucleotides might be introduced into selected sites and, after amplification, used as the basis for incorporation of modified versions that would provide site-specific probes for fluorescence or cross-linking in either DNA or RNA transcripts.[43-45] For these applications, complete fidelity in practical terms is an essential requirement, since PCR products or transcripts containing P or Z could not be readily purified in preparative experiments from contaminants containing G or C. We set up a series of stringent tests of the ability of the P-Z pair to be faithfully propagated during PCR. Various P/Z-containing templates were amplified in the absence of the corresponding modified triphosphate (*i.e.,* dPTP or dZTP). We observed the preparation of full-length PCR amplicons over a pH range below the pK_a of Z (*i.e.,* where Z should be in the fully protonated state), with no indication of significantly reduced efficiency. We therefore conclude that sufficiently faithful propagation of P-Z information is not possible with the current design.

In summary, our study illustrates the complexities designing synthetic base-pairs directed by subtle differences in hydrogen-bonding. The incorporation of the nitro group in the Z nucleotide design was originally introduced in order to reduce the rate of C1′ epimerization.[21] However, this increases the acidity of the N^1-H proton, which contributes to mispairing with G. This subtle interplay highlights the need for new AEGIS designs that are both chemically stable and maintain high pairing fidelity. Furthermore, this study highlights the need for further structural and biochemical studies to be conducted on understanding the mechanisms of the loss of synthetic genetic information. We envisage that this work will assist in the design process of new AEGIS and other synthetic base-pairs that have a reduced propensity to mispair with natural nucleotides.

EXPERIMENTAL PROCDURES

Synthesis and Purification of P-Z-containing ODN2-3

ODN2-3 were synthesized using standard solid phase oligonucleotide synthesis protocols on an ABI 394 synthesizer.[39] Phosphoramidites and CPG supports loaded with standard nucleosides were purchased from LINK Technologies Ltd (Bellshill, UK). AEGIS phosphoramidites (dZ and dP) were purchased from Firebird Biomolecular Sciences, LLC (Alachua, Florida, USA).

Deprotection Protocol for ODN 2-3

DBU (1 M in MeCN, 3 mL/mmol) was added to Z-containing ODNs immobilised on a CPG (1 mmol) support. The suspension was shaken overnight at room temperature to remove the NPE protecting group from the Z nucleobase. The DBU solution was removed and the CPG support was washed once with MeCN. Ammonia (DNA grade, 3 ml/mmol) was added and the suspension was shaken for three hours at 55 °C. The yellow supernatant was removed after three hours. The CPG support was then washed with ammonia (1 mL/mmol) and the combined layers concentrated under reduced pressure to obtain crude ODN2-3 as a yellow solution that was purified by reverse-phase HPLC on a Dionex UltiMate 3000 System. ODN1 and ODN 4 were purchased from Eurofins (Wolverhampton, UK) and Eurogentec (Southampton, UK).

Analytical Data

MALDI-ToF mass spectra were recorded using a Shimadzu Biotech Axima CFR spectrometer, using 3-hydroxypicolinic acid (HPA) as the matrix. All mass spectra were recorded in negative mode (Fig S1-4). Analytical RP-HPLC was performed on a Dionex UltiMate 3000 system using a Clarity 5 µM Oligo-RP column (250 x 4.6 mm) in a triethylammonium acetate (TEAA) Buffer system. (Buffer A: 100 mM TEAA pH 7.5 in water. Buffer B 100 mM TEAA pH 7.5 in 80% MeCN. Flow rate: 1 mL/min. Gradient as described in Table S1) For ODN1 and ODN2 spectra were obtained at a column temperature of 80 °C to minimize peaks caused by secondary structures (duplex formation) (Fig S5-8).

ODN1: MS: $[M-H]^-$ calcd. 3645.4 found 3645.4. – HPLC: retention time 9.7 min (80 °C).
ODN2: MS:$[M-H]^-$ calcd. 3689.4 found 3689.5. – HPLC: retention time 10.3 min (80 °C).
ODN3: $[M-H]^-$ calcd. 3689.4 found 3689.2. – HPLC: retention time 12.1 min (25°C).
ODN4: $[M-H]^-$ calcd. 3661.4 found 3661.3. – HPLC: retention time 12.1 min (25°C).

Thermal UV-Vis Measurements

Cacodylate Buffer (10 mM sodium cacodylate, 10 mM KCl, 10 mM $MgCl_2$, 5 mM $CaCl_2$) was used as the buffer system. The different pH values were adjusted using 1 M HCl and 1 M NaOH. 40-50 mM stock solutions of ODN1-3 were diluted to 4 µM in 10 µM Cacodylate buffer (10 mM sodium cacodylate, 10 mM KCl, 10 mM $MgCl_2$, 5 mM $CaCl_2$) previously set to the desired pH. UV measurements were obtained using a Shimadzu UV-1800 with an 8-series micro multi cell, each cell containing a sample volume of 100 µL. The UV spectra were obtained over a range from 20 to 90 °C for ODN1 and ODN2 and 20 to 70 °C for ODN3 using a ramp speed of 0.5 °C per minute and measurement intervals of 0.2 °C. The measured wavelength was 260 nm. All measurements were performed at least 4 times to obtain an average value (Table S2-3, Fig. S9-28).

Circular Dichroism

For CD analysis, 40-50 mM stock solutions of ODN 1-4 were diluted to 10 µM in 10 µM Cacodylate buffer (10 mM sodium cacodylate, 10 mM KCl, 10 mM $MgCl_2$, 5 mM $CaCl_2$) set to either pH 5.5 or 7.5, the same conditions as for the UV spectroscopy. CD spectra were obtained on a Jasco J-810 instrument at 25 °C, a rate of 20 nm/min and a wavelength increment of 0.2 nm (Fig. 3B). In addition ellipticity (mdeg) was recorded for ODN 1-3 between 320 and 202 nm in intervals of 5 °C over a range from 25 °C to 85 °C at pH 5.5 and 7.5. All spectra were corrected against elliptic readings obtained by the buffer (10 mM sodium cacodylate, 10 mM KCl, 10 mM $MgCl_2$, 5 mM $CaCl_2$) (Fig. S29-34).

Crystallization and Data Collection

Thin, platelike crystals of ODN2 duplex $(CGCPAATTZGCG)_2$ were obtained through sitting drop vapour diffusion by mixing 200 nL of 0.5 mM ODN2 in water with 200 nL precipitant solution, containing 20 mM sodium cacodylate, pH 7.0 and 600 mM calcium chloride. The drop was equilibrated against a reservoir containing 50% hexylene glycol at 298 K. The crystals were flash-frozen in liquid nitrogen and diffracted to 2.17 Å. These crystals were indexed in space group H3 with unit cell $a = b = 41.7$ Å and $c = 99.9$ Å, $\alpha = \beta = 90°$ and $\gamma = 120°$, containing two strands of DNA. Diffraction data were collected on beamline I04-1 at Diamond Light Source, UK and processed using MOSFLM [46].

Small crystals of duplex ODN3 $d(CGCGAATTZGCG)_2$ were obtained using the same method by mixing 350nL of 1 mM ODN3 in 100mM NaCl, 20mM Tris, pH8.5 with 350nL precipitant solution, containing 0.1M Calcium acetate pH 7.8 and 32% hexylene glycol (MPD). The crystals also belonged to space group H3 with unit cell $a = b = 41.5$ Å and $c = 101.7$, $\alpha = \beta = 90°$ and $\gamma = 120°$, containing two strands of DNA, and diffracted to 2.46 Å. Diffraction data were collected on beamline I02 at Diamond Light Source, UK and processed using MOSFLM.[46] Crystals of ODN3 were also grown at pH 8.5 using the same method and data collected on beamline IO3 and processed as above. The crystals had space group H3 with unit cell $a = b = 40.9$ Å and $c = 102.2$ Å, $\alpha = \beta = 90°$ and $\gamma = 120°$, containing two strands of DNA, and diffracted to 2.35 Å (Table S4).

Structure solution and refinement

All three crystal structures were solved with the CCP4 package [47] by molecular replacement using the PHASER program [48] with native DNA oligomer (pdb code 463D [29]) as the starting model. The structures were refined with iterative cycles of manual rebuilding with the COOT program [49] followed by refinement with Refmac5. The P and Z nucleotides were unambiguously identified in the electron density using positive density for the nitro group specific for this non-natural base. The crystallographic data is summarized in Table 1 and the models and structure factors have been deposited in the PDB (accession codes: ODN2 5LJ4, ODN3a 5l4S and ODN3b 5KTV). While not outliers, the Rfree values for two of these structures compare less favourably with averages determined across the entire (mainly protein) PDB. This arises partially from the lack of clear refinement restraints for the artificial base pairs. However, the structures compare very favourably with DNA-only structures averaged across the Nucleic Acid Database (http://ndbserver.rutgers.edu). The final structures were analysed using the 3DNA software[50] and figures were drawn with the PYMOL program[51] (Table S5).

Polymerase Chain Reaction (PCR)

Initial PCR studies were conducted in MOPS (pH 7.34) or Tris (pH 8.30) with 50 ng of GloC plasmid DNA template,[52] primers containing P and Z nucleotides as well as the corresponding reverse primer at 0.5 and 0.3 µM respectively. In the experiments with primer Z concentrations of the dNTP were 0.5 mM dA, dG, dT and 0.6 mM for dC and dPTP. In the experiments with Primer-P the concentrations of the dNP were 0.1 mM dA, dG, dT, dC and 0.6 mM for dZTP. Each PCR contained 0.25 and 0.3 units REDTaq polymerase (Sigma) respectively. Amplification was for the indicated number of cycles. Samples were analysed by electrophoresis on a 1.5% agarose gel and visualized by staining with ethidium bromide (Fig. S35-16).

Further PCR studies (Figs. 4 and S17) were done in 10 mM sodium cacodylate (pH adjusted with HCl), 50 mM KCl, 10 mM $MgCl_2$, 5 mM $CaCl_2$, 0.1 mM DTT and 0.1 mg/ml BSA. Reactions were done in 20 µl with 40 fmol. of template DNA (ODNs 3P-Temp and 3Z-Temp, and Bsp-Z and Bsp-P [20]) and primers at 0.5 µM. All dNTPs were at 0.25 mM in the experiment shown, but the concentrations and ratios were varied in other experiments. Each PCR contained 0.5 units REDTaq polymerase (Sigma). Amplification was for 30 cycles. Samples were analysed by electrophoresis on a 3% agarose gel, staining with ethidium bromide and quantitative imaging on a Typhoon imager. The control lanes, lacking either template or amplification, were blank (Figs. 4 and S37). Sequences for all the Z and P containing templates and primers are described in Table S6.

ACCESSION CODES

The supplemental crystallography data reported in the paper have been deposited in the RSCB Protein Data Bank under accession numbers: 5LJ4 (ODN2), 5l4S (ODN3a), 5KTV (ODN3b). These data are freely available via http://www.rcsb.org/pdb/home/home.do

SUPPLEMENTARY INFORMATION

Supplemental Information includes figures for MALDI-ToF and HPLC analysis, tables containing full data from the UV-vis measurements, including further figures for comparison, thermal UV-vis analysis of ODN1-3, Data collection for X-ray crystallographic analysis and figures for the PCR amplification of Z-P DNA templates as well as a table detailing all the ZP containing template and primers used.

AUTHOR CONTRIBUTIONS

Conceptualization, N.R.Z., E.C.I., R.L.B. and G.A.B.; Methodology, L.F.R., N.R.Z., C.A., L.P.E., S.D.; Investigation, L.F.R., A.A.S., C.A., N.B., L.P.E., S.O.; Writing – Original Draft,.L.F.R., N.R.Z., I.C.E., R.L.B., G.A.B.; Writing – Review & Editing, L.F.R., N.R.Z., I.C.E., R.L.B., G.A.B.; Funding Acquisition, I.C.E., R.L.B., G.A.B.; Resources, I.C.E., R.L.B., G.A.B.; Supervision, I.C.E., R.L.B., G.A.B.

ACKNOWLEDGMENTS

This work was supported by the BBSRC (BB/J020087/1 to R.L.B. and G.A.B. BB/J02080X/1 to G.A.B. and E.C.I.) and the Leverhulme Trust (RPG-2014-001 to E.C.I. and G.A.B.). We thank Dr Sharon Kelly (University of Glasgow) for measurement of the CD spectra. The authors thank Diamond Light Source for beamtime (proposals mx8922 and mx12342), and the staff of beamlines I02, I03 and I04-1 for assistance with crystal testing and data collection.

REFERENCES AND NOTES

1. Kunkel, T.A., and Bebenek, K. (2000). DNA Replication Fidelity. Annu. Rev. Biochem *69*, 497-529.
2. Watson, J.D., and Crick, F.H.C. (1953). Molecular structure of nucleic acids - A structure for deoxyribose nucleic acid. Nature *171*, 737-738.
3. Watson, J.D., and Crick, F.H.C. (1953). Genetical implications of the structure of deoxyribonucleic acid. Nature *171*, 964-967.
4. Malyshev, D.A., and Romesberg, F.E. (2015). The Expanded Genetic Alphabet. Angew. Chem.-Int. Edit. *54*, 11930-11944.
5. Hirao, I., Kimoto, M., and Yamashige, R. (2012). Natural versus Artificial Creation of Base Pairs in DNA: Origin of Nucleobases from the Perspectives of Unnatural Base Pair Studies. Acc. Chem. Res. *45*, 2055-2065.
6. Benner, S.A., and Sismour, A.M. (2005). Synthetic biology. Nat. Rev. Genet. *6*, 533-543.
7. Moran, S., Ren, R.X.-F., and Kool, E.T. (1997). A thymidine triphosphate shape analog lacking Watson–Crick pairing ability is replicated with high sequence selectivity. Proc. Natl. Acad. Sci. USA *94*, 10506-10511.
8. Guckian, K.M., Krugh, T.R., and Kool, E.T. (1998). Solution structure of a DNA duplex containing a replicable difluorotoluene-adenine pair. Nat. Struct. Mol. Biol. *5*, 954-959.
9. Kool, E.T., and Sintim, H.O. (2006). The difluorotoluene debate - a decade later. Chem. Commun., 3665-3675.
10. Hirao, I., and Kimoto, M. (2012). Unnatural base pair systems toward the expansion of the genetic alphabet in the central dogma. Proc. Jpn. Acad. Ser. B-Phys. Biol. Sci. *88*, 345-367.
11. Kaul, C., Muller, M., Wagner, M., Schneider, S., and Carell, T. (2011). Reversible bond formation enables the replication and amplification of a crosslinking salen complex as an orthogonal base pair. Nat. Chem. *3*, 794-800.
12. Malyshev, D.A., Dhami, K., Quach, H.T., Lavergne, T., Ordoukhanian, P., Torkamani, A., and Romesberg, F.E. (2012). Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. Proc. Natl. Acad. Sci. USA *109*, 12005-12010.
13. Dhami, K., Malyshev, D.A., Ordoukhanian, P., Kubelka, T., Hocek, M., and Romesberg, F.E. (2014). Systematic exploration of a class of hydrophobic unnatural base pairs yields multiple new candidates for the expansion of the genetic alphabet. Nucleic Acids Res. *42*, 10235-10244.
14. Kimoto, M., Yamashige, R., Matsunaga, K., Yokoyama, S., and Hirao, I. (2013). Generation of high-affinity DNA aptamers using an expanded genetic alphabet. Nat. Biotechnol. *31*, 453-458.
15. Yamashige, R., Kimoto, M., Takezawa, Y., Sato, A., Mitsui, T., Yokoyama, S., and Hirao, I. (2012). Highly specific unnatural base pair systems as a third base pair for PCR amplification. Nucleic Acids Res. *40*, 2793-2806.
16. Malyshev, D.A., Dhami, K., Lavergne, T., Chen, T.J., Dai, N., Foster, J.M., Correa, I.R., and Romesberg, F.E. (2014). A semi-synthetic organism with an expanded genetic alphabet. Nature *509*, 385-388.
17. Kimoto, M., Kawai, R., Mitsui, T., Yokoyama, S., and Hirao, I. (2009). An unnatural base pair system for efficient PCR amplification and functionalization of DNA molecules. Nucleic Acids Res. *37*, e14.
18. Hirao, I., Mitsui, T., Kimoto, M., and Yokoyama, S. (2007). An efficient unnatural base pair for PCR amplification. J. Am. Chem. Soc. *129*, 15549-15555.
19. Li, L.J., Degardin, M., Lavergne, T., Malyshev, D.A., Dhami, K., Ordoukhanian, P., and Romesberg, F.E. (2014). Natural-like Replication of an Unnatural Base Pair for the Expansion of the Genetic Alphabet and Biotechnology Applications. J. Am. Chem. Soc. *136*, 826-829.

20. Yang, Z., Chen, F., Alvarado, J.B., and Benner, S.A. (2011). Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System. J. Am. Chem. Soc. *133*, 15105-15112.
21. Yang, Z.Y., Sismour, A.M., Sheng, P.P., Puskar, N.L., and Benner, S.A. (2007). Enzymatic incorporation of a third nucleobase pair. Nucleic Acids Res. *35*, 4238-4249.
22. Yang, Z.Y., Hutter, D., Sheng, P.P., Sismour, A.M., and Benner, S.A. (2006). Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern. Nucleic Acids Res. *34*, 6095-6101.
23. Zhang, L.Q., Yang, Z.Y., Sefah, K., Bradley, K.M., Hoshika, S., Kim, M.J., Kim, H.J., Zhu, G.Z., Jimenez, E., Cansiz, S., *et al.* (2015). Evolution of Functional Six-Nucleotide DNA. J. Am. Chem. Soc. *137*, 6734-6737.
24. Leal, N.A., Kim, H.J., Hoshika, S., Kim, M.J., Carrigan, M.A., and Benner, S.A. (2015). Transcription, Reverse Transcription, and Analysis of RNA Containing Artificial Genetic Components. ACS Synth. Biol. *4*, 407-413.
25. Georgiadis, M.M., Singh, I., Kellett, W.F., Hoshika, S., Benner, S.A., and Richards, N.G.J. (2015). Structural Basis for a Six Nucleotide Genetic Alphabet. J. Am. Chem. Soc. *137*, 6947-6955.
26. Yang, Z., Chen, F., Chamberlin, S.G., and Benner, S.A. (2010). Expanded Genetic Alphabets in the Polymerase Chain Reaction. Angew. Chem. Int. Ed. *49*, 177-180.
27. Sefah, K., Yang, Z.Y., Bradley, K.M., Hoshika, S., Jimenez, E., Zhang, L.Q., Zhu, G.Z., Shanker, S., Yu, F.H., Turek, D., *et al.* (2014). In vitro selection with artificial expanded genetic information systems. Proc. Natl. Acad. Sci. U. S. A. *111*, 1449-1454.
28. Sheng, P., Yang, Z., Kim, Y., Wu, Y., Tan, W., and Benner, S.A. (2008). Design of a novel molecular beacon: modification of the stem with artificially genetic alphabet. Chem. Commun., 5128-5130.
29. Liu, J., and Subirana, J.A. (1999). Structure of d(CGCGAATTCGCG) in the presence of Ca2+ ions. J. Biol. Chem. *274*, 24749-24752.
30. Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R.E. (1980). Crystal structure analysis of a complete turn of B-DNA. Nature *287*, 755-758.
31. Drew, H.R., Wing, R.M., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R.E. (1981). Structure of a B-DNA dodecamer: conformation and dynamics. Proc. Natl. Acad. Sci. USA *78*, 2179-2183.
32. Patel, D.J., Kozlowski, S.A., Marky, L.A., Broka, C., Rice, J.A., Itakura, K., and Breslauer, K.J. (1982). Premelting and melting transitions in the d(CGCGAATTCGCG) self-complementary duplex in solution. Biochemistry *21*, 428-436.
33. DeRose, E.F., Perera, L., Murray, M.S., Kunkel, T.A., and London, R.E. (2012). Solution Structure of the Dickerson DNA Dodecamer Containing a Single Ribonucleotide. Biochemistry *51*, 2407-2416.
34. Renciuk, D., Blacque, O., Vorlickova, M., and Spingler, B. (2013). Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. Nucleic Acids Res. *41*, 9891-9900.
35. Hunter, W.N., Brown, T., Kneale, G., Anand, N.N., Rabinovich, D., and Kennard, O. (1987). The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution. J. Biol. Chem. *262*, 9962-9970.
36. Krishnamurthy, R. (2012). Role of pKa of Nucleobases in the Origins of Chemical Evolution. Acc. Chem. Res. *45*, 2035-2044.
37. Wilcox, J.L., and Bevilacqua, P.C. (2013). A Simple Fluorescence Method for pKa Determination in RNA and DNA Reveals Highly Shifted pKa's. J. Am. Chem. Soc. *135*, 7390-7393.
38. Kypr, J., Kejnovská, I., Renčiuk, D., and Vorlíčková, M. (2009). Circular dichroism and conformational polymorphism of DNA. Nucleic Acids Res. *37*, 1713-1725.
39. Yang, Z., Hutter, D., Sheng, P., Sismour, A.M., and Benner, S.A. (2006). Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern. Nucleic Acids Res. *34*, 6095-6101.
40. Geyer, C.R., Battersby, T.R., and Benner, S.A. (2003). Nucleobase Pairing in Expanded Watson-Crick-like Genetic Information Systems. Structure *11*, 1485-1498.
41. Szulik, M.W., Pallan, P.S., Nocek, B., Voehler, M., Banerjee, S., Brooks, S., Joachimiak, A., Egli, M., Eichman, B.F., and Stone, M.P. (2015). Differential Stabilities and Sequence-Dependent Base Pair Opening Dynamics of Watson-Crick Base Pairs with 5-Hydroxymethylcytosine, 5-Formylcytosine, or 5-Carboxylcytosine. Biochemistry *54*, 1294-1305.
42. Kim, H.J., Chen, F., and Benner, S.A. (2012). Synthesis and Properties of 5-Cyano-Substituted Nucleoside Analog with a Donor-Donor-Acceptor Hydrogen-Bonding Pattern. J. Org. Chem. *77*, 3664-3669.

43. Someya, T., Ando, A., Kimoto, M., and Hirao, I. (2015). Site-specific labeling of RNA by combining genetic alphabet expansion transcription and copper-free click chemistry. Nucleic Acids Res. *43*, 6665-6676.
44. Ishizuka, T., Kimoto, M., Sato, A., and Hirao, I. (2012). Site-specific functionalization of RNA molecules by an unnatural base pair transcription system via click chemistry. Chem. Commun. *48*, 10835-10837.
45. Seo, Y.J., Malyshev, D.A., Lavergne, T., Ordoukhanian, P., and Romesberg, F.E. (2011). Site-Specific Labeling of DNA and RNA Using an Efficiently Replicated and Transcribed Class of Unnatural Base Pairs. J. Am. Chem. Soc. *133*, 19878-19888.
46. Leslie, A.G.W., and Powell, H.R. (2007). Processing diffraction data with mosflm. In Evolving Methods for Macromolecular Crystallography: The Structural Path to the Understanding of the Mechanismof Action of CBRN Agents, R.J. Read, and J.L. Sussman, eds. (Springer Netherlands), pp. 41-51.
47. Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A.*, et al.* (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. Sect. D-Biol. Crystallogr. *67*, 235-242.
48. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J. Appl. Crystallogr. *40*, 658-674.
49. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. Sect. D-Biol. Crystallogr. *66*, 486-501.
50. Lu, X.-J., and Olson, W.K. (2008). 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat. Protocols *3*, 1213-1227.
51. Schrodinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.
52. Hodson, M.J., Hudson, A.J., Cherny, D., and Eperon, I.C. (2012). The transition in spliceosome assembly from complex E to complex A purges surplus U1 snRNPs from alternative splice sites. Nucleic Acids Res. *40*, 6850-6862.

Figure 1. Base-pairs of (A) naturally-occurring G-C, (B) P-Z and (C) the pH-dependent G-Z mispair.

Figure 2. (A) Overlay of ODN1 (magenta), ODN2 (orange), ODN3a (green) and ODN3b (cyan) crystal structures. Base-pairs for the 4-position (B) G-C (ODN1), (C) P-Z (ODN2 at pH 7.0), (D) G-Z (ODN3a at pH 7.8), and (E) G-Z (ODN3b at pH 8.5). Hydrogen-bond distances shown in Ångstroms. The $2F_o$-F_c electron density is contoured at 1.0σ.

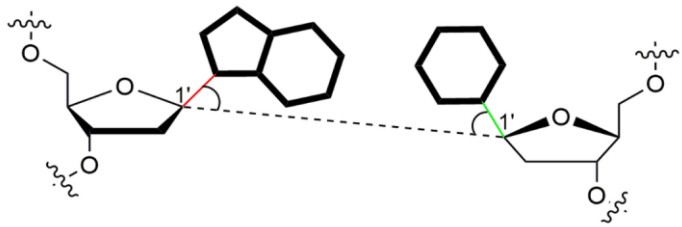
Figure 3. Comparative analysis of duplex stability and conformation of ODN1-4. (A) UV-vis melting temperatures of ODN1-3 duplexes between pH 5.5 and 9.5 measured at 260 nm. (See also Fig. S9-28, Table S2-3) (B) CD spectra of ODN1-3 at pH 7.5 exhibit characteristic right-handed B-type DNA spectra. (See also Fig. S29-34)

Figure 4. The effect of pH on PCR amplification of a DNA template containing a single P-Z base pair compared to a DNA template containing no P and Z nucleotides. Each PCR was conducted in cacodylate buffer. (See also Fig. 35-37)

Figure 5. Putative resonance structures of the Z anion.

Table 1. Schematic representation of a purine-pyrimidine base-pair in position 4.

Table 1. Schematic representation of a purine-pyrimidine base-pair in position 4.



	C-1′-C-1′ Distance (Å)	N-9-C-1′-C-1′ Angle (°)	N-1(C-1)-C-1′-C-1′ Angle (°)	Angle of C-1′-C-1′ vector (relative to ODN1) (°)
ODN1[a]	10.4	56.5	55.9	0
ODN2	10.8	55.5	52.3	2.3
ODN3a	10.7	44.5	63.0	9.6
ODN3b	10.8	57.5	54.7	1.0
ODN4[b]	10.4	42.5	69.9	14.0

Table legend.
[a]PDB ID 463D
[b]PDB ID 113D