Published in final edited form as:

J Am Chem Soc. 2011 September 28; 133(38): 15105–15112. doi:10.1021/ja204910n.

Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System

Zunyi Yang^{1,2,*}, Fei Chen^{1,2,*}, J. Brian Alvarado^{1,2}, and Steven A. Benner^{1,2}

¹Foundation for Applied Molecular Evolution (FfAME), 720 SW 2nd Avenue, Suite 201, Gainesville, FL 32601

²The Westheimer Institute for Science and Technology (TWIST), 720 SW 2nd Avenue, Suite 208, Gainesville, FL 32601

Abstract

The next goals in the development of a synthetic biology that use artificial genetic systems will require chemistry-biology combinations that allow the amplification of DNA containing any number of sequential and non-sequential non-standard nucleotides. This amplification must ensure that the non-standard nucleotides are not unidirectionally lost during PCR amplification (unidirectional loss would cause the artificial system to revert to an all-natural genetic system). Further, technology is needed to sequence artificial genetic DNA molecules. The work reported here meets all three of these goals for a six-letter artificially expanded genetic information system (AEGIS) that comprises four standard nucleotides (G, A, C, and T) and two additional non-standard nucleotides (Z and P). We report polymerases and PCR conditions that amplify a wide range of GACTZP DNA sequences having multiple consecutive unnatural synthetic genetic components with low (0.2% per theoretical cycle) levels of mutation. We demonstrate that residual mutation processes both introduce *and* remove unnatural nucleotides, allowing the artificial genetic system to evolve as such, rather than revert to a wholly natural system. We then show that mechanisms for these residual mutation processes can be exploited in a strategy to sequence “six-letter” GACTZP DNA. These are all not yet reported for any other synthetic genetic system.

Introduction

Efforts to develop analogs of nucleic acids to support various synthetic biologies^{1,2} are characterized by chemical successes that now must confront biological realities. Over a dozen non-standard genetic molecules have been prepared,³ including many with nucleotide modifications and expansions. Proposed additions to DNA alphabets include pairing based on different hydrogen bonding patterns,^{4,5} steric complementarity,^{6,7} and even no hydrogen bonds at all.^{8,9}

The ability of these analogs to bind specifically and orthogonally to natural DNA makes several of them important in clinical diagnostics. For example, an alternative genetic system

Correspondence and requests for materials should be addressed to S.A.B. (sbenner@ffame.org).

*These authors contributed equally to this work.

Author Contributions Z.Y., F.C. and S.A.B. planned the experiments. Z.Y., F.C. and J.B.A. performed the six-letter GACTZP PCR, sequenced the GACTZP DNA, and measured the mutation between standard and non-standard nucleotides. Z.Y., S.A.B. and F.C. analyzed the data. S.A.B. and Z.Y. wrote the paper. The authors declare that they are inventors on various patent applications covering various of the compounds and methods reported here.

Supporting Information. Polymerase extension reading through multiple consecutive nonstandard nucleobases, PCR amplification of the GACTZP DNA and the sequencing results are available via the Internet at <http://pubs.org>.

that expands the number of nucleotides in DNA by rearranging hydrogen bonding patterns to give a six-letter artificially expanded genetic information systems (AEGIS)¹⁰ today personalizes the care of 400,000 patients infected with viral diseases such as HIV and hepatitis.¹¹

However, these chemical successes have forced synthetic biologists to encounter the realities of natural biochemistry, which has evolved for billions of years to accept G, A, C, and T. Although, many natural polymerases can accept some components of unnatural genetic systems, all synthetic genetic systems examined to date suffer unidirectional *losses* when amplifying unnatural components.¹² In addition, their performance is generally limited to the acceptance of single unnatural nucleotides or nonadjacent unnatural pairs;¹³ template-based synthesis of DNA containing runs of consecutive non-standard base pairs is largely unknown.

Clinical diagnostics can use architectures where polymerases add single non-standard nucleotides. For example, such single AEGIS addition supports architectures that diagnose respiratory diseases¹⁴ and detect cystic fibrosis mutations.¹⁵ However, a fully functioning synthetic genetic system also needs to support PCR amplification of *any* oligonucleotide sequence built from six or more different building blocks.

Further, the tools that we take for granted in standard molecular biology are unavailable to most synthetic genetic systems. In particular, rapid sequencing methods are not available for any artificial genetic alphabet. Instead, most artificial genetic sequences are analyzed on individual exemplars using difficult methods.¹⁶

To overcome these barriers for synthetic genetics, we focused on what is known about the chemical details of molecular interactions between DNA polymerases and DNA. For example, some time ago, Joyce, Steitz and others noted that all four standard nucleotides (G, A, C, and T, or GACT) present electron density to the minor groove,¹⁷ either from N3 of the purines or from the exocyclic oxygen of the pyrimidines (green lobes, Figure 1). Little else in common is presented by G, A, C, and T to either groove. Therefore, Joyce and Steitz suggested their “minor groove scanning hypothesis”, which holds that polymerases seek this electron density as a way of achieving uniform acceptance of their four substrates. Although we and others have successfully copied single, non-consecutive unnatural nucleotides lacking this electron density, often with mutated polymerases,¹⁸ we reasoned that if we were to move forward in our development of synthetic genetic systems, the easiest path would be to accommodate this need of natural polymerases arising from billions of years of evolution.

Two matched components of the AEGIS synthetic genetic system present this electron density to the minor groove (Figure 1), at the same time having rearranged hydrogen bonding, permitting them to pair exclusively with each other, and not with GACT nucleobases. These are 6-amino-5-nitro-3-(1'- β -D-2'-deoxyribofuranosyl)-2(1H)-pyridone and 2-amino-8-(1'- β -D-2'-deoxyribofuranosyl)-imidazo[1,2-*a*]-1,3,5-triazin-4(8H)-one), respectively implementing hydrogen bond *donor-donor-acceptor* and *acceptor-acceptor-donor* patterns on small and large nucleobases. We trivially named these **Z** and **P**. These particular implementations were also chosen as the fifth and sixth nucleobase^{19,20} because, in addition to their presenting the scanned electron density, they are also insensitive to both oxidation²¹ and epimerization.²²

We report here polymerases that support PCR-amplification of six-letter GACTZP DNA having essentially any sequence, including sequences containing as many as four consecutive non-standard nucleotides. Error under optimized conditions was shown to be ca. 0.002 per theoretical cycle for both gain and loss of **Z:P** pairs. Furthermore, the chemical

mechanism for the residual mutation was studied, and found to resemble mechanisms for mutation in standard DNA. The resulting understanding of mechanisms of mutation allowed us to manipulate mutation rates to maximize the *loss* of Z and P during copying and PCR amplification. This loss allowed us to develop procedures to exploit standard sequencing methods to sequence GACTZP DNA. This provides a critical analytical tool to advance this particular synthetic biology.

Methods and Materials

Phosphoramidites, Triphosphates, and Polymerases

Protected phosphoramidites of the non-standard nucleosides dZ (protected as the O-NPE ether) and dP, and the triphosphates dZTP and dPTP were obtained from the Foundation for Applied Molecular Evolution (www.ffame.org, Nitroaminopyridone-dZ (Cat. # DZ-ZY001), Imidazoaminopyrimidone-dP (Cat. # DP-ZY001), dZTP (Cat. # DZTP-ZY101), dPTP (Cat. # DPTP-ZY102)). Polymerases were obtained from New England Biolabs. GACT DNA was obtained from IDT (Coralville IA). Other reagents were obtained from Promega and Sigma-Aldrich, and used as received.

GACTZP Oligonucleotide Synthesis

Oligonucleotides containing dZ and dP were synthesized using standard phosphoramidite chemistry on an ABI 394 DNA synthesizer on controlled pore glass supports. Following synthesis, oligonucleotides containing dZ were first treated with 1 M of DBU in anhydrous acetonitrile to remove the O-protection group (O-NPE ether), then, following the standard deprotection procedure in aqueous concentrated ammonia overnight at 55°C. Deprotection procedure for oligonucleotides containing dP is the same as the standard procedure.¹⁹

Polymerase Extension Reading Through Multiple Consecutive Non-standard Nucleotides

5'-³²P-Labeled primer (Primer-F1 or Primer-R1, 0.2 pmole of radio-labeled primer plus 4 pmole of non radio-labeled primer, final concentration 70 nM) was annealed to a template containing multiple consecutive non-standard nucleobases (dP or dZ, 6 pmole, final concentration 100 nM) in 1x ThermoPol polymerase reaction buffer (pH = 8.0 at room temperature) or 1x HF Phusion buffer (pH = 8.3 at room temperature) by heating at 96 °C for 5 min and then slow cooling (0.5 h) to room temperature. dNTPs (final 0.1 mM for each), or dNTPs, dZTP and dPTP (final 0.1 mM for each) were added at room temperature. The mixture was pre-heated at 72 °C for 30 seconds. Extension was initiated by adding *Taq* (2.5 units), Deep Vent (exo⁺, 2 units), or Phusion (1 unit) DNA polymerase to give a final volume of 60 µL. The primer was extended at 72 °C and aliquots (7 µL) were taken from each reaction at time intervals (1, 2, 4, 8, and 16 min), quenched by PAGE loading/quench buffer (10 µL, 10 mM EDTA in formamide). Samples were resolved by electrophoresis using a 16% PAGE (7 M urea). The gel was analyzed using MolecularImager software. See Supplementary Table S1 for the sequences of primer and template, and Supplementary Figure S1 for gel images.

PCR Amplification of the Synthetic GACTZP DNA (Figure 2)

Six-letter PCR amplification of GACTZP DNA containing multiple consecutive non-standard nucleobases (2P-Temp, 3P-Temp, and 4P-Temp, final 0.5 nM for each, Table 1) was carried out in 1x ThermoPol reaction buffer (pH = 8.0 at room temperature, for Deep Vent (exo⁺) and *Taq* DNA polymerase, respectively), or 1x HF Phusion buffer (pH = 7.0 at room temperature, for Phusion DNA polymerase), 0.5 µM of each Primer-F1 and Primer-R1, four standard dNTPs (each 0.1 mM), dZTP (0.05 mM), dPTP (0.6 mM), and 0.05 unit/µl of DNA polymerase (*Taq*) or 0.02 unit/µL (Deep Vent (exo⁺) and Phusion, respectively)

on the DNAEngine Peltier Thermal Cycler (Bio- Rad) in a total volume of 50 μ L. The following PCR conditions were used: one cycle of 95 $^{\circ}$ C for 2 min; followed by 21 cycles of (95 $^{\circ}$ C for 20 s, 58 $^{\circ}$ C for 25 s, 72 $^{\circ}$ C for 3 min); and finally 72 $^{\circ}$ C for 10 min. Upon the completion of PCR, samples (10 μ L) were taken from each PCR mixture, mixed with 6x agarose loading dye (2 μ L, Promega), and analyzed on a 3% agarose gel. See Figure 2 for gel images.

Mutation interconverting Z:P, C:G and T:A Pairs (Figure 3)

- a. The “forward” mutation converting C:G pairs into Z:P pairs using digestion with the restriction endonuclease (Bsp120I).

Eight parallel PCRs were performed in 1x ThermoPol buffer at two different pHs (8.8 and 8.0 at 25 $^{\circ}$ C). The PCR mixture contained identical amounts of primers, (Primer-F3 (5 pmole) and Primer-R3 (1 pmole of 5'- 32 P-labeled primer and 4 pmole of non- 32 P-labeled primer), each 250 nM final), template (Bsp-G, 0.25 nM final), four standard dNTPs (0.2 mM each), and JumpStart *Taq* DNA polymerase (0.075 unit/ μ L, Sigma). Two non-standard nucleotide triphosphates, dZTP and dPTP (each 0.2 mM), were absence or present in each PCR mixture (see Figure 3a for details). The PCR mixture (20 μ L of total volume) were cycled using the following conditions: one cycle of 95 $^{\circ}$ C for 1 min; followed by 26 cycles of (95 $^{\circ}$ C for 30 s, 55 $^{\circ}$ C for 30 s, 72 $^{\circ}$ C for 1 min); and finally 72 $^{\circ}$ C for 10 min. After PCR amplification, samples (5 μ L) were taken from each PCR mixture, mixed with PAGE loading/quench buffer (7 μ L, 10 mM EDTA in formamide), and resolved by electrophoresis using 10% PAGE (7 M urea). The gel was analyzed using MolecularImager software. The results shown all primers were consumed and PCR amplicon was produced with the expected length. Then, another 1 μ L of the PCR mixture was digested with Bsp120I (0.5 μ L, final 0.5 units/ μ L) in 1x Buffer B (10 mM Tris-HCl, 10 mM MgCl₂, 0.1 mg/mL BSA, pH 7.5) at 37 $^{\circ}$ C for 20 hours (10 μ L of reaction volume). Additional 0.5 μ L of Bsp120I was added to the digestion mixture and incubated for another 20 hours. The digestion products were resolved on 10 % PAGE gel (7 M urea) and visualized by autoradiography (see Figure 3a for results).

- b. Measuring the “reverse” mutation of Z:P pair to give C:G and T:A pair using restriction endonuclease (Bsp120I).

Six parallel PCRs were performed in 1x ThermoPol buffer at two different pHs (8.8 and 8.0, measured at 25 $^{\circ}$ C). The PCR mixture contained identical amounts of primers, (Primer-F3 (5 pmole) and Primer-R3 (1 pmole of 5'- 32 P-labeled primer and 4 pmole of non- 32 P-labeled primer), each 250 nM final), synthetic templates (Bsp-Z and Bsp-P, Table 1, each 0.25 nM final), four standard dNTPs (200 μ M each), various amount of dZTP and dPTP (20 μ M (lane 1), 10 μ M (lane 2), and 5 μ M (lane 3), respectively, Figure 3b), and JumpStart *Taq* DNA polymerase (0.075 unit/ μ L, Sigma). The PCR mixture (20 μ L of total volume) was cycled (26 rounds of 95 $^{\circ}$ C for 30 s, 55 $^{\circ}$ C for 30 s, 72 $^{\circ}$ C for 1 min). Upon the consumption of all of the primers, the PCR mixture (1 μ L) was digested with Bsp120I (0.5 μ L, final 0.5 units/ μ L) in 1x Buffer B (10 mM Tris-HCl, 10 mM MgCl₂, 0.1mg/mL BSA, pH 7.5) at 37 $^{\circ}$ C for 20 hours (10 μ L of reaction volume). Additional 0.5 μ L of Bsp120I was added to the digestion mixture and incubated for another 20 hours. The digestion products were resolved on 10 % PAGE gel (7 M urea) and visualized by autoradiography (see Figure 3b for results).

- c. Measuring the conversion of Z into C and T and P into A and G using the restriction endonuclease Bsp120I.

In 1x ThermoPol reaction buffer (pH 8.8 at 25 °C) and 0.2 mM of each four standard dNTPs (without dZTP and dPTP), single-stranded synthetic template Bsp-Z (Figure 3c, lane 1) or Bsp-P (Figure 3c, lane 2) was 1000 fold amplified with primers (Primer-F3 and Primer-R3) using JumpStart *Taq* DNA polymerase (0.08 unit/ μ L, Sigma). Upon the completion of PCR amplification, 1 μ L of PCR mixture was digested with Bsp120I (1 μ L, final 1 units/ μ L) in 1x Buffer B at 37 °C for 20 hours (10 μ L of reaction volume). Then, additional 0.5 μ L of Bsp120I was added to the digestion mixture and incubated for another 20 hours. The digestion products were resolved on 10% PAGE gel (7 M urea) and visualized by autoradiography (see Figure 3c for results).

Measuring the Retention and Mutation of Z:P Pair in Optimized Six-letter PCR (Figure 4)

In 1x ThermoPol reaction buffer (pH 8.0 measured at 25 °C), synthetic template (Bsp-P, Table 1) or standard template (Bsp-G, Table 1) was amplified (1000 to 100000 fold, respectively) with primers (250 nM final concentration of Primer-F3 and Primer-R3) and dA,T,G/TPs = 0.1 mM, dCTP = 0.4 mM, dZTP = 0.05 mM, and dPTP = 0.6 mM using JumpStart *Taq* DNA polymerase (0.08 unit/ μ L, Sigma). The PCR mixture were cycled using the following conditions: one cycle of 95 °C for 1 min; followed by 31 cycles of (95 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min); and finally 72 °C for 10 min. Upon the consumption of primers, 1 μ L of PCR mixture was digested with Bsp120I (0.5 μ L, final 0.5 units/ μ L) in 1x Buffer B at 37 °C for 20 hours (10 μ L of reaction volume). Additional 0.5 μ L of Bsp120I was added to the digestion mixture and incubated for another 20 hours. The digestion products were resolved on 10 % PAGE gel (7 M urea) and visualized by autoradiography (see Figure 4 for results).

PCR Amplification of the GACTZP DNA and Sanger Sequencing of the PCR Products (Figure 5 and Supplementary Table S2)

Synthetic GACTZP DNA containing various numbers of **Z** and **P** nucleotides incorporated at various positions, adjacent and spaced apart (final 0.04 nM of each, Supplementary Table S1) were amplified in 1x ThermoPol reaction buffer (pH = 8.0, measured at room temperature) containing primers (0.4 μ M each of Primer-F1 and Primer-R1, or Primer-F2 and Primer-R2, or Primer-F3 and Primer-R3), dA,T,G/TPs (each 0.1 mM), dCTP (0.2 mM to 0.4 mM), dZTP (0.05 mM), dPTP (0.6 mM), and 0.05 unit/ μ L of JumpStart *Taq* DNA polymerase in a total volume of 50 μ L. The following PCR conditions were used: one cycle of 95 °C for 1 min; followed by 21 cycles of (95 °C for 20 s, 58 °C for 25 s, 72 °C for 3 min); and finally 72 °C for 10 min. Upon the completion of the PCR, samples (10 μ L) were taken from each PCR mixture, mixed with 6x agarose loading dye (2 μ L, Promega), and analyzed on agarose gel.

As a first step towards sequencing, the remaining single stranded primers and excess triphosphates were degraded in the amplicon mixture by incubating aliquots (20 μ L) of the PCR mixture with ExoSAP-IT (8 μ L, USB, Cleveland, Ohio) at 37 °C for 30 min, and then at 80 °C for 15 min. Double stranded amplicons were then recovered by Qiaquick Nucleotide Remove Kit (Qiagen, Valencia, CA). The GACTZP DNA was eluted from the spin column using EB buffer (200 μ L, 10 mM Tris⁺Cl, pH 8.5) and sequenced following the strategies and protocols described below.

The purified GACTZP DNA was further amplified using JumpStart *Taq* DNA polymerase (0.05 unit/ μ L) in 1x ThermoPol reaction buffer (pH = 8.8 at room temperature), 0.25 μ M of each Primer (Primer-F1 and Primer-R1, or Primer-F2 and Primer-R2, or Primer-F3 and Primer-R3, Supplementary Table S1), four standard dNTPs (final 0.2 mM of each) and dPTP (final 0.2 mM, Figure 5b). The following PCR conditions were used: one cycle of 95

°C for 1 min; followed by 25 cycles of (95 °C for 20 s, 58 °C for 25 s, 72 °C for 1.5 min); and finally 72 °C for 15 min. Upon the completion of PCR, PCR products were analyzed by agarose gel electrophoresis.

Fresh PCR products were cloned into the pCR[®]2.1-TOPO[®] vector and transformed the recombinant vector into One Shot[®] DH5 α [™]-T1^R chemically competent cells using the TOPO TA Cloning[®] Kits (Invitrogen, Carlsbad, CA). Blue-white screening gave 24 to 40 colonies that were submitted for Sanger sequencing (BioBasic, Canada). The sequence results were shown in Supplementary Table S2.

Results

Polymerases to amplify GACTZP six-letter DNA were screened using templates containing various numbers of **Z** and **P** nucleotides adjacent and apart (Table 1). Consistent with the minor groove scanning hypothesis, and different from experiments with AEGIS components that do not present electron density to the minor groove,¹⁸ no polymerase rejected the **Z:P** pairs entirely. Further, polymerases accepting **Z:P** pairs at single site also accepted **Z:P** pairs when spaced apart (data not shown).

However, some polymerases had difficulty accepting multiple consecutive **Z** and **P** nucleotides. For example, Deep Vent (exo⁺) accepted two consecutive template-**P**'s and **Z**'s but not three. *Taq* and Phusion, in contrast, incorporated dZTP opposite three and four consecutive **P**'s, and dPTP opposite three and four consecutive **Z**'s (Supplementary Figure S1). Phusion and *Taq* DNA polymerases also PCR-amplified templates containing three and four consecutive **Z** and **P** nucleotides with efficiencies only slightly lower than with standard DNA (Figure 2, lanes 2, 3, 5, 7). Here, the retention of artificial bases in the PCR products obtained from *Taq* was verified by sequencing (see below, Supplementary Table S1 and S2). Performance at this level has not been seen with any other artificial genetic system, including those developed in this laboratory.^{18, 25} The relative facility with which this performance is obtained might be viewed as being consistent with the Steitz-Joyce scanning hypothesis.

But what about fidelity? To follow the misincorporation of **Z** and **P** into standard sequences and the loss of **Z:P** pairs from GACTZP DNA, we exploited our observation that the Bsp120I restriction endonuclease does not cleave its recognition sequence (5'-GGGCCC-3') if any C or G is replaced by **Z** or **P**.²³ A DNA molecule containing a 5'-GGGCCC-3' sequence was amplified by *Taq* PCR (at pH 8.8 or 8.0) with and without dZTP and dPTP. The amplicons were then treated with Bsp120I. In the presence of 0.2 mM of both dZTP and dPTP (Figure 3a, lane 4), after 1000 fold PCR amplification, 16% and 6% of the amplicon obtained at pH 8.8 and 8.0 (respectively) resisted digestion (see also Supplementary Fig. S2). These results indicate *Taq* only slowly replaces C:G by **Z:P** pairs, with the error at high "error-generating" pH 8.8 (ca. 0.25% per theoretical cycle per site; the pH is measured at room temperature) dropping to less than 0.1% per theoretical cycle per site at the lower pH of 8.0. The observed pH dependency suggested that mismatching arises predominantly as a result of deprotonated dZTP pairing with G, which becomes significant at high pH (Figure 3a, left lane 3). In contrast, dPTP pairing with C is negligible at both pH's (Figure 3a, lane 2).

These fidelity results also demonstrated the existence of small amounts of "forward" mutation, where *non*-standard components *enter* a sequence during copying, rather than being lost (Figure 3a). In artificial genetic systems in general, polymerases show only a natural propensity to *lose* unnatural components. This is, we believe, the first example of forward mutation in any synthetic genetic system.

Mutation of T:A pairs to **Z:P** pairs was even rarer. To identify rare substitutions of this kind, we exploited our observation (unpublished) that *DraI* does not cut at its recognition sequence (5'-TTTAAA-3') if any site contains **Z** or **P**. Here, DNA containing a 5'-TTTAAA-3' sequence was amplified 1000 fold with and without d**ZTP** and d**P**TP at pH 8.8 and 8.0. Here, no detectable fraction of the amplicon became resistant to cleavage (data not shown). This showed that any T:A to **Z:P** forward mutation was less facile than mutation of C:G to **Z:P**, and did not occur to less than one part in ca. 50,000.

To measure the rates of "reverse" mutation that convert **Z:P** pairs into C:G or T:A pairs, PCR amplification was performed on a template that contained the *Bsp120I* recognition sequences disrupted by **Z** and **P** nucleotides under forcing conditions. Here, double-strand GACTZP DNA (Bsp-Z and Bsp-P, Table 1), containing 5'-GGGCC**Z**-3' and 3'-CCCGGP-5', were 1000 fold amplified with low concentrations of d**ZTP** and d**P**TP (5 μ M to 20 μ M each) and the four standard dNTPs (200 μ M each). Followed by endonuclease digestion, the fraction of cleavable amplicons was used as a metric to quantitate **Z:P** loss. As shown in Figure 3b, most amplicon was digested by *Bsp120I*, indicating reverse mutation of the **Z:P** pair to a C:G pair, recreating the recognition sequence. At pH 8.0, loss was less than 7% per theoretical cycle at 20 μ M of d**P**TP and d**ZTP** (Figure 3b, right lane 1).

To drive the loss of **Z:P** pairs more forcibly, DNA containing 5'-GGGCC**Z**-3' (Bsp-Z, Table 1) or 3'-CCCGGP-5' (Bsp-P, Table 1) sequences were amplified without *any* d**ZTP** or d**P**TP. Here, 95% of PCR product was digested (Figure 3c, lane 1), indicating that *Taq* incorporates both dGTP (95%) and dATP (5%) opposite template **Z** in the absence of d**P**TP, mutating **Z** into C or T. In contrast, 70% of PCR product resisted digestion (Figure 3c, lane 2), indicating that *Taq* incorporates both dTTP (70%) and dCTP (30%) opposite template **P** in the absence of d**ZTP**, mutating **P** into A or G.

The observed mutation under forcing conditions and pH dependency suggested mechanisms for mutation (Figure 3d) and conditions that might maximize the fidelity of copying GACTZP DNA. At pH 8.0 (measured at room temperature), decreasing the concentration of d**ZTP** from 0.2 mM to 0.05 mM significantly reduced the "forward mutation" (converting C:G pairs into **Z:P** pairs). This d**ZTP** concentration (0.05 mM) is also sufficient to faithfully incorporate d**ZTP** opposite template-P, and also prevents mispairing of dCTP and dTTP with template-P. This was also verified by the subsequent sequencing results in Supplementary Figure S4.

Next, increasing the concentration of dCTP (from 0.2 to 0.6 mM) essentially eliminates mispairing of d**ZTP** with G (Supplementary Fig. S2b). Likewise, increasing the concentration of d**P**TP to 0.6 mM ensures d**P**TP pairing with template-**Z** in competition with dGTP. Last, decreasing the concentration of dA,T,G/TPs to 0.1 mM and adjusting the ratio of standard to non-standard triphosphates, finishes the optimization process. Under these optimized conditions, retention of **Z:P** pairs averaged 99.8% per theoretical PCR cycle, while the loss and gain of **Z:P** pairs is ca. 0.2% per theoretical PCR cycle (Figure 4 and Supplementary Fig. 2b). In contrast, under normal triphosphate concentrations (0.2 mM, without optimizing the concentrations), the retention of one Z-P pair is 99.2% per theoretical PCR cycle, and about 0.6% per theoretical cycle from natural to artificial base pair (for all Z/Ps in the recognition sequence) (Supplementary Fig. S2a).

Any biotechnology based on an evolvable genetic molecule built from six nucleotide letters needs analytical tools to determine its sequence. Accordingly, we developed such a tool for GACTZP DNA, which we describe here and use to show that amplicons arising from known initial sequences retain **Z** and **P** at their proper positions.

This sequencing tool exploited both the power of high throughput DNA sequencing technologies²⁴ and the understanding of how **Z:P** pairs in duplex DNA might evolve to give C:G and/or T:A pairs during PCR amplification (Figure 3d). *Opposite* of our goal while developing high fidelity six-letter PCR, which sought to *minimize* the loss of **Z** and **P**, our goal in developing sequencing tools was to *maximize* the loss of **Z** and **P**. The sequencing procedure developed had these steps (Figure 5 and Supplementary Figure S3):

- a. A sample of duplex GACTZP DNA is PCR-amplified using dNTPs (0.2 mM each) and *just* dPTP (0.2 mM), *without* dZTP. Lacking dZTP, any **P** in the template directs addition of either dTTP or dCTP to the primer. The presence of some dPTP allows any **Z** in any template to direct the incorporation of **P** in a product strand; the derived **P** subsequently directs incorporation of either T or C in the next copying step (Supplementary Figure S3b).
- b. The products of the “conversion” PCR reaction are then shotgun cloned.
- c. Individual DNA molecules from the clones are sequenced.
- d. The resulting sequences are aligned and compared.
- e. Sites in the alignment that hold both C and T in various aligned sequences are inferred to have arisen from sites in the parent sequence that held **Z**; sites in the alignment that hold both G and A are inferred to have arisen from sites in the parent sequence that held **P**. Sites in the alignment that consistently hold G, A, C, and T in *all* of the aligned sequences are inferred to have arisen from sites in the parent sequence that held G, A, C, and T, respectively.

In more detail, sites that originally held **P** in the precursor would hold either G or A in the converted sequence as a result of steps that involved **P:C** and **P:T** mispairing (respectively) in the absence of dZTP. If the mismatching is balanced, the result will generate a “G” call in half of the sequences, and a “A” call the other half. Similarly, sites that originally held **Z** will generate either a “C” call or a “T” call, through a first step involving **Z:P** pairing, and a second involving **P:C** and **P:T** mispairing (respectively). Sites that originally held G, A, C, and T will give uniform calls in all of the sequences returned though consistent G:C and T:A pairing (*pace* an occasional PCR error). Thus, the sequence of the precursor and the positions of **Z** and **P** in that sequence can be inferred (Figure 5 and Supplementary Figure S3).

We found that *Taq* DNA polymerase supports the needed level of mismatching of template **P** against T or C at these concentrations in the total absence of dZTP (Figure 3c). In contrast, in conditions that we examined, template-**Z** in the total absence of dPTP directed overwhelmingly the incorporation of G (leading to amplicons where **Z** is replaced by C), not the balanced mixture of G and A that would be most useful to infer a sequence.

To demonstrate the use of “conversion PCR” to sequence GACTZP DNA, DNA molecules containing various consecutive and non-consecutive **Z**’s and **P**’s (Supplementary Table S1) were first amplified under optimized six-letter PCR conditions (Supplementary Figure S3a). To convert **Z:P** pairs in PCR amplicon to a mixture of T:A and C:G pairs (Supplementary Figure S3b), a second PCR was performed in 1x ThermoPol reaction buffer (pH = 8.8, measured at room temperature) with *Taq*, standard dNTPs (0.2 mM each), no dZTP, and dPTP (0.2 mM) to further amplify the **Z:P** containing PCR amplicon. The second PCR products were then cloned into the pCR[®]-2.1-TOPO[®] plasmid and transformed into *E. coli* (DH5α), colonies were picked, plasmids were isolated and Sanger sequenced, and the separate sequencing results compared (Supplementary Table S2).

As expected, at sites in the amplicon that originated as A:T or G:C pairs, all Sanger sequences concurred (Supplementary Table S2). However, at sites in the amplicon that originated as **Z:P** pairs, the sequences differed and showed a mixture of T:A and C:G pairs at those sites. Thus, the positions of the **Z:P** pairs in the parent amplicons could be inferred, and were found to be where they were placed in the original template that was PCR amplified. Control experiments amplifying targets that contained no **Z:P** pairs in a GACT sequence (Bsp-C, Table 1) showed negligible false calls of T:A and C:G pairs.

DISCUSSION

These results show that both *Taq* and Phusion polymerases support six-letter GACTZP PCR for sequences containing up to (and including) four consecutive non-standard synthetic nucleotides. This represents a considerable advance over the current “art”, where only single or nonadjacent non-standard nucleotides can be part of a PCR amplification. This confirms the value of the “minor groove scanning” hypothesis to guide the design of at least one synthetic genetic system.

When constructing random sequences from a six letter alphabet, with each nucleotide being present in equal amounts (16.7%), a longer run of five (or more) consecutive d**Z** or d**P** nucleotides is expected to occur only twice in ca. every 7800 sites. As a typical plasmid contains ca. 3000 nucleotides, the level of performance demonstrated here should be sufficient to ensure the replication of a plasmid-sized DNA molecule containing entirely random sequences. Work is now in progress to engineer strains of *E. coli* that accept plasmids containing **Z:P** pairs.

With *Taq*, the minimized error after optimization is ca. 0.002 per theoretical cycle in both directions, the “forward” direction that gains **Z:P** pairs and the “reverse” direction that loses **Z:P** pairs. Absent selection pressure, this implies that GACTZP DNA would “evolve” to randomize their sequence with respect to G, C, **Z**, and **P**, rather than lose **Z** and **P** and gradually revert to natural DNA. To date, all other artificial genetic systems evolve to lose their unnatural components.²⁵

The mechanism for mutation resembles mechanisms for mutation in standard DNA, as evidenced by the dependence on pH of the mutation rate. Before deprotonation, **Z** presents to its partner a hydrogen bond donor-donor-acceptor pattern, proceeding from the major to the minor groove, a pattern that is complementary to the acceptor-acceptor-donor pattern of **P**. After deprotonation, with a $pK_a \approx 7.8$ free in solution, **Z** presents a donor-acceptor-acceptor hydrogen bonding pattern, as does standard C, and is complementary to standard G. The predominant mutation processes can be explained by **Z**:G mismatching from deprotonated **Z**.

This allows mutation to meet two goals of necessary to support Darwinian evolution. The level of mutation is sufficient to allow mutation, at the same time as not being so high as to cause an “error catastrophe” in a small genome. Further, the mutation is bidirectional; because it allows **Z** to be introduced as well as lost, the unnatural system does not trivially evolve to return to an entirely natural DNA molecule.

Of special importance in this work are the procedures that allow us to sequence GACTZP DNA. Sequencing strategies were developed using various hypothetical mechanisms for mutation. These allowed us to manipulate mutation rates to controllably convert d**Z** and d**P** nucleotides to standard nucleotides that could be cloned or directly sequenced using any next generation sequencing technologies. These should allow GACTZP DNA to support SELEX experiments. Work to develop these is also underway.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are indebted to the Defense Threat Reduction Agency (HDTRA1-08-1- 0052), the National Human Genome Research Institute (R01HG004831), the National Institute of General Medical Sciences (R01GM081527), and Nucleic Acids Licensing LLC for the support of this work.

References

1. Szybalski, W. *In vivo* and *in vitro* Initiation of transcription. In: Kohn, A.; Shatkey, A., editors. *Control of Gene Expression*. New York: Plenum Press; 1974. p. 23-24,p. 404-405.p. 411-412.p. 415-417.
2. Benner SA, Yang Z, Chen F. *Comptes Rendus*. 2010; 14:372–387.
3. Henry AA, Romesberg FE. *Current Opin Chem Biol*. 2003; 7:727–733.
4. Piccirilli JA, Krauch T, Moroney SE, Benner SA. *Nature*. 1990; 343:33–37. [PubMed: 1688644]
5. Bain JD, Chamberlin AR, Switzer CY, Benner SA. *Nature*. 1992; 356:537–539. [PubMed: 1560827]
6. Hikida Y, Kimoto M, Yokoyama S, Hirao I. *Nat Protoc*. 2010; 5:1312–1323. [PubMed: 20595959]
7. Hirao I, Mitsui T, Kimoto M, Yokoyama S. *J Am Chem Soc*. 2007; 129:15549– 15555. [PubMed: 18027940]
8. Delaney JC, Henderson PT, Helquist SA, Morales JC, Essigmann JM, Kool ET. *Proc Nat Acad Sci USA*. 2003; 100:4469–4473. [PubMed: 12676985]
9. Malyshev DA, Seo YJ, Ordoukhanian P, Romesberg FE. *J Am Chem Soc*. 2009; 131:14620–14621. [PubMed: 19788296]
10. Benner SA. *Acc Chem Res*. 2004; 37:784–797. [PubMed: 15491125]
11. (a) Arens MG, Buller RS, Rankin A, Mason S, Whetsell A, Agapov E, Lee WM, Storch GA. *J Clin Microbiol*. 2010; 48:2387–2395. [PubMed: 20484608] (b) Elbeik T, Markowitz N, Nassos P, Kumar U, Beringer S, Haller B, Ng V. *J Clin Microbiol*. 2004; 42:3120–3127. [PubMed: 15243070] (c) Elbeik T, Surtihadi J, Destree M, Gorlin J, Holodniy M, Jortani SA, Kuramoto K, Ng V, Valdes R, Valsamakis A, Terrault NA. *J Clin Microbiol*. 2004; 42:563–569. [PubMed: 14766817]
12. Johnson SC, Sherrill CB, Marshall DJ, Moser MJ, Prudent JR. *Nucl Acids Res*. 2004; 32:1937–1941. [PubMed: 15051811]
13. Kimoto M, Kawai R, Mitsui T, Ypkoyama S, Hirao I. *Nucl Acids Res*. 2009; 37:e14. [PubMed: 19073696]
14. Lee WM, Grindle K, Pappas T, Marshall DJ, Moser MJ, Beaty EL, Shult PA, Prudent JR, Gern JE. *J Clin Microbiol*. 2007; 45:2626–2634. [PubMed: 17537928]
15. Johnson SC, et al. *J Clin Chem*. 2004; 50:2019–2027.
16. Hirao I, Kimoto M, Mitsui T, Fujiwara T, Kawai R, Sato A, Harada Y, Yokoyama S. *Nat Methods*. 2006; 3:729–735. [PubMed: 16929319]
17. (a) Joyce CM, Steitz TA. *Ann Rev Biochem*. 1994; 63:777–822. [PubMed: 7526780] (b) Kool ET, Morales JC, Guckian KM. *Angew Chem Int Ed Engl*. 2000; 39:990–1009. [PubMed: 10760910] (c) Kim TW, Delaney JC, Essigmann JM, Kool ET. *Proc Natl Acad Sci U S A*. 2005; 102:15803–8. [PubMed: 16249340] (d) Matsuda S, Leconte AM, Romesberg FE. *J Am Chem Soc*. 2007; 129:5551–5557. [PubMed: 17411040]
18. Sismour AM, Lutz S, Park JH, Lutz MJ, Boyer PL, Hughes SH, Benner SA. *Nucl Acids Res*. 2004; 32:728–735. [PubMed: 14757837]
19. Yang Z, Hutter D, Sheng P, Sismour AM, Benner SA. *Nucl Acids Res*. 2006; 34:6095–6101. [PubMed: 17074747]
20. Yang Z, Sismour AM, Sheng P, Puskar NL, Benner SA. *Nucl Acids Res*. 2007; 35:4238–4249. [PubMed: 17576683]
21. Krosigk UV, Benner SA. *J Am Chem Soc*. 1995; 117:5361–5362.

22. Hutter D, Benner SA. *J Org Chem.* 2003; 68:9839–9842. [PubMed: 14656120]
23. Chen F, Yang ZY, Yan MC, Brian JA, Wang GG, Benner SA. *Nucleic Acid Res.* 2011; 39:3949–3961. [PubMed: 21245035]
24. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang XH, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV. *Nature Biotech.* 2009; 27:1013–1023.
25. Johnson SC, Sherrill CB, Marshall DJ, Moser MJ, Prudent JR. *Nucl Acids Res.* 2004; 32:1937–1941. [PubMed: 15051811]

Figure 1. The expanded GACTZP genetic system

Left column: Matched C:G, Z:P, and T:A pairs all fit the Watson-Crick geometry (a small pyrimidine analog with one ring complements in size a large purine analog with two rings and are all joined by three hydrogen bonds). Electron density presented to the minor groove is represented as shaded green lobes. Right column: Mismatched C:P, Z:G, and T:P pairs. Note clashes between electrons (gray lobes) or hydrogens, which can be mitigated by protonation/deprotonation, respectively.

Figure 2. Agarose gel (3%) resolving amplicons from “six-letter” GACTZP PCR with standard templates and synthetic templates containing multiple consecutive dPs

Lane 1: Control without template; Lane 2: Amplification of standard template, without dZTP and dPTP; Lane 3: Amplification of standard template, with dZTP and dPTP; Lane 4 and 6: Amplification of synthetic template, without dZTP and dPTP; Lane 5 and 7: Amplification of synthetic template, with dZTP and dPTP. dNTPs (0.1 mM for each), dZTP (0.05 mM) and dPTP (0.6 mM). M: 25 bp marker. See Methods and Materials for PCR conditions.

a **b**

d
Transversion of Base Pairs Through Misincorporation

Figure 3. Mutation interconverting Z:P, C:G and T:A Pairs

a. Measuring the “forward” mutation converting C:G pairs into Z:P pairs using digestion with the Bsp120I restriction endonuclease. Standard oligonucleotide (Bsp-G, Table 1) containing the Bsp120I recognition sequence (5'-GGGCCC-3') were 1000 fold amplified using *Taq* DNA polymerase at pH = 8.8 or 8.0 with standard dNTPs (0.2 mM of each) with or without dZTP and dPTP. Then, PCR amplicon were digested by endonuclease (Bsp120I). Lane 1: In the absence of both dZTP and dPTP; Lane 2: With dPTP (0.2 mM); Lane 3: With dZTP (0.2 mM); Lane 4: With both dZTP and dPTP (0.2 mM for each), Not digested: indicates the fraction of PCR product resisted endonuclease digestion; Digested: indicates the fraction of PCR product was digested. See Methods and Materials for PCR conditions.

b. Measure of the “reverse” mutation of Z:P to give C:G and T:A using restriction endonuclease. Two complementary synthetic templates (Bsp-Z and Bsp-P, Table 1) containing 5'-GGGCCZ-3' and 3'-CCCGGP-5', were 1000 fold amplified using *Taq* with standard dNTPs (200 μM), dZTP and dPTP (with various concentration, lane 1 (20 μM), lane 2 (10 μM), lane 3 (5 μM)), then, PCR amplicon were digested by endonuclease (Bsp120I). See Methods and Materials for details.

c. Measuring the mutation of Z into C and T (left) and P into A and G (right) using restriction endonuclease digestion. Single-stranded synthetic oligonucleotide containing either 5'-GGGCCZ-3' (left, lane 1, Bsp-Z) or 3'-CCCGGP-5' (right, lane 2, Bsp-P), was 1000 fold amplified using *Taq* with only four standard dNTPs (0.2 mM) in 1x ThermoPol reaction buffer (pH 8.8 at room

temperature). Then, PCR amplicon was digested by endonuclease (Bsp120I). See Methods and Materials for details.

d. Observed pathways of mutation between non-standard nucleotides and standard nucleotides. Conversion of C:G pair to Z:P pair (forward mutation) involves mis-incorporation of dZTP opposite template-G to form Z:G mismatch at high pH (8.8) (Fig. 3a);

Conversion of Z:P pair to C:G pair (reverse mutation) involves two pathways: 1) the most likely pathway, mis-incorporation of dGTP opposite template-Z to form G:Z mismatch at high pH (8.8) (Fig. 3b); 2) mis-incorporation of dCTP opposite template-P to form C:P mismatch (Fig. 3c, lane 2);

Conversion of Z:P pair to T:A pair (reverse mutation) involves two pathways: 1) the most likely pathway, mis-incorporation of dTTP opposite template-P to form T:P mismatch; 2) misincorporation of dATP opposite template-Z to form A:Z mismatch (Fig. 3c).

See Figure 1 for corresponding matched and mismatched base pairs.

Figure 4. Measuring the retention and mutation of Z:P pair in optimized six-letter PCR

Standard template (Bsp-G, Table 1) and synthetic template (Bsp-P, Table 1) were amplified using *Taq* DNA polymerase under 1x Thermopol buffer (pH 8.0), followed by endonuclease digestion (Bsp120I). dA,T,G/TPs = 0.1 mM, dCTP = 0.4 mM, dZTP = 0.05 mM, and dPTP = 0.6 mM.

Lane 1 and 2: Standard template was amplified 10^4 fold using *Taq*, without (lane 1) and with (lane 2) dZTP and dPTP;

Lane 3, 4, and 5: Synthetic template, 10^3 (lane 3), 10^4 (lane 4), and 10^5 (lane 5) fold amplification, with both dZTP and dPTP;

Not digested: indicates the fraction of PCR product retained Z:P pair, therefore, resisted endonuclease digestion;

Digested: indicates the fraction of PCR product was digested. See Methods and Materials for details.

Figure 5. Strategy for sequencing GACTZP DNA

(a) The positions of **Z** and **P** in an amplicon are inferred by a process that converts **Z:P** pairs into a mixture of T:A pairs and C:G pairs, followed by standard Sanger sequencing.

Comparison of the resulting sequences shows only T:A or C:G pairs at sites where T:A or C:G pairs were present in the initial amplicon, but mixtures of T:A and C:G pairs at sites where **Z:P** pairs were present in the initial amplicon.

(b) Manipulation of the concentrations of dPTP without dZTP allows stepwise conversion of **Z:P** pairs into C:G pairs or into T:A pairs.

Table 1

Oligonucleotides used in “six-letter” GACTZP PCR and sequencing.

2P-Temp:	5'- <u>GCGTAATACGACTCACTATAGACGA</u> PP CTACTTTAGTGAGGGTTAATTCGC-3'
2Z-Temp:	3'- <u>CGCATTATGCTGAGTGATATCTGCT</u> ZZ GATGAAATCACTCCCAATTAAGCG-5'
3P-Temp:	5'- <u>GCGTAATACGACTCACTATAGACACT</u> PPP TACTCACTTTAGTGAGGGTTAATTCGC-3'
3Z-Temp:	3'- <u>CGCATTATGCTGAGTGATATCTGTGA</u> ZZZ ATGAGTGAAATCACTCCCAATTAAGCG-5'
4P-Temp:	5'- <u>GCGTAATACGACTCACTATAGACACT</u> PPPP TACTCACTTTAGTGAGGGTTAATTCGC-3'
4Z-Temp:	3'- <u>CGCATTATGCTGAGTGATATCTGTGA</u> ZZZZ ATGAGTGAAATCACTCCCAATTAAGCG-5'
2G-Temp:	5'- <u>GCGTAATACGACTCACTATAGACGA</u> GG CTACTTTAGTGAGGGTTAATTCGC-3'
4G-Temp:	5'- <u>GCGTAATACGACTCACTATAGACACT</u> GGGG TACTCACTTTAGTGAGGGTTAATTCGC-3'
ZZ-2P:	5'-GACACTAGTAGCACTCACTATACGTGACT PTC AC ZZ AGTG CP ACTACGGTCATAGCTGTTTCCTGTGTGCGA-3'
PP-2Z:	3'-CTGTGATCATCGTGAGTGATATGCACTGAG Z AGTG PP TCAC Z TGATGCCAGTATCGACAAAGGACACACGCT-5'
Bsp-Z:	5'-CTAGGACGACGGACTGCCTATGAGAGACATGAGGGCC Z GGTACCATCGATACGTTGCGATCGCTCCTTCCTG-3'
Bsp-P:	3'-GATCCTGCTGCCTGACGGATACTCTCTGTACTCCCG P CCATGGTAGCTATGCAACGCTAGCGAGGAAGGAC-5'
Bsp-C:	5'-CTAGGACGACGGACTGCCTATGAGAGACATGAGGGCC C GGTACCATCGATACGTTGCGATCGCTCCTTCCTG-3'
Bsp-G:	3'-GATCCTGCTGCCTGACGGATACTCTCTGTACTCCCG G CCATGGTAGCTATGCAACGCTAGCGAGGAAGGAC-5'

Underline sequences are either primers or primer binding sequences. The recognition sequences of restriction endonuclease Bsp120I is shown in underlined italic letters.